

Database Assignment Part-2

Given That:

Field Description - flights.csv

Date

AirlineCode

FlightNum

Origin

Destination

DepartureTime

DepartureDelay

ArrivalTime

ArrivalDelay

Airtime

Distance

Field Description - airlines.csv

AirlineCode

Description

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

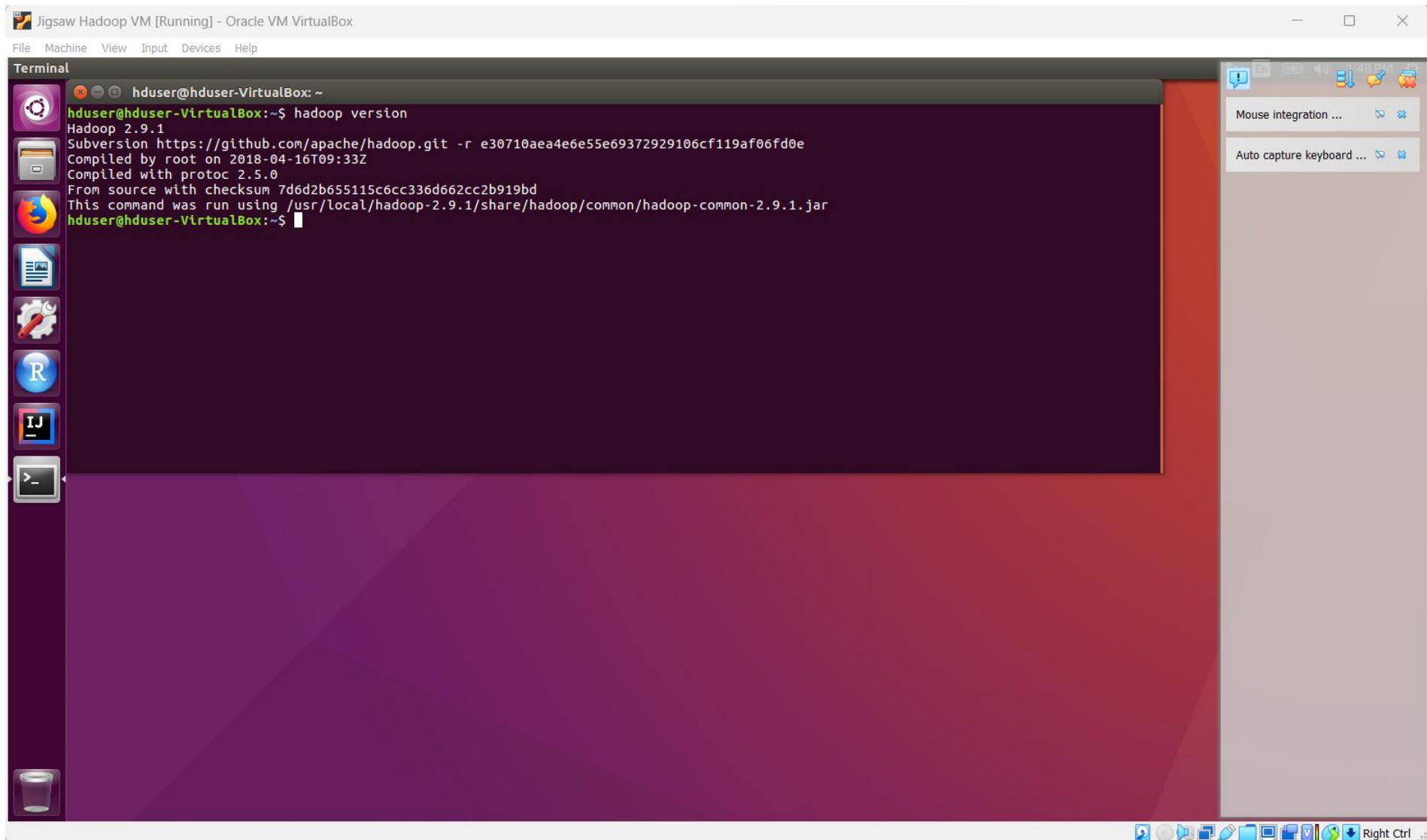
find the average distance travelled by a flight

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

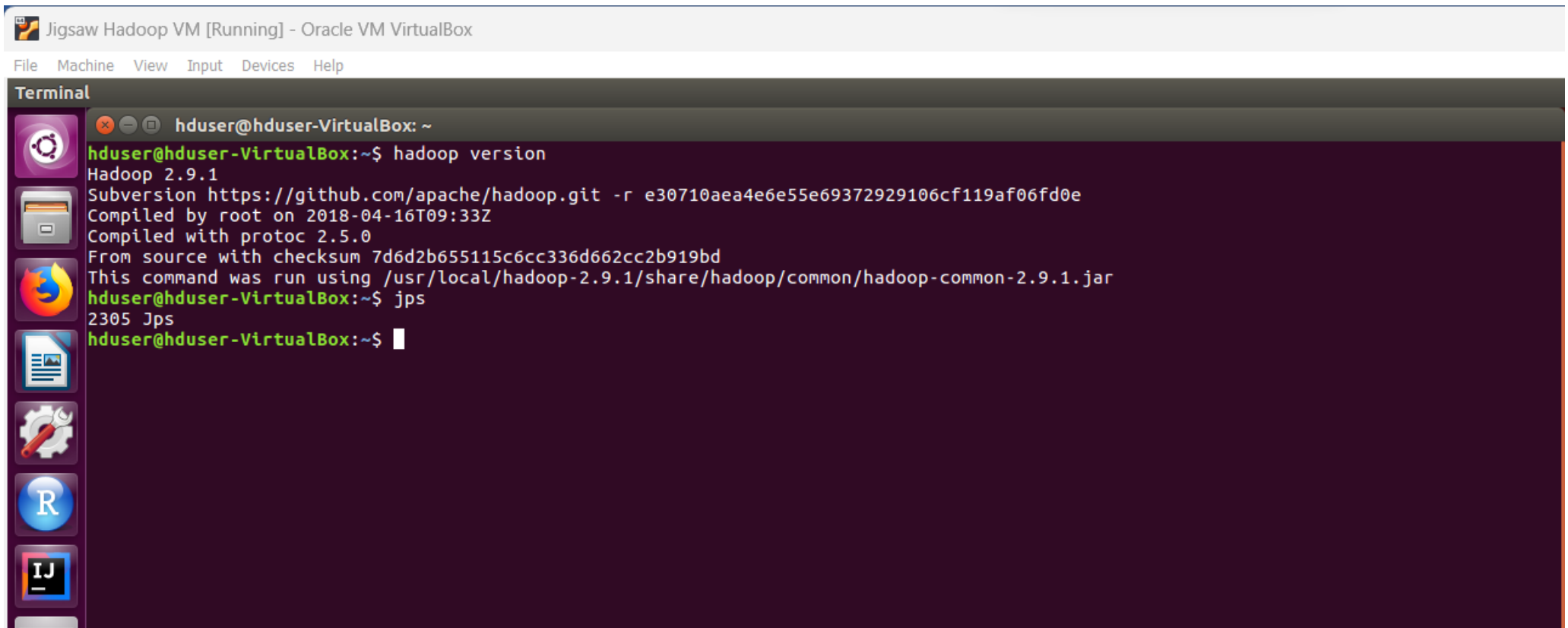
Attach screenshots of the outputs along with the query executed in the solutions file

Solutions/Answer:

1st) Checking the Hadoop Version



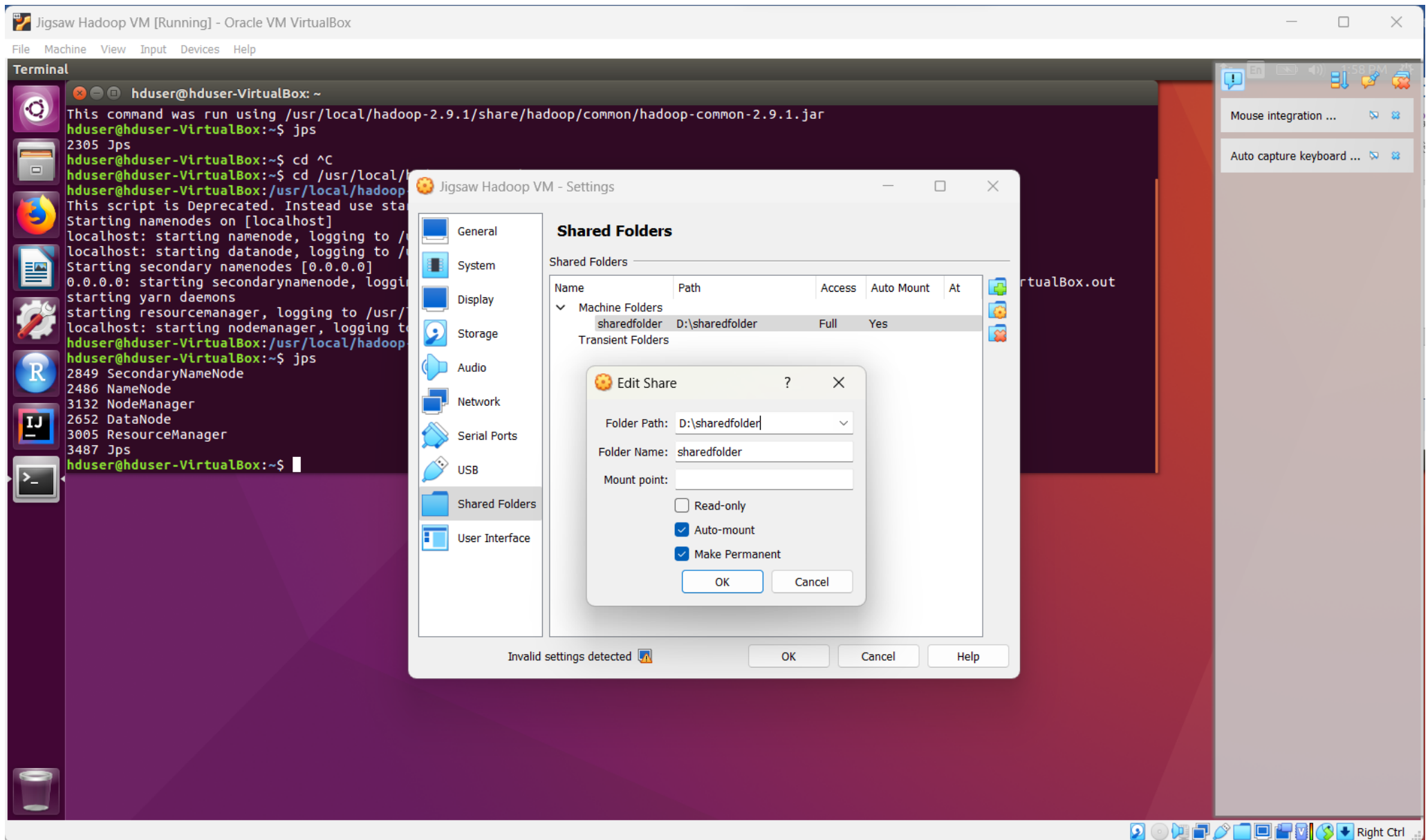
2nd) Checking Jps (Java Process Status) running or not as not



3rd) Starting jps

```
hduser@hduser-VirtualBox:~$ cd /usr/local/hadoop-2.9.1/sbin
hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-namenode-hduser-VirtualBox.out
localhost: starting datanode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-datanode-hduser-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-secondarynamenode-hduser-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.9.1/logs/yarn-hduser-resourcemanager-hduser-VirtualBox.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.9.1/logs/yarn-hduser-nodemanager-hduser-VirtualBox.out
hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$ cd
hduser@hduser-VirtualBox:~$ jps
2849 SecondaryNameNode
2486 NameNode
3132 NodeManager
2652 DataNode
3005 ResourceManager
3487 Jps
hduser@hduser-VirtualBox:~$ █
```

4th) Now Create a folder in the D: drive of the host OS (and copy the path)



Path: D:\sharedfolder

Two .csv files airlines & flights are available on D drive (Host File system)

sharedfolder

New

✂

📄

📄

📄

📄

🗑

↕ Sort

☰ View

...

← → ↕ ↑

📁 > This PC > Local Disk (D:) > sharedfolder >

🔍 Search sharedfolder

🏠 Home

> 👤 Jyoti - Personal

🖥 Desktop

📁 Downloads

📁 Documents

🖼 Pictures

🎵 Music

📺 Videos

📁 Camera

📁 New folder

📁 A1LITE BACKUP

📁 Camera

> 💻 This PC

> 🌐 Network

Name	Date modified	Type	Size
📁 retail_db	22-12-2022 08:55	File folder	
📄 airlines	03-01-2023 13:34	Microsoft Excel Co...	44 KB
📄 flights	03-01-2023 13:34	Microsoft Excel Co...	30,918 KB
📄 sample_stocks_data	21-12-2022 09:13	Microsoft Excel Co...	186 KB
📄 Data Testing Assignment - Part 1	03-01-2023 13:34	Microsoft Word D...	17 KB
📄 UseCases	03-01-2023 13:34	Text Document	2 KB

Flights.csv

Airlines.csv

File Home Insert Page Layout Formulas Data Review View Developer Help

Paste Copy Format Painter

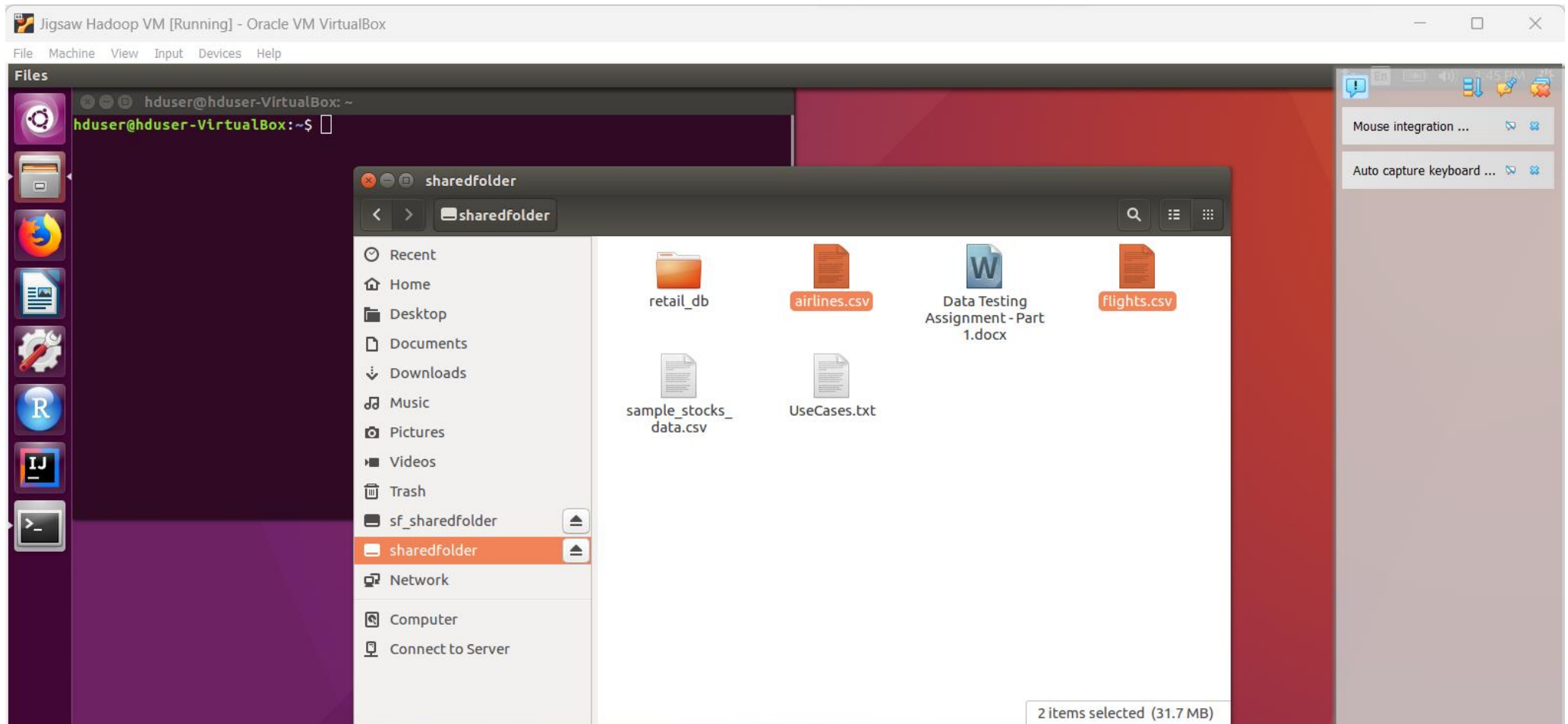
Clipboard Font Alignment

	A	B	C
1	19031	Mackey International Inc.	
2	19032	Munz Northern Airlines Inc.	
3	19033	Cochise Airlines Inc.	
4	19034	Golden Gate Airlines Inc.	
5	19035	Aeromech Inc.	
6	19036	Golden West Airlines Co.	
7	19037	Puerto Rico Intl Airlines	
8	19038	Air America Inc.	
9	19039	Swift Aire Lines Inc.	
10	19040	American Central Airlines	
11	19041	Valdez Airlines	
12	19042	Southeast Alaska Airlines	
13	19043	Altair Airlines Inc.	
14	19044	Chitina Air Service	
15	19045	Marco Island Airways Inc.	
16	19046	Caribbean Air Services Inc.	
17	19047	Sundance Airlines	
18	19048	Seair Alaska Airlines Inc.	
19	19049	Southeast Airlines Inc.	
20	19050	Alaska Aeronautical Indust.	
21	19051	Imperial Airlines Inc.	
22	19052	Trans Western Airlines Utah	
23	19053	Wright Airlines Inc.	
24	19054	Presidential Express	
25	19055	Mississippi Valley Airlines	
26	19056	Channel Flying Inc.	
27	19057	Rocky Mountain Airways Inc.	
28	19058	Midstate Airlines Inc.	
29	19059	Sedalia Marshall Boonvl Stg	
30	19060	Inlet Airlines	
31	19061	Command Airways Inc.	

airlines

Ready Accessibility: Unavailable

5TH) Now files are available on Ubuntu shared folder



6TH) Now Putting Data into HDFS

create an **airlines_dir** and put **airlines.csv** and **flights.csv** data into it

- /user/hduser is the working directory in the VM

Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Terminal

```
hduser@hduser-VirtualBox: ~
hduser@hduser-VirtualBox:~$ sh Downloads/mountsf
[sudo] password for hduser:
hduser@hduser-VirtualBox:~$ jps
2849 SecondaryNameNode
2486 NameNode
3132 NodeManager
2652 DataNode
4316 Jps
3005 ResourceManager
hduser@hduser-VirtualBox:~$ hadoop fs -ls /user/hduser
Found 3 items
drwxr-xr-x  - hduser supergroup          0 2023-01-03 14:09 /user/hduser/airlines_dir
drwxr-xr-x  - hduser supergroup          0 2022-12-22 09:01 /user/hduser/retail_db
drwxr-xr-x  - hduser supergroup          0 2022-12-21 09:36 /user/hduser/sample_stocks_dir
hduser@hduser-VirtualBox:~$
```

Now reading the airlines.csv file

```
hduser@hduser-VirtualBox:~$ hadoop fs -cat /user/hduser/airlines_dir/airlines.csv
19031,Mackey International Inc.
19032,Munz Northern Airlines Inc.
19033,Cochise Airlines Inc.
19034,Golden Gate Airlines Inc.
19035,Aeromech Inc.
19036,Golden West Airlines Co.
19037,Puerto Rico Intl Airlines
19038,Air America Inc.
19039,Swift Aire Lines Inc.
19040,American Central Airlines
19041,Valdez Airlines
19042,Southeast Alaska Airlines
19043,Altair Airlines Inc.
19044,Chittina Air Service
19045,Marco Island Airways Inc.
19046,Caribbean Air Services Inc.
19047,Sundance Airlines
19048,Seair Alaska Airlines Inc.
19049,Southeast Airlines Inc.
19050,Alaska Aeronautical Indust.
19051,Imperial Airlines Inc.
19052,Trans Western Airlines Utah
19053,Wright Airlines Inc.
19054,Presidential Express
19055,Mississippi Valley Airlines
19056,Channel Flying Inc.
19057,Rocky Mountain Airways Inc.
19058,Midstate Airlines Inc.
19059,Sedalia Marshall Boonvl Stg
19060,Inlet Airlines
19061,Command Airways Inc.
19062,Emerald Airlines
19063,Orion Air Inc.
19064,Pan Am Express
19065,Air Micronesia Inc.
19066,Sunbird Inc.
19067,Wings West Airlines
19068,Southcentral Air Inc.
19069,Air Cargo Express Inc.
19070,Associated Aviation Act.
19071,Antilles Air Boats Inc.
19072,AAA Airlines
19073,Argosy Air Lines Inc.
19074,Air Bahia
```



Now reading the flights.csv file

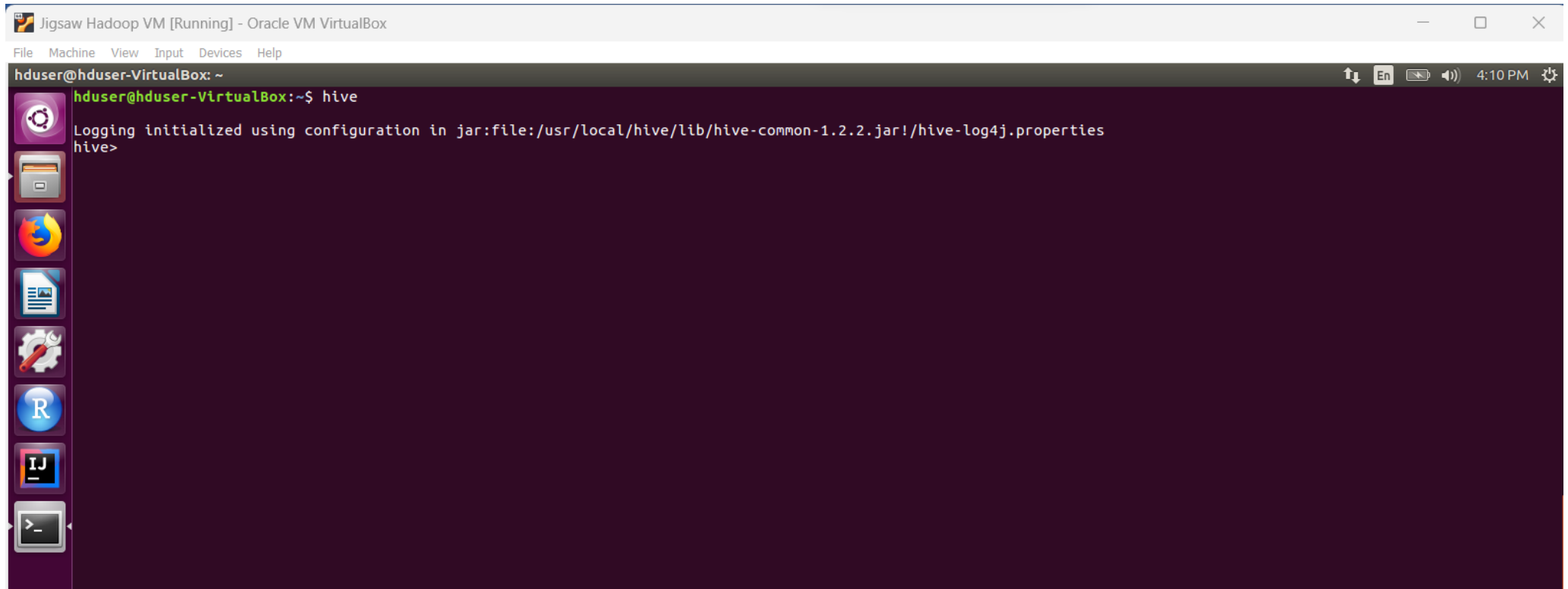
21034,Atlanta ATL Cargo
hduser@hduser-VirtualBox:~\$ hadoop fs -cat /user/hduser/airlines_dir/flights.csv

File Machine View Input Devices Help

hduser@hduser-VirtualBox: ~
2014-04-30,19393,558,SAT,DAL,1736,11.00,1831,1.00,42.00,248.00
2014-04-30,19393,655,SAT,DAL,0657,-3.00,0754,-6.00,47.00,248.00
2014-04-30,19393,2158,SAT,DAL,1014,-1.00,1112,-3.00,44.00,248.00
2014-04-30,19393,2350,SAT,DAL,1435,-5.00,1530,-15.00,42.00,248.00
2014-04-30,19393,2420,SAT,DAL,1254,-1.00,1350,-5.00,42.00,248.00
2014-04-30,19393,4215,SAT,DAL,2109,119.00,2201,111.00,41.00,248.00
2014-04-30,19393,528,SAT,DEN,1530,0.00,1639,-6.00,111.00,794.00
2014-04-30,19393,922,SAT,DEN,0653,3.00,0801,-4.00,113.00,794.00
2014-04-30,19393,1073,SAT,DEN,1942,22.00,2051,16.00,110.00,794.00
2014-04-30,19393,3385,SAT,ELP,1742,42.00,1814,44.00,78.00,496.00
2014-04-30,19393,4557,SAT,ELP,0933,-2.00,1009,-1.00,86.00,496.00
2014-04-30,19393,264,SAT,HOU,0647,-3.00,0732,-13.00,32.00,192.00
2014-04-30,19393,395,SAT,HOU,1656,26.00,1749,24.00,36.00,192.00
2014-04-30,19393,441,SAT,HOU,1413,-2.00,1500,-10.00,36.00,192.00
2014-04-30,19393,549,SAT,HOU,1031,-4.00,1133,3.00,44.00,192.00
2014-04-30,19393,2515,SAT,HOU,1835,30.00,1917,17.00,33.00,192.00
2014-04-30,19393,4809,SAT,HRL,1222,12.00,1312,7.00,39.00,233.00
2014-04-30,19393,498,SAT,LAS,1803,168.00,1911,186.00,165.00,1069.00
2014-04-30,19393,505,SAT,LAS,0613,-2.00,0716,11.00,166.00,1069.00
2014-04-30,19393,580,SAT,LAS,1827,52.00,1937,72.00,168.00,1069.00
2014-04-30,19393,711,SAT,LAS,1139,14.00,1240,25.00,168.00,1069.00
2014-04-30,19393,3860,SAT,LAX,1939,24.00,2044,19.00,176.00,1211.00
2014-04-30,19393,4737,SAT,LAX,1054,-1.00,1206,-4.00,176.00,1211.00
2014-04-30,19393,2524,SAT,MCO,0647,-3.00,1028,3.00,147.00,1041.00
2014-04-30,19393,2170,SAT,MDW,1737,17.00,1959,4.00,127.00,1036.00
2014-04-30,19393,2791,SAT,MDW,0544,-1.00,0808,-12.00,133.00,1036.00
2014-04-30,19393,387,SAT,PHX,1238,48.00,1258,48.00,130.00,843.00
2014-04-30,19393,575,SAT,PHX,2128,128.00,2148,123.00,130.00,843.00
2014-04-30,19393,3125,SAT,PHX,0746,16.00,0810,15.00,133.00,843.00
2014-04-30,19393,503,SAT,SAN,1845,20.00,1938,18.00,164.00,1129.00
2014-04-30,19393,1016,SAT,SAN,1201,-4.00,1301,-4.00,166.00,1129.00
2014-04-30,19393,173,SAT,STL,0552,-3.00,0746,-14.00,102.00,786.00
2014-04-30,19393,615,SAT,STL,1434,-1.00,1631,-9.00,104.00,786.00
2014-04-30,19393,230,SAT,TPA,1549,-1.00,1905,-5.00,123.00,972.00
2014-04-30,19393,3058,SDF,ATL,0556,-4.00,0715,0.00,60.00,321.00
2014-04-30,19393,617,SDF,BWI,1226,-4.00,1352,-8.00,72.00,495.00
2014-04-30,19393,1993,SDF,BWI,1839,19.00,2007,12.00,74.00,495.00
2014-04-30,19393,3681,SDF,BWI,0749,-6.00,0921,-9.00,74.00,495.00
2014-04-30,19393,4277,SDF,BWI,1626,6.00,1805,10.00,77.00,495.00
2014-04-30,19393,587,SDF,DEN,0926,-4.00,1011,-9.00,152.00,1024.00
2014-04-30,19393,1078,SDF,LAS,1910,145.00,2004,139.00,222.00,1624.00
2014-04-30,19393,728,SDF,MCO,1721,96.00,1924,99.00,104.00,719.00
2014-04-30,19393,133,SDF,MDW,2147,117.00,2146,106.00,43.00,271.00
2014-04-30,19393,462,SDF,MDW,1250,10.00,1249,-1.00,46.00,271.00
2014-04-30,19393,505,SDF,MDW,1801,16.00,1813,23.00,45.00,271.00
2014-04-30,19393,595,SDF,MDW,1037,17.00,1038,8.00,46.00,271.00
2014-04-30,19393,4126,SDF,MDW,0557,-3.00,0604,-6.00,45.00,271.00
hduser@hduser-VirtualBox:~\$

7TH) Use Cases: 1. Create Hive tables for the above datasets by identifying the right datatypes

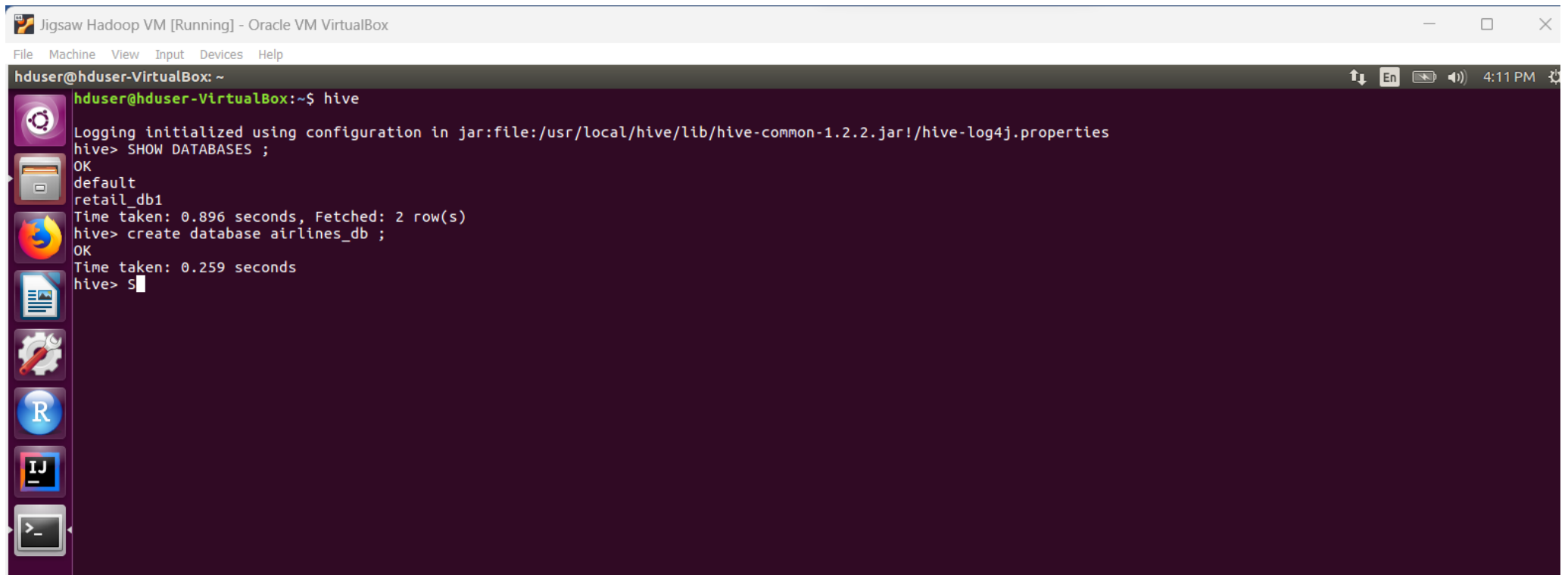
Starting hive



The screenshot shows a VirtualBox window titled "Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox". Inside the window is a terminal window with a dark purple background. The terminal prompt is "hduser@hduser-VirtualBox: ~". The user has entered the command "hive". The output of the command is "Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties" followed by a "hive>" prompt. On the left side of the terminal window, there is a vertical dock with several application icons: a gear, a folder, a Firefox browser, a document, a settings gear, a blue circle with a white 'R', a terminal icon, and a terminal icon with a green cursor.

```
hduser@hduser-VirtualBox: ~  
hduser@hduser-VirtualBox:~$ hive  
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties  
hive>
```

Creating airlines_db



```
hduser@hduser-VirtualBox: ~  
hduser@hduser-VirtualBox:~$ hive  
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties  
hive> SHOW DATABASES ;  
OK  
default  
retail_db1  
Time taken: 0.896 seconds, Fetched: 2 row(s)  
hive> create database airlines_db ;  
OK  
Time taken: 0.259 seconds  
hive> s
```

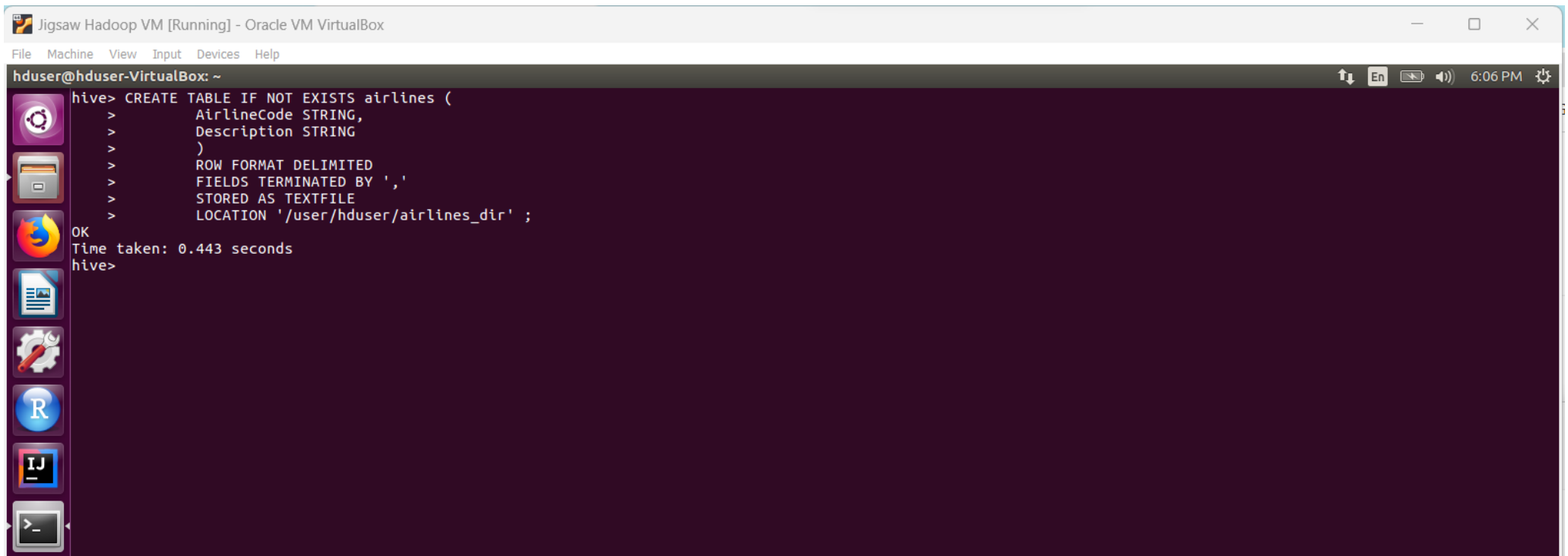
Now checking existing db & Tables

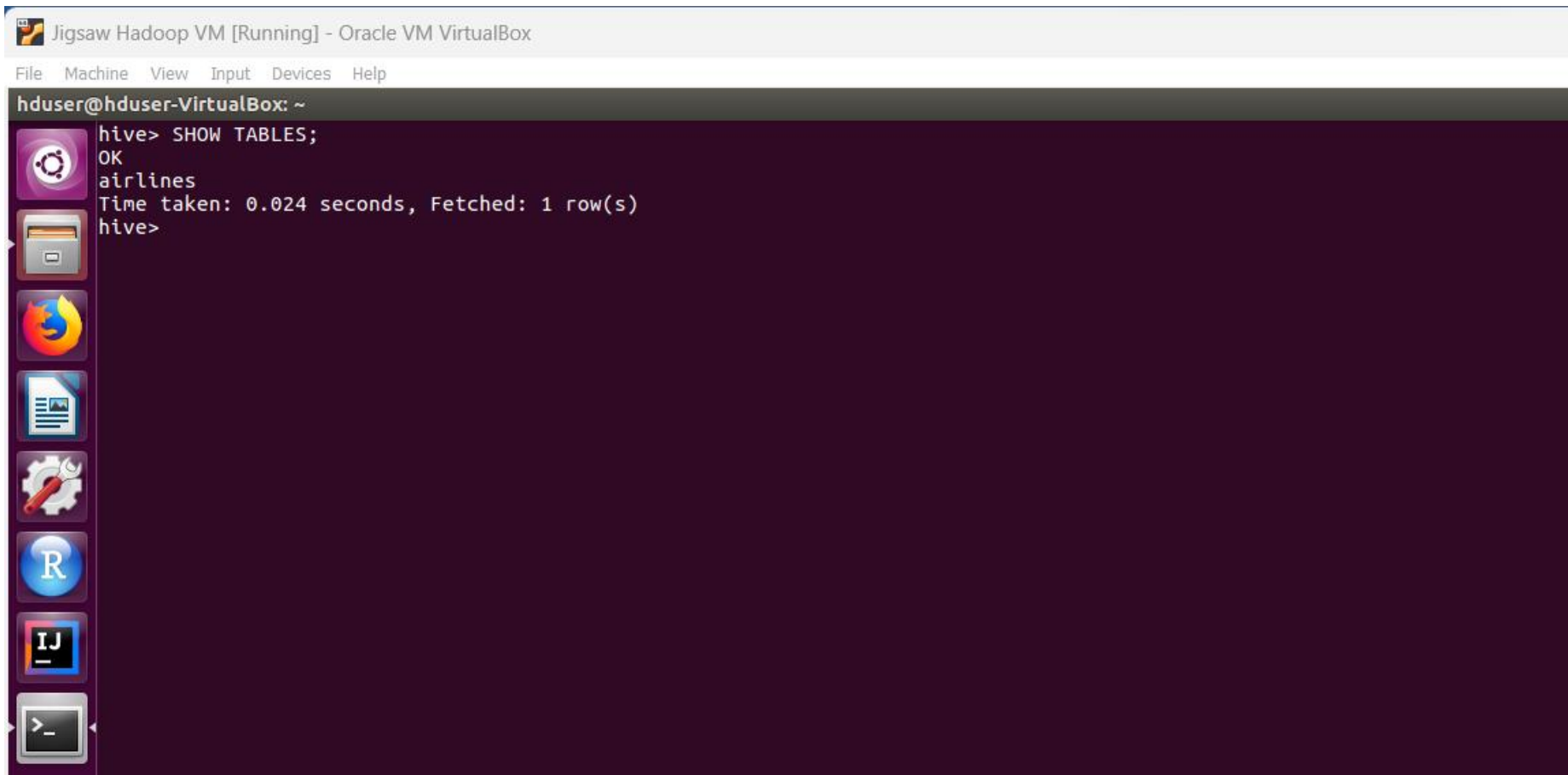
Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

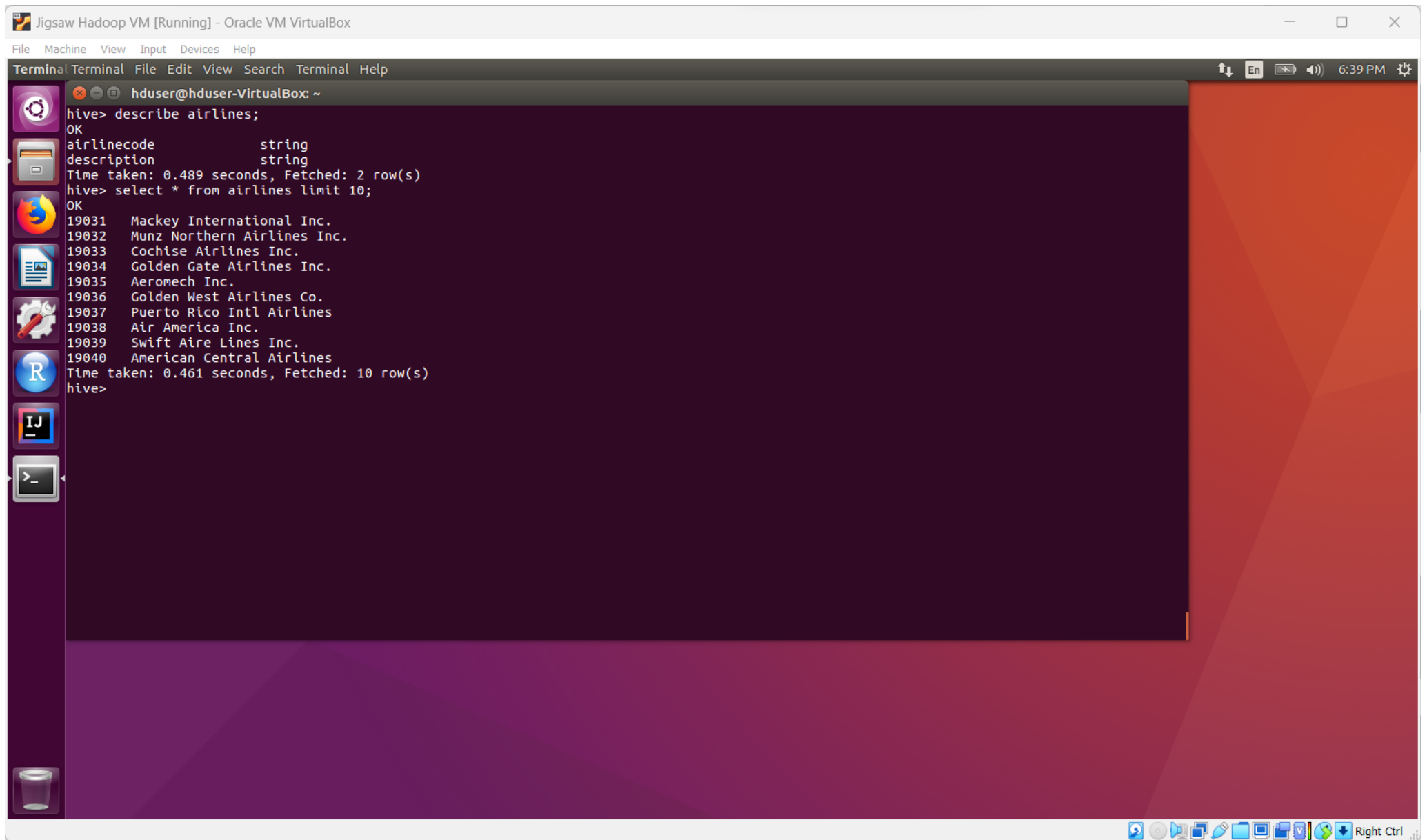
```
hduser@hduser-VirtualBox: ~  
hduser@hduser-VirtualBox:~$ hive  
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties  
hive> SHOW DATABASES ;  
OK  
default  
retail_db1  
Time taken: 0.896 seconds, Fetched: 2 row(s)  
hive> create database airlines_db ;  
OK  
Time taken: 0.259 seconds  
hive> USE airlines_db ;  
OK  
Time taken: 0.062 seconds  
hive> SHOW TABLES ;  
OK  
Time taken: 0.077 seconds  
hive>
```

Now creating airlines table in hive





Airlines data has been loaded.



Now Creating another directory for flights

```
hduser@hduser-VirtualBox:~$ hadoop fs -ls /user/hduser
Found 4 items
drwxr-xr-x   - hduser supergroup          0 2023-01-03 15:57 /user/hduser/airlines_dir
drwxr-xr-x   - hduser supergroup          0 2023-01-03 18:18 /user/hduser/flights_dir
drwxr-xr-x   - hduser supergroup          0 2022-12-22 09:01 /user/hduser/retail_db
drwxr-xr-x   - hduser supergroup          0 2022-12-21 09:36 /user/hduser/sample_stocks_dir
hduser@hduser-VirtualBox:~$
```

Now loading flights data from .csv file

```
hduser@hduser-VirtualBox: ~
2014-04-30,19393,580,SAT,LAS,1827,52.00,1937,72.00,168.00,1069.00
2014-04-30,19393,711,SAT,LAS,1139,14.00,1240,25.00,168.00,1069.00
2014-04-30,19393,3860,SAT,LAX,1939,24.00,2044,19.00,176.00,1211.00
2014-04-30,19393,4737,SAT,LAX,1054,-1.00,1206,-4.00,176.00,1211.00
2014-04-30,19393,2524,SAT,MCO,0647,-3.00,1028,3.00,147.00,1041.00
2014-04-30,19393,2170,SAT,MDW,1737,17.00,1959,4.00,127.00,1036.00
2014-04-30,19393,2791,SAT,MDW,0544,-1.00,0808,-12.00,133.00,1036.00
2014-04-30,19393,387,SAT,PHX,1238,48.00,1258,48.00,130.00,843.00
2014-04-30,19393,575,SAT,PHX,2128,128.00,2148,123.00,130.00,843.00
2014-04-30,19393,3125,SAT,PHX,0746,16.00,0810,15.00,133.00,843.00
2014-04-30,19393,503,SAT,SAN,1845,20.00,1938,18.00,164.00,1129.00
2014-04-30,19393,1016,SAT,SAN,1201,-4.00,1301,-4.00,166.00,1129.00
2014-04-30,19393,173,SAT,STL,0552,-3.00,0746,-14.00,102.00,786.00
2014-04-30,19393,615,SAT,STL,1434,-1.00,1631,-9.00,104.00,786.00
2014-04-30,19393,230,SAT,TPA,1549,-1.00,1905,-5.00,123.00,972.00
2014-04-30,19393,3058,SDF,ATL,0556,-4.00,0715,0.00,60.00,321.00
2014-04-30,19393,617,SDF,BWI,1226,-4.00,1352,-8.00,72.00,495.00
2014-04-30,19393,1993,SDF,BWI,1839,19.00,2007,12.00,74.00,495.00
2014-04-30,19393,3681,SDF,BWI,0749,-6.00,0921,-9.00,74.00,495.00
2014-04-30,19393,4277,SDF,BWI,1626,6.00,1805,10.00,77.00,495.00
2014-04-30,19393,587,SDF,DEN,0926,-4.00,1011,-9.00,152.00,1024.00
2014-04-30,19393,1078,SDF,LAS,1910,145.00,2004,139.00,222.00,1624.00
2014-04-30,19393,728,SDF,MCO,1721,96.00,1924,99.00,104.00,719.00
2014-04-30,19393,133,SDF,MDW,2147,117.00,2146,106.00,43.00,271.00
2014-04-30,19393,462,SDF,MDW,1250,10.00,1249,-1.00,46.00,271.00
2014-04-30,19393,505,SDF,MDW,1801,16.00,1813,23.00,45.00,271.00
2014-04-30,19393,595,SDF,MDW,1037,17.00,1038,8.00,46.00,271.00
2014-04-30,19393,4126,SDF,MDW,0557,-3.00,0604,-6.00,45.00,271.00
```

Used below query to create flights table:

```
CREATE TABLE IF NOT EXISTS flights (
```

```
    Flight_Date DATE,
```

```
    AirlineCode STRING,
```

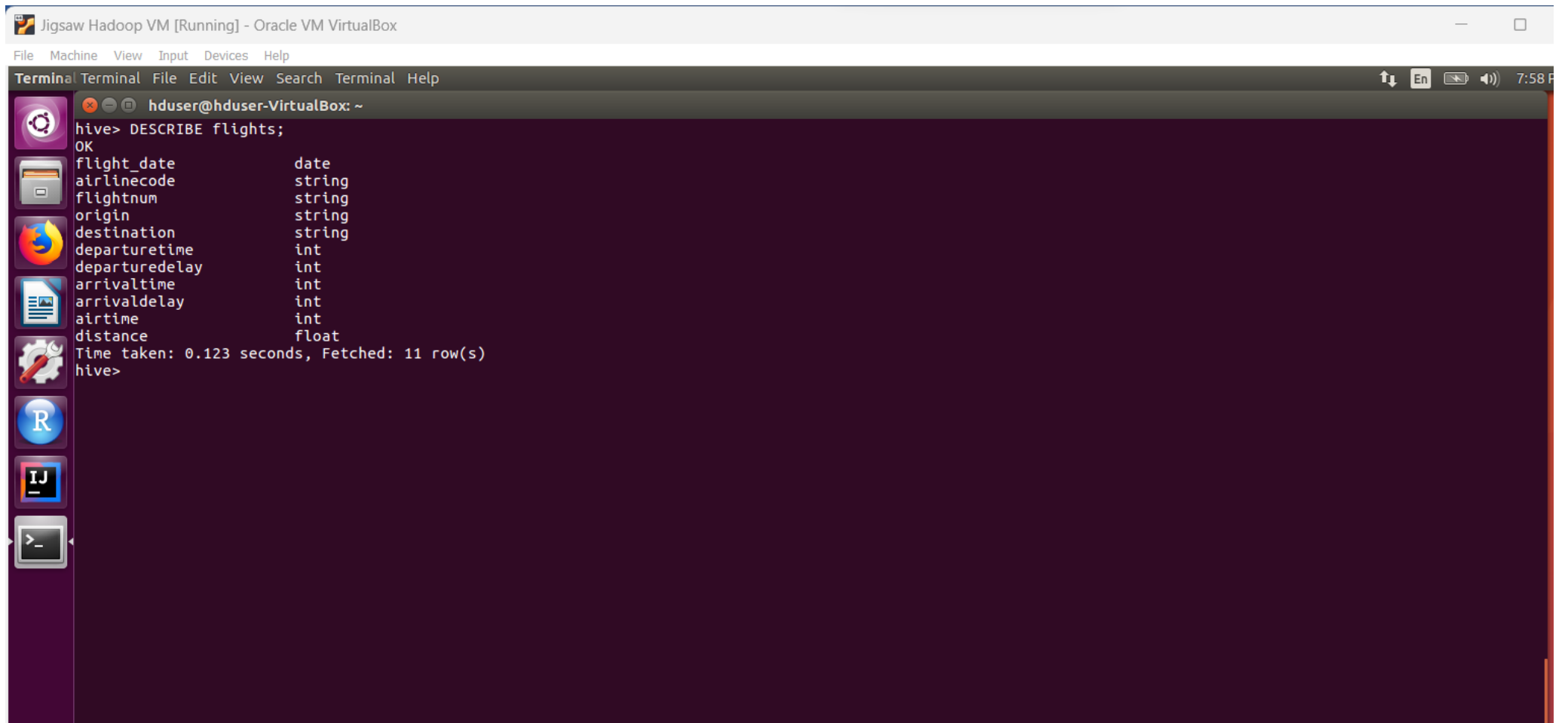
```
    FlightNum STRING,
```

```
Origin STRING,  
Destination STRING,  
DepartureTime INT,  
DepartureDelay TIMESTAMP,  
ArrivalTime TIMESTAMP,  
ArrivalDelay TIMESTAMP,  
Airtime TIMESTAMP,  
Distance FLOAT  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
LOCATION '/user/hduser/flights_dir' ;
```

```
Time taken: 0.002 seconds, Fetched: 2 row(s)
hive> CREATE TABLE IF NOT EXISTS flights (
>     Flight_Date DATE,
>     AirlineCode STRING,
>     FlightNum STRING,
>     Origin STRING,
>     Destination STRING,
>     DepartureTime INT,
>     DepartureDelay TIMESTAMP,
>     ArrivalTime TIMESTAMP,
>     ArrivalDelay TIMESTAMP,
>     Airtime TIMESTAMP,
>     Distance FLOAT
> )
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> STORED AS TEXTFILE
> LOCATION '/user/hduser/flights_dir' ;
OK
Time taken: 0.119 seconds
hive>
```

```
hive> DESCRIBE flights;
OK
flight_date          date
airlinecode          string
flightnum            string
origin               string
destination           string
departuretime        int
departuredelay       timestamp
arrivaltime          timestamp
arrivaldelay         timestamp
airtime              timestamp
distance             float
Time taken: 0.404 seconds, Fetched: 11 row(s)
hive>
```

Now Altered table



The screenshot shows a terminal window titled "Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox". The terminal is running Hive and has executed the command `DESCRIBE flights;`. The output lists the columns and their data types for the 'flights' table. The columns are: `flight_date` (date), `airlinecode` (string), `flightnum` (string), `origin` (string), `destination` (string), `departuretime` (int), `departuredelay` (int), `arrivaltime` (int), `arrivaldelay` (int), `airtime` (int), and `distance` (float). The terminal also shows the execution time as 0.123 seconds and that 11 rows were fetched. The prompt `hive>` is visible at the end of the output.

```
hive> DESCRIBE flights;
OK
flight_date      date
airlinecode      string
flightnum        string
origin           string
destination       string
departuretime    int
departuredelay   int
arrivaltime      int
arrivaldelay     int
airtime          int
distance         float
Time taken: 0.123 seconds, Fetched: 11 row(s)
hive>
```

Now again loading data from

Load data local INPATH '/home/hduser/Downloads/sharedfolder/flights.csv'

OVERWRITE into TABLE flights;


```

hive> Load data local INPATH '/home/hduser/Downloads/sharedfolder/flights.csv'
> OVERWRITE into TABLE flights;
Loading data to table airlines_db.flights
Table airlines_db.flights stats: [numFiles=0, numRows=0, totalSize=0, rawDataSize=0]
OK
Time taken: 1.219 seconds
hive> SELECT * FROM flights LIMIT 10 ;
OK
2014-04-01      19805   1      JFK    LAX    854    -6     1217    2      355    2475.0
2014-04-01      19805   2      LAX    JFK    944    14     1736   -29     269    2475.0
2014-04-01      19805   3      JFK    LAX    1224   -6     1614   39     371    2475.0
2014-04-01      19805   4      LAX    JFK    1240   25     2028  -27     264    2475.0
2014-04-01      19805   5      DFW    HNL    1300   -5     1650   15     510    3784.0
2014-04-01      19805   6      OGG    DFW    1901  126     640    95     385    3711.0
2014-04-01      19805   7      DFW    OGG    1410  125     1743   138    497    3711.0
2014-04-01      19805   8      HNL    DFW    1659    4     458   -22     398    3784.0
2014-04-01      19805   9      JFK    LAX    648   -7     1029   19     365    2475.0
2014-04-01      19805  10      LAX    JFK    2156   21     556    1     265    2475.0
Time taken: 0.123 seconds, Fetched: 10 row(s)
hive>

```

Enabling column names

set hive.cli.print.header=true;

set hive.resultset.use.unique.column.names=false;

```

hive> SELECT * FROM flights LIMIT 10 ;
OK
flight_date      airlinecode      flightnum      origin  destination      departuretime      departuredelay      arrivaltime      arrivaldelay      airtime distance
2014-04-01      19805   1      JFK    LAX    854    -6     1217    2      355    2475.0
2014-04-01      19805   2      LAX    JFK    944    14     1736   -29     269    2475.0
2014-04-01      19805   3      JFK    LAX    1224   -6     1614   39     371    2475.0
2014-04-01      19805   4      LAX    JFK    1240   25     2028  -27     264    2475.0
2014-04-01      19805   5      DFW    HNL    1300   -5     1650   15     510    3784.0
2014-04-01      19805   6      OGG    DFW    1901  126     640    95     385    3711.0
2014-04-01      19805   7      DFW    OGG    1410  125     1743   138    497    3711.0
2014-04-01      19805   8      HNL    DFW    1659    4     458   -22     398    3784.0
2014-04-01      19805   9      JFK    LAX    648   -7     1029   19     365    2475.0
2014-04-01      19805  10      LAX    JFK    2156   21     556    1     265    2475.0
Time taken: 0.082 seconds, Fetched: 10 row(s)
hive>

```

Now,

2. Solve the following use cases

Find count of flights that had arrival delay

SELECT COUNT(ArrivalDelay) as Arrived_Delay_Flights_Count FROM flights where ArrivalDelay>0;

```
hive> SELECT COUNT(ArrivalDelay) as Arrived_Delay_Flights_Count FROM flights where ArrivalDelay>0;
Query ID = hduser_20230103202301_d39d2dd0-fb2a-4821-9f54-9f03c77bf297
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0002, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0002/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-03 20:23:09,252 Stage-1 map = 0%, reduce = 0%
2023-01-03 20:23:16,557 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.89 sec
2023-01-03 20:23:21,808 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.01 sec
MapReduce Total cumulative CPU time: 3 seconds 10 msec
Ended Job = job_1672734213680_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.01 sec HDFS Read: 31668108 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 10 msec
OK
arrived_delay_flights_count
185651
Time taken: 21.248 seconds, Fetched: 1 row(s)
hive> █
```

arrived_delay_flights_count

185651

Find count of flights that had departure delay

SELECT COUNT(DepartureDelay) as Departure_Delay_Flights_Count FROM flights where DepartureDelay>0;

Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Terminal

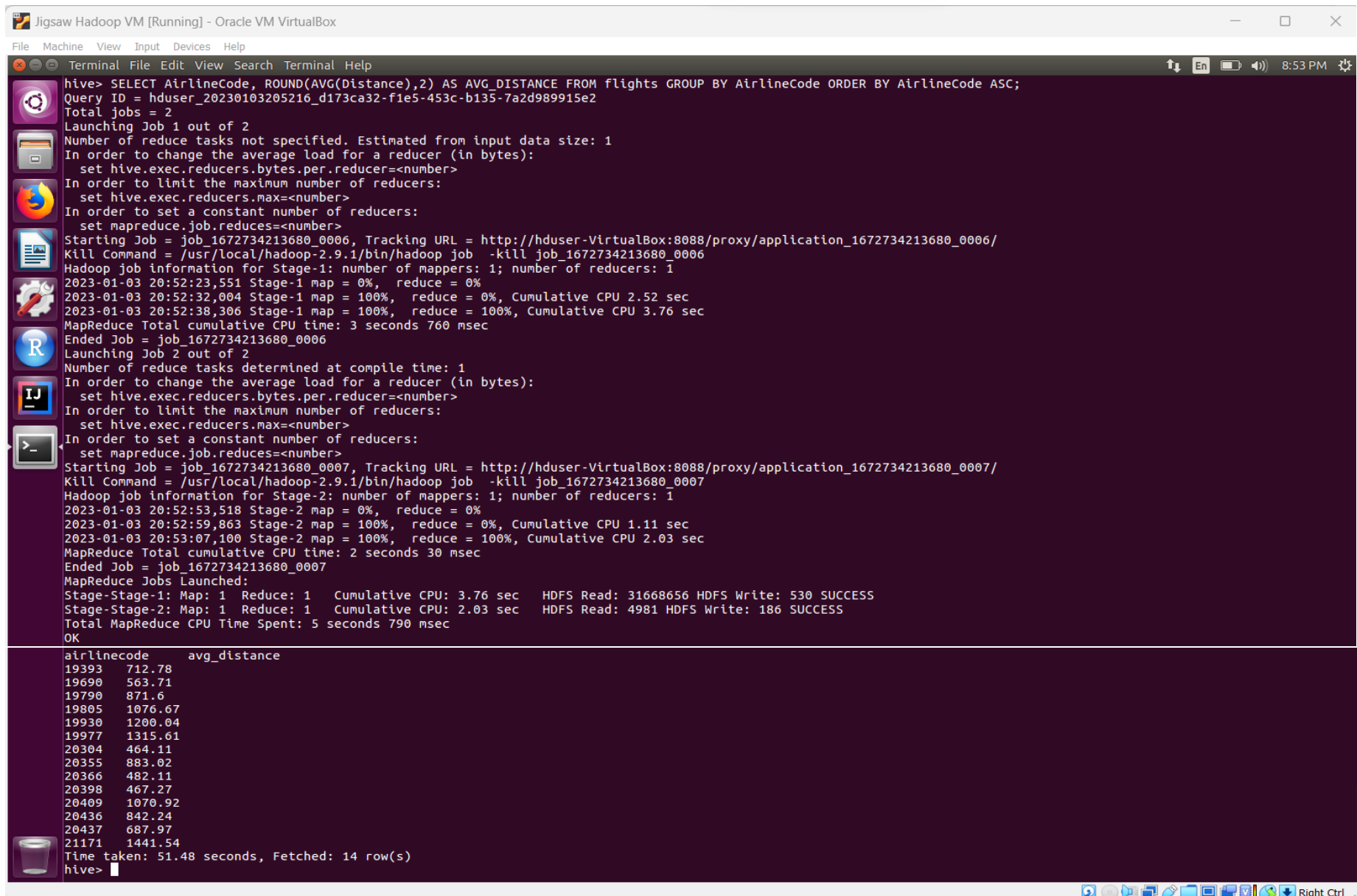
```
hduser@hduser-VirtualBox: ~
hive> SELECT COUNT(DepartureDelay) as Departure_Delay_Flights_Count FROM flights where DepartureDelay>0;
Query ID = hduser_20230103202709_501655ad-d46c-45a2-8c70-310863713593
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0003, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0003/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-03 20:27:17,141 Stage-1 map = 0%, reduce = 0%
2023-01-03 20:27:24,456 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.85 sec
2023-01-03 20:27:30,728 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.88 sec
MapReduce Total cumulative CPU time: 2 seconds 880 msec
Ended Job = job_1672734213680_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.88 sec HDFS Read: 31668122 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 880 msec
OK
departure_delay_flights_count
179015
Time taken: 21.958 seconds, Fetched: 1 row(s)
hive> S
```

departure_delay_flights_count

179015

find the average distance travelled by a flight

SELECT AirlineCode, AVG(Distance) FROM flights GROUP BY AirlineCode ORDER BY ASC;



The screenshot shows a terminal window titled "Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox". The terminal displays the execution of a Hive query: `SELECT AirlineCode, ROUND(AVG(Distance),2) AS AVG_DISTANCE FROM flights GROUP BY AirlineCode ORDER BY AirlineCode ASC;`. The output shows the query ID, total jobs (2), and the execution of two jobs. Job 1 completes successfully. Job 2 is also shown, including its progress and completion. The final output is a table with two columns: `airlinecode` and `avg_distance`. The table contains 14 rows of data, showing the average distance for each airline code. The terminal also displays the time taken (51.48 seconds) and the number of rows fetched (14).

```
hive> SELECT AirlineCode, ROUND(AVG(Distance),2) AS AVG_DISTANCE FROM flights GROUP BY AirlineCode ORDER BY AirlineCode ASC;
Query ID = hduser_20230103205216_d173ca32-f1e5-453c-b135-7a2d989915e2
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0006, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0006/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-03 20:52:23,551 Stage-1 map = 0%, reduce = 0%
2023-01-03 20:52:32,004 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.52 sec
2023-01-03 20:52:38,306 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.76 sec
MapReduce Total cumulative CPU time: 3 seconds 760 msec
Ended Job = job_1672734213680_0006
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0007, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0007/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0007
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-01-03 20:52:53,518 Stage-2 map = 0%, reduce = 0%
2023-01-03 20:52:59,863 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.11 sec
2023-01-03 20:53:07,100 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.03 sec
MapReduce Total cumulative CPU time: 2 seconds 30 msec
Ended Job = job_1672734213680_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.76 sec HDFS Read: 31668656 HDFS Write: 530 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.03 sec HDFS Read: 4981 HDFS Write: 186 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 790 msec
OK
airlinecode      avg_distance
19393            712.78
19690            563.71
19790            871.6
19805            1076.67
19930            1200.04
19977            1315.61
20304            464.11
20355            883.02
20366            482.11
20398            467.27
20409            1070.92
20436            842.24
20437            687.97
21171            1441.54
Time taken: 51.48 seconds, Fetched: 14 row(s)
hive>
```

airlinecode	avg_distance
-------------	--------------

19393	712.78
-------	--------

19690	563.71
-------	--------

19790	871.6
-------	-------

19805	1076.67
-------	---------

19930	1200.04
-------	---------

19977	1315.61
-------	---------

20304	464.11
-------	--------

20355	883.02
-------	--------

20366	482.11
-------	--------

20398	467.27
-------	--------

20409	1070.92
-------	---------

20436	842.24
-------	--------

20437	687.97
-------	--------

21171	1441.54
-------	---------

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

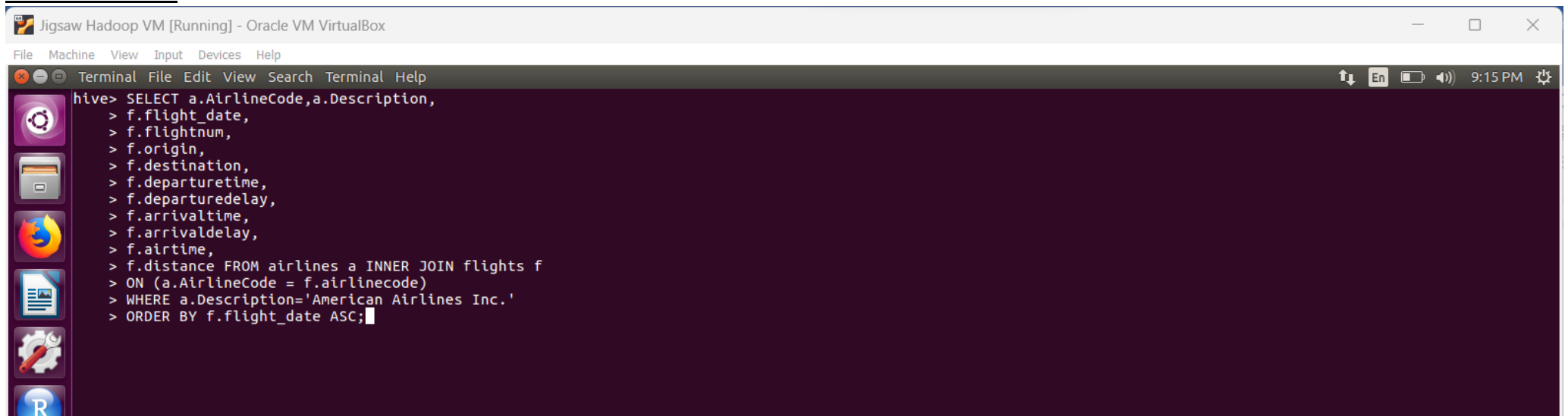
SELECT a.AirlineCode,a.Description,

f.flight_date,

f.flightnum,

f.origin,
f.destination,
f.departuretime,
f.departuredelay,
f.arrivaltime,
f.arrivaldelay,
f.airtime,
f.distance FROM airlines a INNER JOIN flights f
ON (a.AirlineCode = f.airlinecode)
WHERE a.Description='American Airlines Inc.'
ORDER BY f.flight_date ASC;

RUNNING QUERY

A screenshot of a terminal window titled "Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox". The terminal shows a Hive SQL query being executed. The query selects various flight attributes from the 'airlines' and 'flights' tables, filtering for 'American Airlines Inc.' and ordering by flight date. The terminal interface includes a menu bar (File, Machine, View, Input, Devices, Help) and a status bar at the bottom showing system icons and the time 9:15 PM. On the left side of the terminal, there is a vertical toolbar with icons for various applications like a file manager, web browser, and terminal.

```
hive> SELECT a.AirlineCode,a.Description,  
> f.flight_date,  
> f.flightnum,  
> f.origin,  
> f.destination,  
> f.departuretime,  
> f.departuredelay,  
> f.arrivaltime,  
> f.arrivaldelay,  
> f.airtime,  
> f.distance FROM airlines a INNER JOIN flights f  
> ON (a.AirlineCode = f.airlinecode)  
> WHERE a.Description='American Airlines Inc.'  
> ORDER BY f.flight_date ASC;
```

Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Terminal File Edit View Search Terminal Help

```
hive> SELECT a.AirlineCode,a.Description,
> f.flight_date,
> f.flightnum,
> f.origin,
> f.destination,
> f.departuretime,
> f.departuredelay,
> f.arrivaltime,
> f.arrivaldelay,
> f.airtime,
> f.distance FROM airlines a INNER JOIN flights f
> ON (a.AirlineCode = f.airlinecode)
> WHERE a.Description='American Airlines Inc.'
> ORDER BY f.flight_date ASC;

Query ID = hduser_20230103211608_b764e085-f83e-4239-b69e-770da722d5fe
Total jobs = 2
Stage-1 is selected by condition resolver.
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0012, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0012/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0012
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2023-01-03 21:16:18,258 Stage-1 map = 0%, reduce = 0%
2023-01-03 21:16:31,853 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 2.51 sec
2023-01-03 21:16:34,987 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 7.52 sec
2023-01-03 21:16:42,294 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.36 sec
MapReduce Total cumulative CPU time: 11 seconds 360 msec
Ended Job = job_1672734213680_0012
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1672734213680_0013, Tracking URL = http://hduser-VirtualBox:8088/proxy/application_1672734213680_0013/
Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1672734213680_0013
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-01-03 21:16:59,486 Stage-2 map = 0%, reduce = 0%
2023-01-03 21:17:06,822 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.13 sec
```


Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Terminal File Edit View Search Terminal Help

```
19805 American Airlines Inc. 2014-04-30 1110 DFW LGA 836 -4 1242 -17 168 1389.0
19805 American Airlines Inc. 2014-04-30 1109 JAX DFW 1453 -7 1631 -19 137 918.0
19805 American Airlines Inc. 2014-04-30 1109 DFW JAX 1059 -6 1413 -7 114 918.0
19805 American Airlines Inc. 2014-04-30 1108 DFW LGA 741 -4 1143 -22 161 1389.0
19805 American Airlines Inc. 2014-04-30 1107 LGA DFW 635 -10 934 -1 216 1389.0
19805 American Airlines Inc. 2014-04-30 1107 DFW LGA 1252 57 1712 62 165 1389.0
19805 American Airlines Inc. 2014-04-30 1106 DFW BWI 1426 -4 1836 6 156 1217.0
19805 American Airlines Inc. 2014-04-30 1106 BWI DFW 1920 10 2212 37 192 1217.0
19805 American Airlines Inc. 2014-04-30 1105 DFW SFO 2135 15 2310 0 198 1464.0
19805 American Airlines Inc. 2014-04-30 1104 DFW LGA 700 -5 1054 -21 161 1389.0
19805 American Airlines Inc. 2014-04-30 1103 LGA DFW 555 -5 855 0 223 1389.0
19805 American Airlines Inc. 2014-04-30 1103 DFW LGA 1219 84 1623 73 169 1389.0
19805 American Airlines Inc. 2014-04-30 1102 AUS DFW 538 -7 637 -8 38 190.0
19805 American Airlines Inc. 2014-04-30 1101 PHX DFW 1150 10 1551 -9 102 868.0
19805 American Airlines Inc. 2014-04-30 1101 DFW PHX 1020 -5 1100 0 135 868.0
19805 American Airlines Inc. 2014-04-30 1100 DFW LGA 611 -4 1010 -20 157 1389.0
19805 American Airlines Inc. 2014-04-30 1099 ORF DFW 740 0 952 -18 177 1212.0
19805 American Airlines Inc. 2014-04-30 1098 MSP MIA 549 -1 1040 10 209 1501.0
19805 American Airlines Inc. 2014-04-30 1095 PHX DFW 1035 0 1446 1 106 868.0
19805 American Airlines Inc. 2014-04-30 1061 MIA ATL 1123 -2 1317 -8 83 594.0
19805 American Airlines Inc. 2014-04-30 1061 ATL MIA 1403 -7 1556 -19 86 594.0
19805 American Airlines Inc. 2014-04-30 1060 LAX ORD 538 -7 1142 -8 225 1744.0
19805 American Airlines Inc. 2014-04-30 1059 MIA MSY 2146 -4 2302 7 102 675.0
19805 American Airlines Inc. 2014-04-30 1057 MIA SEA 1726 1 2106 -4 351 2724.0
19805 American Airlines Inc. 2014-04-30 1055 TUS DFW 1138 -2 1530 -25 94 813.0
19805 American Airlines Inc. 2014-04-30 1055 DFW TUS 1032 -3 1059 -1 122 813.0
19805 American Airlines Inc. 2014-04-30 1054 DFW BOS 1244 9 1707 -3 180 1562.0
19805 American Airlines Inc. 2014-04-30 1054 BOS DFW 1750 -5 2110 0 240 1562.0
19805 American Airlines Inc. 2014-04-30 1053 DFW ELP 2303 43 2345 45 86 551.0
19805 American Airlines Inc. 2014-04-30 1052 SFO DFW 17 -3 520 -15 168 1464.0
19805 American Airlines Inc. 2014-04-30 1051 MIA ORD 836 -4 1050 0 162 1197.0
19805 American Airlines Inc. 2014-04-30 1357 SJU JFK 1257 -8 1723 -2 213 1598.0
19805 American Airlines Inc. 2014-04-30 1357 JFK SJU 802 2 1202 7 197 1598.0
19805 American Airlines Inc. 2014-04-30 1356 AUS DFW 852 -3 1001 -4 37 190.0
19805 American Airlines Inc. 2014-04-30 1355 IAD DFW 1627 7 1859 14 189 1172.0
19805 American Airlines Inc. 2014-04-30 1355 DFW IAD 1144 -6 1549 9 161 1172.0
19805 American Airlines Inc. 2014-04-30 1354 MIA LGA 910 -5 1151 -14 140 1096.0
19805 American Airlines Inc. 2014-04-30 1352 DFW BOS 1527 -3 1958 -7 179 1562.0
19805 American Airlines Inc. 2014-04-30 1351 BDL LAX 851 -9 1204 -16 335 2527.0
19805 American Airlines Inc. 2014-04-30 1350 STT MIA 836 1 1114 -21 144 1107.0
19805 American Airlines Inc. 2014-04-30 1349 RDU DFW 624 -6 838 -2 169 1061.0
19805 American Airlines Inc. 2014-04-30 1347 PIT DFW 653 -7 912 2 174 1067.0
19805 American Airlines Inc. 2014-04-30 1346 ORD MIA 905 -5 1327 17 171 1197.0
19805 American Airlines Inc. 2014-04-30 1346 MIA ORD 1531 1 1840 45 168 1197.0
19805 American Airlines Inc. 2014-04-30 1345 JFK MIA 727 -3 1027 -13 151 1089.0
19805 American Airlines Inc. 2014-04-30 1344 SAN DFW 1302 17 1750 5 142 1171.0
```

Time taken: 68.616 seconds, Fetched: 43256 row(s)

hive>

Thus, the above query displays the 43,256 records details for 'American Airlines Inc.' from airlines and flights both tables.