# **MySQL**

**Table creation in MySQL database:**

CREATE TABLE flight\_input3(s\_no integer,

Year integer, Month LONGTEXT, DayofMonth integer,DayofWeek integer,

DepTime integer, CRSDepTime integer, ArrTime integer, CRSArrTime integer,

UniqueCarrier LONGTEXT, FlightNum integer, TailNum LONGTEXT,

ActualElapsedTime integer, CRSElapsedTime integer,

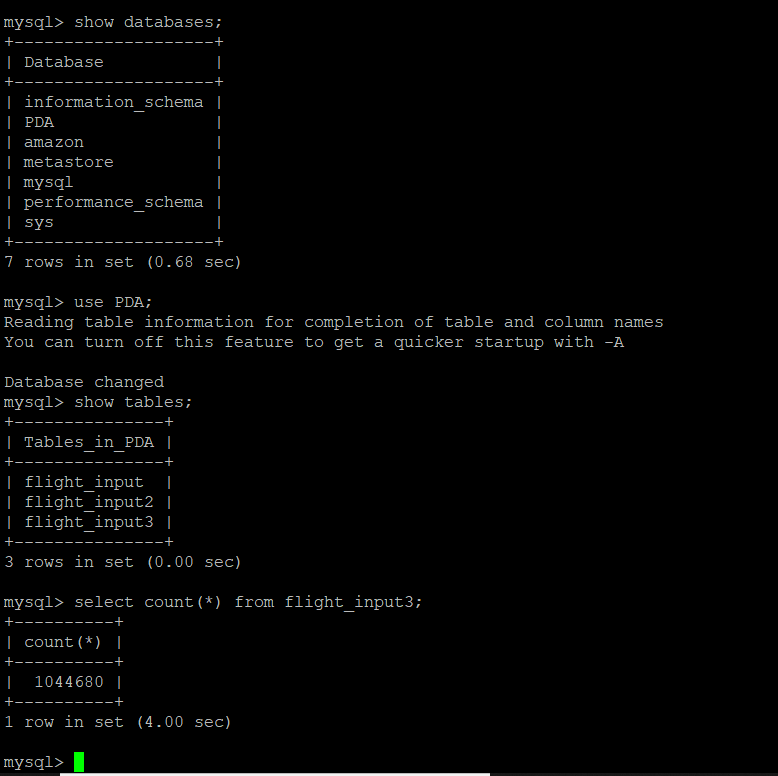
AirTime integer, ArrDelay integer,DepDelay integer, Origin LONGTEXT,

Dest LONGTEXT, Distance integer,TaxiIn integer, TaxiOut integer,

Cancelled integer, CancellationCode LONGTEXT);

**Loading of data from local into MySQL table.**

mysql -u hive -p --local\_infile=1 PDA -e "LOAD DATA LOCAL INFILE '/home/hduser/Downloads/DelayedFlights.csv' INTO TABLE flight\_input3 FIELDS TERMINATED BY ','";



# **Hive**

# Table creation:

CREATE TABLE PDA.flight\_input3

(s\_no int, Year int, Month string, DayofMonth int, DayofWeek int, DepTime int, CRSDepTime int,

ArrTime int, CRSArrTime int, UniqueCarrier string, FlightNum int, TailNum string, ActualElapsedTime int,

CRSElapsedTime int, AirTime int, ArrDelay int, DepDelay int, Origin string, Dest string, Distance int,

TaxiIn int, axiOut int, Cancelled int, CancellationCode string )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",", "quoteChar" = "\"")

stored as textfile

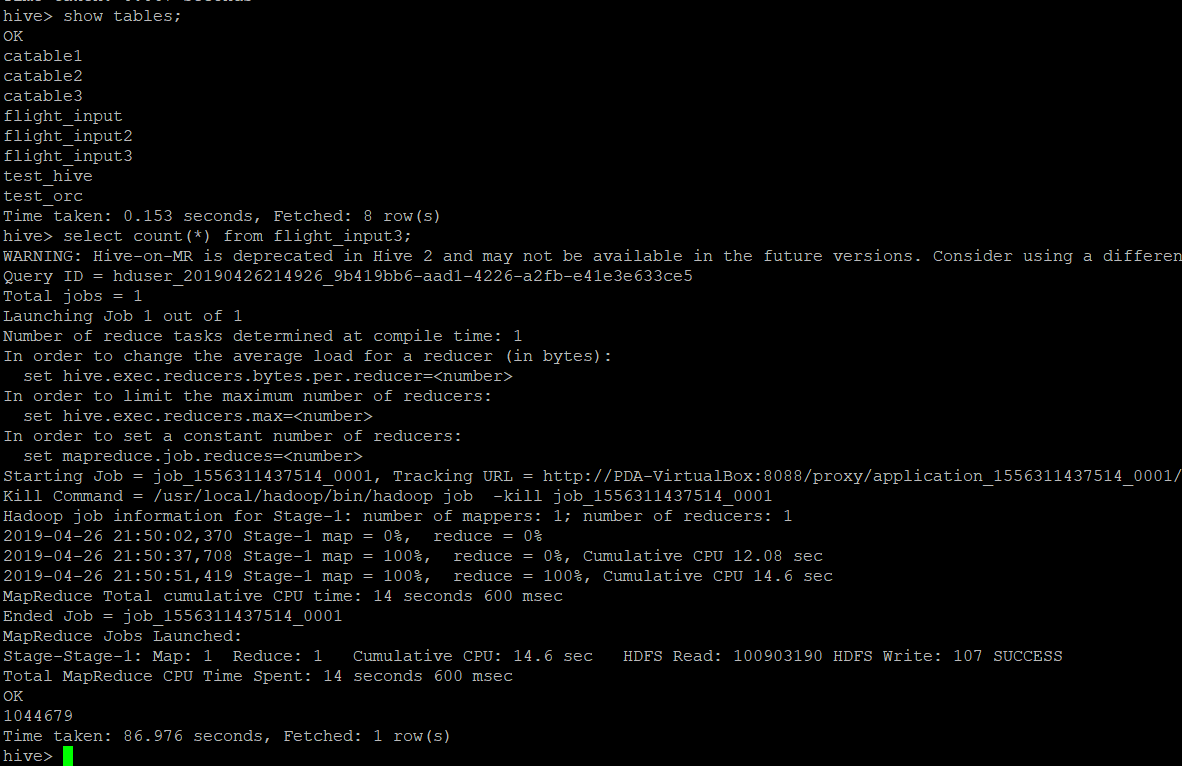
tblproperties ("skip.header.line.count"="1");

## Loading data from MySQL into HDFS:

load data inpath '/flight\_input3/flight\_input3.csv' overwrite into table flight\_input3;

# Loading data from HDFS into hive using Sqoop:

sqoop import --connect jdbc:mysql://127.0.0.1/PDA --username hive --password admin --table flight\_input3 --target-dir /flight\_input3 -m 1;



# Hive Queries:

***Query1: Flights that got cancelled maximum number of times during each month.***

**INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_Q1'**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY '\t'**

**STORED AS TEXTFILE**

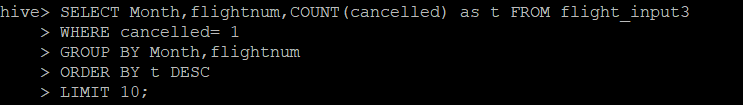
**SELECT Month,flightnum,COUNT(cancelled) as t FROM flight\_input3**

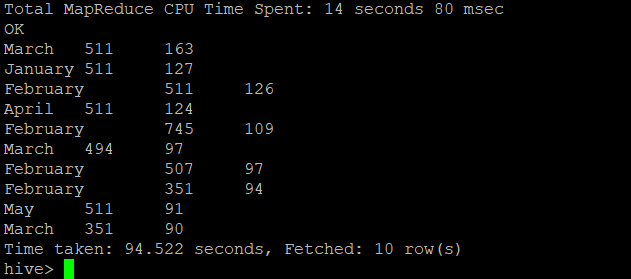
**WHERE cancelled= 1**

**GROUP BY Month,flightnum**

**ORDER BY t DESC**

**LIMIT 10;**





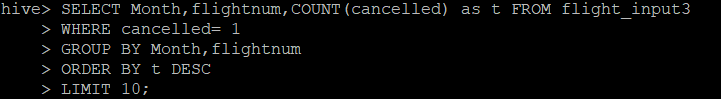
***Query2: Flights that remained in air more time during each month.***

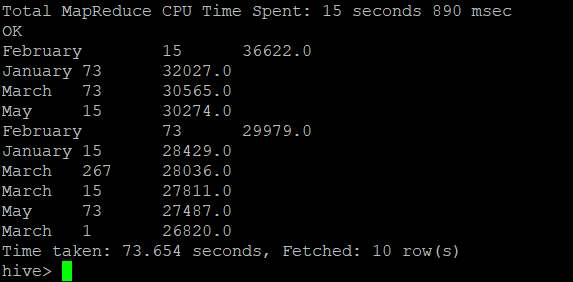
**SELECT Month,flightnum, SUM(AirTime) as d FROM flight\_input3**

**GROUP BY Month, flightnum**

**ORDER BY d DESC**

**LIMIT 10;**





***Query 3: Top 20 most visited airports in Unites States***

**INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_Q3'**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY '\t'**

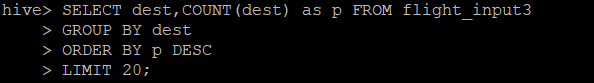
**STORED AS TEXTFILE**

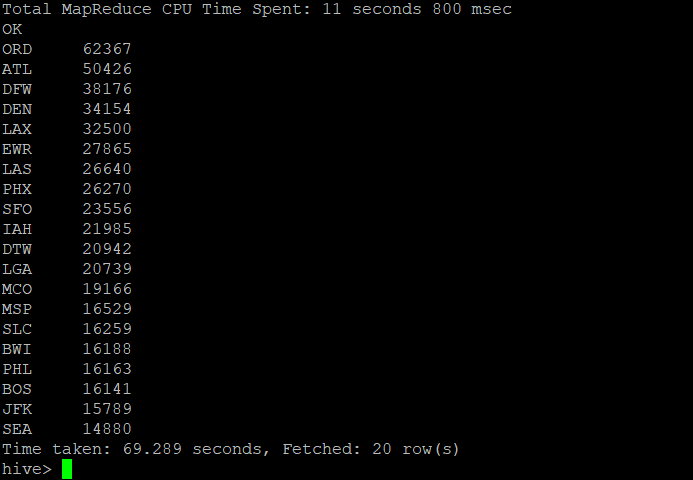
**SELECT dest,COUNT(dest) as p FROM flight\_input3**

**GROUP BY dest**

**ORDER BY p DESC**

**LIMIT 20;**





***Query 4: Top carriers in US airports***

**INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_Q4'**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY '\t'**

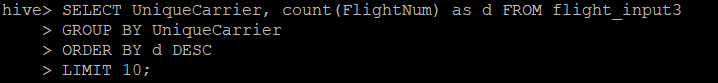
**STORED AS TEXTFILE**

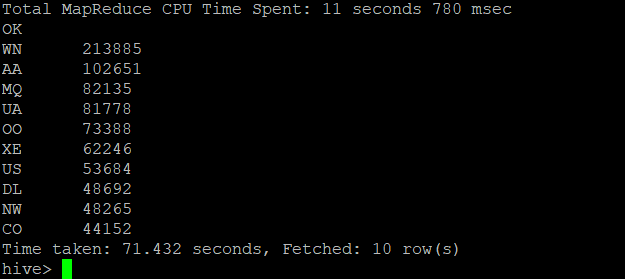
**SELECT UniqueCarrier, count(FlightNum) as d FROM flight\_input3**

**GROUP BY UniqueCarrier**

**ORDER BY d DESC**

**LIMIT 10;**

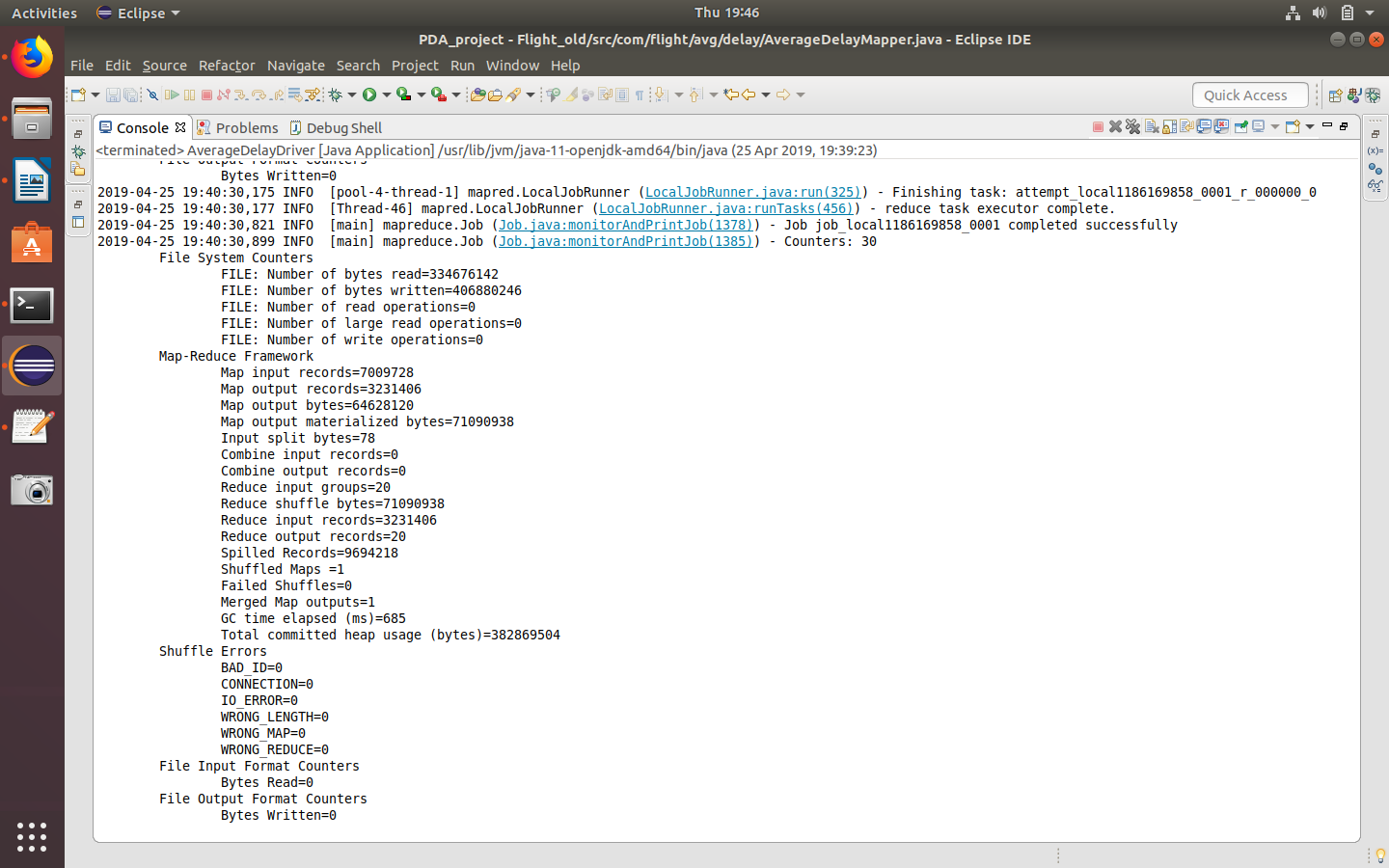




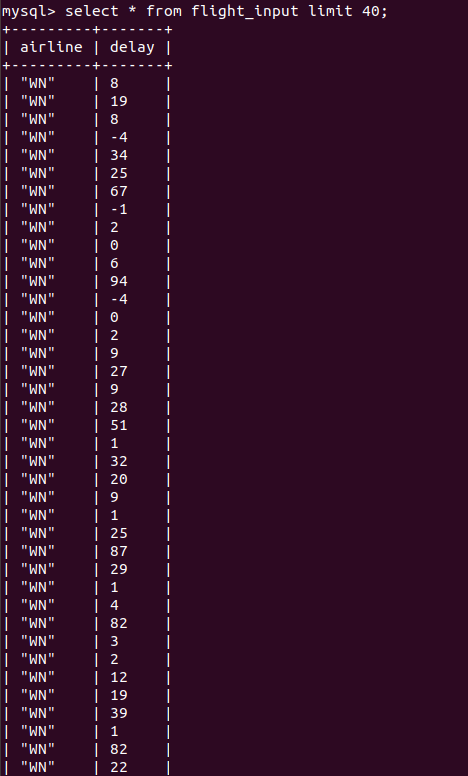


**MapReduce**

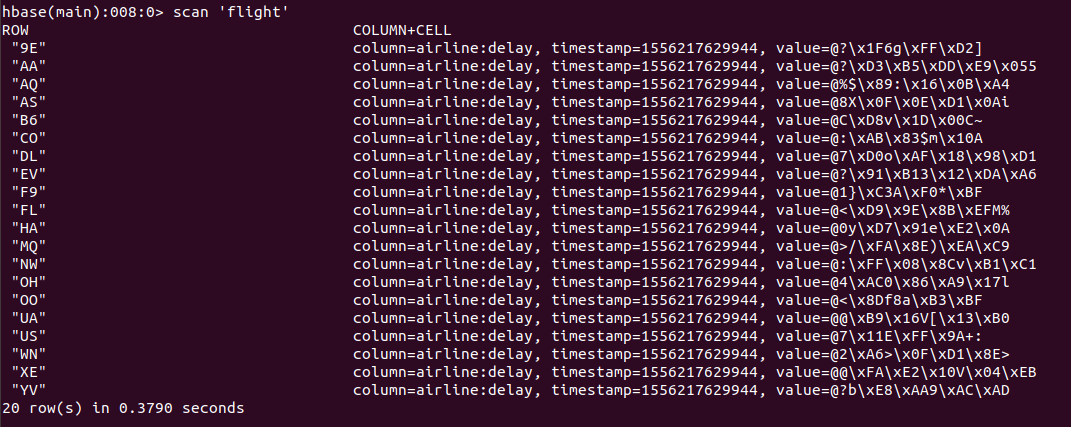
# BI Query1: Average delay time caused by airline carriers.



# Input- Extracted from MySQL database through MapR by programmatically accessing MySQL.



# Output - MapR output generated in Hbase through MapR program.

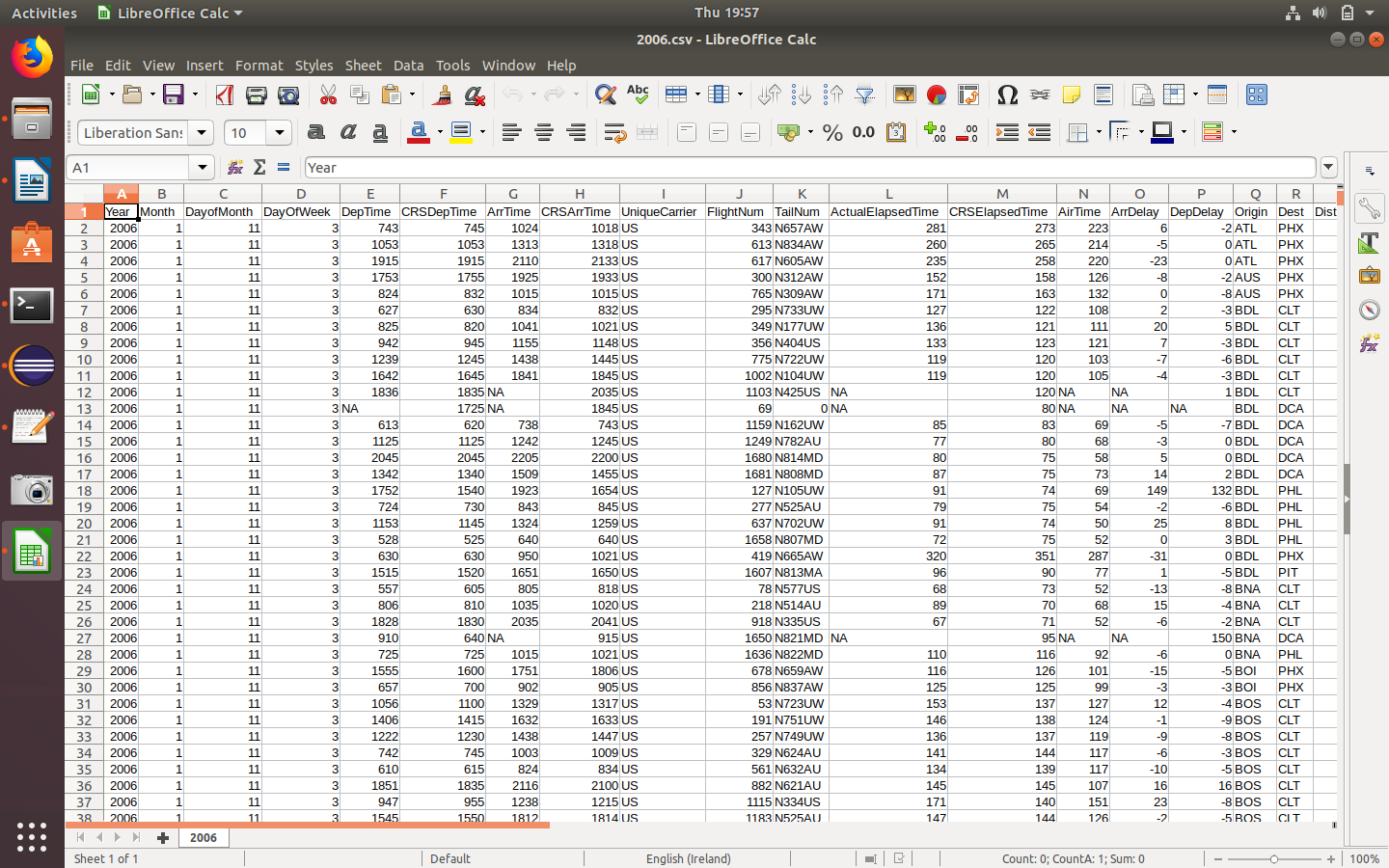


**Java Files**

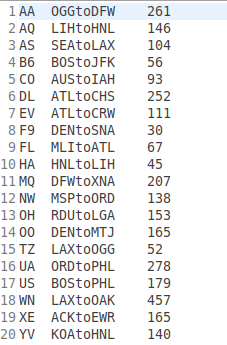


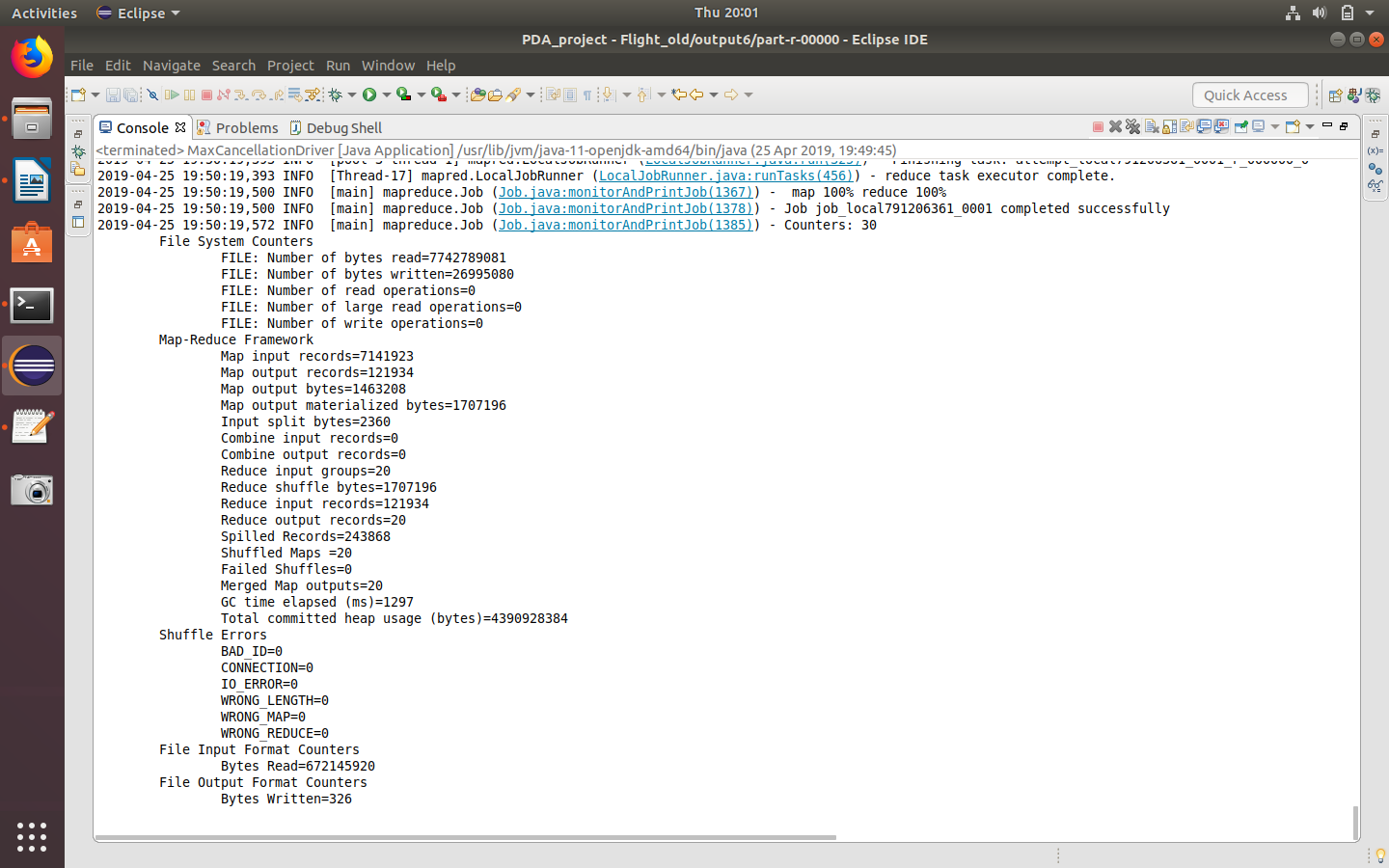


BI Query2: Maximum number of flights that got cancelled in each route.



Output:





**Java Files**



****

**Link to the dataset.**

<https://www.dropbox.com/s/wuja65yb2mo769a/DelayedFlights.csv?dl=0>