
Field Description - flights.csv

Date
AirlineCode
FlightNum
Origin
Destination
DepartureTime
DepartureDelay
ArrivalTime
ArrivalDelay
Airtime
Distance

Field Description - airlines.csv

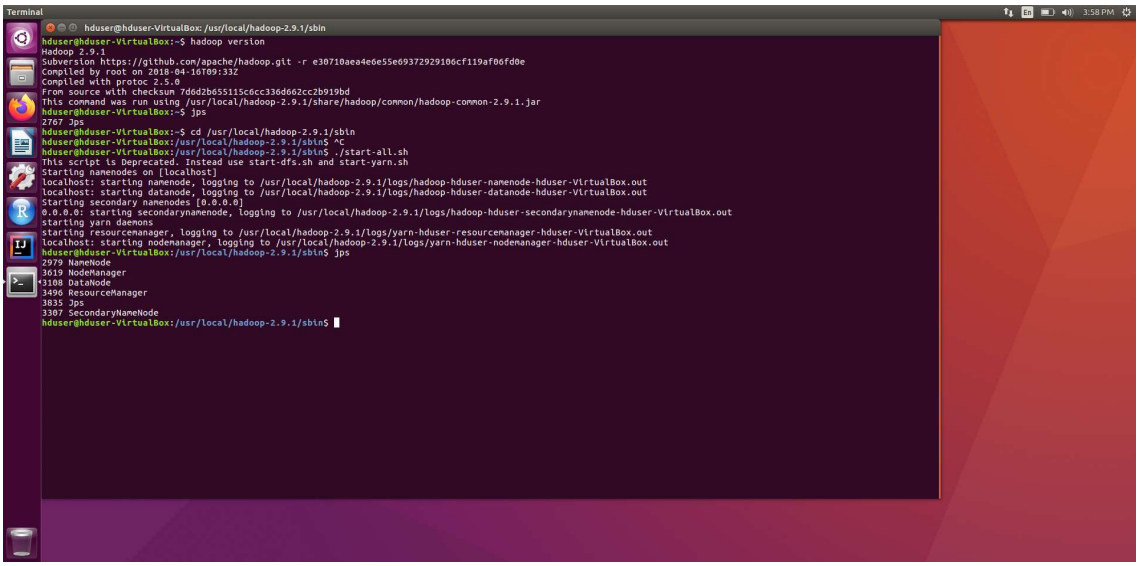
AirlineCode
Description

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

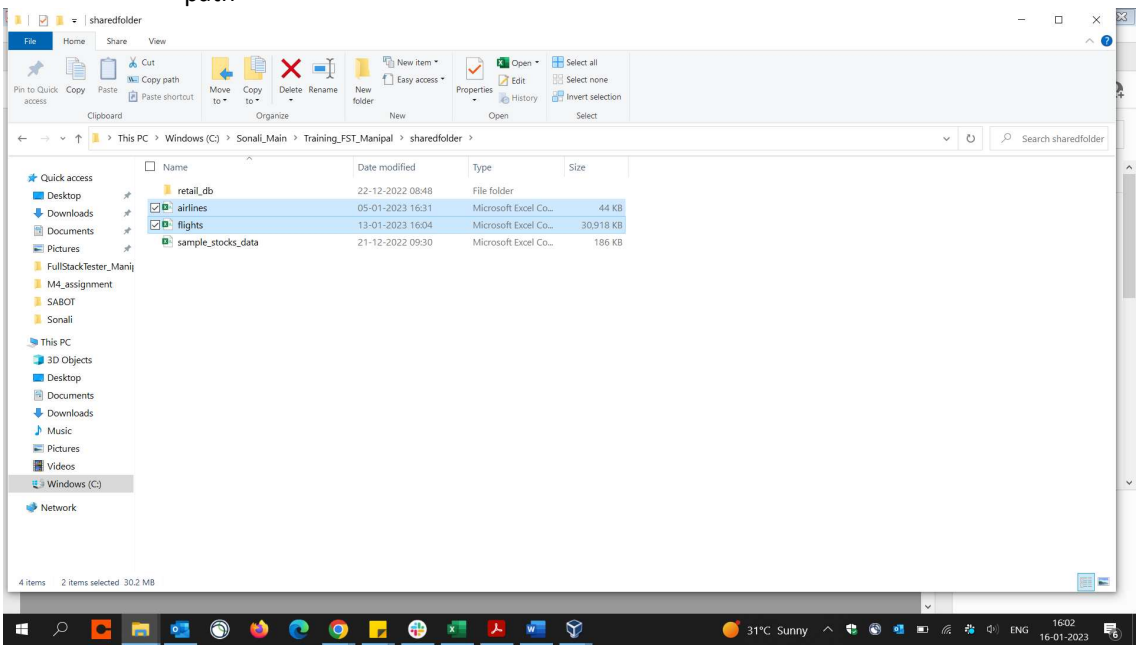
---
Attach screenshots of the outputs along with the query executed in the solutions file

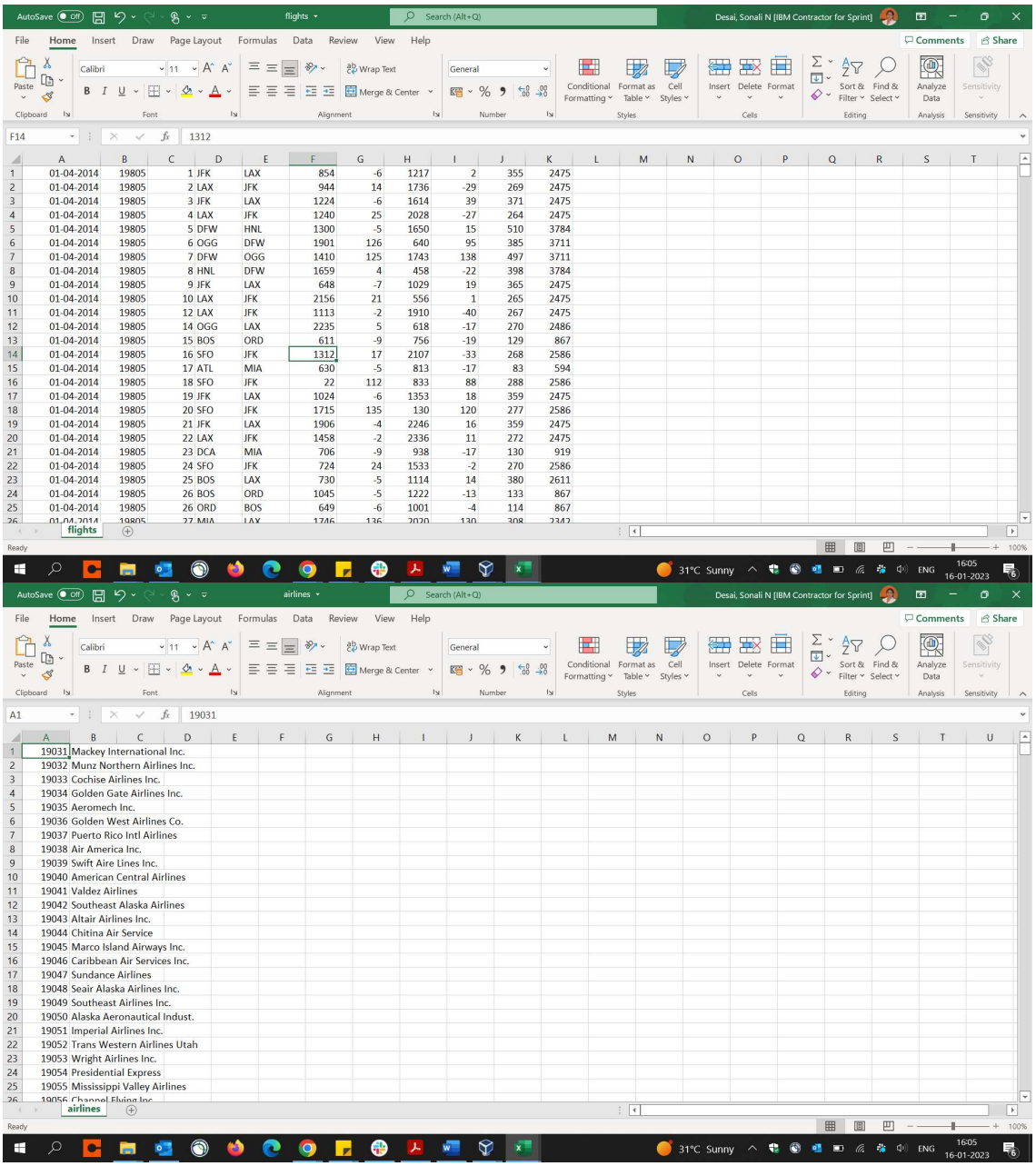
SOLUTION:

1. Create Hive tables for the above datasets by identifying the right datatypes
    a. Open VM & login with Username-hduser, Password-hadoop
    b. Open terminal
    c. Check Hadoop version - $ hadoop version
    d. Check Java Process Status - $ jps
    e. Change directory - $ cd /usr/local/Hadoop-2.9.1/sbin
    f. Start Hadoop service - $ ./start-all.sh
    g. Check Java Process Status - $ jps (Hadoop should be running now)



h. Configure ‘sharedfolder’ on local system, copy the CSV files to it, copy the folder path

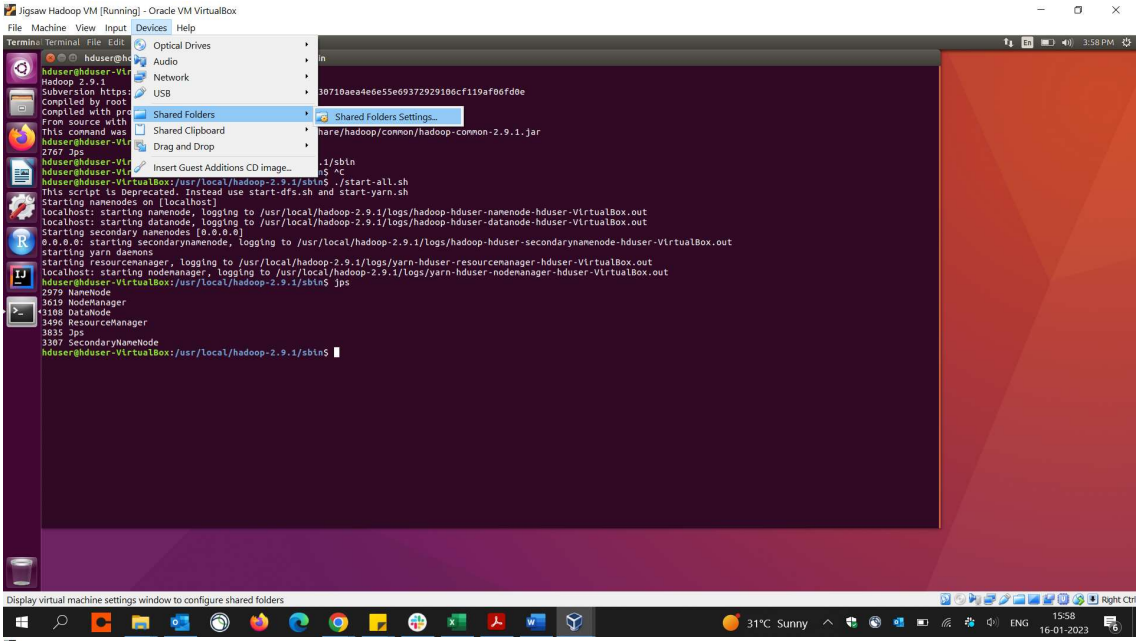


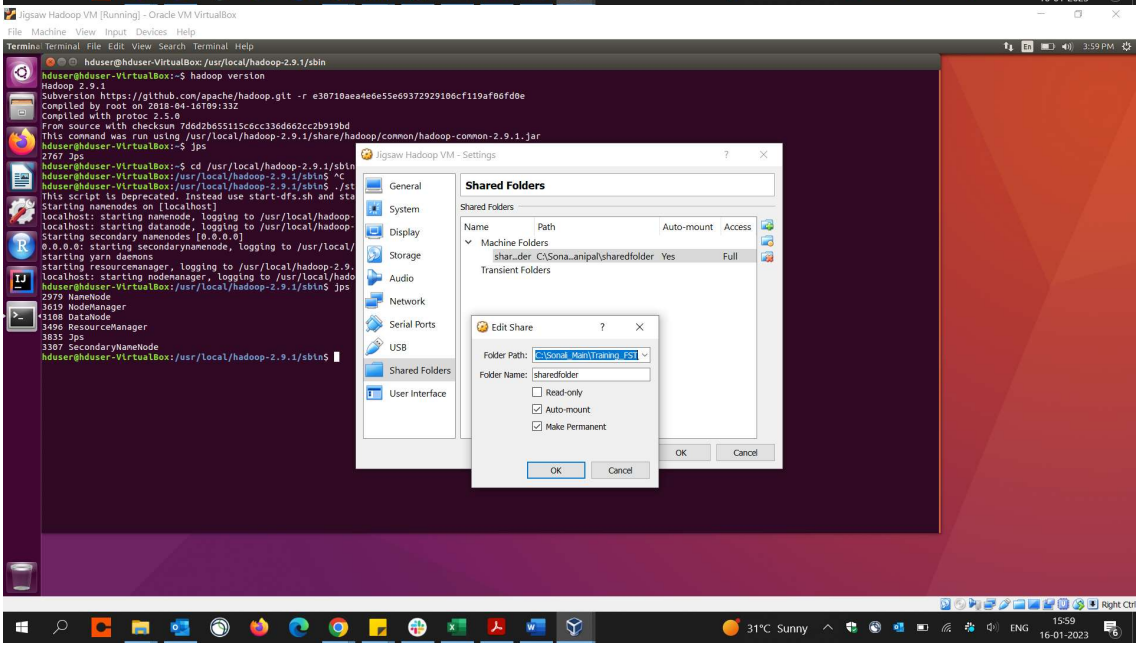


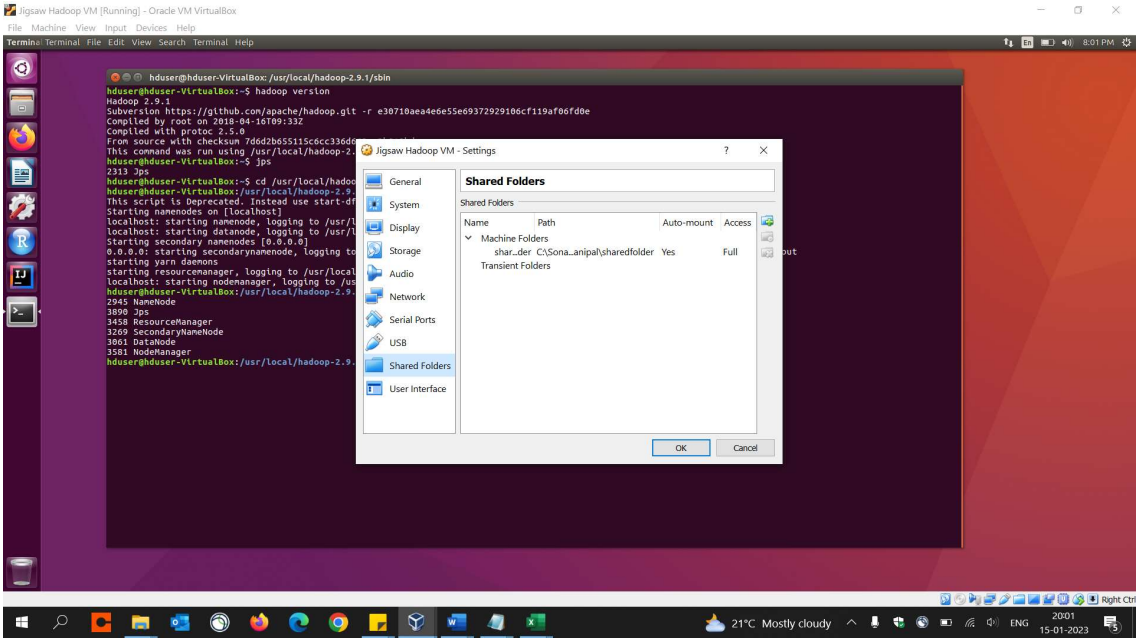
	A	B	C	D	E	F	G	H	I	J	K
1	01-04-2014	19805	1	JFK	LAX	854	-6	1217	2	355	2475
2	01-04-2014	19805	2	LAX	JFK	944	14	1736	-29	269	2475
3	01-04-2014	19805	3	JFK	LAX	1224	-6	1614	39	371	2475
4	01-04-2014	19805	4	LAX	JFK	1240	25	2028	-27	264	2475
5	01-04-2014	19805	5	DFW	HNL	1300	-5	1650	15	510	3784
6	01-04-2014	19805	6	OGG	DFW	1901	126	640	95	385	3711
7	01-04-2014	19805	7	DFW	OGG	1410	125	1743	138	497	3711
8	01-04-2014	19805	8	HNL	DFW	1659	4	458	-22	398	3784
9	01-04-2014	19805	9	JFK	LAX	648	-7	1029	19	365	2475
10	01-04-2014	19805	10	LAX	JFK	2156	21	556	1	265	2475
11	01-04-2014	19805	12	LAX	JFK	1113	-2	1910	-40	267	2475
12	01-04-2014	19805	14	OGG	LAX	2235	5	618	-17	270	2486
13	01-04-2014	19805	15	BOS	ORD	611	-9	756	-19	129	867
14	01-04-2014	19805	16	SFO	JFK	1312	17	2107	-33	268	2586
15	01-04-2014	19805	17	ATL	MIA	630	-5	813	-17	83	594
16	01-04-2014	19805	18	SFO	JFK	22	112	833	88	288	2586
17	01-04-2014	19805	19	JFK	LAX	1024	-6	1353	18	359	2475
18	01-04-2014	19805	20	SFO	JFK	1715	135	130	120	277	2586
19	01-04-2014	19805	21	JFK	LAX	1906	-4	2246	16	359	2475
20	01-04-2014	19805	22	LAX	JFK	1458	-2	2336	11	272	2475
21	01-04-2014	19805	23	DCA	MIA	706	-9	938	-17	130	919
22	01-04-2014	19805	24	SFO	JFK	724	24	1533	-2	270	2586
23	01-04-2014	19805	25	BOS	LAX	730	-5	1114	14	380	2611
24	01-04-2014	19805	26	BOS	ORD	1045	-5	1222	-13	133	867
25	01-04-2014	19805	26	ORD	BOS	649	-6	1001	-4	114	867

	A	B
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

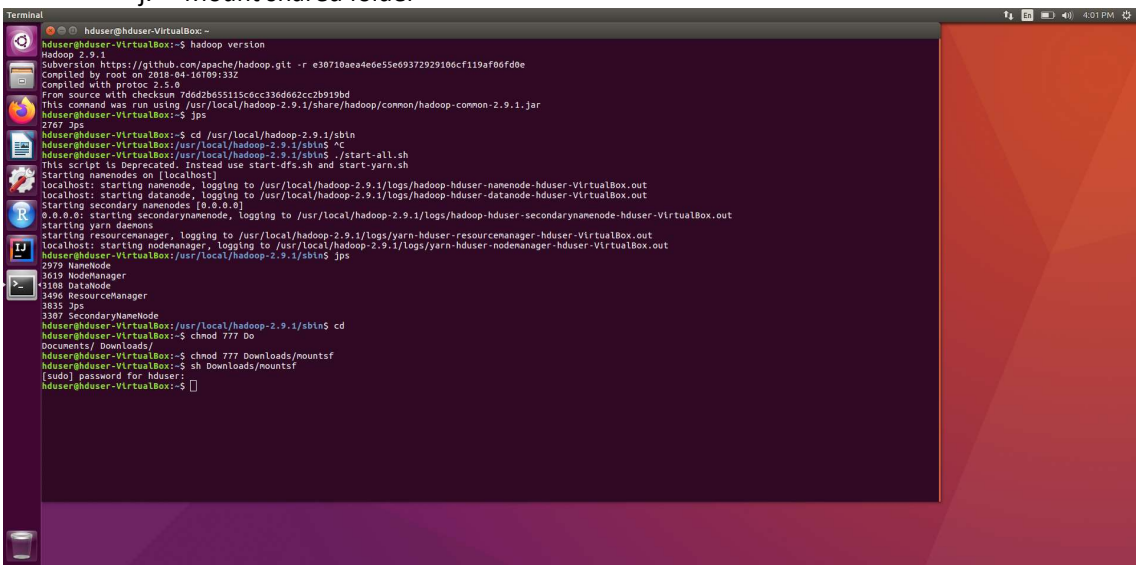
i. Configure Devices -> Shared folders -> Shared folders settings -> local shared folder path on the VM.


Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Optical Drives
Audio
Network
USB
Shared Folders
Shared Clipboard
Drag and Drop
Insert Guest Additions CD image...
Shared Folders Settings...
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$ jps
2979 NameNode
3619 NodeManager
3108 DataNode
3496 ResourceManager
3835 Jps
3307 SecondaryNameNode
hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$
Display virtual machine settings window to configure shared folders
31°C Sunny
ENG
15:58
16-01-2023

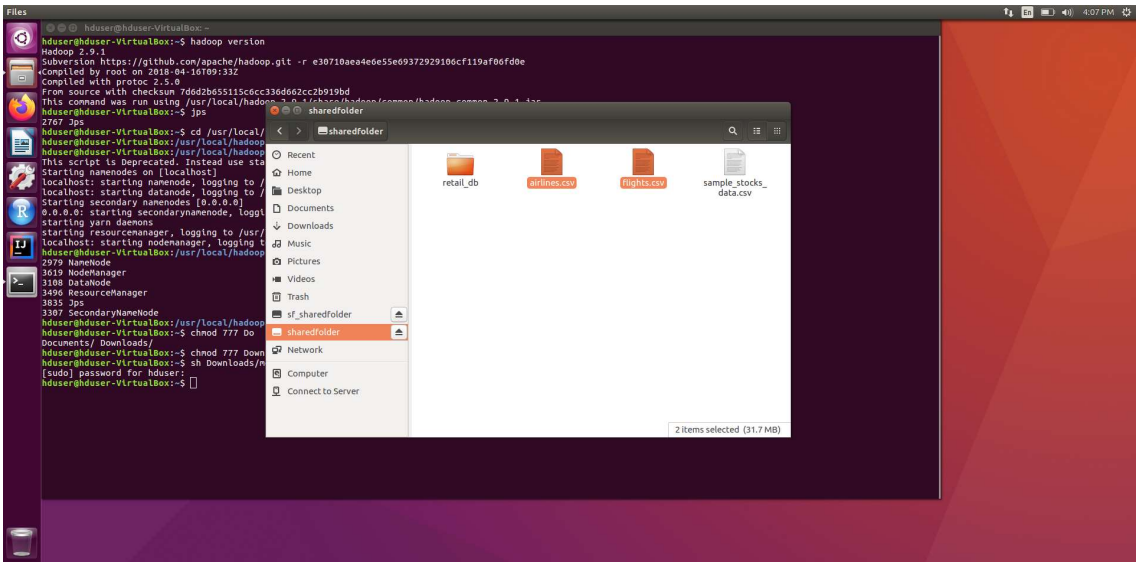

Jigsaw Hadoop VM [Running] - Oracle VM VirtualBox
hduser@hduser-VirtualBox: /usr/local/hadoop-2.9.1/sbin
hduser@hduser-VirtualBox:~$ hadoop version
Hadoop 2.9.1
Subversion https://github.com/apache/hadoop.git -r e30710aea4e6e55e69372929106cf119af06fd0e
Compiled by root on 2018-04-16T09:33Z
Compiled with protoc 2.5.0
From source with checksum 7d6d2b655115c6cc336d662cc2b919bd
This command was run using /usr/local/hadoop-2.9.1/share/hadoop/common/hadoop-common-2.9.1.jar
Jigsaw Hadoop VM - Settings
General
System
Display
Storage
Audio
Network
Serial Ports
USB
Shared Folders
User Interface
Shared Folders
Shared Folders
Name
Path
Auto-mount
Access
Machine Folders
shar...der
C:\Sona...anipal\sharedfolder
Yes
Full
Transient Folders
Edit Share
Folder Path:
C:\Sonali_Main\Training_FST
Folder Name:
sharedfolder
Read-only
Auto-mount
Make Permanent
OK
Cancel
OK
Cancel
31°C Sunny
ENG
15:59
16-01-2023



j. Mount shared folder



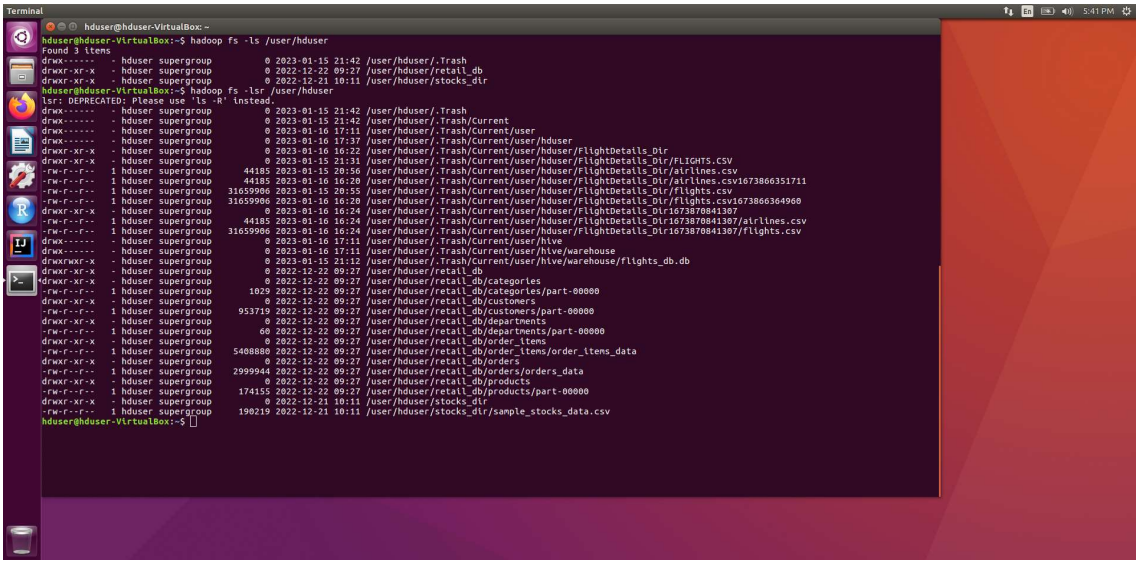
k. Check that files are available on VM



/user/hduser as it is the working directory

$ hadoop fs -ls /user/hduser

Check files

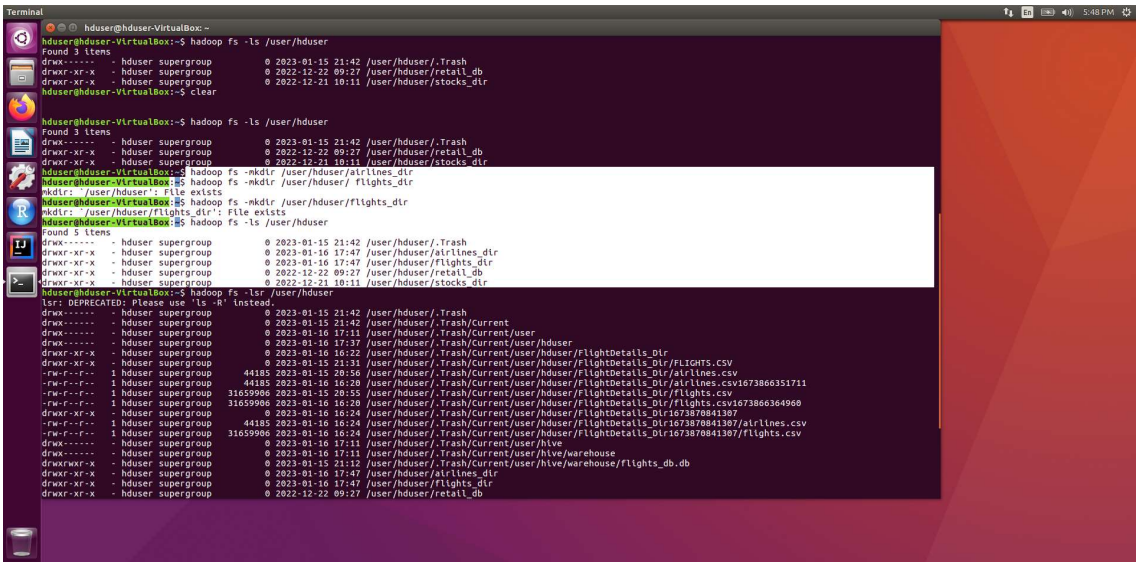


l. Put data on HDFS - Create directory ‘airlines_dir’ & put ‘airlines.csv’ , Create directory ‘flights_dir’ and put ‘flights.csv’ files in it.

$ hadoop fs -mkdir /user/hduser/airlines_dir

$ hadoop fs -mkdir /user/hduser/ flights_dir

$ hadoop fs -lsr /user/hduser

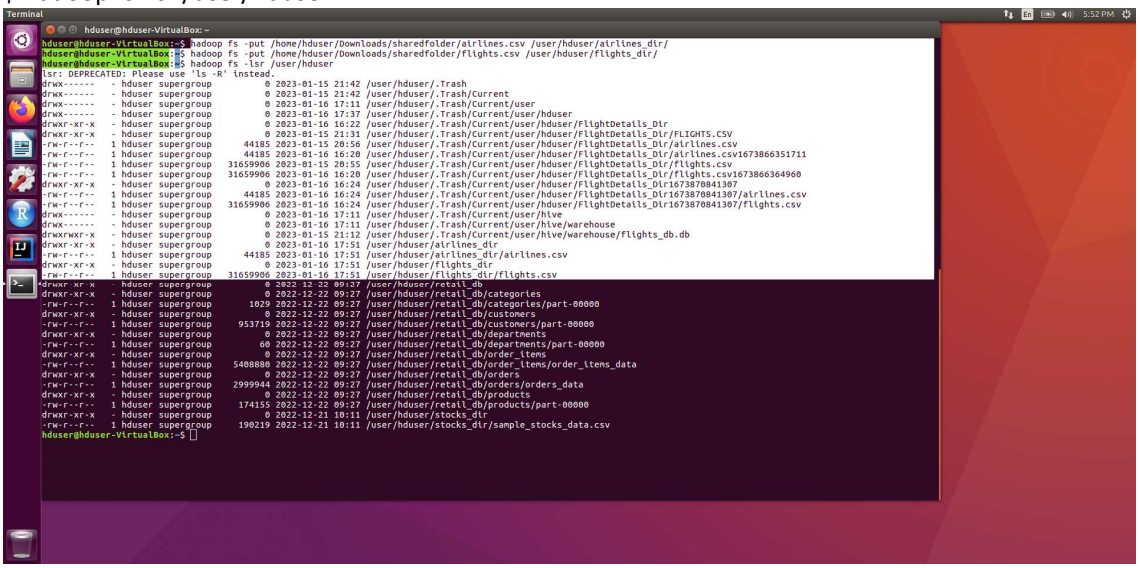


Put ‘airelines.csv’ to airlines_Dir & ‘flights.csv’ to ‘flights_dir’

$ hadoop fs -put /home/hduser/Downloads/sharedfolder/airlines.csv /user/hduser/airlines_dir/

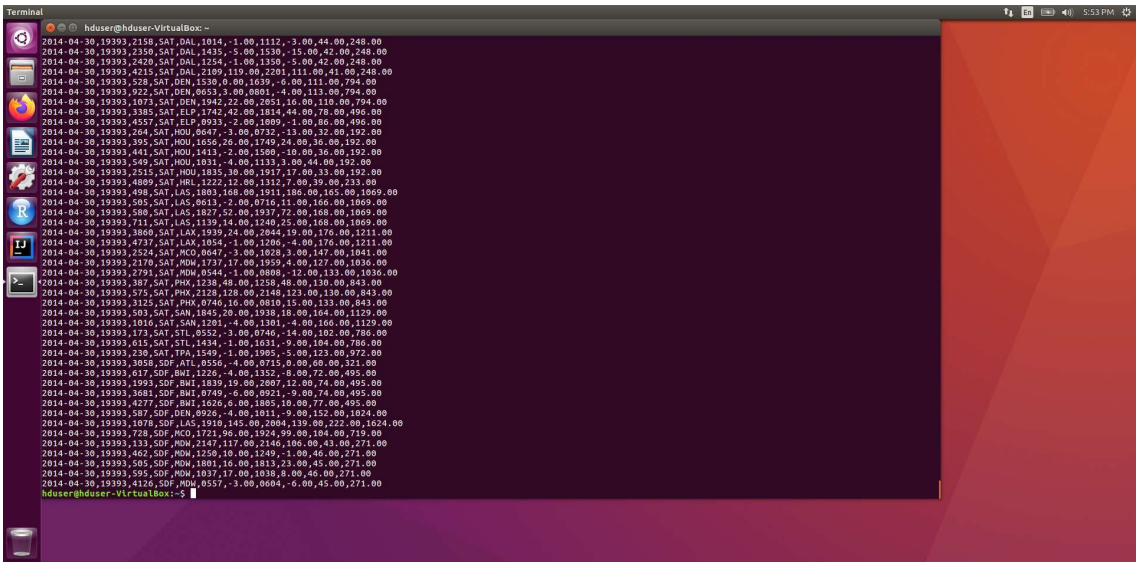
$ hadoop fs -put /home/hduser/Downloads/sharedfolder/flights.csv /user/hduser/flights_dir/

$ hadoop fs -lsr /user/hduser

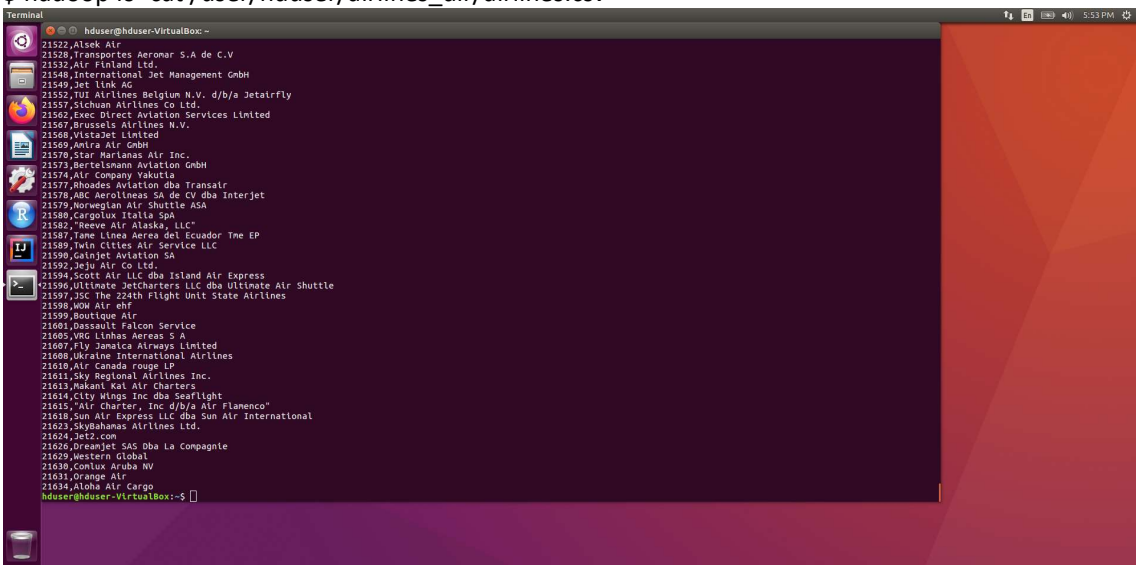


m. Read flights.csv & airlines.csv

$ hadoop fs -cat /user/hduser/flights_dir/flights.csv

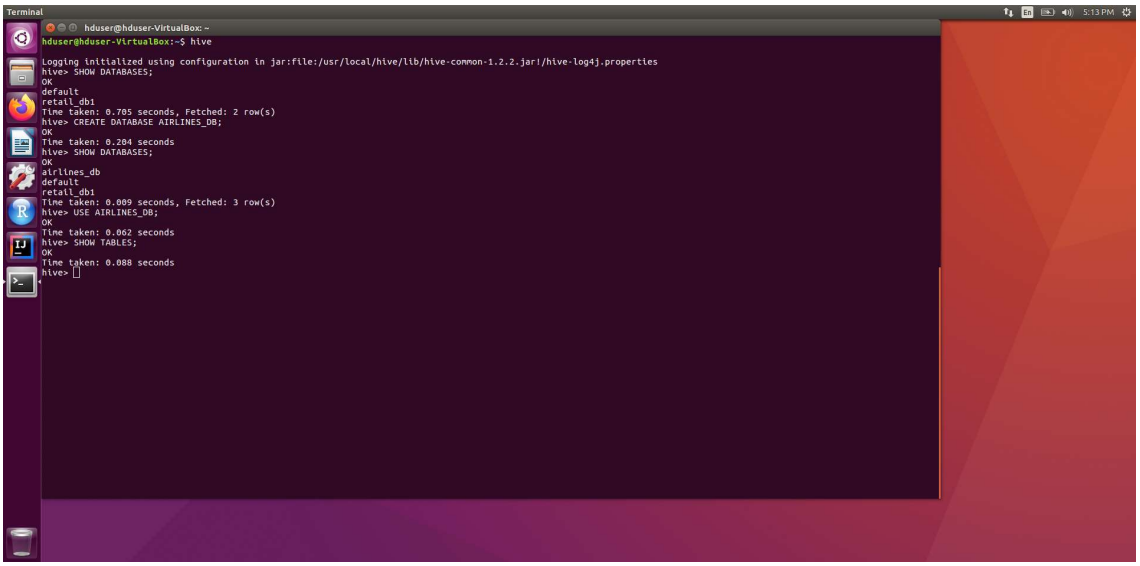


$ hadoop fs -cat /user/hduser/airlines_dir/airlines.csv



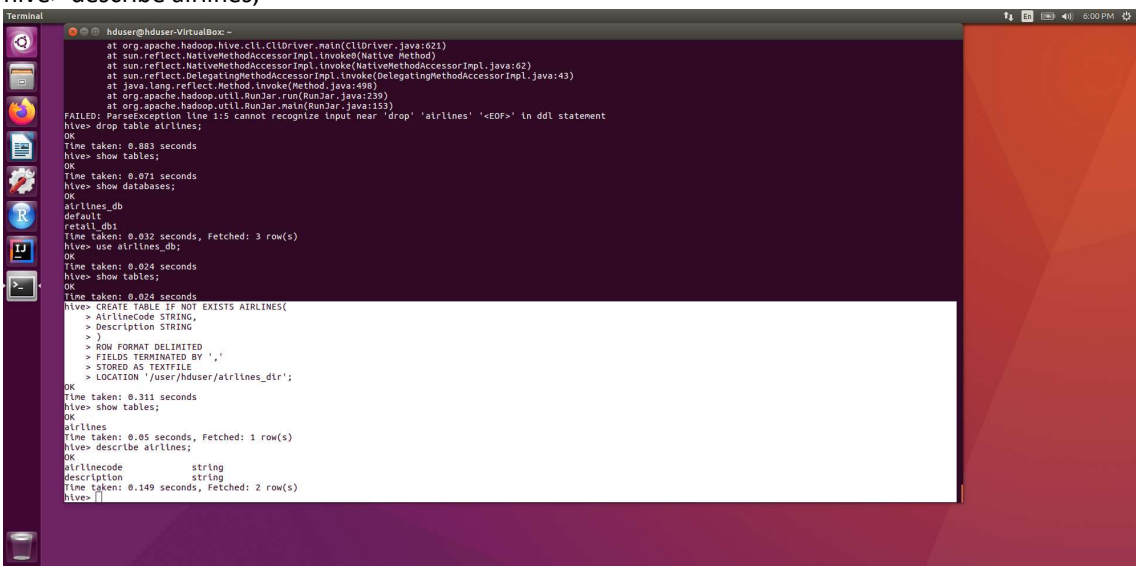
n. Create database & tables

- open hive => $ hive
- see existing data bases => hive> SHOW DATABASES;
- create database => hive> CREATE DATABASE AIRLINES_DB;
- use created database => hive> USE AIRLINES_DB;
- view tables in AIRLINES_DB => hive> SHOW TABLES;



o. create table for AIRLINES

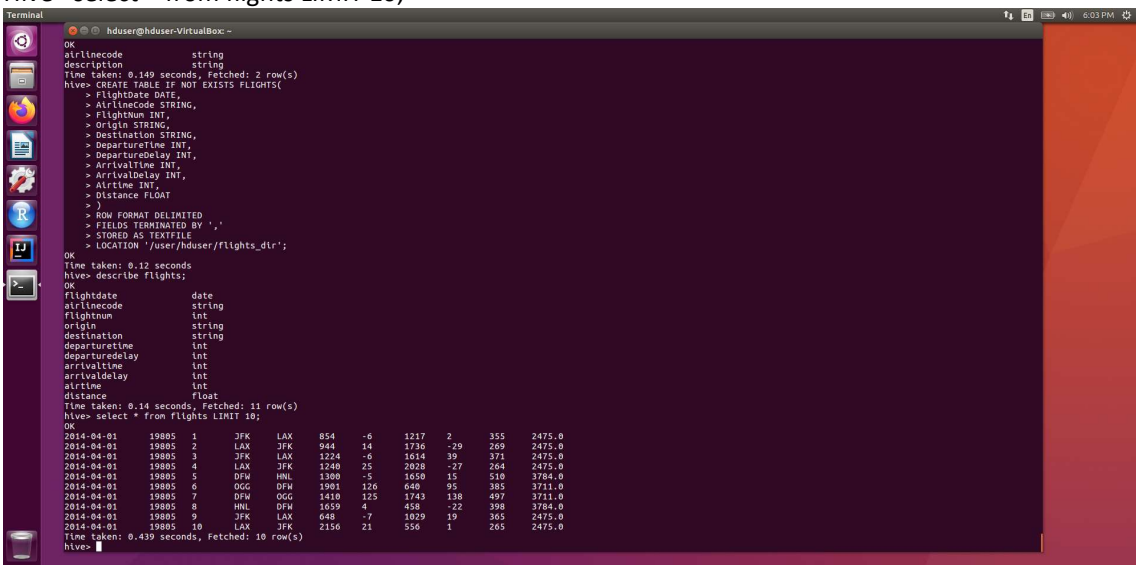
```
hive>CREATE TABLE IF NOT EXISTS AIRLINES(
AirlineCode STRING,
Description STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/hduser/airlines_dir';
hive> describe airlines;
```



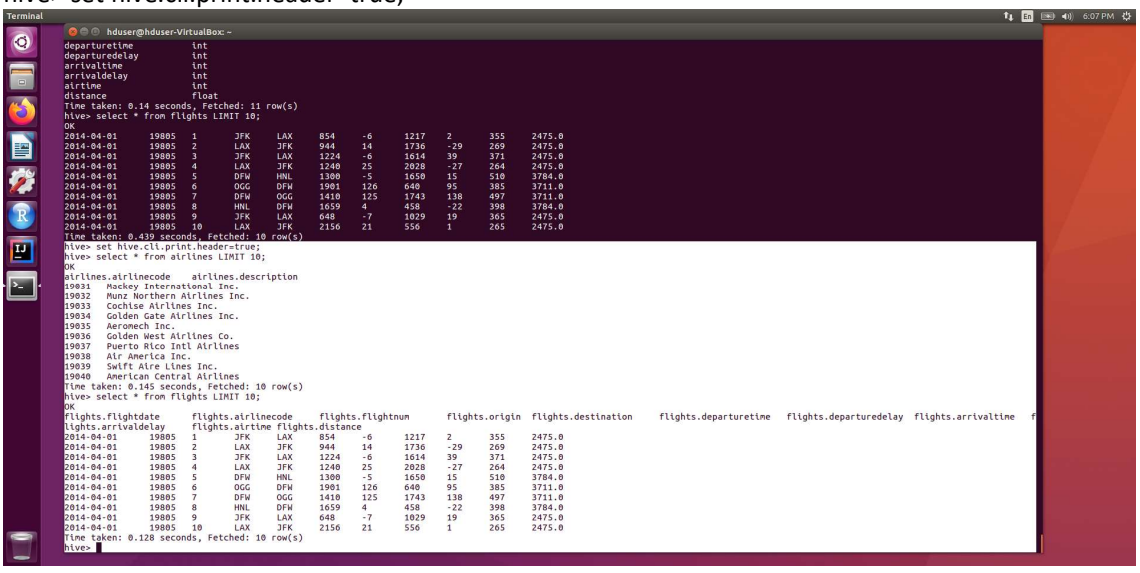
p. create table for FLIGHTS

```
hive> CREATE TABLE IF NOT EXISTS FLIGHTS(
FlightDate DATE,
AirlineCode STRING,
FlightNum INT,
Origin STRING,
Destination STRING,
DepartureTime INT,
```

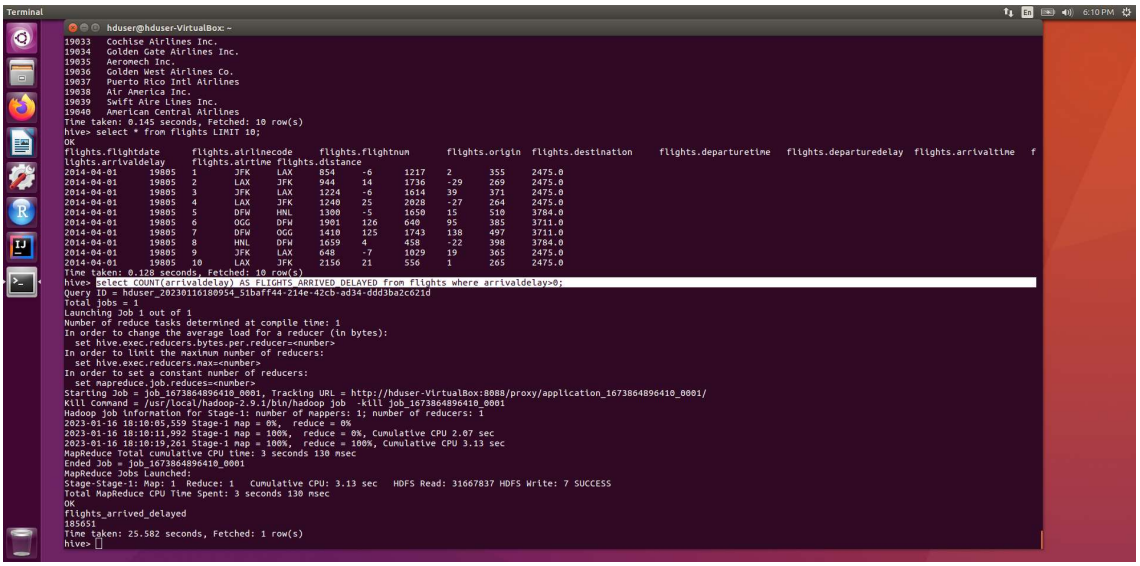
DepartureDelay INT,
ArrivalTime INT,
ArrivalDelay INT,
Airtime INT,
Distance FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/hduser/flights_dir';
Hive> describe flights;
Hive> select * from flights LIMIT 10;



Enable column names
hive> set hive.cli.print.header=true;



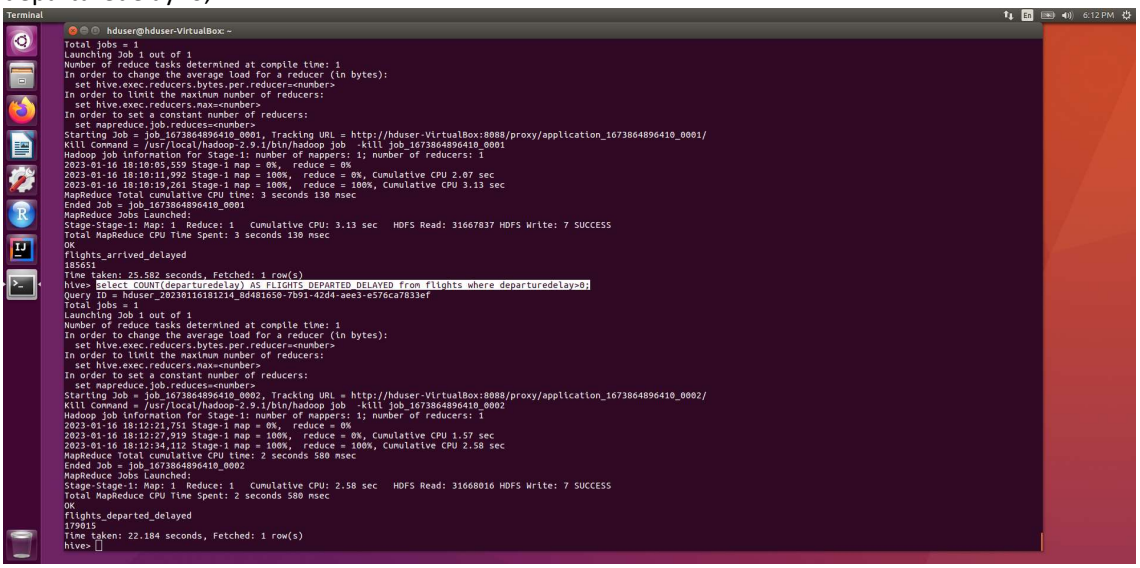
2. Solve the following use cases
Find count of flights that had arrival delay
Hive> select COUNT(arrivaldelay) AS FLIGHTS_ARRIVED_DELAYED from flights where arrivaldelay>0;



flights_arrived_delayed
185651

Find count of flights that had departure delay

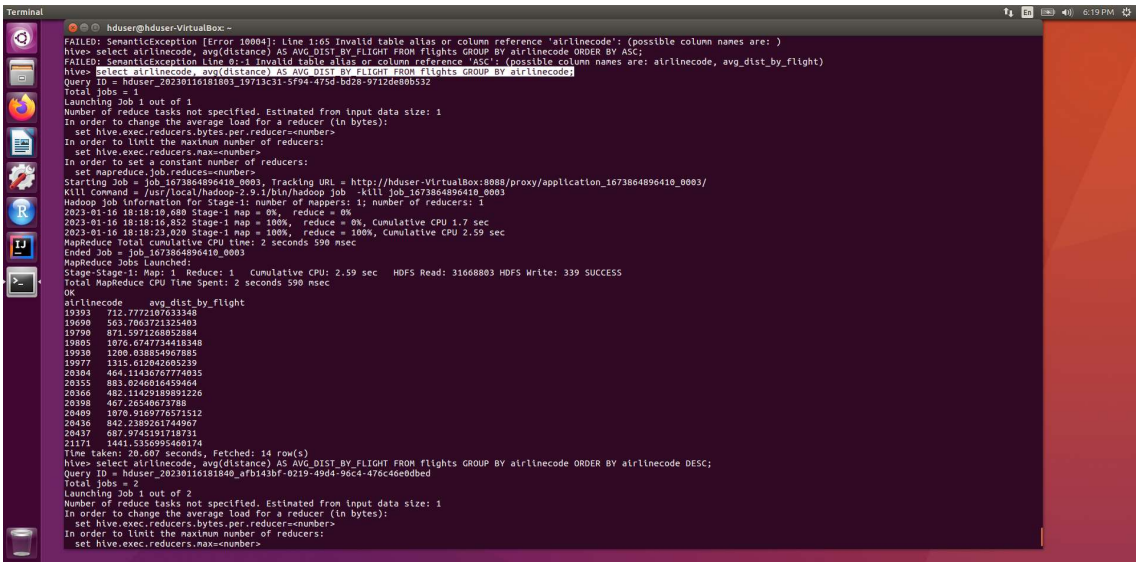
Hive> select COUNT(departuredelay) AS FLIGHTS_DEPARTED_DELAYED from flights where departuredelay>0;



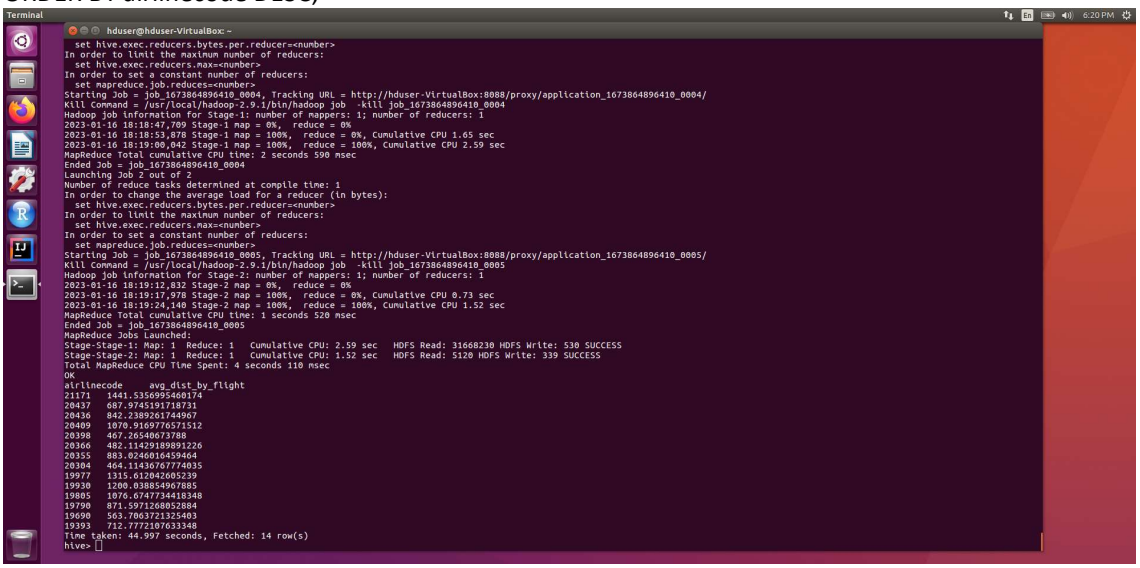
flights_departed_delayed
179015

find the average distance travelled by a flight

hive> select airlinecode, avg(distance) AS AVG_DIST_BY_FLIGHT FROM flights GROUP BY airlinecode;

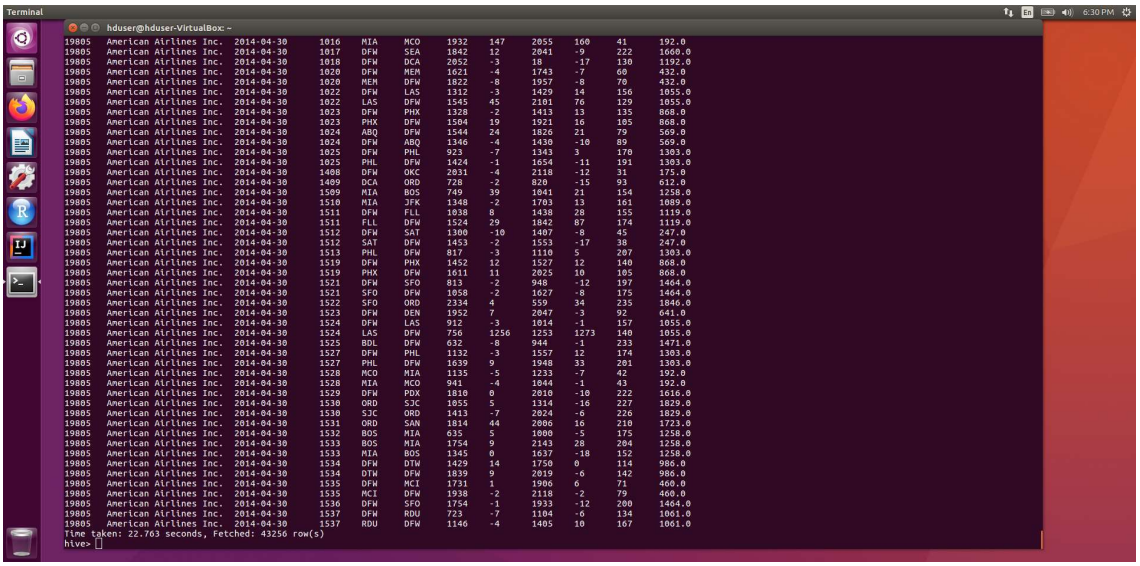


Hive> select airlinecode, avg(distance) AS AVG_DIST_BY_FLIGHT FROM flights GROUP BY airlinecode ORDER BY airlinecode DESC;



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

Hive> SELECT a.airlinecode,a.description,
f.flightdate,
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.';




```
SELECT a.airlinecode,a.description,
f.flightdate,
f.flightnum,
f.origin,
f.destination,
f.departuretime,
f.departuredelay,
f.arrivaltime,
f.arrivaldelay,
f.airtime,
f.distance FROM flights f  LEFT JOIN airlines a
ON (a.airlinecode = f.airlinecode)
WHERE a.description='American Airlines Inc.'
```