
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:

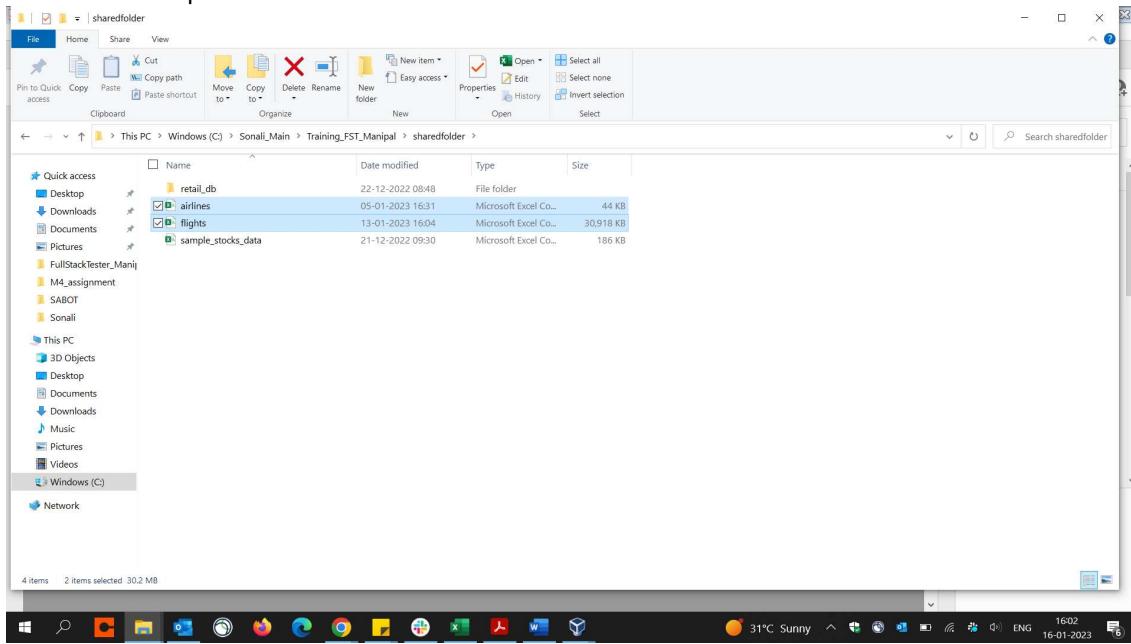
1. **Create Hive tables for the above datasets by identifying the right datatypes**
 - a. Open VM & login with Username-hduser, Password-hadoop
 - b. Open terminal
 - c. Check Hadoop version - \$ hadoop version
 - d. Check Java Process Status - \$ jps
 - e. Change directory - \$ cd /usr/local/Hadoop-2.9.1/sbin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

```

Terminal
hduser@hduser-VirtualBox:~/usr/local/hadoop-2.9.1/sbin$ hadoop version
Compiled by root on 2018-04-16T09:33Z
Compiled with protoc 3.5.0
From source at revision 7dddb5515cc336d62cc2b91bd
This command was run using /usr/local/hadoop-2.9.1/share/hadoop/common/hadoop-common-2.9.1.jar
hduser@hduser-VirtualBox:~$ jps
2248 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
[HDUser] Starting namenodes on [localhost]
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-namenode-hduser-VirtualBox.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-secondarynamenode-hduser-VirtualBox.out
Starting secondarynamenodes [0.0.0.0]
localhost: starting yarn daemons
localhost: 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-nodenanager-hduser-VirtualBox.out
hduser@hduser-VirtualBox:~/usr/local/hadoop-2.9.1/sbin$ jps
2248 Jps
3019 NodeManager
3108 DataNode
3455 ResourceManager
3455 SecondaryNameNode
3307 NodeManager
hduser@hduser-VirtualBox:~/usr/local/hadoop-2.9.1/sbin$ 

```

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



The image shows two Microsoft Excel windows side-by-side.

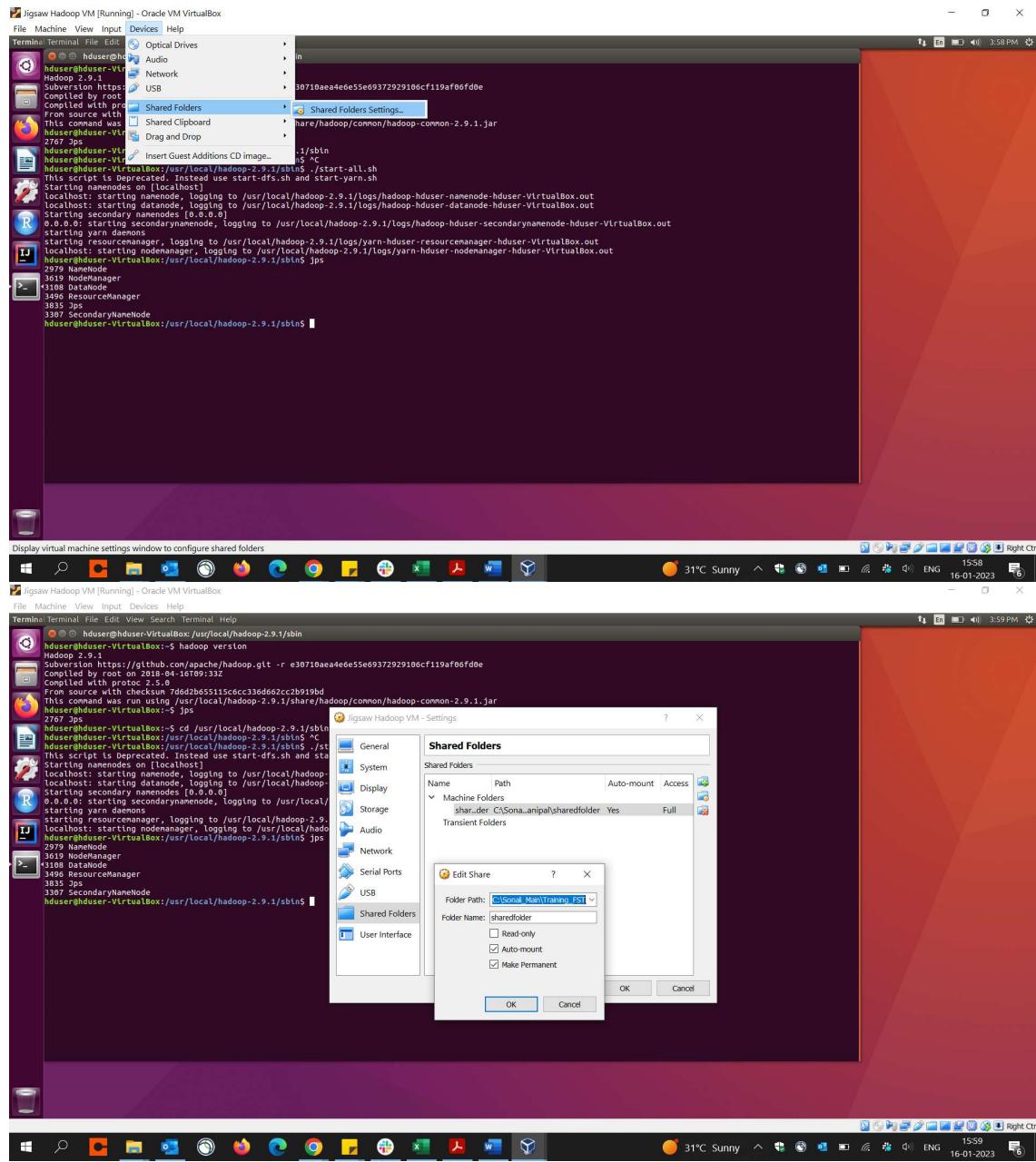
Top Window (flights):

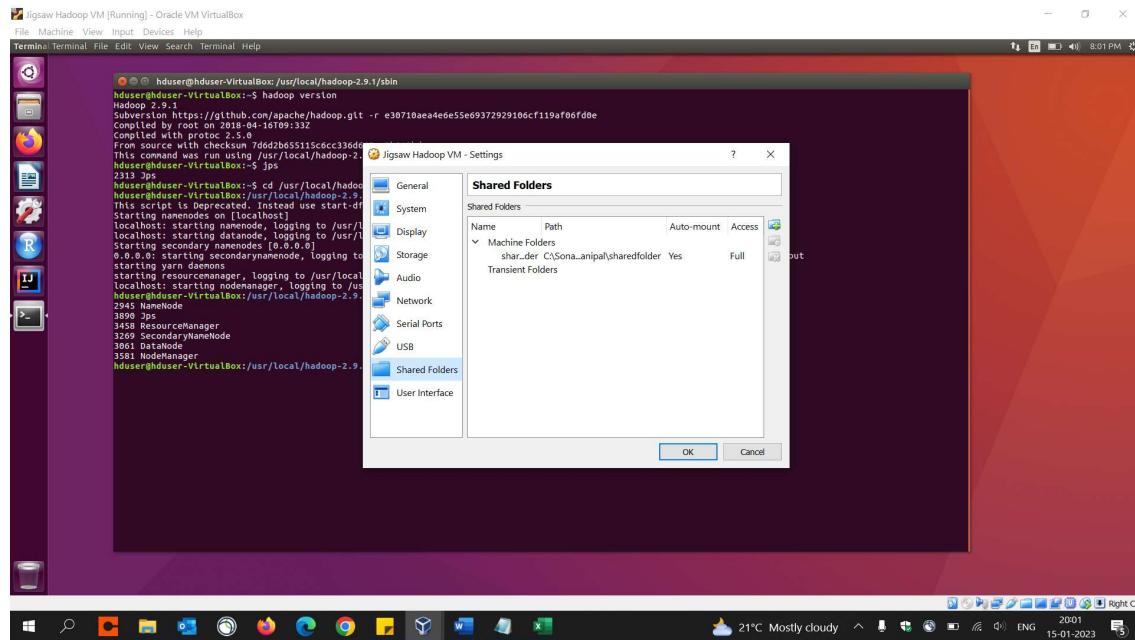
- Sheet: flights
- Cells A1 to K26 contain flight data for January 2014. Column A shows dates (e.g., 01-04-2014), columns B-D show origin and destination airports, and columns E-K show various numerical values (e.g., 854, -6, 1217).
- Row 14 highlights the range F14:K14.
- Row 26 is labeled 'flights'.

Bottom Window (airlines):

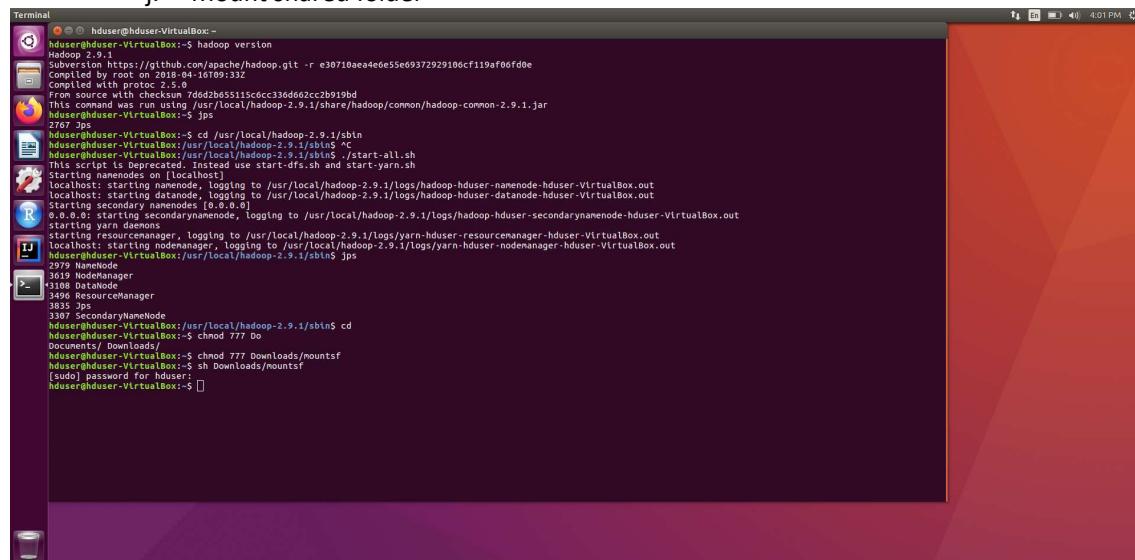
- Sheet: airlines
- Cells A1 to U26 list airline names, starting with 'Mackey International Inc.' at A1.
- Row 26 is labeled 'airlines'.

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

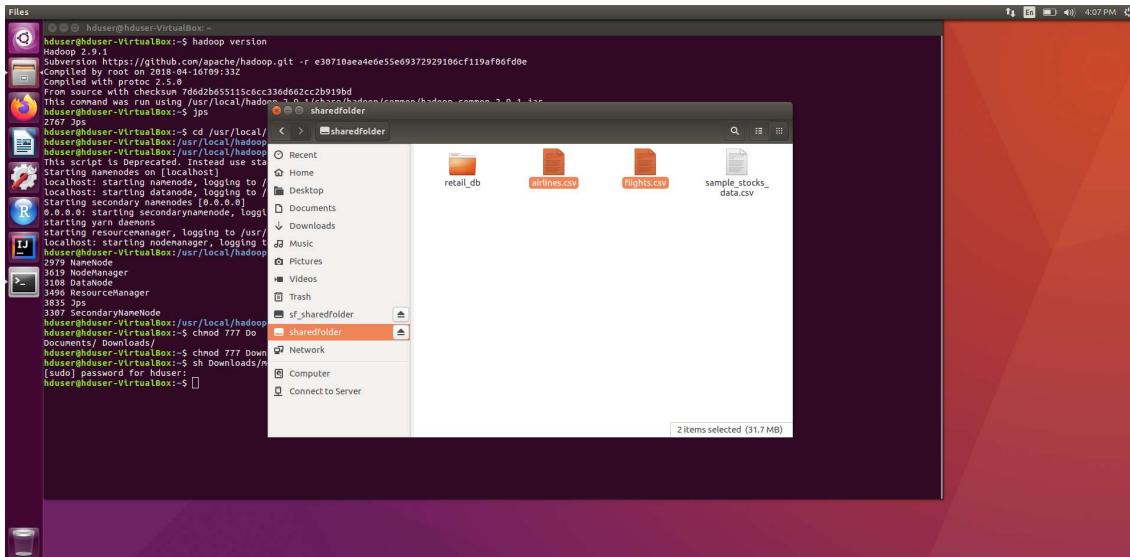




j. Mount shared folder



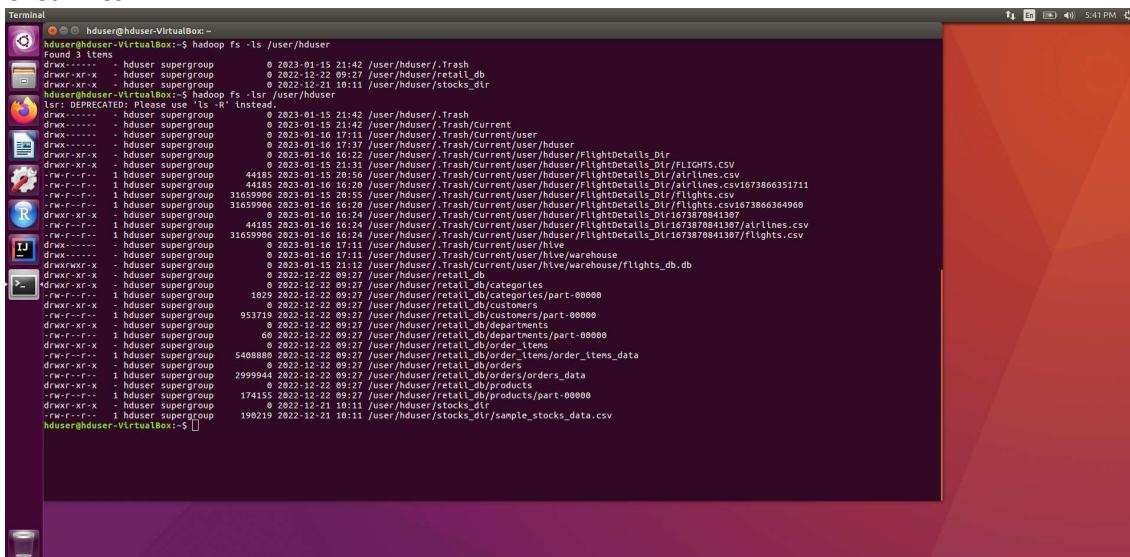
k. Check that files are available on VM



/user/hduser as it is the working directory

\$ hadoop fs -ls /user/hduser

Check files



- i. 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

```

hduser@hduser-VirtualBox:~$ hadoop fs -ls /user/hduser
Found 3 items
drwx----- 1 hduser supergroup 0 2023-01-15 21:42 /user/hduser/.Trash
drwxr-xr-x 1 hduser supergroup 0 2022-12-22 09:27 /user/hduser/retail_db
drwxr-xr-x 1 hduser supergroup 0 2022-12-21 10:11 /user/hduser/stocks_dir
hduser@hduser-VirtualBox:~$ clear

hduser@hduser-VirtualBox:~$ hadoop fs -ls /user/hduser
Found 3 items
drwx----- 1 hduser supergroup 0 2023-01-15 21:42 /user/hduser/.Trash
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:47 /user/hduser/airlines_dir
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 16:22 /user/hduser/flights_dir
hduser@hduser-VirtualBox:~$ hadoop fs -mkdir /user/hduser/airlines_dir
hduser@hduser-VirtualBox:~$ hadoop fs -mkdir /user/hduser/ flights_dir
hduser@hduser-VirtualBox:~$ hadoop fs -mkdir /user/hduser/stocks_dir
hduser@hduser-VirtualBox:~$ hadoop fs -ls /user/hduser
ls: DEPRECATED: Please use 'ls -R' instead.
drwx----- 1 hduser supergroup 0 2023-01-15 21:42 /user/hduser/.Trash
drwx----- 1 hduser supergroup 0 2023-01-15 21:42 /user/hduser/.Trash/Current
drwx----- 1 hduser supergroup 0 2023-01-16 17:37 /user/hduser/.Trash/Current/user/hduser
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 16:22 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir
drwxr-xr-x 1 hduser supergroup 0 2023-01-15 20:56 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/FLIGHTS.CSV
drwxr-xr-x 1 hduser supergroup 44185 2023-01-16 16:28 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/airlines.csv
drwxr-xr-x 1 hduser supergroup 31659986 2023-01-15 20:55 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/flights.csv
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/airlines.csv
drwxr-xr-x 1 hduser supergroup 44185 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/flights.csv
drwxr-xr-x 1 hduser supergroup 31659986 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/FlightDetails_Dir/airlines.csv
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:11 /user/hduser/.Trash/Current/user/hive/warehouse
drwxr-xr-x 1 hduser supergroup 0 2023-01-15 21:31 /user/hduser/.Trash/Current/user/hive/warehouse/flights_db
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:47 /user/hduser/.Trash/stocks_dir
drwxr-xr-x 1 hduser supergroup 0 2022-12-22 09:27 /user/hduser/retail_db

```

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

```

```

hduser@hduser-VirtualBox:~$ hadoop fs -put /home/hduser/Downloads/sharedFolder/airlines.csv /user/hduser/airlines_dir/
hduser@hduser-VirtualBox:~$ hadoop fs -put /home/hduser/Downloads/sharedFolder/flights.csv /user/hduser/flights_dir/
ls: DEPRECATED: Please use 'ls -R' instead.
drwx----- 1 hduser supergroup 0 2023-01-15 21:42 /user/hduser/.Trash
drwx----- 1 hduser supergroup 0 2023-01-16 22:42 /user/hduser/.Trash/Current
drwx----- 1 hduser supergroup 0 2023-01-16 22:42 /user/hduser/.Trash/Current/user/hduser
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:37 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir
drwxr-xr-x 1 hduser supergroup 0 2023-01-15 21:31 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/FLIGHTS.CSV
drwxr-xr-x 1 hduser supergroup 44185 2023-01-16 16:28 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/airlines.csv
drwxr-xr-x 1 hduser supergroup 31659986 2023-01-16 16:28 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/flights.csv
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/airlines.csv
drwxr-xr-x 1 hduser supergroup 44185 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/flights.csv
drwxr-xr-x 1 hduser supergroup 31659986 2023-01-16 16:24 /user/hduser/.Trash/Current/user/hduser/FlightDetails_Dir/1673878841307/FlightDetails_Dir/airlines.csv
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:11 /user/hduser/.Trash/Current/user/hive/warehouse
drwxr-xr-x 1 hduser supergroup 0 2023-01-15 21:31 /user/hduser/.Trash/Current/user/hive/warehouse/flights_db
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:51 /user/hduser/.Trash/stocks_dir
drwxr-xr-x 1 hduser supergroup 44185 2023-01-16 17:51 /user/hduser/airlines_dir/airlines.csv
drwxr-xr-x 1 hduser supergroup 0 2023-01-16 17:51 /user/hduser/flights_dir/flights.csv
drwxr-xr-x 1 hduser supergroup 0 2022-12-22 09:27 /user/hduser/retail_db
drwxr-xr-x 1 hduser supergroup 1029 2022-12-22 09:27 /user/hduser/retail_db/categories/part-00000
drwxr-xr-x 1 hduser supergroup 0 2022-12-22 09:27 /user/hduser/retail_db/customers/part-00000
drwxr-xr-x 1 hduser supergroup 953719 2022-12-22 09:27 /user/hduser/retail_db/customers/part-00000
drwxr-xr-x 1 hduser supergroup 68 2022-12-22 09:27 /user/hduser/retail_db/departments/part-00000
drwxr-xr-x 1 hduser supergroup 548888 2022-12-22 09:27 /user/hduser/retail_db/order_items
drwxr-xr-x 1 hduser supergroup 99 2022-12-22 09:27 /user/hduser/retail_db/order_items/_SUCCESS
drwxr-xr-x 1 hduser supergroup 2999944 2022-12-22 09:27 /user/hduser/retail_db/orders/orders_data
drwxr-xr-x 1 hduser supergroup 0 2022-12-22 09:27 /user/hduser/retail_db/products
drwxr-xr-x 1 hduser supergroup 17418 2022-12-22 09:27 /user/hduser/retail_db/products/part-00000
drwxr-xr-x 1 hduser supergroup 0 2022-12-21 10:11 /user/hduser/stocks_dir
drwxr-xr-x 1 hduser supergroup 190219 2022-12-21 10:11 /user/hduser/stocks_dir/sample_stocks_data.csv

```

m. Read flights.csv & airlines.csv

```

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

```

```

Terminal
hduser@hduser-VirtualBox: ~
2014-04-30_19393,2158,SAT,DAL,1014,-1.00,1112,3.00,44.00,248.00
2014-04-30_19393,2159,SAT,DAL,1014,-1.00,1112,3.00,44.00,248.00
2014-04-30_19393,2420,SAT,DAL,1254,-1.00,1330,5.00,42.00,248.00
2014-04-30_19393,4215,SAT,DAL,1109,-19.00,2201,111.00,41.00,248.00
2014-04-30_19393,1039,SAT,DEN,1139,0.00,1639,-6.00,111.00,794.00
2014-04-30_19393,1040,SAT,DEN,1139,0.00,1639,-6.00,111.00,794.00
2014-04-30_19393,1073,SAT,DEN,1942,22.00,2051,16.00,118.00,794.00
2014-04-30_19393,3885,SAT,ELP,1742,42.00,1814,44.00,78.00,496.00
2014-04-30_19393,3886,SAT,ELP,1742,42.00,1814,44.00,78.00,496.00
2014-04-30_19393,264,SAT,HOU,00447,-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,-2.00,1580,-1.00,36.00,192.00
2014-04-30_19393,442,SAT,HOU,1413,-2.00,1580,-1.00,36.00,192.00
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_19393,505,SAT,LAS,00113,-2.00,0716,11.00,5.00,1669.00
2014-04-30_19393,580,SAT,LAS,1827,52.00,1937,72.00,168.00,1669.00
2014-04-30_19393,711,SAT,LAS,1134,14.00,1240,25.00,168.00,1669.00
2014-04-30_19393,712,SAT,LAS,1134,14.00,1240,25.00,168.00,1669.00
2014-04-30_19393,4737,SAT,LAX,1054,-1.00,1260,-4.00,176.00,1211.00
2014-04-30_19393,2524,SAT,MCO,0047,-3.00,1023.00,0.00,147.00,1041.00
2014-04-30_19393,2525,SAT,MCO,0047,-3.00,1023.00,0.00,147.00,1041.00
2014-04-30_19393,2791,SAT,MOW,0544,-1.00,8898,-12.00,113.00,1036.00
2014-04-30_19393,387,SAT,PHX,1238,49.00,1258,48.00,130.00,843.00
2014-04-30_19393,388,SAT,PHX,1238,49.00,1258,48.00,130.00,843.00
2014-04-30_19393,312,SAT,PHX,1740,46.00,0818,15.00,110.00,843.00
2014-04-30_19393,503,SAT,SAN,1845,26.00,1938,18.00,164.00,1129.00
2014-04-30_19393,1016,SAT,SAN,1261,-4.00,1381,-4.00,168.00,1129.00
2014-04-30_19393,1017,SAT,SAN,1261,-4.00,1381,-4.00,168.00,1129.00
2014-04-30_19393,619,SAT,STL,1034,-1.00,1631,-9.00,104.00,756.00
2014-04-30_19393,238,SAT,TPA,1549,-1.00,1985,-5.00,123.00,972.00
2014-04-30_19393,250,SAT,TPA,1549,-1.00,1985,-5.00,123.00,972.00
2014-04-30_19393,537,SOF,BMI,1839,-4.00,1357,-1.00,104.00,493.00
2014-04-30_19393,1993,205,SOF,BMI,1839,-4.00,1357,-1.00,104.00,493.00
2014-04-30_19393,3681,SOF,BMI,1849,-6.00,0921,-9.00,74.00,495.00
2014-04-30_19393,3682,SOF,BMI,1849,-6.00,0921,-9.00,74.00,495.00
2014-04-30_19393,587,SOF,DEN,0926,-4.00,1011,-9.00,152.00,1024.00
2014-04-30_19393,1078,SOF,LAS,1919,145.00,2004,130.00,222.00,1624.00
2014-04-30_19393,1079,SOF,LAS,1919,145.00,2004,130.00,222.00,1624.00
2014-04-30_19393,1133,SOF,ACO,1707,11.00,0924,99.00,104.00,1029.00
2014-04-30_19393,1134,SOF,ACO,1707,11.00,0924,99.00,104.00,1029.00
2014-04-30_19393,462,SOF,MOW,1258,16.00,1249,-1.00,46.00,271.00
2014-04-30_19393,505,SOF,MOW,1601,16.00,1813,23.00,45.00,271.00
2014-04-30_19393,506,SOF,MOW,1601,16.00,1813,23.00,45.00,271.00
2014-04-30_19393,4126,SOF,MOW,0557,-3.00,0901,-0.00,45.00,271.00
hduser@hduser-VirtualBox: ~
```



```

Terminal
hduser@hduser-VirtualBox: ~
21522,Alsek Air
21523,Aerolineas Aeromar S.A de C.V
21532,Air Finland Ltd.
21548,International Jet Management Gmbh
21549,Interjet
21552,TUI Airlines Belgium N.V. d/b/a Jetairfly
21557,Sichuan Airlines Co Ltd.
21562,Exel Direct Aviation Services Limited
21563,Embraer Executive Services N.V.
21568,Wistazet Limited
21569,Amira Air GmbH
21570,Aviation Airline Air Inc.
21573,Berlinsvenska Aviation GmbH
21574,Alr Company Yakutia
21577,Rhodes Aviation Services Transair
21578,SA de CV Interjet SA de CV bba Interjet
21579,Norwegian Air Shuttle ASA
21580,Cargolux Italia SpA
21581,Delta Air Lines LLC
21587,Tame Linea Aerea del Ecuador Tme EP
21589,Twin Cities Air Service LLC
21590,Cathay Aviation SA
21591,Delta Air Lines LLC
21594,Scott Air LLC dba Island Air Express
21596,Ultimate JetCharters LLC dba Ultimate Air Shuttle
21597,Ultimate Flight Unit South Airlines
21598,Now Air ehr
21599,Boutique Air
21601,Dassault Falcon Service
21602,Delta Air Lines Arrear S A
21607,Fly Jamaica Airways Limited
21609,Ukraine International Airlines
21610,Allegiant Air
21611,sky Regional Airlines Inc.
21613,Makan Kal Air Charters
21614,City Wings Inc dba Seawright
21615,Allegiant Air dba Allegiant Air Flamenco"
21616,Sun Air Express LLC dba Sun Air International
21623,Skylahamas Airlines Ltd.
21624,Shuttle Air
21626,Dreamjet SAS dba La Compagnie
21629,Western Global
21630,Contiux Aruba NV
21631,Allegiant Air
21634,Aloha Air Cargo
hduser@hduser-VirtualBox: ~
```

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;

```
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_db
Time taken: 0.705 seconds, Fetched: 2 row(s)
hive> CREATE DATABASE AIRLINES_DB;
OK
Time taken: 0.204 seconds
hive> SHOW DATABASES;
OK
AIRLINES_DB
airlines_db
default
retail_db
Time taken: 0.009 seconds, Fetched: 3 row(s)
hive> USE AIRLINES_DB;
OK
Time taken: 0.062 seconds
hive> SHOW TABLES;
OK
Time taken: 0.088 seconds
hive> [ ]
```

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;
```

```
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:621)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:5 cannot recognize input near ' drop' 'airlines' '<EOF>' in ddl statement
hive> drop table airlines;
OK
Time taken: 0.883 seconds
hive> show tables;
OK
Time taken: 0.071 seconds
hive> show databases;
OK
airlines_db
default
retail_db
Time taken: 0.032 seconds, Fetched: 3 row(s)
hive> use airlines_db;
OK
Time taken: 0.024 seconds
hive> show tables;
OK
Time taken: 0.024 seconds
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';
OK
Time taken: 0.311 seconds
hive> show tables;
OK
airlines
Time taken: 0.05 seconds, Fetched: 1 row(s)
hive> describe airlines;
OK
AirlineCode          string
description          string
Time taken: 0.149 seconds, Fetched: 2 rows
hive> [ ]
```

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;

```

```

Terminal ① hduser@hduser-VirtualBox: ~
OK
+-----+-----+
| airlinecode | string |
| description | string |
+-----+-----+
Time taken: 0.10 seconds
hive> CREATE TABLE IF NOT EXISTS FLIGHTS(
  >   _FlightDate DATE,
  >   AirlineCode STRING,
  >   Description STRING,
  >   Origin STRING,
  >   Destination STRING,
  >   DepartureTime INT,
  >   ArrivalTime INT,
  >   ArrivalDelay INT,
  >   Airtime INT,
  >   Distance FLOAT
  > )
  > ROW FORMAT DELIMITED
  >   FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE
  > LOCATION '/user/hduser/flights_dir';
OK
Time taken: 0.12 seconds
hive> describe flights;
OK
+-----+-----+
| flightdate | date |
| airlinecode | string |
| flightnum | int |
| origin | string |
| destination | string |
| departuretime | int |
| arrivaltime | int |
| arrivaldelay | int |
| airtime | int |
| distance | float |
+-----+-----+
Time taken: 0.14 seconds, Fetched: 11 row(s)
hive> select * from flights LIMIT 10;
OK
+-----+-----+
| 2014-04-01 | 19885 | 1 | JFK | LAX | 854 | -6 | 1217 | 2 | 355 | 2475.0 |
| 19885 | 2 | LAX | JFK | 944 | 14 | 1736 | -29 | 269 | 2475.0 |
| 2014-04-01 | 19885 | 3 | JFK | LAX | 1224 | -6 | 1614 | 39 | 371 | 2475.0 |
| 2014-04-01 | 19885 | 4 | LAX | JFK | 1240 | 25 | 2028 | -27 | 264 | 2475.0 |
| 2014-04-01 | 19885 | 5 | DFW | HNL | 1380 | -5 | 1650 | 15 | 510 | 3784.0 |
| 2014-04-01 | 19885 | 6 | OGG | DFW | 1981 | 126 | 640 | 95 | 385 | 3711.0 |
| 2014-04-01 | 19885 | 7 | DFW | OGG | 1410 | 125 | 1743 | 138 | 497 | 3711.0 |
| 2014-04-01 | 19885 | 8 | HNL | DFW | 1659 | 4 | 458 | -22 | 398 | 3784.0 |
| 2014-04-01 | 19885 | 9 | JFK | LAX | 648 | -7 | 1029 | 19 | 365 | 2475.0 |
| 2014-04-01 | 19885 | 10 | LAX | JFK | 2156 | 21 | 556 | 1 | 265 | 2475.0 |
+-----+-----+
Time taken: 0.439 seconds, Fetched: 10 row(s)
hive>

```

Enable column names

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

```

Terminal ① hduser@hduser-VirtualBox: ~
OK
+-----+-----+
| departuretime | int |
| arrivaldelay | int |
| arrivaltime | int |
| arrivaldelay | int |
| airtime | int |
| distance | float |
+-----+-----+
Time taken: 0.14 seconds, Fetched: 11 row(s)
hive> select * from flights LIMIT 10;
OK
+-----+-----+
| 2014-04-01 | 19885 | 1 | JFK | LAX | 854 | -6 | 1217 | 2 | 355 | 2475.0 |
| 19885 | 2 | LAX | JFK | 944 | 14 | 1736 | -29 | 269 | 2475.0 |
| 2014-04-01 | 19885 | 3 | JFK | LAX | 1224 | -6 | 1614 | 39 | 371 | 2475.0 |
| 2014-04-01 | 19885 | 4 | LAX | JFK | 1240 | 25 | 2028 | -27 | 264 | 2475.0 |
| 2014-04-01 | 19885 | 5 | DFW | HNL | 1380 | -5 | 1650 | 15 | 510 | 3784.0 |
| 2014-04-01 | 19885 | 6 | OGG | DFW | 1981 | 126 | 640 | 95 | 385 | 3711.0 |
| 2014-04-01 | 19885 | 7 | DFW | OGG | 1410 | 125 | 1743 | 138 | 497 | 3711.0 |
| 2014-04-01 | 19885 | 8 | HNL | DFW | 1659 | 4 | 458 | -22 | 398 | 3784.0 |
| 2014-04-01 | 19885 | 9 | JFK | LAX | 648 | -7 | 1029 | 19 | 365 | 2475.0 |
| 2014-04-01 | 19885 | 10 | LAX | JFK | 2156 | 21 | 556 | 1 | 265 | 2475.0 |
+-----+-----+
Time taken: 0.439 seconds, Fetched: 10 row(s)
hive> set hive.cli.print.header=true;
hive> select * from airlines LIMIT 10;
OK
+-----+-----+
| airlines.airlinecode | airlines.description |
| 19831 | Mackay International Inc. |
| 19832 | Midwest Airlines Inc. |
| 19833 | Cochise Airlines Inc. |
| 19834 | Golden Gate Airlines Inc. |
| 19835 | Aeromexico Inc. |
| 19836 | Gol Caribe Central Airlines Co. |
| 19837 | Puerto Rico Intl Airlines |
| 19838 | At! America Inc. |
| 19839 | San Juan Express Inc. |
| 19840 | American Central Airlines |
+-----+-----+
Time taken: 0.145 seconds, Fetched: 10 row(s)
hive> select * from flights LIMIT 10;
OK
+-----+-----+
| flights.flightdate | flights.airlinecode | flights.distance | flights.flightnum | flights.origin | flights.destination | flights.departuretime | flights.departuredelay | flights.arrivaltime | f | |
| 2014-04-01 | 19885 | 1 | JFK | LAX | 854 | -6 | 1217 | 2 | 355 | 2475.0 |
| 2014-04-01 | 19885 | 2 | LAX | JFK | 944 | 14 | 1736 | -29 | 269 | 2475.0 |
| 2014-04-01 | 19885 | 3 | JFK | LAX | 1224 | -6 | 1614 | 39 | 371 | 2475.0 |
| 2014-04-01 | 19885 | 4 | LAX | JFK | 1240 | 25 | 2028 | -27 | 264 | 2475.0 |
| 2014-04-01 | 19885 | 5 | DFW | HNL | 1380 | -5 | 1650 | 15 | 510 | 3784.0 |
| 2014-04-01 | 19885 | 6 | OGG | DFW | 1981 | 126 | 640 | 95 | 385 | 3711.0 |
| 2014-04-01 | 19885 | 7 | DFW | OGG | 1410 | 125 | 1743 | 138 | 497 | 3711.0 |
| 2014-04-01 | 19885 | 8 | HNL | DFW | 1659 | 4 | 458 | -22 | 398 | 3784.0 |
| 2014-04-01 | 19885 | 9 | JFK | LAX | 648 | -7 | 1029 | 19 | 365 | 2475.0 |
| 2014-04-01 | 19885 | 10 | LAX | JFK | 2156 | 21 | 556 | 1 | 265 | 2475.0 |
+-----+-----+
Time taken: 0.128 seconds, Fetched: 10 row(s)
hive>

```

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;
```

```

Terminal ④ hduser@hduser-VirtualBox:~ 
hive> select COUNT(arrivaldelay) AS FLIGHTS_ARRIVED_DELAYED from flights where arrivaldelay>0;
Query ID = hduser_20230116180954_51baff44-214e-42cd-a3d4-ddd3ba2c621d
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 mapred.job.reduces=<number>
Starting Job = job_1673864896410_0001, Tracking URL: http://hduser-VirtualBox:8080/proxy/application_1673864896410_0001/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1673864896410_0001
Hadoop Job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-16 18:10:05,559 Stage-1 map = 0%, reduce = 0%
2023-01-16 18:10:05,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.07 sec
2023-01-16 18:10:19,261 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.13 sec
MapReduce Total cumulative CPU time: 3 seconds 130 msec
Ended Job = job_1673864896410_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.13 sec HDFS Read: 31667837 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 130 msec
OK
flights_arrived_delayed
185651
Time taken: 0.128 seconds, Fetched: 10 row(s)
hive> []

```

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;

```

Terminal ④ hduser@hduser-VirtualBox:~ 
hive> select COUNT(departuredelay) AS FLIGHTS_DEPARTED_DELAYED from flights where departuredelay>0;
Query ID = hduser_20230116181214_80481658-7b51-42d4-aee5-6576ca833ef
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 mapred.job.reduces=<number>
Starting Job = job_1673864896410_0001, Tracking URL: http://hduser-VirtualBox:8080/proxy/application_1673864896410_0001/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1673864896410_0001
Hadoop Job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-16 18:16:59,559 Stage-1 map = 0%, reduce = 0%
2023-01-16 18:16:59,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.07 sec
2023-01-16 18:18:11,992 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.13 sec
MapReduce Total cumulative CPU time: 3 seconds 130 msec
Ended Job = job_1673864896410_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.13 sec HDFS Read: 31667837 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 130 msec
OK
flights_departed_delayed
179015
Time taken: 25.582 seconds, Fetched: 1 row(s)
hive> []

```

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;

```

Terminal ② huser@huser-VirtualBox:-
FAILED: SemanticException [Error 10004]: line 1:65 Invalid table alias or column reference 'airlinecode': (possible column names are: )
    at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.analyzeFromQuery(SemanticAnalyzer.java:145)
    at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.analyzeFromQuery(SemanticAnalyzer.java:140)
FAILED: SemanticException line 0:1 Invalid table alias or column reference 'ASC': (possible column names are: airlinecode, avg_dist_by_flight)
hive> select airlinecode, avg(distance) AS AVG_DIST_BY_FLIGHT FROM flights GROUP BY airlinecode;
Query ID = hduser_202301161818001_19713c31-5f94-475d-bd28-9712de80b532
Total jobs = 1
Launching Job 1 out of 1
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_1673864896410_0003, Tracking URL = http://hduser-VirtualBox:8080/proxy/application_1673864896410_0003/
KILL Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1673864896410_0003
Hadoop Job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-16 18:18:19,688 Stage-1 map = 0%, reduce = 0%
2023-01-16 18:18:19,809 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.7 sec
2023-01-16 18:18:19,820 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.59 sec
MapReduce Total cumulative CPU time: 2 seconds 590 msec
Ended Job = job_1673864896410_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1, Reduce: 1, Cumulative CPU: 2.59 sec HDFS Read: 31668803 HDFS Write: 339 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 590 msec
OK
+-----+
| airlinecode | avg_dist_by_flight |
+-----+
| 19393     | 712.777210763348   |
| 19690     | 563.7663721325493  |
| 19700     | 377.653721325493   |
| 19885     | 1076.677334418348   |
| 19930     | 1208.03854967885   |
| 19977     | 1315.612042652325   |
| 20384     | 467.265486737888   |
| 20385     | 883.0246016459464   |
| 20386     | 482.1142919899126   |
| 20389     | 467.265486737888   |
| 20409     | 1076.9169776571512   |
| 20436     | 842.2389261744967   |
| 20437     | 887.97451917873   |
| 21171     | 1441.5356995460174   |
Time taken: 20.007 seconds, Fetched: 14 row(s)
hive> select airlinecode, avg(distance) AS AVG_DIST_BY_FLIGHT FROM flights GROUP BY airlinecode ORDER BY airlinecode DESC;
Query ID = hduser_202301161818044_afb143bf-9219-49d4-96c4-476c46edbed
Total jobs = 1
Launching Job 1 out of 1
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_1673864896410_0004, Tracking URL = http://hduser-VirtualBox:8080/proxy/application_1673864896410_0004/
KILL Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1673864896410_0004
Hadoop Job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-16 18:18:47,709 Stage-1 map = 0%, reduce = 0%
2023-01-16 18:18:47,729 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.65 sec
2023-01-16 18:19:00,042 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.59 sec
MapReduce Total cumulative CPU time: 2 seconds 590 msec
Ended Job = job_1673864896410_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1, Reduce: 1, Cumulative CPU: 2.59 sec HDFS Read: 31668230 HDFS Write: 530 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 590 msec
OK
+-----+
| airlinecode | avg_dist_by_flight |
+-----+
| 21171     | 1441.5356995460174   |
| 20437     | 887.97451917873   |
| 20389     | 467.265486737888   |
| 20409     | 1076.9169776571512   |
| 20385     | 883.0246016459464   |
| 20384     | 464.1143676774035   |
| 19977     | 1315.612042652329   |
| 19390     | 1208.03854967885   |
| 19885     | 1076.677334418348   |
| 19790     | 871.5971268052884   |
| 19900     | 363.7663721325493   |
| 19393     | 712.777210763348   |
Time taken: 44.997 seconds, Fetched: 14 row(s)
hive> 
```

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

```

Terminal ② huser@huser-VirtualBox:-
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_1673864896410_0004, Tracking URL = http://hduser-VirtualBox:8080/proxy/application_1673864896410_0004/
KILL Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job_1673864896410_0004
Hadoop Job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-16 18:19:12,912 Stage-2 map = 0%, reduce = 0%
2023-01-16 18:19:17,978 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.73 sec
2023-01-16 18:19:24,111 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.52 sec
MapReduce Total cumulative CPU time: 1 seconds 520 msec
Ended Job = job_1673864896410_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1, Reduce: 1, Cumulative CPU: 2.59 sec HDFS Read: 31668230 HDFS Write: 530 SUCCESS
Stage-Stage-2: Map: 1, Reduce: 1, Cumulative CPU: 1.52 sec HDFS Read: 5120 HDFS Write: 339 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 110 msec
OK
+-----+
| airlinecode | avg_dist_by_flight |
+-----+
| 21171     | 1441.5356995460174   |
| 20437     | 887.97451917873   |
| 20389     | 467.265486737888   |
| 20409     | 1076.9169776571512   |
| 20385     | 883.0246016459464   |
| 20384     | 464.1143676774035   |
| 19977     | 1315.612042652329   |
| 19390     | 1208.03854967885   |
| 19885     | 1076.677334418348   |
| 19790     | 871.5971268052884   |
| 19900     | 363.7663721325493   |
| 19393     | 712.777210763348   |
Time taken: 44.997 seconds, Fetched: 14 row(s)
hive> 
```

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

Hive> SELECT a.airlinecode,a.description,
f.fligdate,
f.flighthnum,
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.flighthdate,
f.flighthnum,
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.'
```