# 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

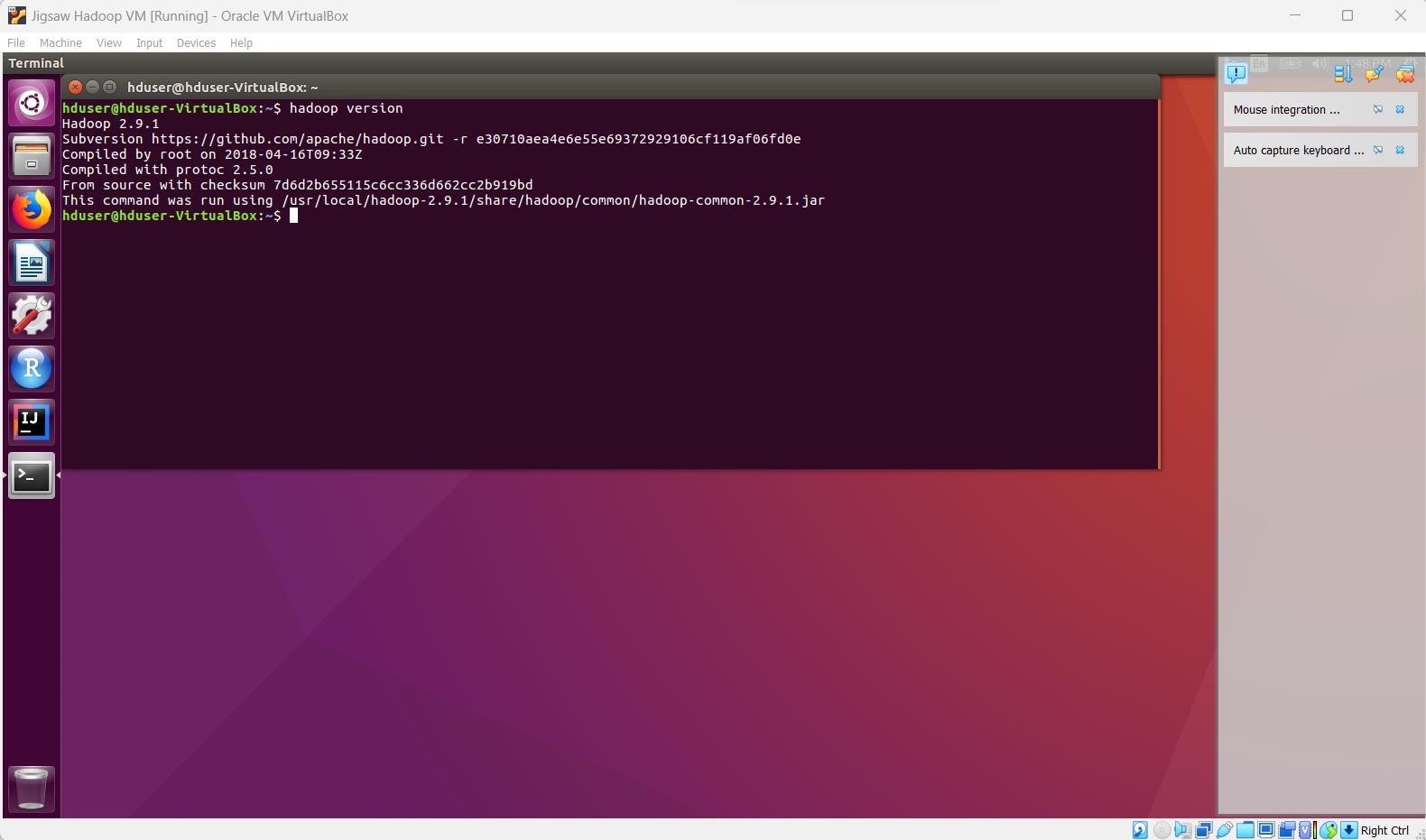
Find count of flights that had arrival delay Find count of flights that had departure delay find the average distance travelled by a flight

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

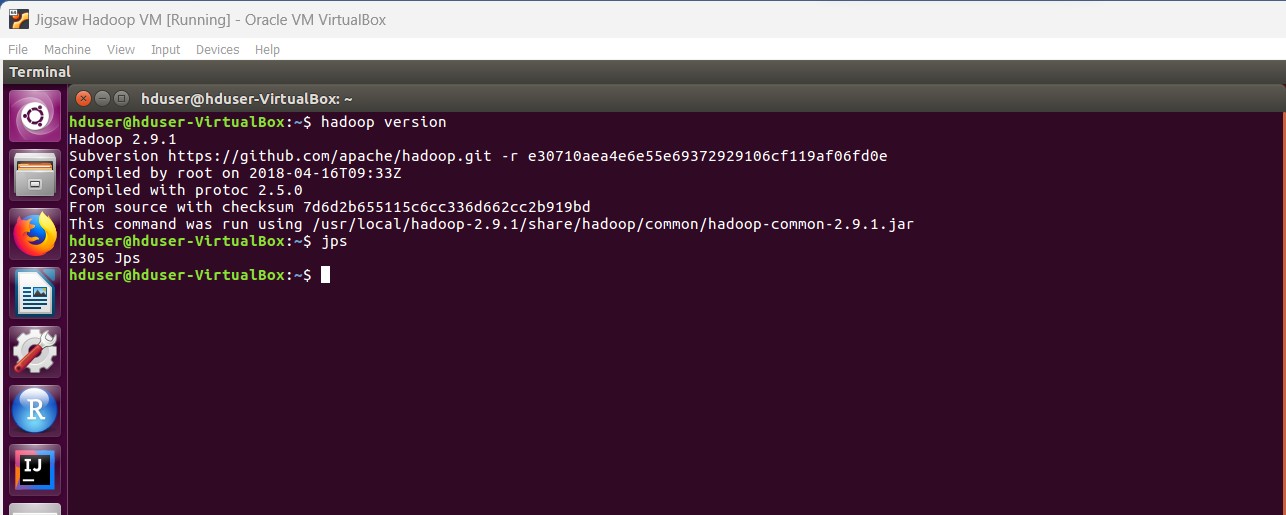
Attach screenshots of the outputs along with the query executed in the solutions file

Solutions/Answer:

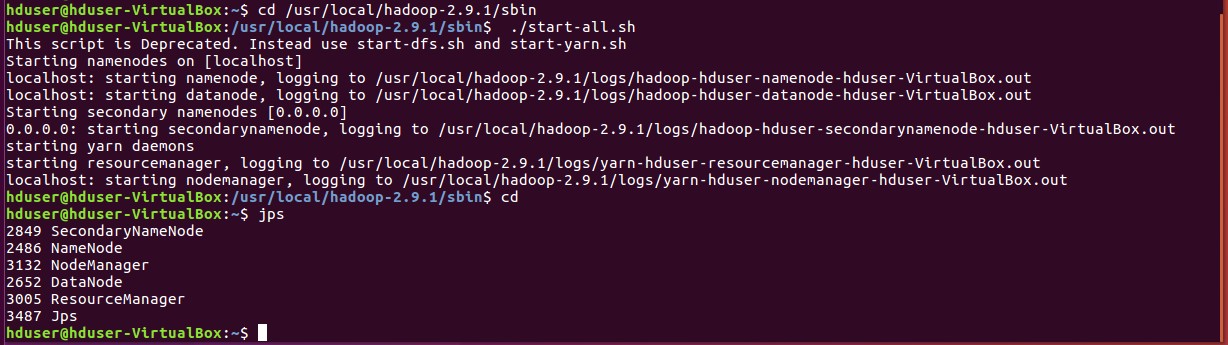
1st) Checking the Hadoop Version



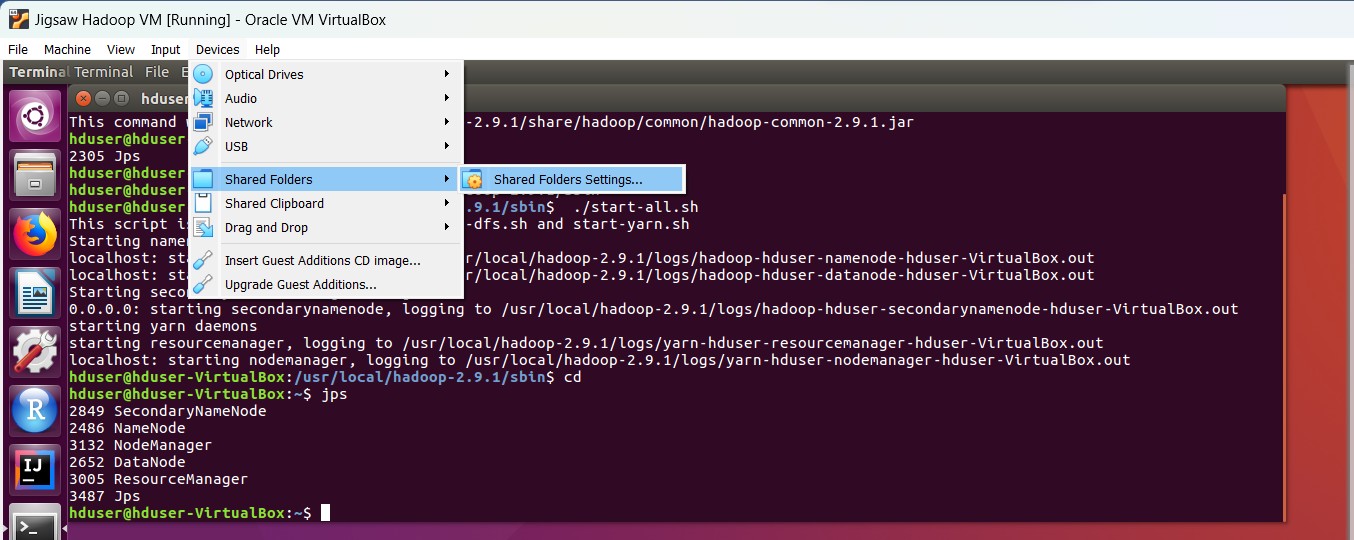
2nd) Checking Jps (Java Process Status) running or not as not

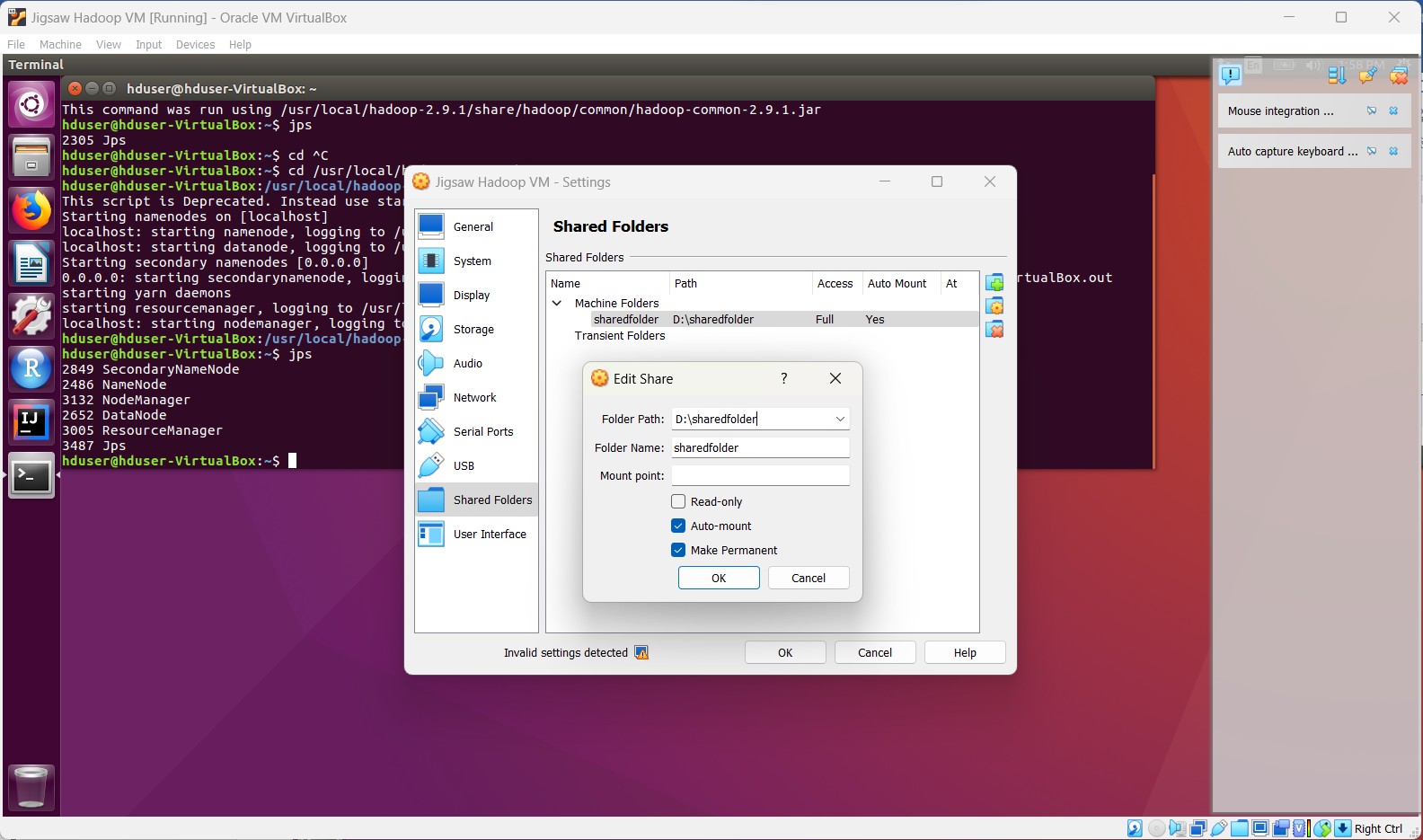


3rd) Starting jps

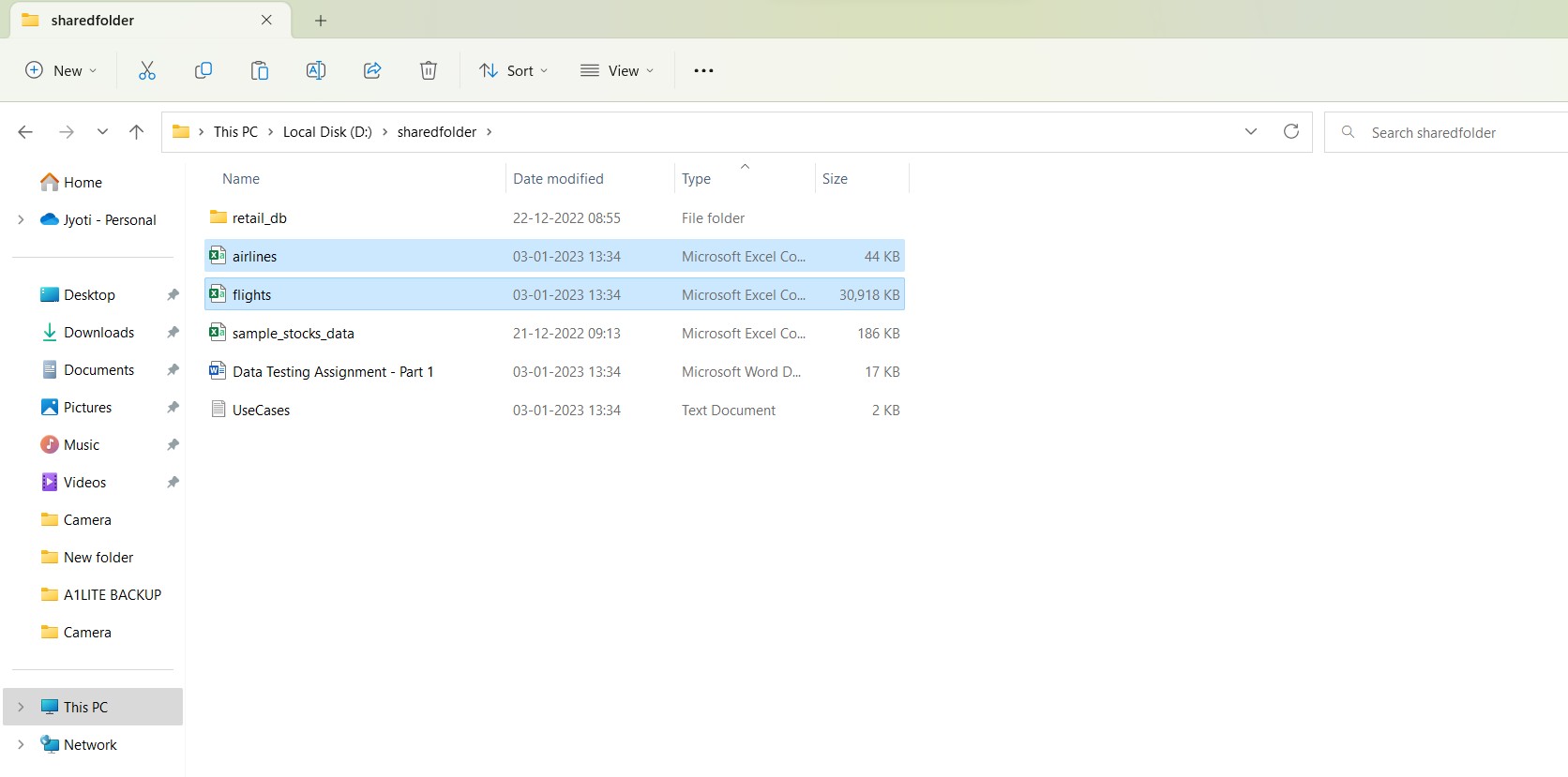


4th) Now Create a folder in the D: drive of the host OS (and copy the path)

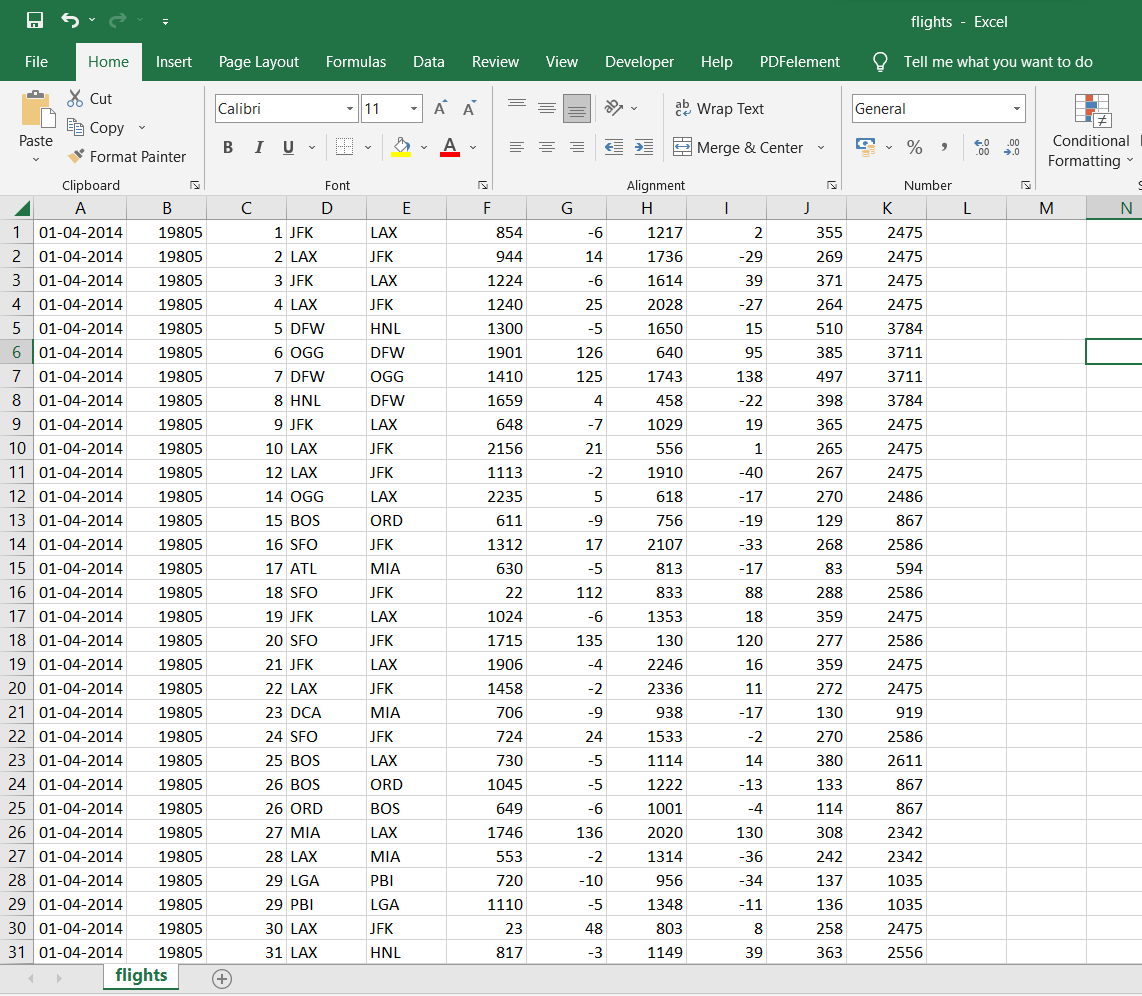




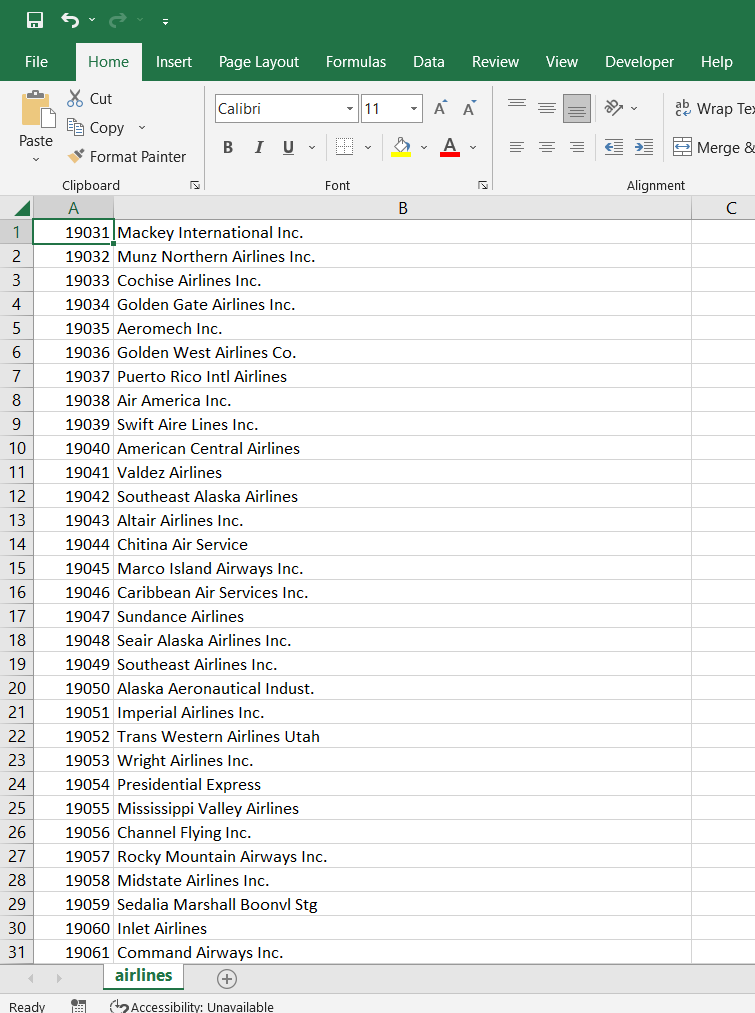
Path: D:\sharedfolder

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

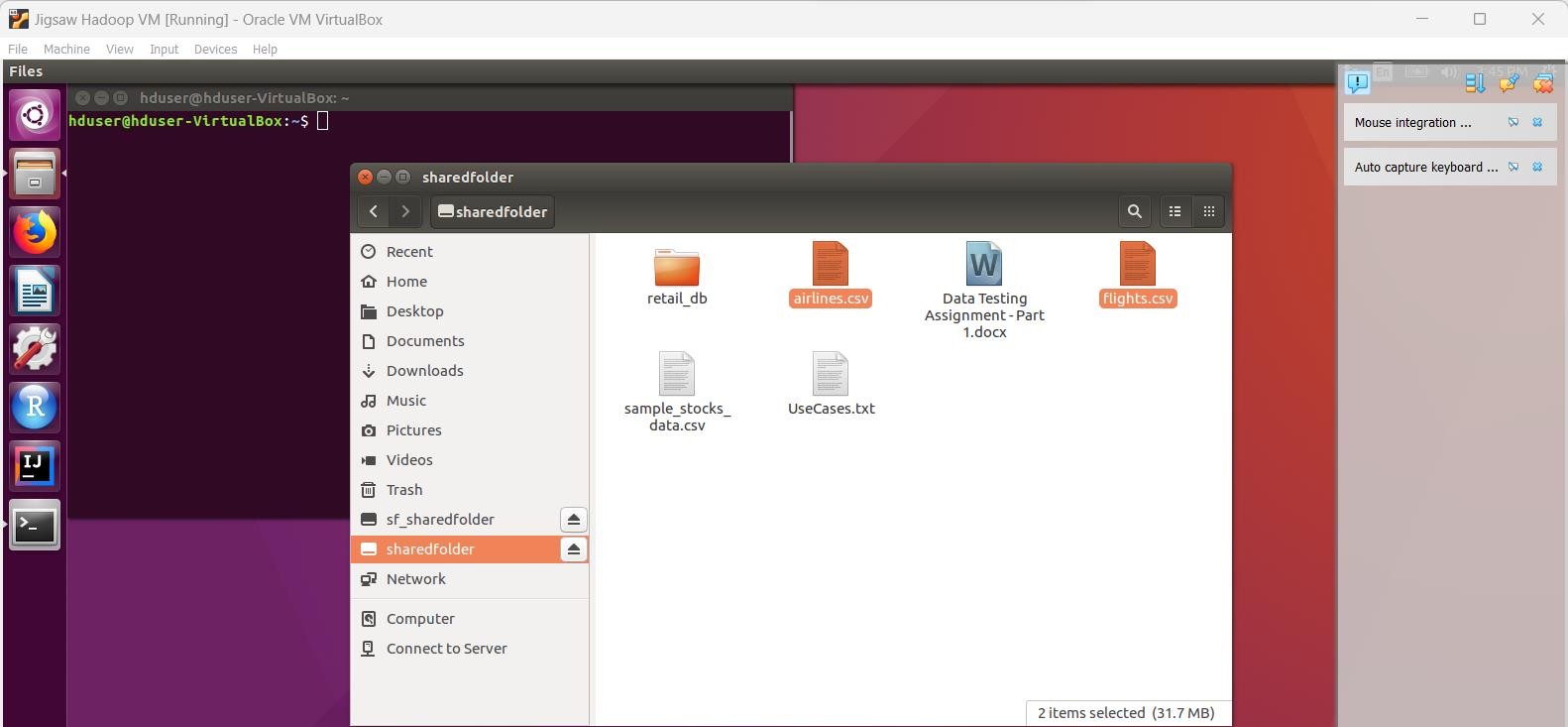
Flights.csv



Airlines.csv



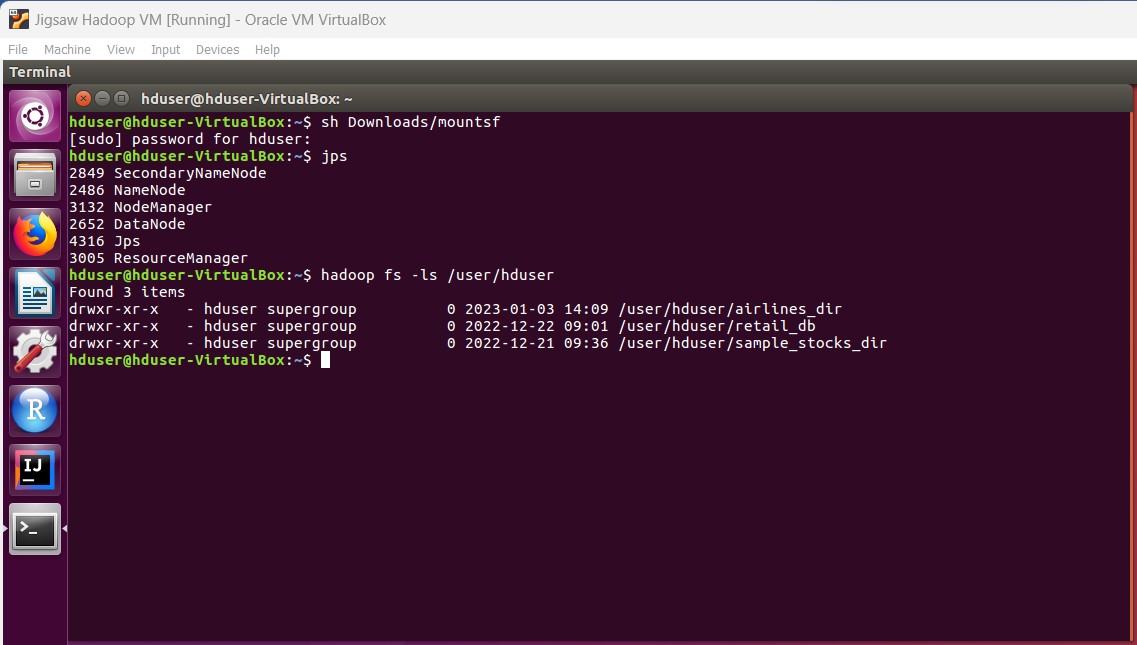
5TH) Now files are available on Ubuntu shared folder



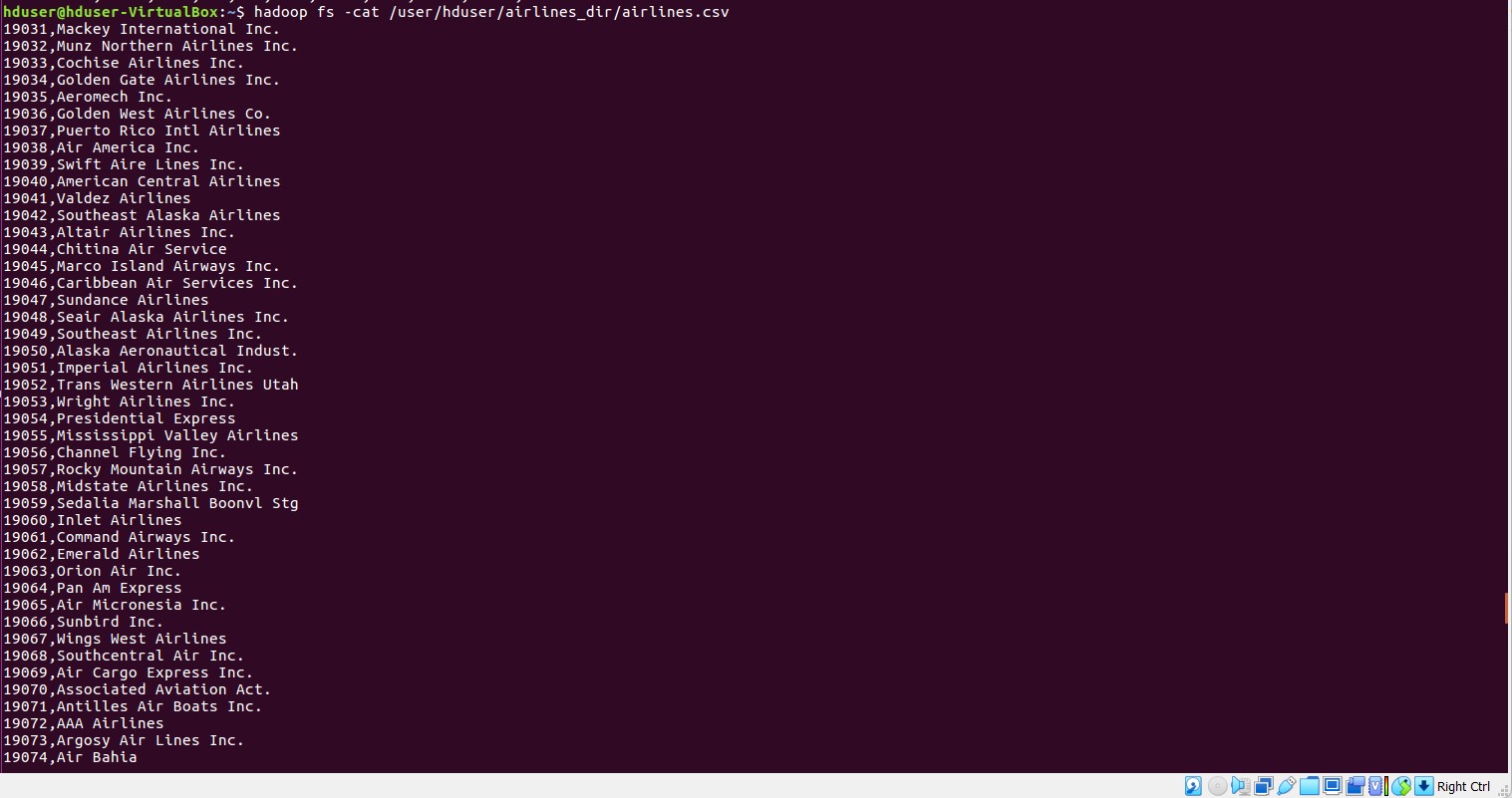
6TH) Now Putting Data into HDFS

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

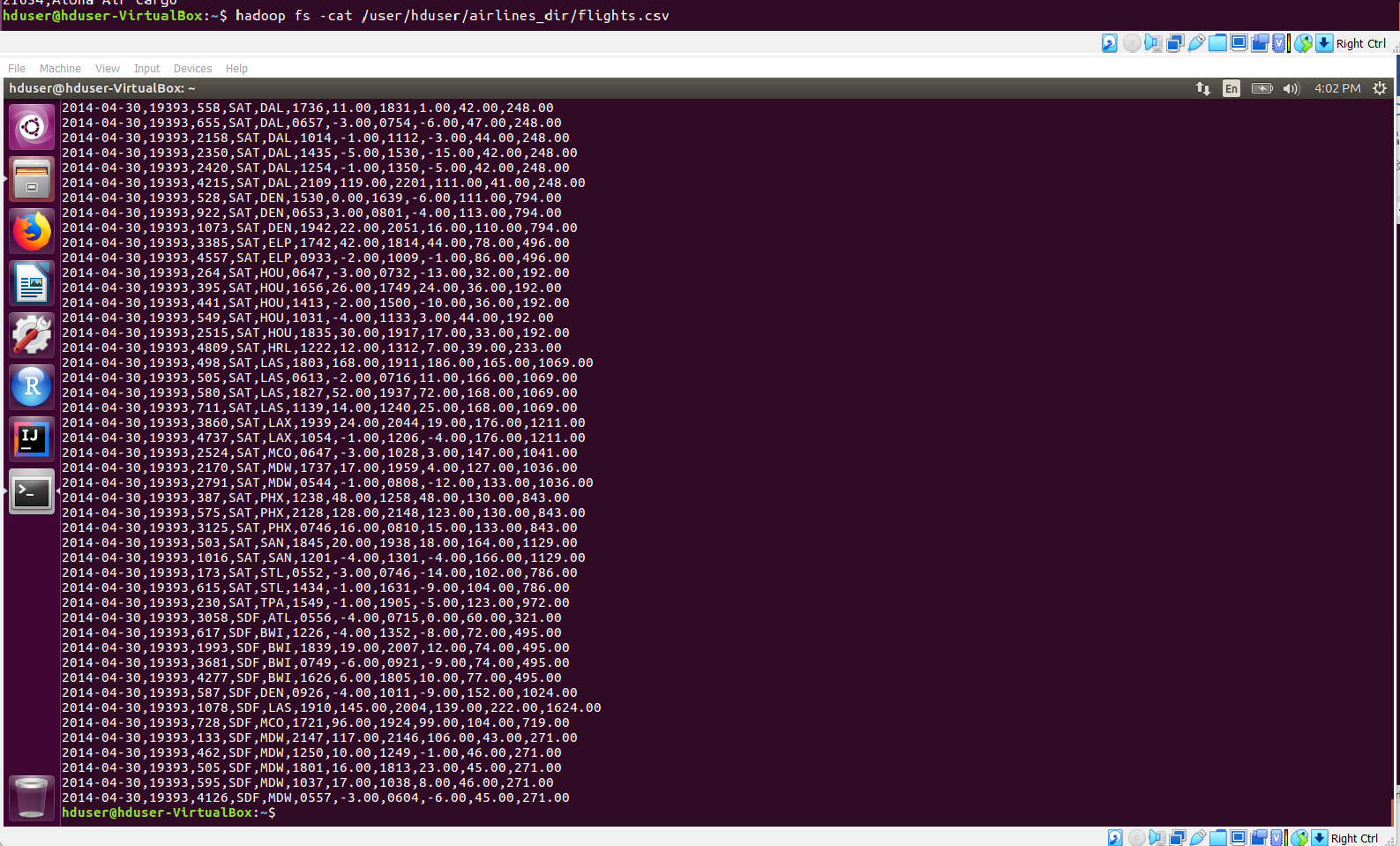
- /user/hduser is the working directory in the VM



Now reading the airlines.csv file

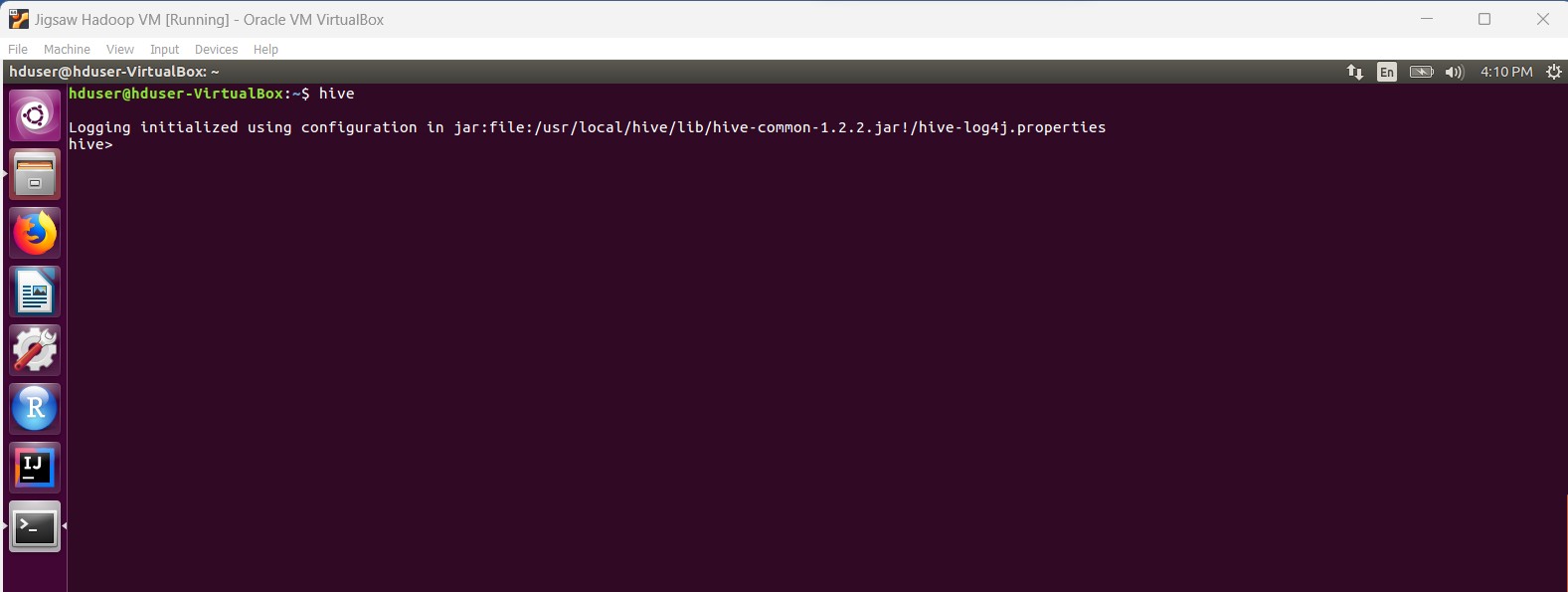


Now reading the flights.csv file

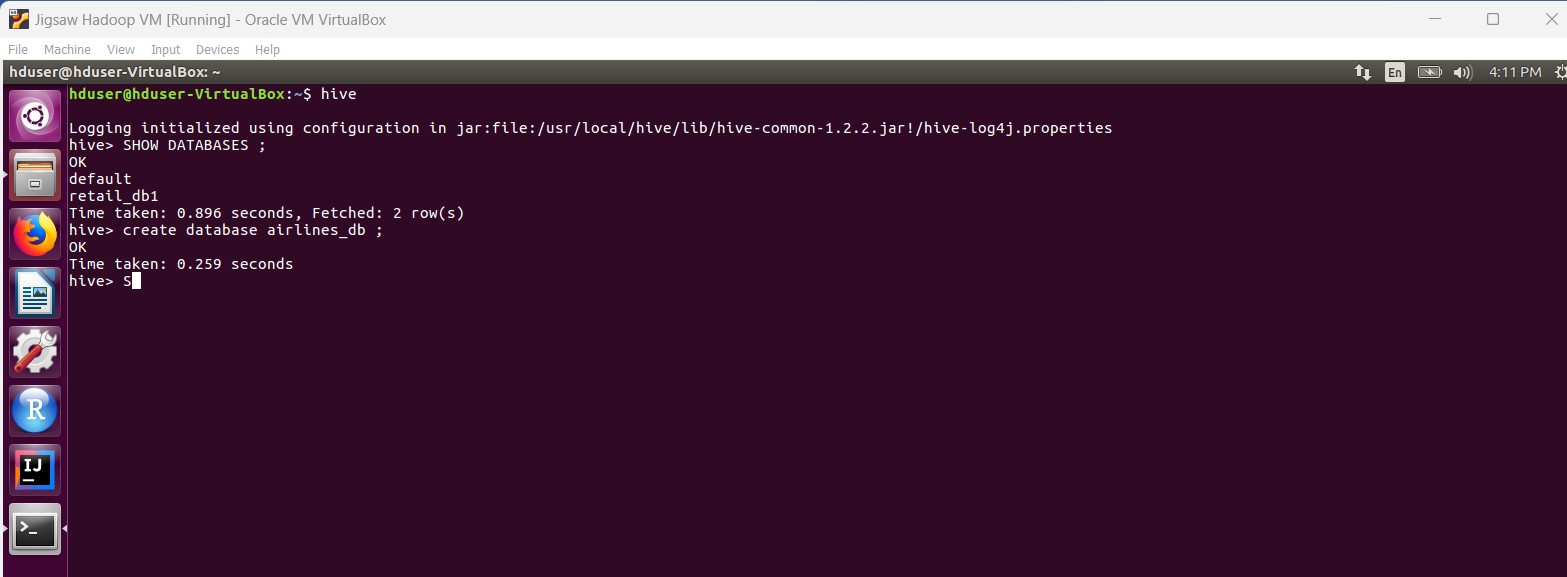


7TH) Use Cases: 1. Create Hive tables for the above datasets by identifying the right datatypes

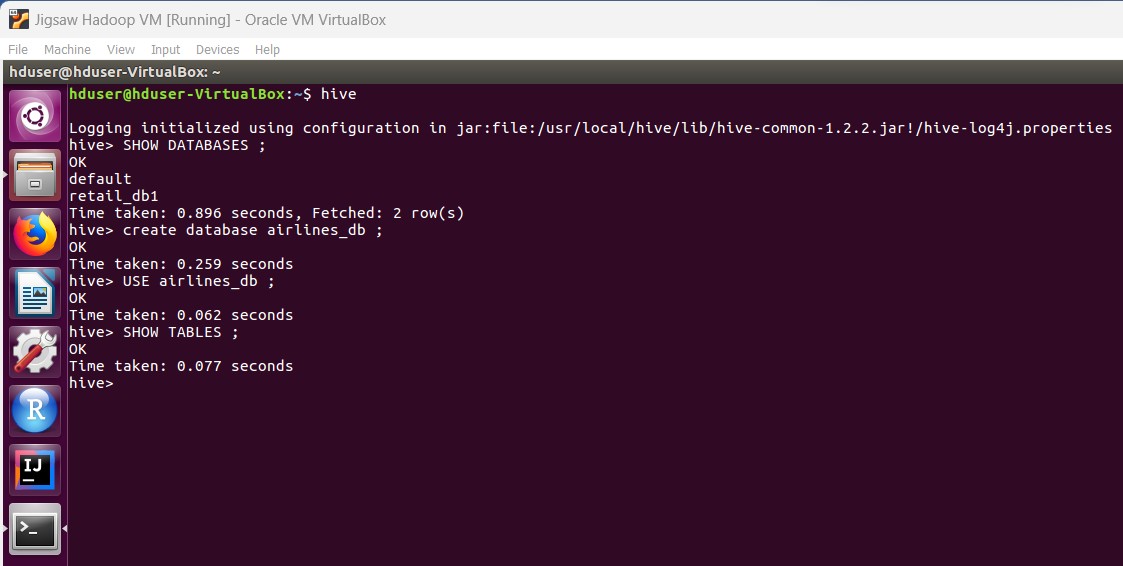
Starting hive



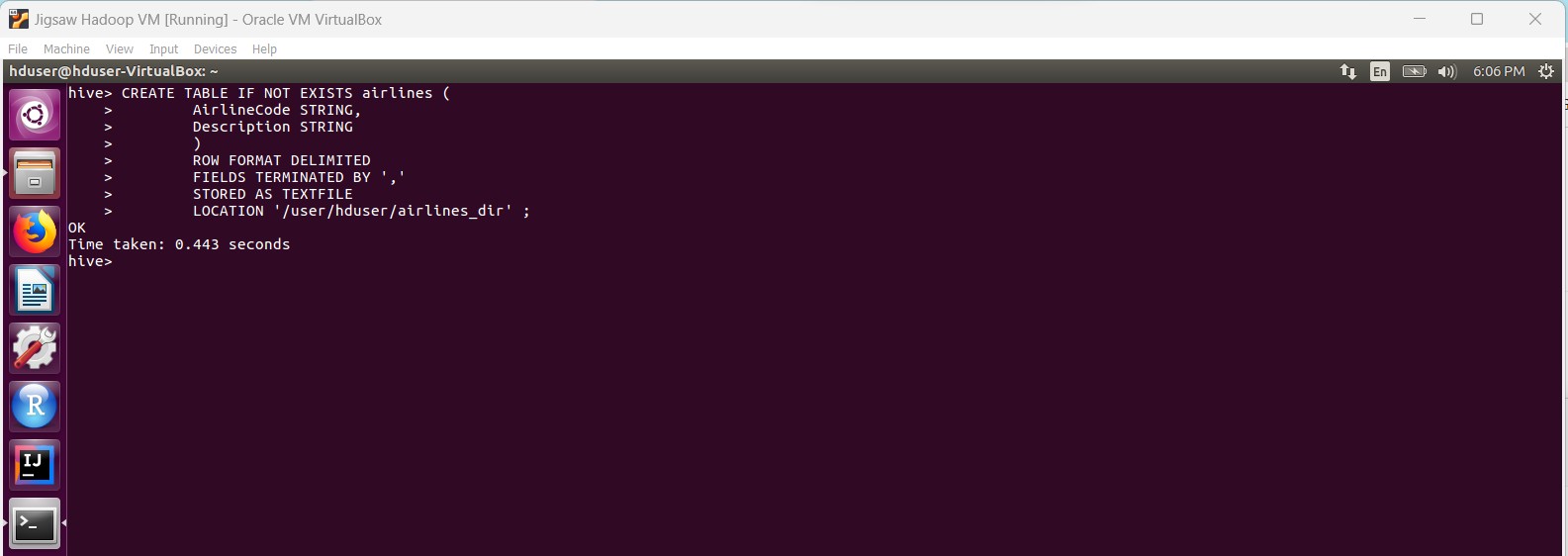
**Creating airlines\_db**

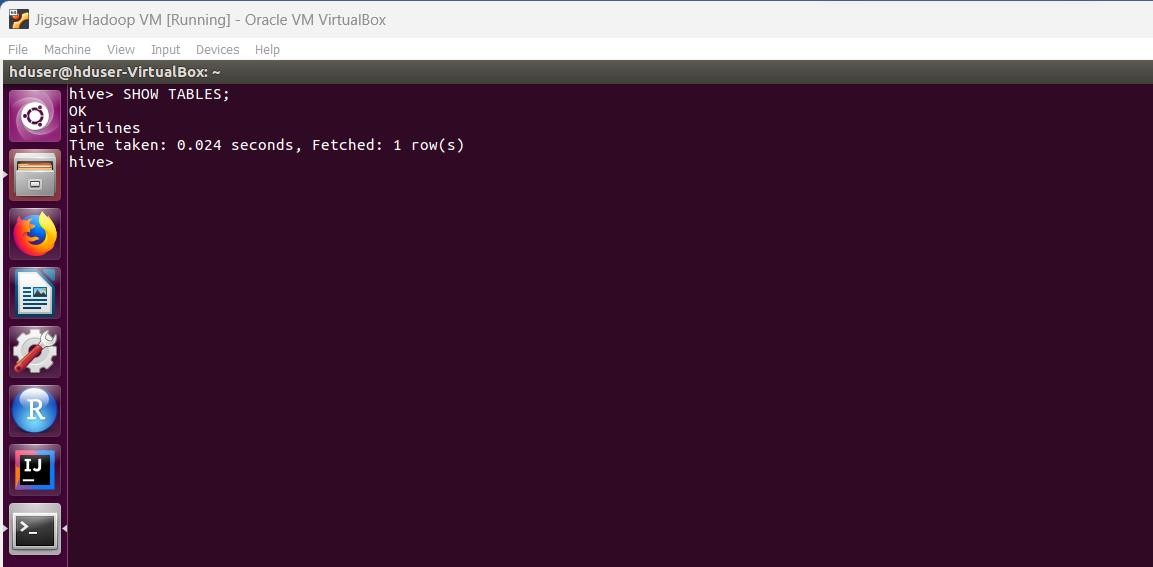


Now checking existing db & Tables

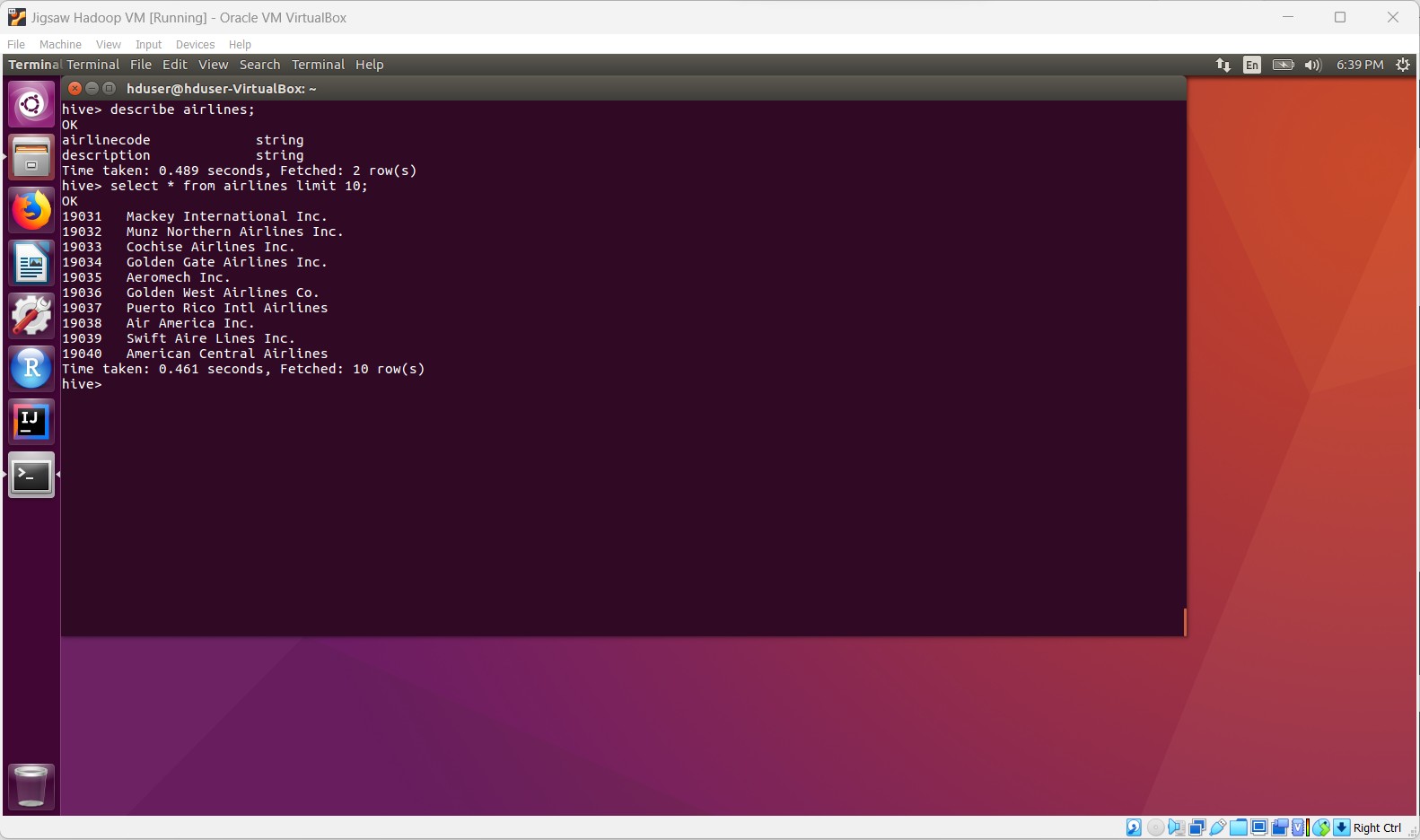


Now creating airlines table in hive

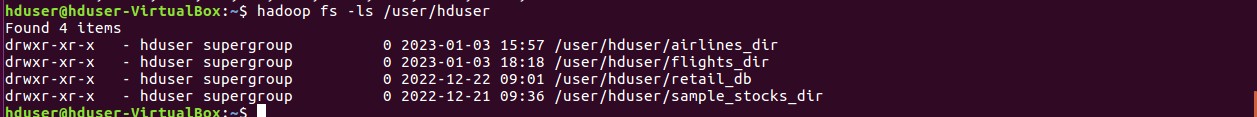




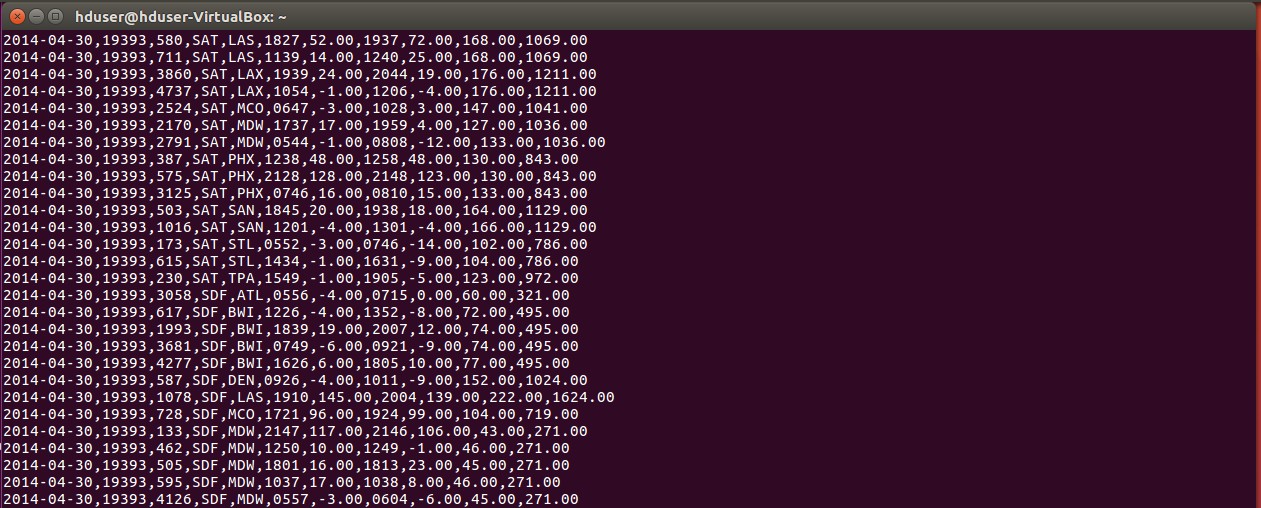
Airlines data has been loaded.



Now Creating another directory for flights



Now loading flights data from .csv file



Used below query to create flights table:

CREATE TABLE IF NOT EXISTS flights (

Flight\_Date DATE, AirlineCode STRING, FlightNum STRING,

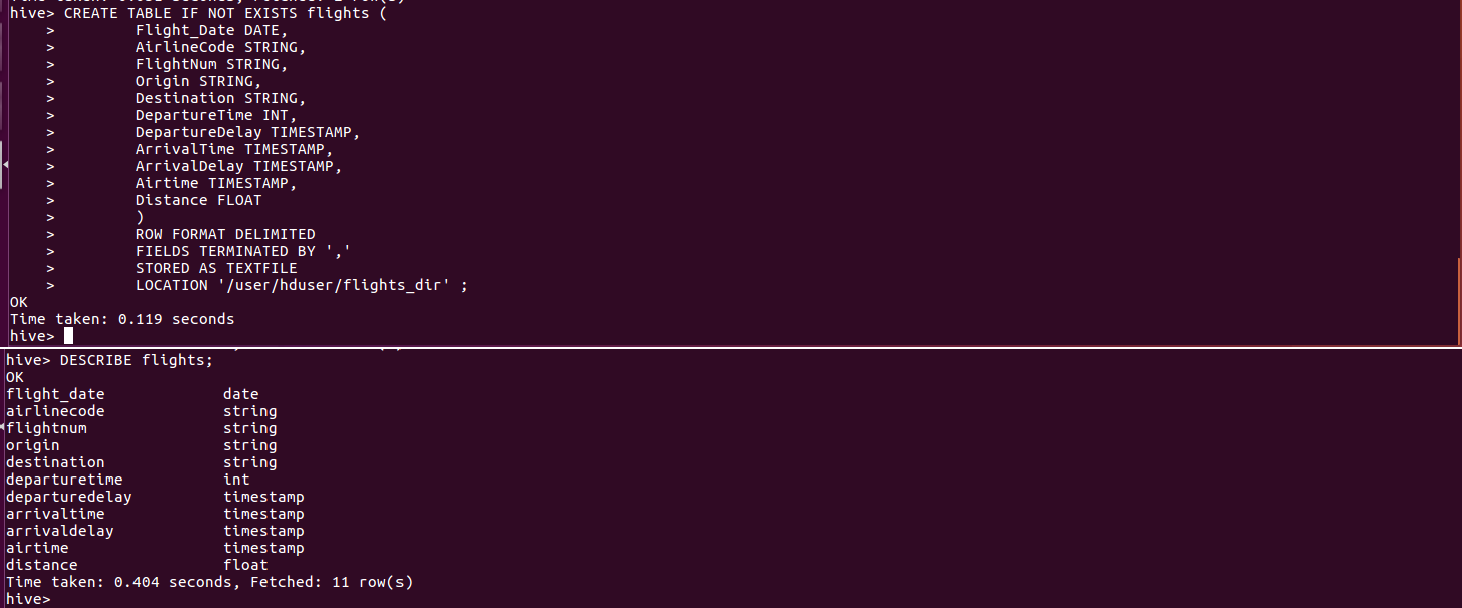
Origin STRING, Destination STRING, DepartureTime INT,

DepartureDelay TIMESTAMP, ArrivalTime TIMESTAMP, ArrivalDelay TIMESTAMP, Airtime TIMESTAMP, Distance FLOAT

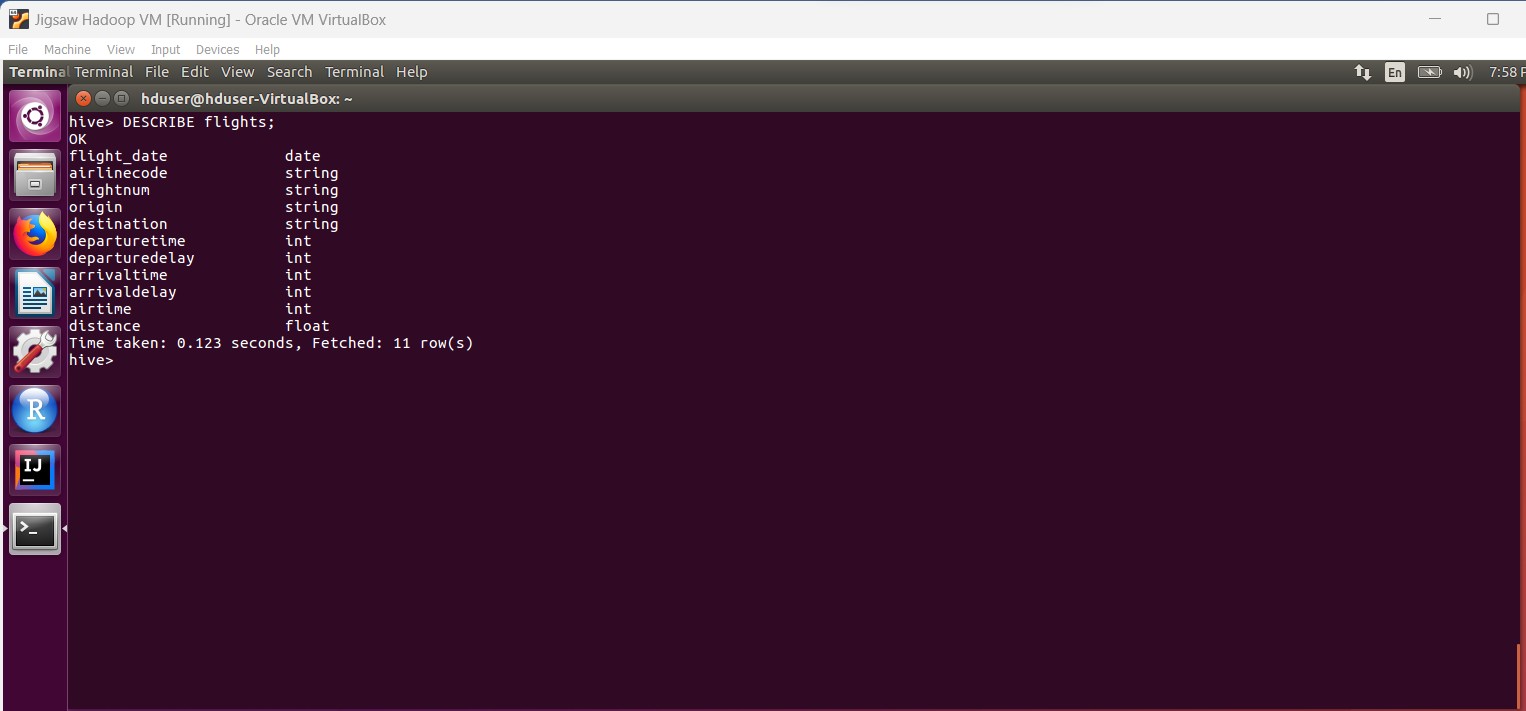
)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE

LOCATION '/user/hduser/flights\_dir' ;

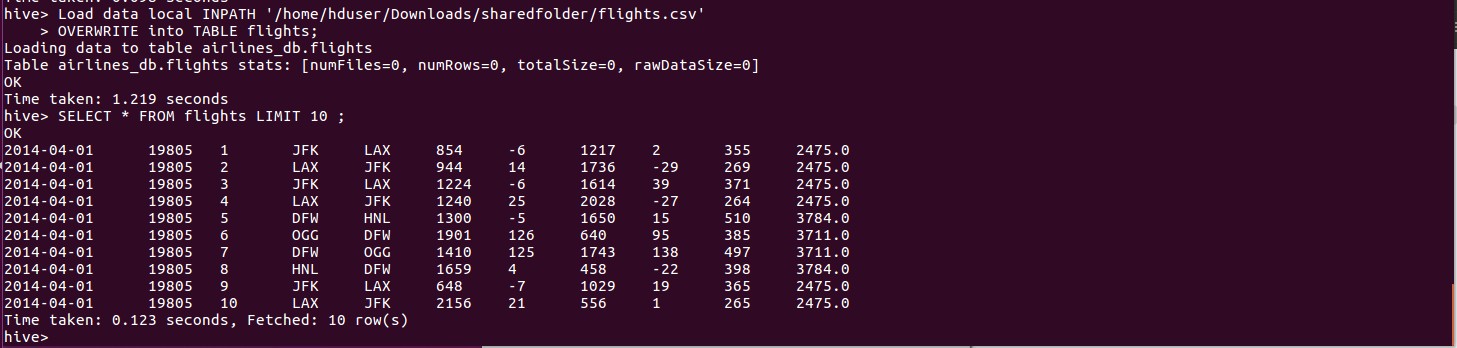


Now Altered table



Now again loading data from

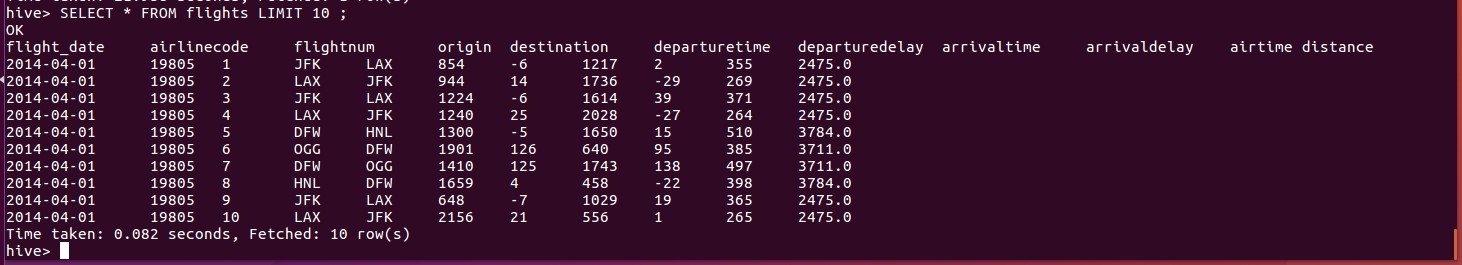
Load data local INPATH '/home/hduser/Downloads/sharedfolder/flights.csv' OVERWRITE into TABLE flights;



# Enabling column names

set hive.cli.print.header=true;

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

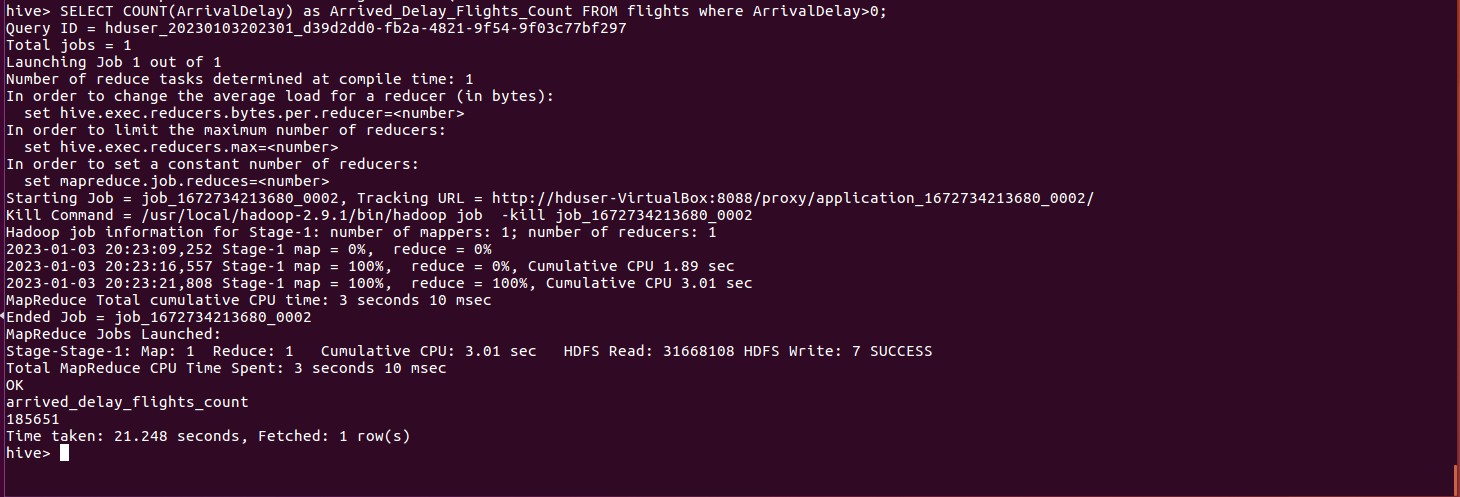


Now,

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;

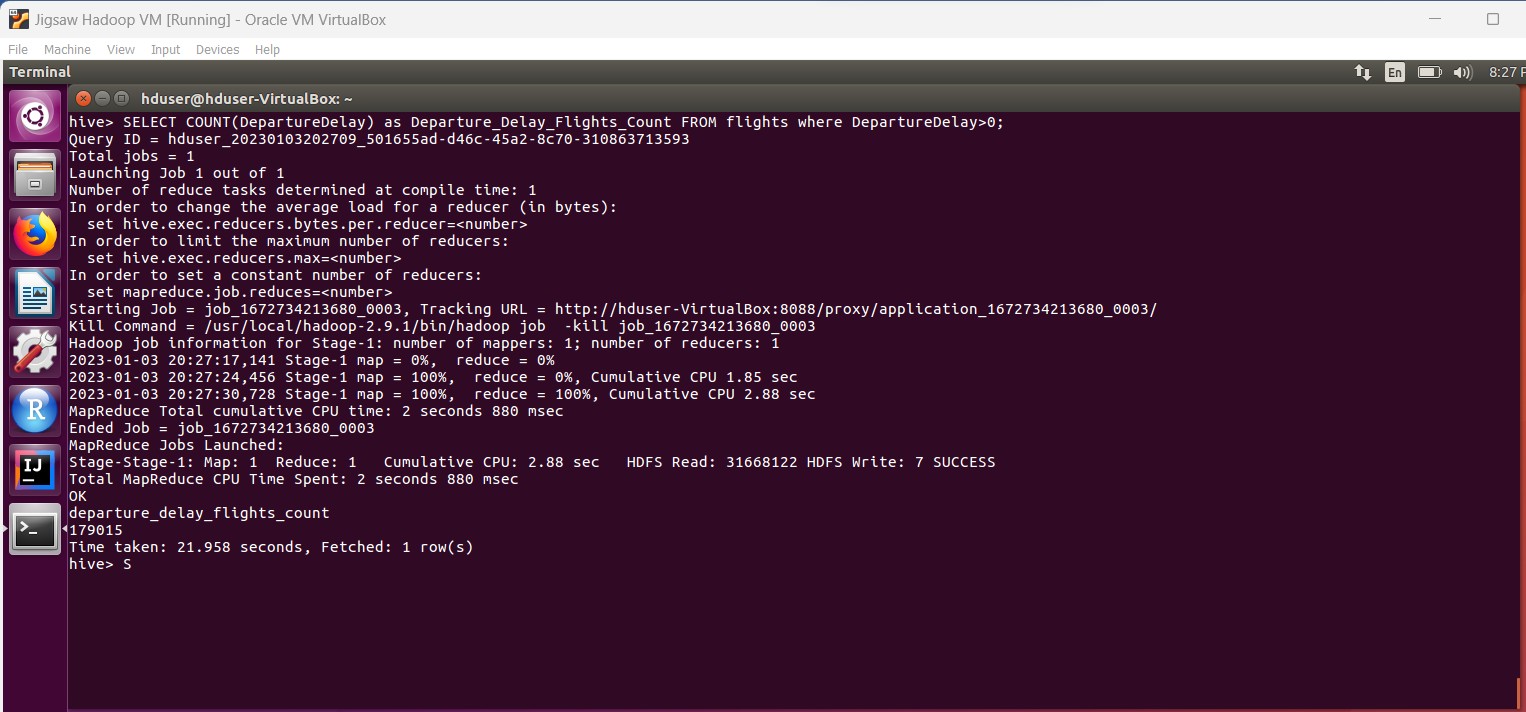


***arrived\_delay\_flights\_count***

# 185651

Find count of flights that had departure delay

SELECT COUNT(DepartureDelay) as Departure\_Delay\_Flights\_Count FROM flights where DepartureDelay>0;

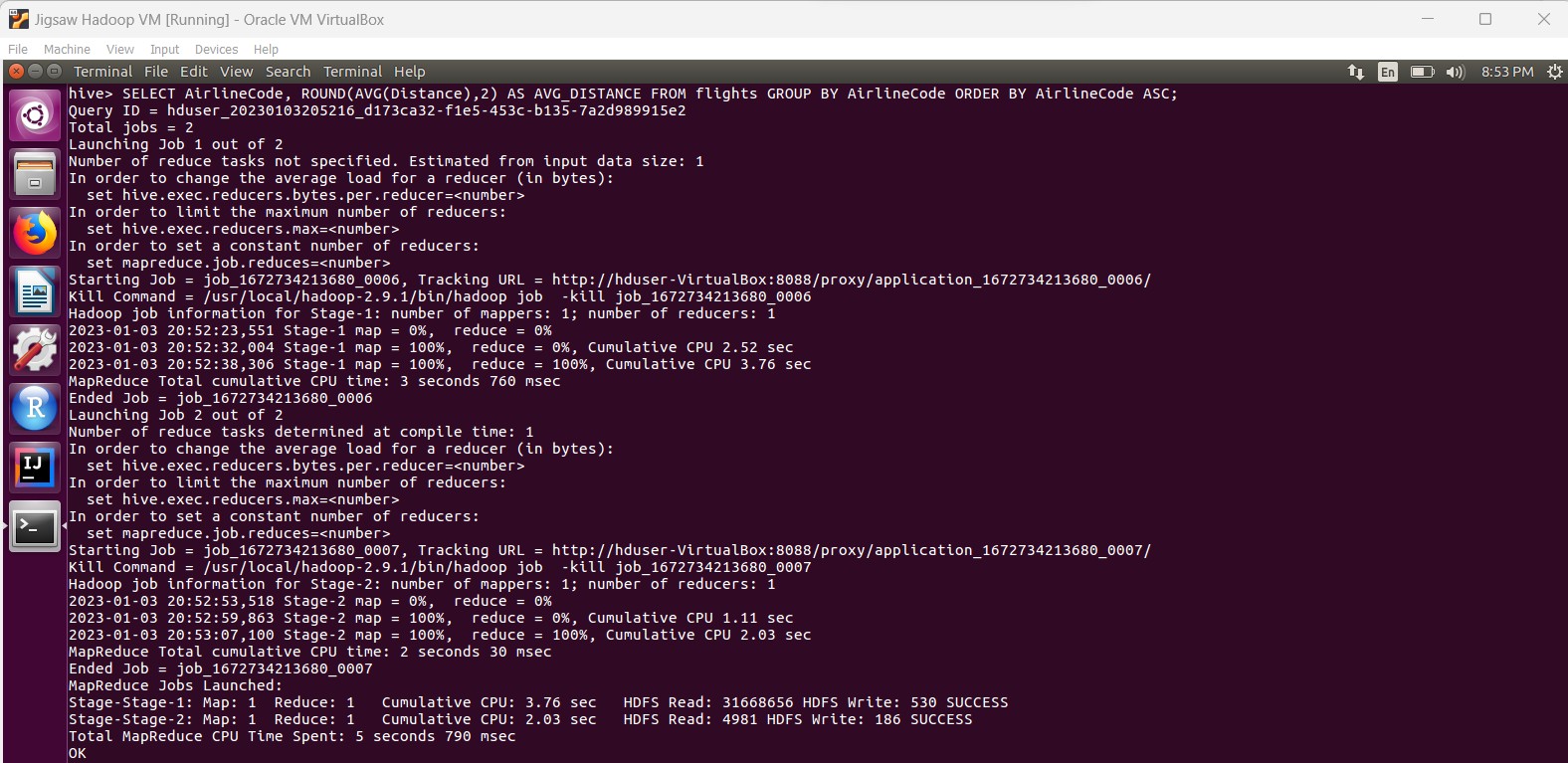


***departure\_delay\_flights\_count***

**179015**

find the average distance travelled by a flight

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



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

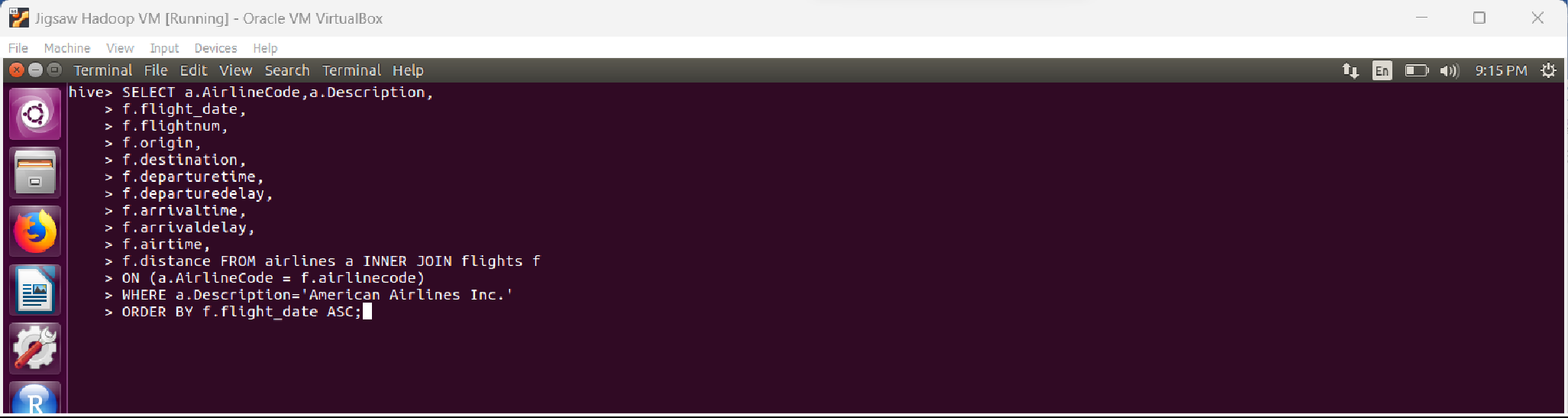
**SELECT a.AirlineCode,a.Description, f.flight\_date,**

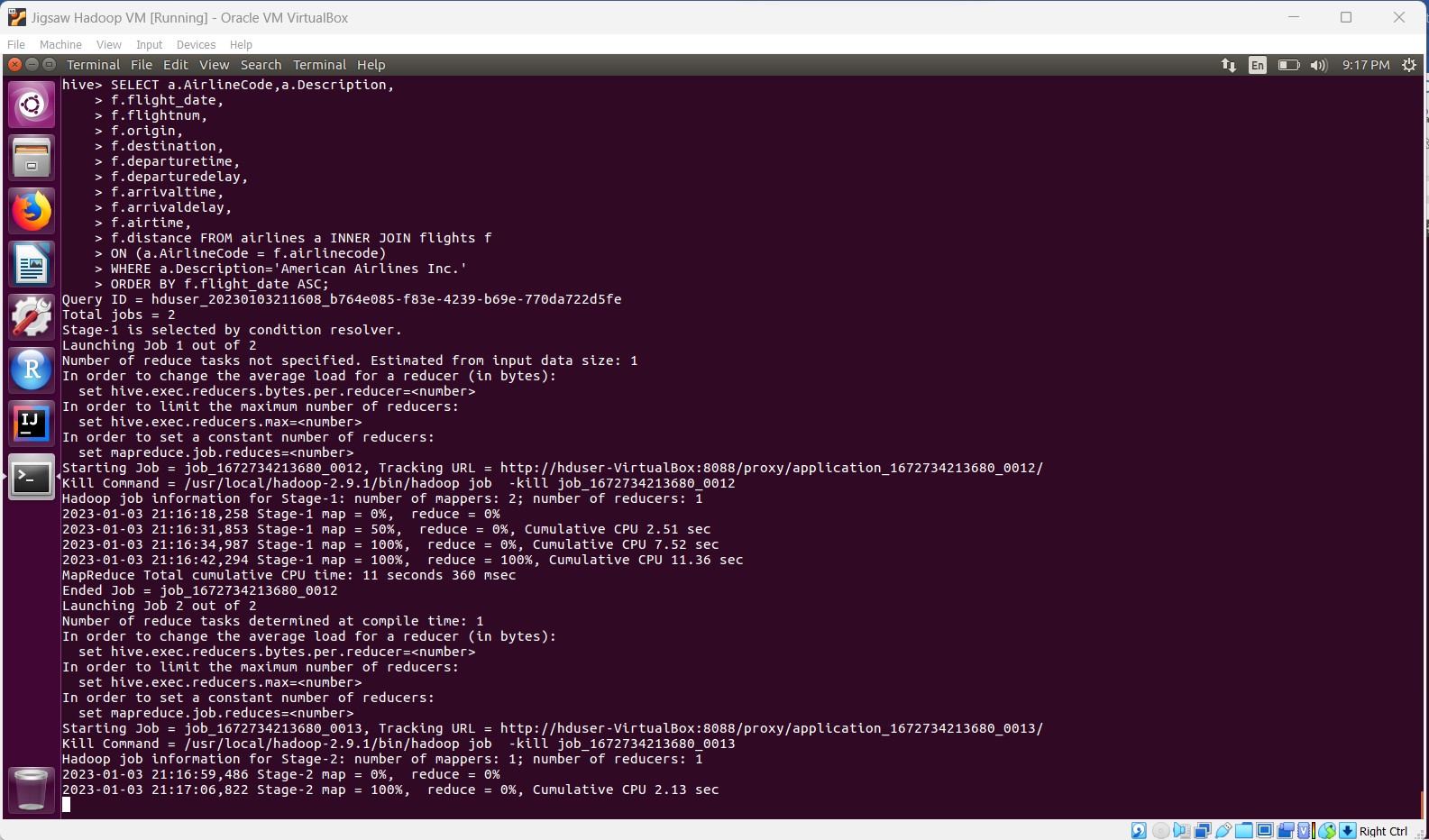
**f.flightnum,**

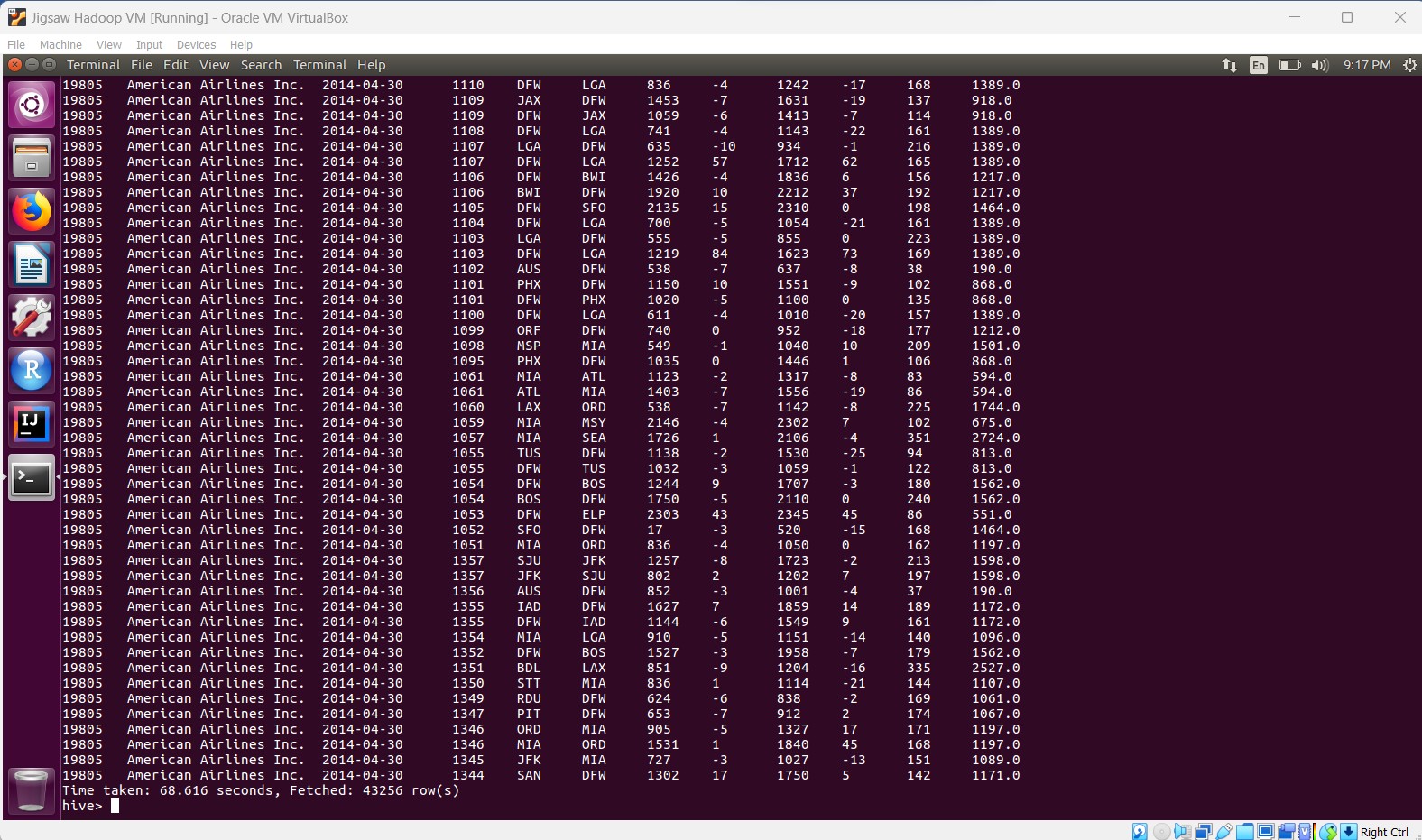
**f.origin, f.destination, f.departuretime, f.departuredelay, f.arrivaltime, f.arrivaldelay, f.airtime,**

**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





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