-------------------------------------------------------------------------------------

Field Description - flights.csv

Date

AirlineCode

FlightNum

Origin

Destination

DepartureTime

DepartureDelay

ArrivalTime

ArrivalDelay

Airtime

Distance

-------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------

Field Description - airlines.csv

AirlineCode (5 Numbers)

Description (80 Char)

-------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------

Use Cases:

1. Create Hive tables for the above datasets by identifying the right datatypes

2. Solve the following use cases

Find count of flights that had arrival delay

Find count of flights that had departure delay

find the average distance travelled by a flight

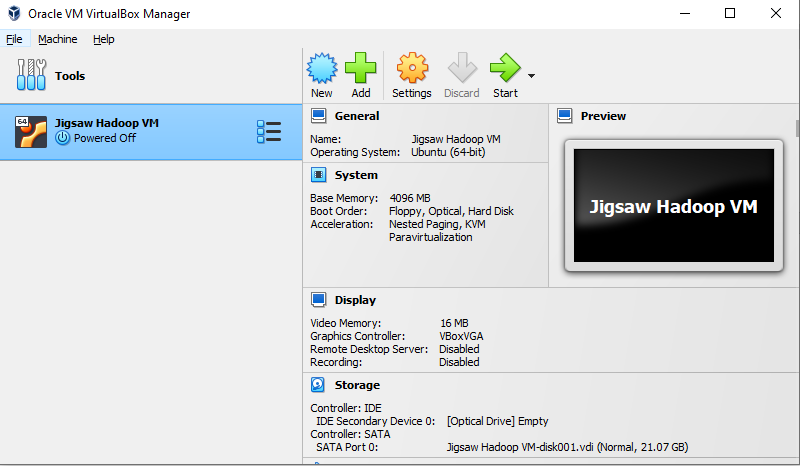
List the data that belong to the airline - American Airlines Inc

-------------------------------------------------------------------------------------

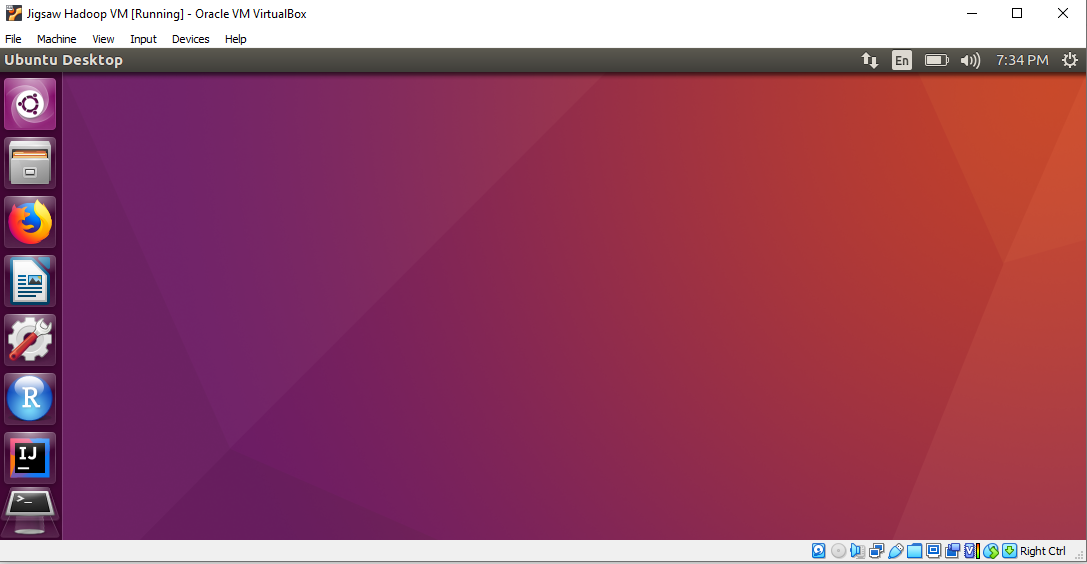
Attach screenshots of the outputs along with the query executed in the solutions file

Assignment:

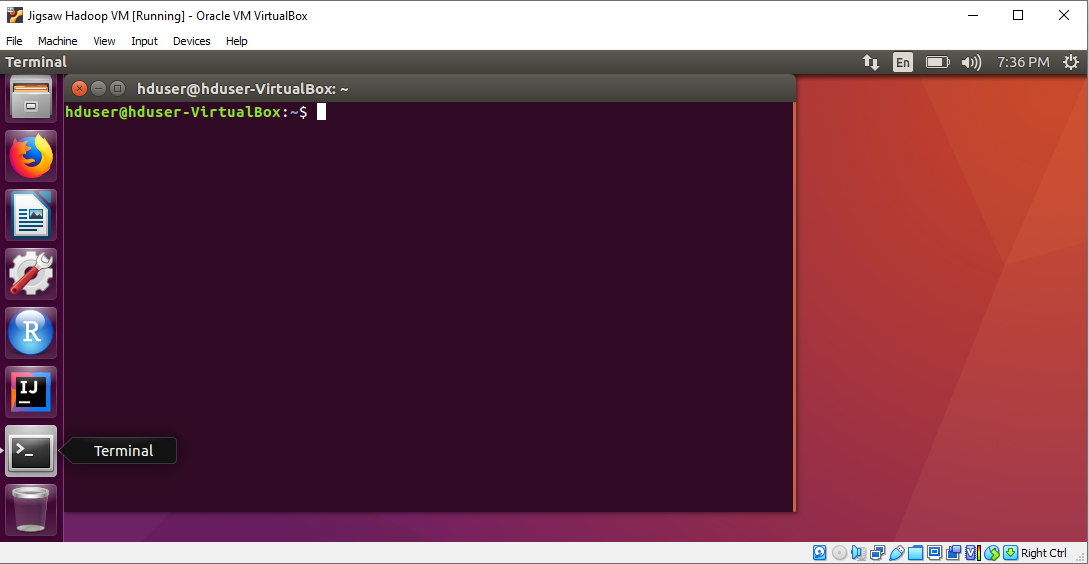
Started the ORACLE VM Virtual Box Manager and started Jigsaw Hadoop VM:



Logged in with valid VM credentials.



Launched a terminal



Checking the Hadoop version:

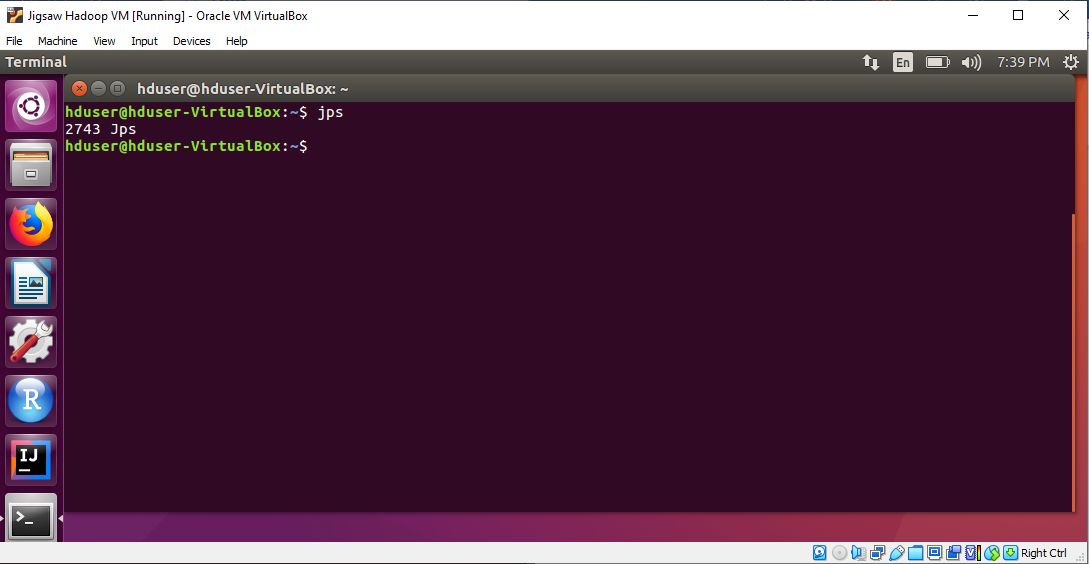
$ Hadoop version

Text

Description automatically generated

Checking if Hadoop services are running:

$ jps



Starting Hadoop services

$ cd /usr/local/hadoop-2.9.1/sbin

$ ./start-all.sh

$ cd

Text

Description automatically generated

Checking if Hadoop services are running:

$ jps

A screenshot of a computer

Description automatically generated

Mounting Shared folder:

C:\sharedfolder location already created in local machine.

VirtualBox Menu --> Devices --> Shared Folders --> Shared Folders Settings --> Check the path is defined in the settings.

Graphical user interface, text

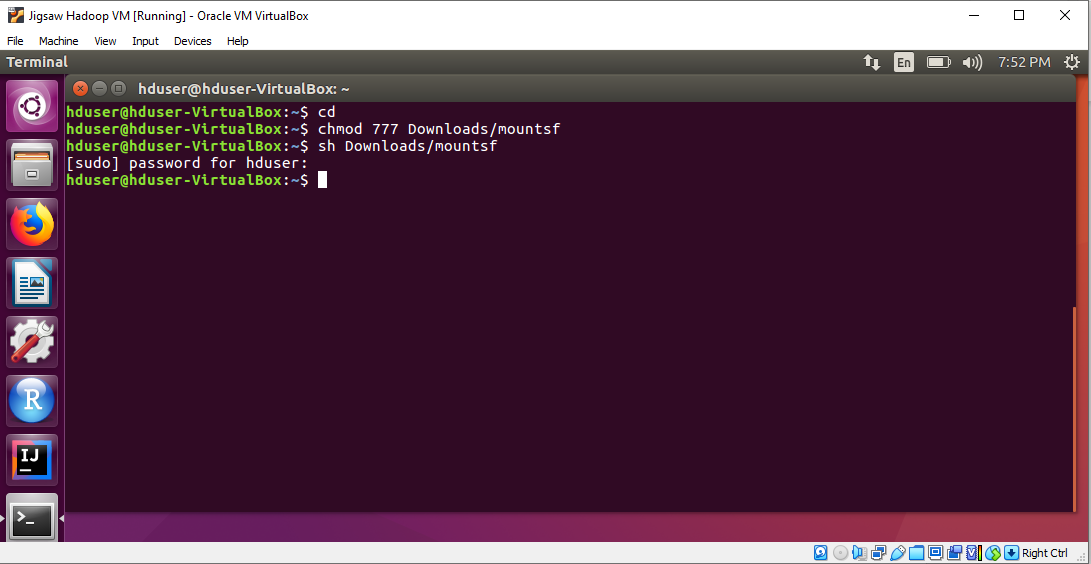
Description automatically generated

In the VM, executed the below commands to mount the shared folder:

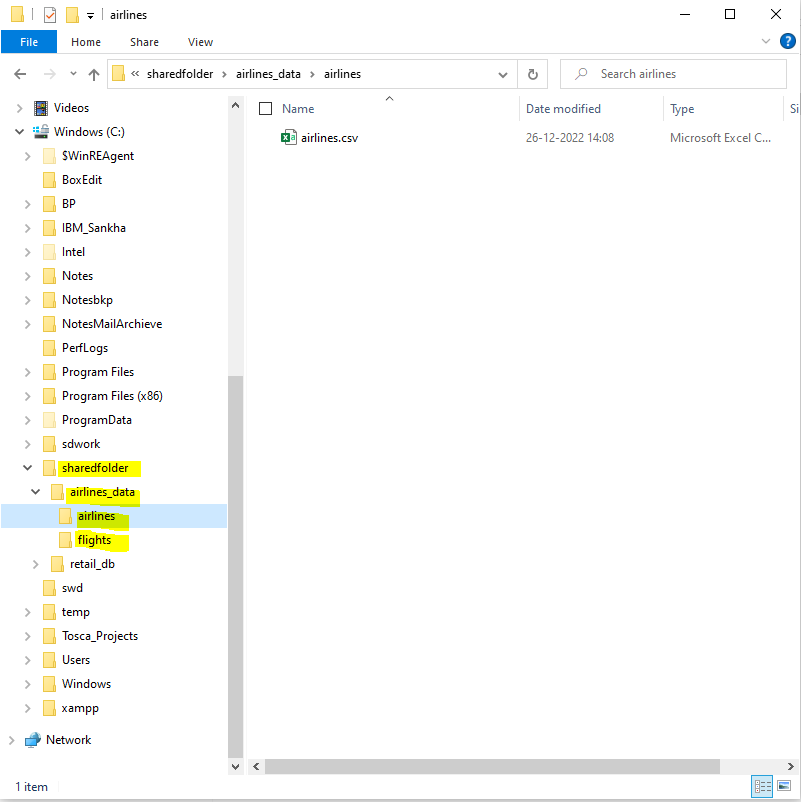
$ cd

$ chmod 777 Downloads/mountsf

$ sh Downloads/mountsf



Flights.csv and airlines.csv is moved inside a folder structure and copied in the c:\sharedfolder of local machine as below:

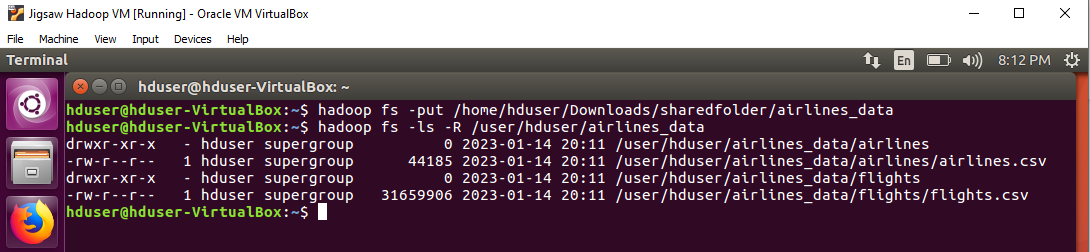


Loading data in the HDFS – executed the below commands:

$ hadoop fs -ls -R /user/hduser

$ hadoop fs -put /home/hduser/Downloads/sharedfolder/airlines\_data

$ hadoop fs -ls -R /user/hduser/airlines\_data



Creating Hive tables representing the data in HDFS:

$ hive

Text

Description automatically generated

hive> SHOW DATABASES;

Shape, rectangle

Description automatically generated

hive> CREATE DATABASE airlines;



hive> SHOW DATABASES;

Shape, rectangle

Description automatically generated

hive> USE airlines;



hive> SHOW TABLES;



hive> CREATE EXTERNAL TABLE IF NOT EXISTS flights(

flDate TIMESTAMP,

AirlineCode INT,

FlightNum INT,

Origin STRING,

Destination STRING,

DepartureTime INT,

DepartureDelay INT,

ArrivalTime INT,

ArrivalDelay INT,

Airtime INT,

Distance INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/hduser/airlines\_data/flights';



hive> SHOW TABLES;

Text

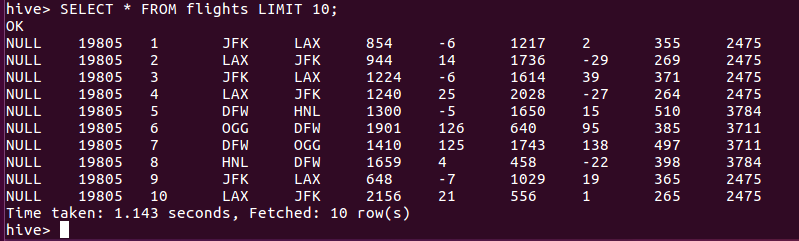
Description automatically generated with medium confidence

hive> DESC flights;

Text

Description automatically generated

Hive> SELECT \* FROM flights LIMIT 10;



hive> CREATE EXTERNAL TABLE IF NOT EXISTS airlines(

AirlineCode INT,

Description STRING

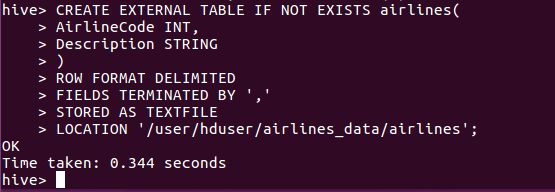
)

ROW FORMAT DELIMITED

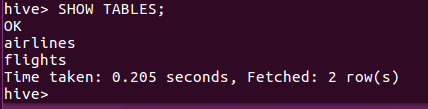
FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/hduser/airlines\_data/airlines';



hive> SHOW TABLES;



hive> DESC airlines;

Text

Description automatically generated

hive> SELECT \* FROM airlines LIMIT 10;

Text

Description automatically generated

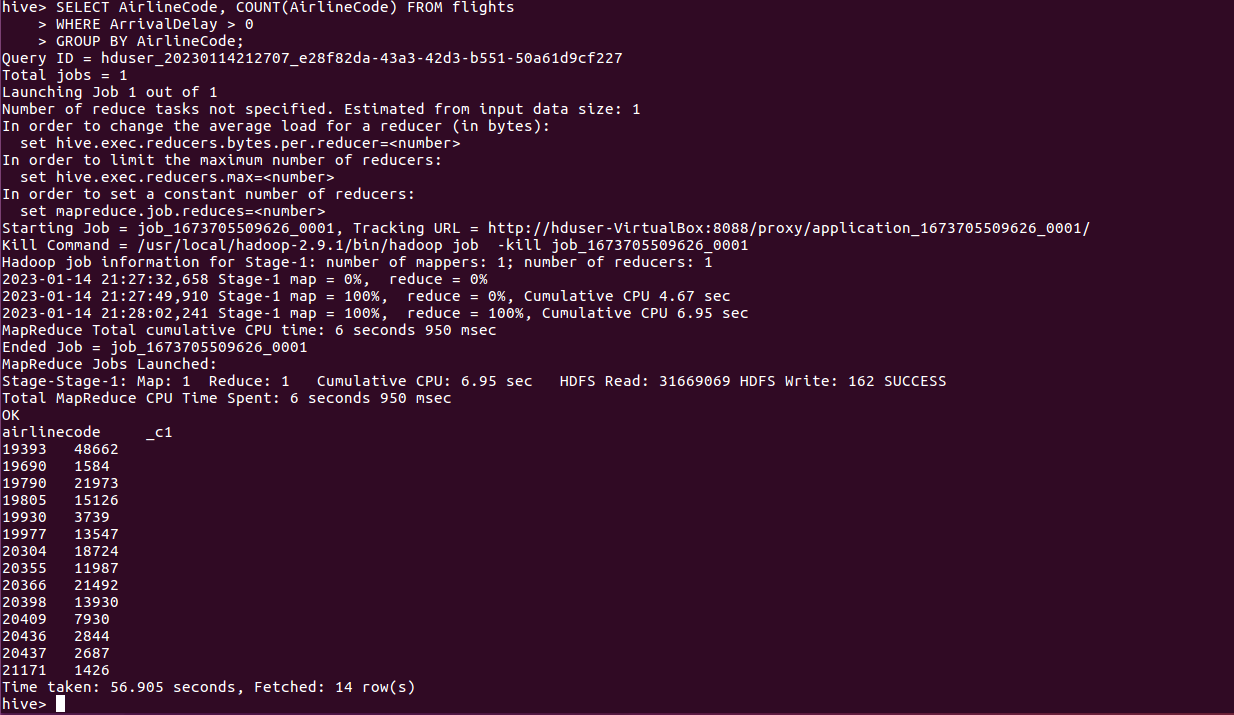
hive> set hive.cli.print.header=true;

**Find count of flights that had arrival delay**

hive> SELECT AirlineCode, COUNT(AirlineCode) FROM flights

WHERE ArrivalDelay > 0

GROUP BY AirlineCode;

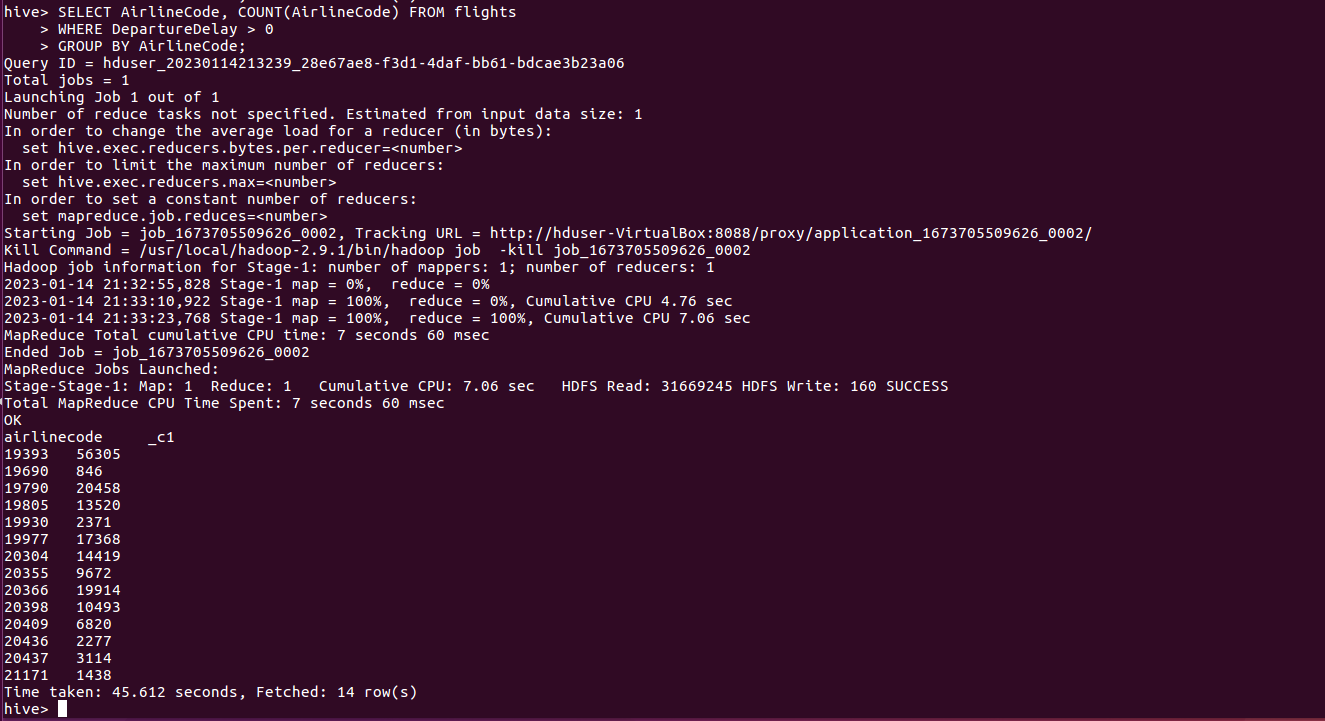


**Find count of flights that had departure delay**:

hive> SELECT AirlineCode, COUNT(AirlineCode) FROM flights

WHERE DepartureDelay > 0

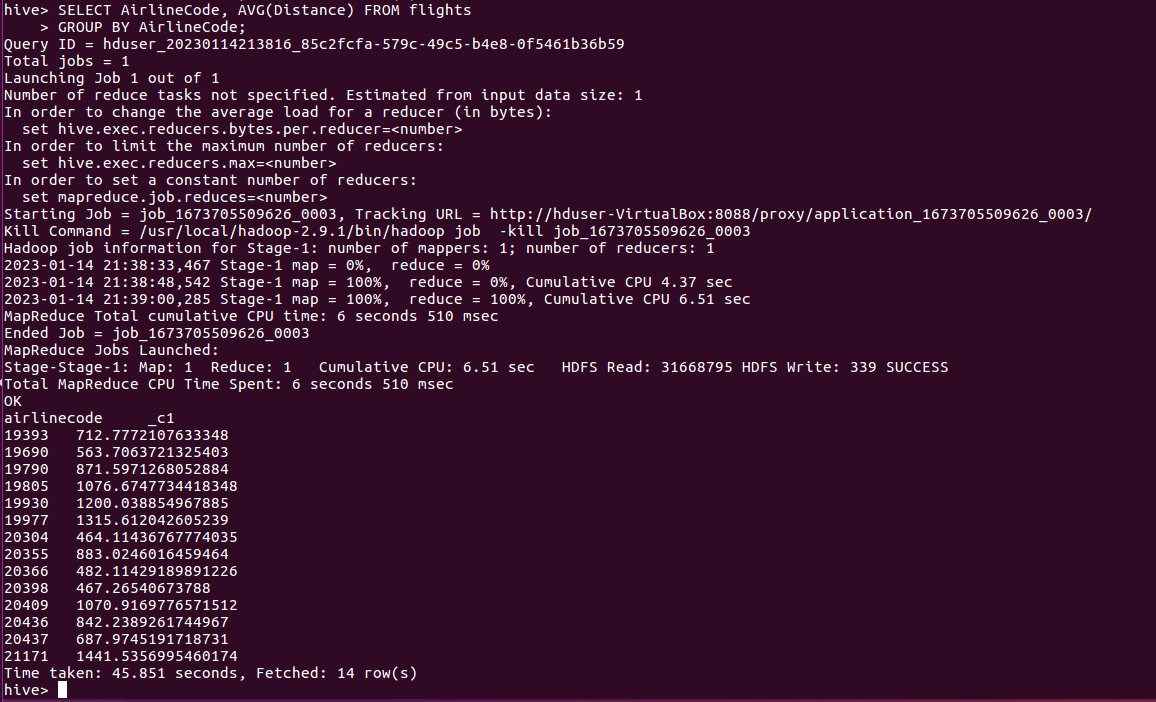
GROUP BY AirlineCode;



**Find the average distance travelled by a flight**:

hive> SELECT AirlineCode, AVG(Distance) FROM flights

GROUP BY AirlineCode;



**List the data that belong to the airline - American Airlines Inc**:

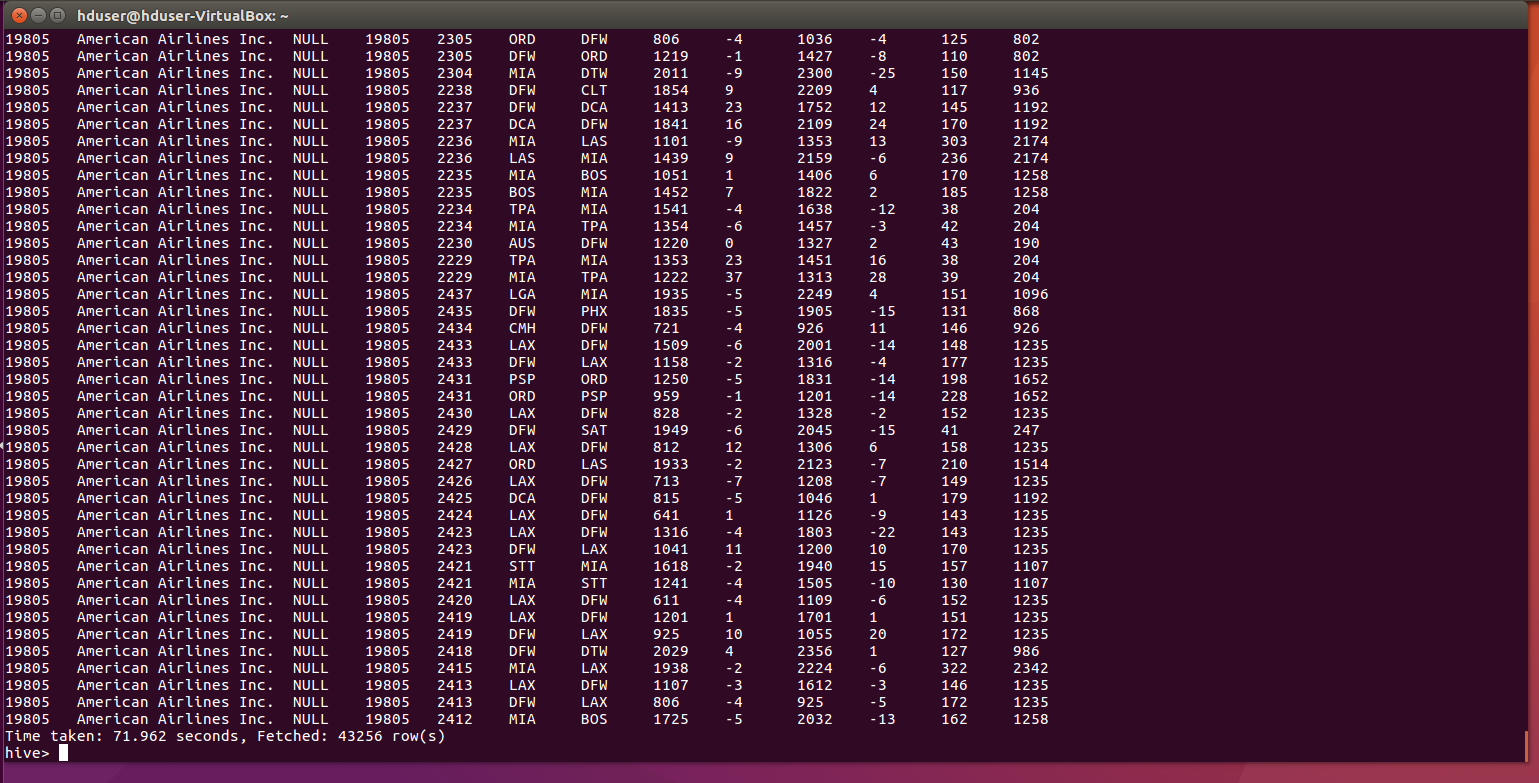
hive> SELECT a.AirlineCode, a.Description, f.flDate, f.AirlineCode, f.FlightNum, f.Origin, f.Destination, f.DepartureTime, f.DepartureDelay, f.ArrivalTime, f.ArrivalDelay, f.Airtime, f.Distance

FROM airlines a

LEFT OUTER JOIN flights f

ON f.AirlineCode = a.AirlineCode

WHERE a.Description LIKE 'American Airlines Inc%';



hive> exit;

Stopping Hadoop services:

$ cd /usr/local/hadoop-2.9.1/sbin

$ ./stop-all.sh

$ jps

Text

Description automatically generated

$ exit [to close the terminal]

Close the VM and exit from Oracle VM VirtualBox.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* END \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*