Analysis of Airport Data Using Hive & Pig

Case Study

19CSE357 – Big Data Analytics

Date: *February 23, 2022*

Group Details:

S. No	Name of the Student	Roll No.
1.	DIVESH KOSURI	CB.EN.U4CSE19422
2.	PENUGONDA KOUSHIK	CB.EN.U4CSE19449
3.	RAVELLA ABHINAV	CB.EN.U4CSE19453
4.	SINGADI SHANTHAN REDDY	CB.EN.U4CSE19459

Dataset Description

The main aim of the dataset is to develop a model for the airline data to provide a platform for new analytics based on the following queries as the problem faced is the existing has the ability to analyze limited data from the following databases

In our case study we are dealing with 3 different datasets named airports_mod, Final_airlines, routes

Fields:

Airports_mod:

- **Sample:** `Goroka,Goroka,Papua New Guinea,GKA,AYGA,-6.081689,145.391881,5282,10,U,Pacific/Port_Moresby`
- Dataset contains mainly unique Airport ID, Name of the airport, City of the respective airport, Country, 3-letter IATA code, Latitude & Longitude, Altitude, Timezone

Final_Airlines:

- **Sample:** `2,135 Airways,\N,,GNL,GENERAL,United States,N`
- In this dataset it contains ID, Name of airline, Shortcut of airline, IATA, ICAO, Callsign, Country

Routes:

- **Sample:** `2B,410,AER,2965,KZN,2990,,0,CR2`
- This dataset contains mainly 3-letter ICAO code, Airline ID, Source airport ID&Code, Destination ID & Code, Halts

Outcome:

We tried to explore detailed analysis on airline datasets such as listing airports operations, list of airlines having no halts etc., Here we mainly focussed on the processing of big datasets using hive component of Hadoop ecosystem in distributed environment.

At last, it will be useful in accessing and processing their user queries.

Loading the Dataset:

Hive:

```
hduser@abhinav:~$ hadoop fs -mkdir /casestudy
2022-03-21 15:23:13,807 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@abhinav:~$ hadoop fs -put /home/hduser/Downloads/airports_mod.dat /casestudy
2022-03-21 15:23:24,228 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@abhinav:~$ hadoop fs -put /home/hduser/Downloads/Final_airlines /casestudy
2022-03-21 15:23:34,139 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@abhinav:~$ hadoop fs -put /home/hduser/Downloads/routes.dat /casestudy
2022-03-21 15:23:45,366 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@abhinav:~$
```

/casestudy Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	308.83 KB	Mar 21 15:23	1	128 MB	Final_airlines
-rw-r--r--	hduser	supergroup	722.18 KB	Mar 21 15:23	1	128 MB	airports_mod.dat
-rw-r--r--	hduser	supergroup	2.27 MB	Mar 21 15:23	1	128 MB	routes.dat

Showing 1 to 3 of 3 entries Previous 1 Next

Pig:

QUERY 1 : Load dataset andusing pig. grunt> airline = LOAD '/user/hduser/proj/airlines.txt' using PigStorage(',') as (airlineID:charArray,airline_name:charArray, airline_alias:charArray, airline_iata:charArray, airline_icao:charArray,callsign:charArray,territory:charArray, active:charArray);

```
hduser@DiveshKosuri:~$ pig
2022-03-23 11:57:20,621 [main] INFO  org.apache.pig.Main - Apache Pig version 0.11.1 (r1459641) compiled Mar 22 2013, 02:13:53
2022-03-23 11:57:20,622 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hduser/pig_1648016840620.log
2022-03-23 11:57:20,645 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hduser/.pigbootup not found
2022-03-23 11:57:21,107 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2022-03-23 11:57:21,120 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-23 11:57:21,413 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-03-23 11:57:22,397 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> airline = LOAD '/home/hduser/Downloads/Final_airlines' using PigStorage(',') as (airlineID:charArray,airline_name:charArray, airline_alias:charArray, airline_iata:charArray, airline_icao:charArray
,callsign:charArray,territory:charArray, active:charArray);
```

Query 2:

Filter all the airlines with territory egypt

```
egy = filter airline by territory == 'Egypt';
dump egy;
```

QUERY 3: GET FIRST 3 ACTIVE AIRLINES WITH NAMES IN UPPER CASE. grunt> up_ar = foreach airline generate UPPER(airline_name),active; grunt> a_3 = limit up_ar 3; grunt> dump a_3;

Query 4:

A = filter filter_airlines by Active == 'Y';
dump A;

Query 5:

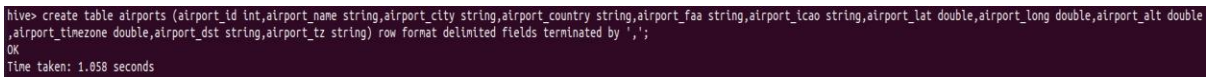
1. active_airlines = filter filter_airlines by active == 'Y';
2. active_airlines_usa = filter active_airlines by territory == 'United States';
3. find_active_airline_names_in_usa = foreach active_airlines_usa generate airline_name;

Queries:

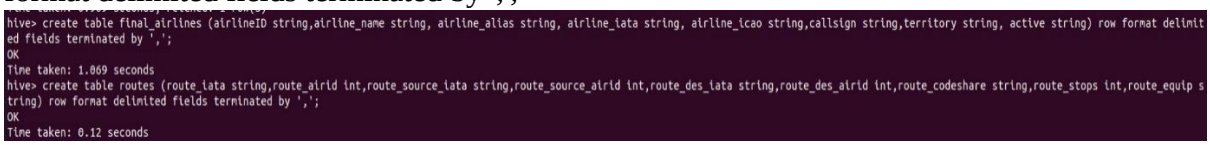
Hive:

1. **Creating table airport for airports_mod dataset:**

   create table airports (airport_id int,airport_name string,airport_city string,airport_country string,airport_faa string,airport_icao string,airport_lat double,airport_long double,airport_alt double,airport_timezone double,airport_dst string,airport_tz string) row format delimited fields terminated by ',';



2. **Creating table final airlines for Final_airlines :**
   create table final_airlines (airlineID string,airline_name string, airline_alias string, airline_iata string, airline_icao string,callsign string,territory string, active string) row format delimited fields terminated by ',';



3. **Creating table route for routes.dat**:
   create table routes (route_iata string,route_airid int,route_source_iata string,route_source_airid int,route_des_iata string,route_des_airid int,route_codeshare string,route_stops int,route_equip string) row format delimited fields terminated by ',';

```
hive> show tables;
OK
airports
final_airlines
routes
Time taken: 0.089 seconds, Fetched: 3 row(s)
hive>
```

4. **loading data into airport table**
   load data inpath '/airports_mod.dat' into table airports;

```
hive> load data inpath '/casestudy/airports_mod.dat' into table airports;
Loading data to table default.airports
OK
Time taken: 1.27 seconds
hive>
```

5. **loading data into final airlines table**
   load data inpath '/Final_airlines' into table final_airlines;

6. **loading data into route table**
   load data inpath '/routes.dat' into table routes;

```
hive> load data inpath '/casestudy/airports_mod.dat' into table airports;
Loading data to table default.airports
OK
Time taken: 1.27 seconds
hive> load data inpath '/casestudy/Final_airlines' into table airports;
Loading data to table default.airports
OK
Time taken: 0.363 seconds
hive> load data inpath '/casestudy/routes.dat' into table airports;
Loading data to table default.airports
OK
Time taken: 0.344 seconds
hive>
```

DIVESH KOSURI – CB.EN.U4CSE19422

Hive:

1)Query should return data of all the airlines that are present in United States territory and are active.

select * from final_airlines where territory="United States" and active="Y";

```
Time taken: 5.801 seconds, Fetched: 1 row(s)
hive> select * from final_airlines where territory="United States" and active="Y";
OK
10      40-Mile Air     NULL    Q5      MLA     MILE-AIR        United States   Y
22      Aloha Airlines  NULL    AQ      AAH     ALOHA   United States   Y
24      American Airlines       NULL    AA      AAL     AMERICAN        United States   Y
35      Allegiant Air   NULL    G4      AAY     ALLEGIANT       United States   Y
109     Alaska Central Express  NULL    KO      AER     ACE AIR United States   Y
149     Air Cargo Carriers      NULL    2Q      SNC     NIGHT CARGO     United States   Y
210     Airlift International   NULL            AIR     AIRLIFT United States   Y
281     America West Airlines   NULL    HP      AWE     CACTUS  United States   Y
282     Air Wisconsin   NULL    ZW      AWI     AIR WISCONSIN   United States   Y
287     Allegheny Commuter Airlines     NULL            ALO     ALLEGHENY       United States   Y
295     Air Sunshine    NULL            RSI     AIR SUNSHINE    United States   Y
315     ATA Airlines    NULL            AMT     AMTRAN  United States   Y
397     Arrow Air       NULL    JW      APW     BIG A   United States   Y
452     Atlantic Southeast Airlines     NULL    EV      ASQ     ACEY    United States   Y
659     American Eagle Airlines NULL    MQ      EGF     EAGLE FLIGHT    United States   Y
792     Access Air      NULL    ZA      CYD     CYCLONE United States   Y
882     Air Florida     NULL    QH      FLZ     AIR FLORIDA     United States   Y
928     Atlas Air       NULL    5Y      GTI     GIANT   United States   Y
1316    AirTran Airways NULL    FL      TRS     CITRUS  United States   Y
1442    Bemidji Airlines        NULL    CH      BMJ     BEMIDJI United States   Y
1472    Bering Air      NULL    8E      BRG     BERING AIR      United States   Y
1629    Cape Air        NULL    9K      KAP     CAIR    United States   Y
1739    Chautauqua Airlines     NULL    RP      CHQ     CHAUTAUQUA      United States   Y
1814    Coastal Air     NULL    DQ              U.S. Virgin Islands     United States   Y
1821    Colgan Air      NULL    9L      CJC     COLGAN  United States   Y
1828    Comair  NULL    OH      COM     COMAIR  United States   Y
1843    CommutAir       NULL    C5      UCA     COMMUTAIR       United States   Y
1860    Compass Airlines        NULL    CP      CPZ     Compass Rose    United States   Y
1881    Continental Airlines    NULL    CO      COA     CONTINENTAL     United States   Y
1883    Continental Express     NULL    CO              JETLINK United States   Y
1884    Continental Micronesia  NULL    CS      CMI     AIR MIKE        United States   Y
1931    Crown Airways   NULL            CRO     CROWN AIRWAYS   United States   Y
```

Airline data required is dependent on two attributes territory and active status . So based on these two attributes with “where” clause we can get the expect output. The output data which is generated has active status=”Y” and territory=”United States”.

2)Query should return territories with maximum airlines.

select count(airlineID),territory from final_airlines group by territory order by count(airlineID) DESC;

```
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 632486 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 632486 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1080    United States
439     Mexico
407     United Kingdom
318     Canada
230     Russia
166     Spain
131     Germany
119     France
93      Australia
91      South Africa
90      Italy
89      Ukraine
85      Nigeria
79      Kazakhstan
70      China
70      Sweden
60      Switzerland
58      Brazil
52      Netherlands
50      Austria
```

3) Query (Partition)

load data inpath '/user/hduser/proj/airlines.txt' into table final_airlines_t partition(territory="United States");

```
hive> load data inpath '/user/hduser/proj/airlines.txt' into table final_airlines_t partition(territory="United States");
Loading data to table default.final_airlines_t partition (territory=United States)
OK
Time taken: 4.799 seconds
hive> select * from final_airlines_t limit 10;
OK
1       Private Flight  NULL    -       N/A                     United States
2       135 Airways     NULL            GNL     GENERAL United States   United States
3       1Time Airline   NULL    1T      RNX     NEXTIME South Africa    United States
4       2 Sqn No 1 Elementary Flying Training School    NULL            WYT             United Kingdom  United States
5       213 Flight Unit NULL            TFU             Russia  United States
6       223 Flight Unit State Airline   NULL            CHD     CHKALOVSK-AVIA  Russia  United States
7       224th Flight Unit       NULL            TTF     CARGO UNIT      Russia  United States
8       247 Jet Ltd     NULL            TWF     CLOUD RUNNER    United Kingdom  United States
9       3D Aviation     NULL            SEC     SECUREX United States   United States
10      40-Mile Air     NULL    Q5      MLA     MILE-AIR        United States   United States
```

RAVELLA ABHINAV – CB.EN.U4CSE19453

Queries:

Hive:

1. Find all the airlines that are active and have an alias names

   Query:
   create table alias_not_null_airlines as SELECT * FROM final_airlines WHERE airline_alias IS NOT NULL AND active="Y";

   Result :

```
hive> create table alias_not_null_airlines as SELECT * FROM final_airlines WHERE airline_alias IS NOT NULL AND active="Y";
Query ID = hduser_20220322100631_7b70eb96-cde3-4913-a9cd-52c2a86c6138
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2022-03-22 10:06:34,839 Stage-1 map = 0%,  reduce = 0%
2022-03-22 10:06:35,878 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local216938443_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/.hive-staging_hive_2022-03-22_10-06-31_052_1734083296723987011-1/-ext-10002
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/alias_not_null_airlines
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 10503019 HDFS Write: 21692 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 5.629 seconds
hive>
```

- **Querying the first 10 rows of the resulted table:**

  Query: SELECT * FROM alias_not_null_airlines LIMIT 10;

```
hive> SELECT * FROM alias_not_null_airlines LIMIT 10;
OK
324     All Nippon Airways      ANA All Nippon Airways  NH      ANA     ALL NIPPON      Japan   Y
576     AirAsia Air Asia        AK      AXM     ASIAN EXPRESS   Malaysia        Y
641     Rossiya-Russian Airlines        Pulkovo Aviation Enterprise     FV      SDM     PULKOVO Russia  Y
1437    bmi     bmi British Midland     BD      BMA     MIDLAND United Kingdom  Y
1531    Brussels Airlines       SN Brussels Airlines    SN      DAT     BEE-LINE        Belgium Y
1879    Contact Air     Contactair      C3      KIS     CONTACTAIR      Germany Y
1946    Czech Airlines  CSA Czech Airlines      OK      CSA     CSA-LINES       Czech Republic  Y
2183    Emirates        Emirates Airlines       EK      UAE     EMIRATES        United Arab Emirates    Y
2297    easyJet EasyJet Airline U2      EZY     EASY    United Kingdom  Y
2417    AirAsia X       FlyAsianXpress  D7      XAX     XANADU  Malaysia        Y
Time taken: 0.411 seconds, Fetched: 10 row(s)
hive>
```

Explaination:

Job is to find list of airlines with alias names and are still operating (Active). This can be achieved by querying using ‘WHERE’ , ‘IS NOT NULL’ and ‘AND’ keywords. In the dataset, all the airlines that has no alias names have ‘NULL’ as the value in their respective cells. So ‘IS’ ‘NOT’ ‘NULL’ keywords are to be used to fetch all the values rows that have alias names and active status can be directly done using ‘WHERE’ clause.

2. Find the count of airlines that choose to have routes with 1 stop.

Query: select count(route_airid) from routes where route_stops like "%1%"

Output:

```
hive> select count(route_airid) from routes where route_stops like "%1%";
Query ID = hduser_20220322101423_0447074f-e6cc-4088-9903-196f3f8ede71
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>
Job running in-process (local Hadoop)
2022-03-22 10:14:25,751 Stage-1 map = 100%,  reduce = 0%
2022-03-22 10:14:26,790 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1580320474_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 25765422 HDFS Write: 43384 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
11
Time taken: 3.636 seconds, Fetched: 1 row(s)
hive>
```

Explanation:
Job here is to find the total count of airlines that has one stop in its routes. So we are to query on routes table we already created using one of the aggregate function “count”.
1. select all the route id’s which have their no of stops equal to 1
2. Add the aggregate function “count” to count the no of ids that are resulted as a result of first query.

3. Find all airports in the world which lie at an altitude greater than 5000 ft.

Query: create table high_alt_airports as select * from airports where airport_alt > 5000;

Result:

```
hive> create table high_alt_airports as select * from airports where airport_alt > 5000;
Query ID = hduser_20220322101742_056c5cee-fd82-4137-9922-1ebec2490a38
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2022-03-22 10:17:45,255 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local2075807615_0003
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/.hive-staging_hive_2022-03-22_10-17-42_629_8769889742244920012-1/-ext-10002
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/high_alt_airports
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 16313974 HDFS Write: 52226 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 3.178 seconds
hive>
```

Subquery:
select * from high_alt_airports limit 10;

```
hive> select * from high_alt_airports limit 10;
OK
1       Goroka  Goroka  Papua New Guinea        GKA     AYGA    -6.081689       145.391881      5282.0  10.0    U       Pacific/Port_Moresby
3       Mount Hagen     Mount Hagen     Papua New Guinea        HGU     AYMH    -5.826789       144.295861      5388.0  10.0    U       Pacific/Port_Moresby
792     Brakpan Brakpan South Africa            FABB    -26.23865       28.301769       5300.0  2.0     U       Africa/Johannesburg
795     Bethlehem       Bethlehem       South Africa            FABM    -28.248392      28.336125       5561.0  2.0     U       Africa/Johannesburg
801     Ermelo  Ermelo  South Africa            FAEO    -26.495644      29.979764       5700.0  2.0     U       Africa/Johannesburg
802     Ficksburg Sentraoes     Ficksburg       South Africa            FAFB    -28.823119      27.9089 5315.0  2.0     U       Africa/Johannesburg
803     Grand Central   Johannesburg    South Africa    GCJ     FAGC    -25.986267      28.140061       5325.0  2.0     U       Africa/Johannesburg
810     Harrismith      Harrismith      South Africa            FAHR    -28.235072      29.106206       5585.0  2.0     U       Africa/Johannesburg
813     Johannesburg Intl       Johannesburg    South Africa    JNB     FAJS    -26.139166      28.246  5558.0  2.0     U       Africa/Johannesburg
816     Krugersdorp     Krugersdorp     South Africa            FAKR    -26.080978      27.725667       5499.0  2.0     U       Africa/Johannesburg
Time taken: 0.44 seconds, Fetched: 10 row(s)
hive>
```

Explaination:
Job is to find out the list of all the airports at higher altitudes (alt > 5000 ft) we use a binary operator “>” to select all the airports that have their airport alt > 5000.

1. We first use the select clause to find all the airports above air_alt > 5000, create a new table high_alt_airports and store the result of above query in that new table.
2. Now we query the table for 10 airports with altitude above 5000 using ‘LIMIT’ keyword.

P. Sai Koushik

1. **Find list of Airports operating in the Country India;**

create table india_opert_airport as select * from airports where airport_country LIKE '%India%';

```
hive> create table india_opert_airport as select * from airports where airport_country LIKE '%India%';
Query ID = hduser_20220322103609_79bca6ed-adca-446e-b81f-3ada52ca92cf
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2022-03-22 10:36:13,652 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local294702227_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/.hive-staging_hive_2022-03-22_10-36-10_002_6773573062564491529-1/-ext-10002
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/india_opert_airport
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 5649832 HDFS Write: 23793 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 3.937 seconds
hive>
```

2. select * from india_opert_airport limit 10; (to show only first 10 values due to large nos of data)

```
hive> select * from india_opert_airport limit 10;
OK
895     Diego Garcia Nsf        Diego Garcia Island     British Indian Ocean Territory          FJDG    -7.313267       72.411089       9.0     6.0     U       I
ndian/Chagos
2994    Ahmedabad       Ahmedabad       India   AMD     VAAH    23.077242       72.63465        189.0   5.5     N       Asia/Calcutta
2995    Akola   Akola   India   AKD     VAAK    20.699006       77.058628       999.0   5.5     N       Asia/Calcutta
2996    Aurangabad      Aurangabad      India   IXU     VAAU    19.862728       75.398114       1911.0  5.5     N       Asia/Calcutta
2997    Chhatrapati Shivaji Intl        Mumbai  India   BOM     VABB    19.088686       72.867919       37.0    5.5     N       Asia/Calcutta
2998    Bilaspur        Bilaspur        India   PAB     VABI    21.9884 82.110983       899.0   5.5     N       Asia/Calcutta
2999    Bhuj    Bhuj    India   BHJ     VABJ    23.287828       69.670147       268.0   5.5     N       Asia/Calcutta
3000    Belgaum Belgaum India   IXG     VABM    15.859286       74.618292       2487.0  5.5     N       Asia/Calcutta
3001    Vadodara        Baroda  India   BDQ     VABO    22.336164       73.226289       129.0   5.5     N       Asia/Calcutta
3002    Bhopal  Bhopal  India   BHO     VABP    23.287467       77.337375       1719.0  5.5     N       Asia/Calcutta
Time taken: 0.146 seconds, Fetched: 10 row(s)
hive>
```

Here we display the first 10 airports operating in India.

3. **Find the list of Airlines having zero stops**

create table stop as select * from routes where route_stops LIKE '%0';

```
hive> create table stop as select * from routes where route_stops LIKE '%0';
Query ID = hduser_20220322103810_16734b74-46df-46ec-9137-7b0f724b97d2
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2022-03-22 10:38:11,757 Stage-1 map = 0%,  reduce = 0%
2022-03-22 10:38:12,763 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local562485070_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/.hive-staging_hive_2022-03-22_10-38-10_093_5663754051071849408-1/-ext-10002
Moving data to directory hdfs://localhost:54310/user/hive/warehouse/stop
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 8029520 HDFS Write: 2331356 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.869 seconds
hive>
```

4. **select * from stop limit 10;**

```
hive> select * from stop limit 10;
OK
2B      410     AER     2965    KZN     2990            0       CR2
2B      410     ASF     2966    KZN     2990            0       CR2
2B      410     ASF     2966    MRV     2962            0       CR2
2B      410     CEK     2968    KZN     2990            0       CR2
2B      410     CEK     2968    OVB     4078            0       CR2
2B      410     DME     4029    KZN     2990            0       CR2
2B      410     DME     4029    NBC     6969            0       CR2
2B      410     DME     4029    TGK     NULL            0       CR2
2B      410     DME     4029    UUA     6160            0       CR2
2B      410     EGO     6156    KGD     2952            0       CR2
Time taken: 0.17 seconds, Fetched: 10 row(s)
hive>
```

SHANTHAN REDDY – CB.EN.U4CSE19459

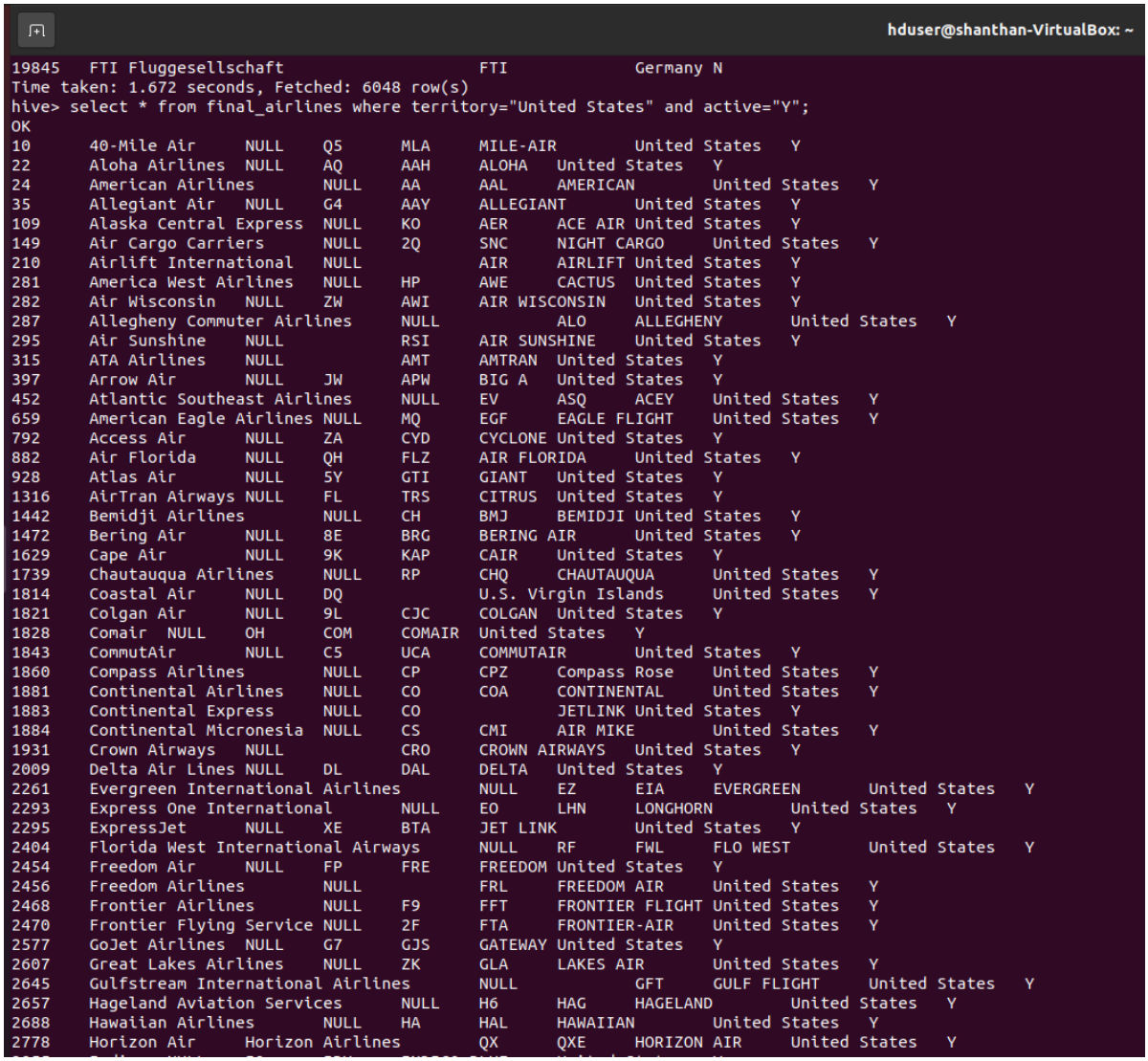
HIVE QUERIES:

1. select * from final airlines where territory="United States" and active="Y";

   Explain: Query should return data of all the airlines that are present in United States territory and are active.

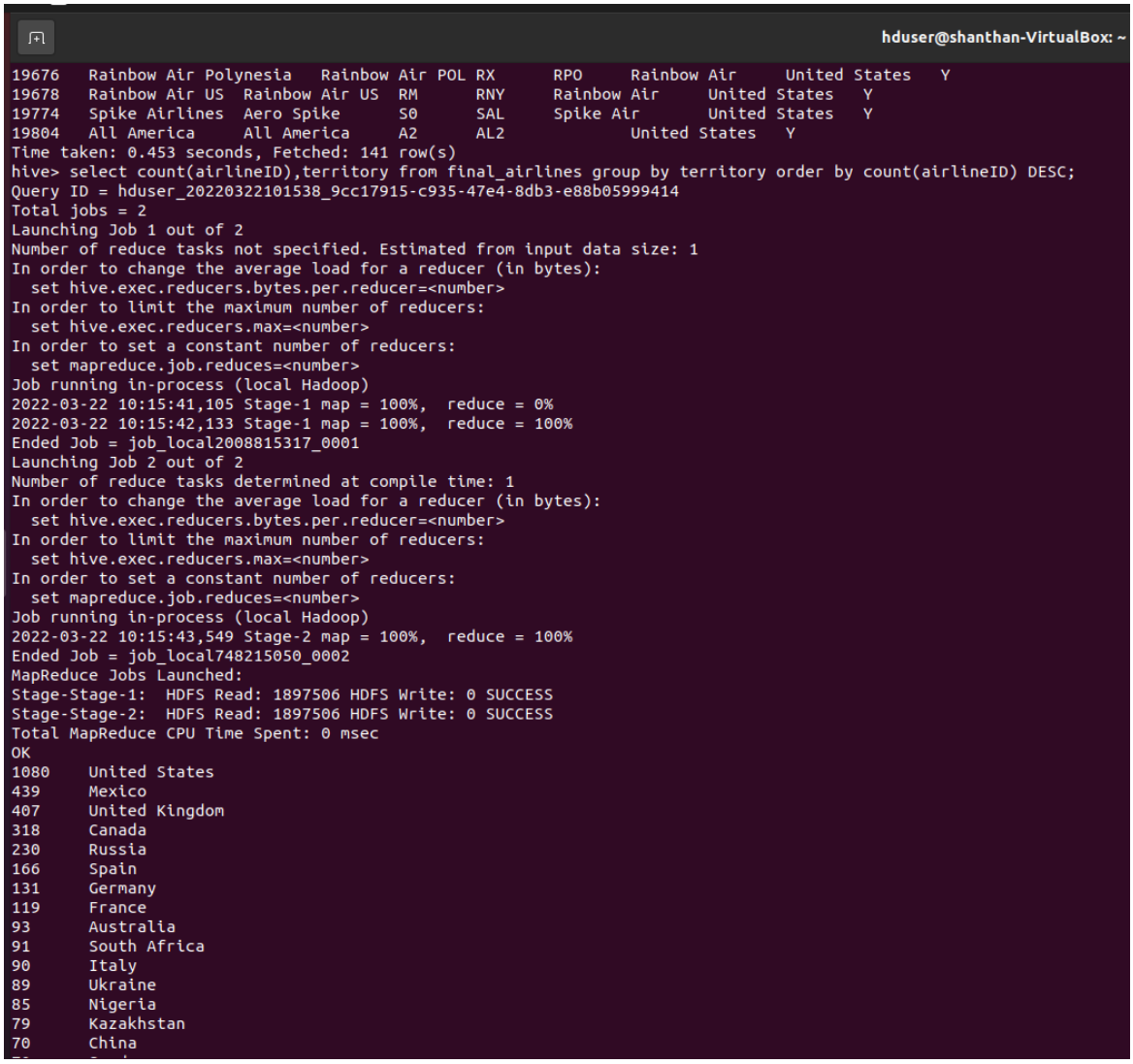
   Airline data required is dependent on two attributes, territory, and active status. So based on these two attributes with “where” clause we can get the expect output.

   The output data which is generated has active status=” Y” and territory=” United States.”

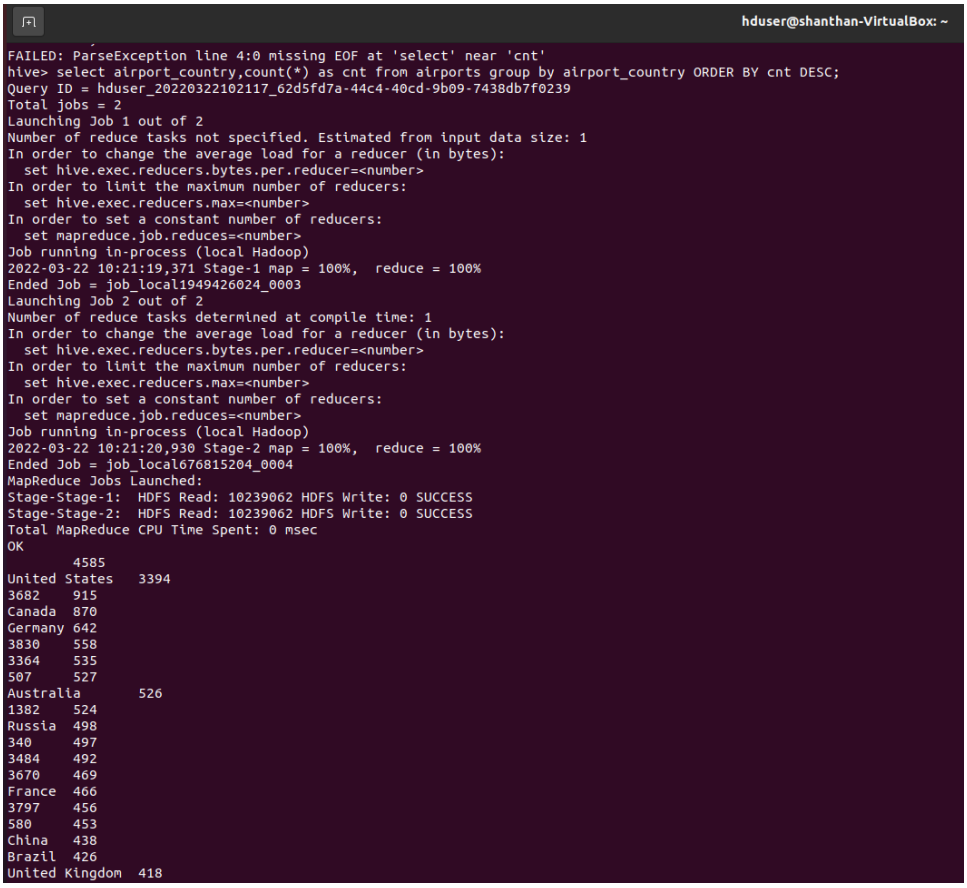


2. select count(airlineID),territory from final_airlines group by territory order by count(airlineID) DESC;

   Explain: To find where most people depend on airways for travel ordering territories based on total number of airlines. The query orders the airlines from highest to least.



4. select airport_country,count(*) as cnt from airports group by airport_country ORDER BY cnt DESC;


hduser@shanthan-VirtualBox: ~
FAILED: ParseException line 4:0 missing EOF at 'select' near 'cnt'
hive> select airport_country,count(*) as cnt from airports group by airport_country ORDER BY cnt DESC;
Query ID = hduser_20220322102117_62d5fd7a-44c4-40cd-9b09-7438db7f0239
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>
Job running in-process (local Hadoop)
2022-03-22 10:21:19,371 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1949426024_0003
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>
Job running in-process (local Hadoop)
2022-03-22 10:21:20,930 Stage-2 map = 100%, reduce = 100%
Ended Job = job_local676815204_0004
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 10239062 HDFS Write: 0 SUCCESS
Stage-Stage-2: HDFS Read: 10239062 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
4585
United States 3394
3682 915
Canada 870
Germany 642
3830 558
3364 535
507 527
Australia 526
1382 524
Russia 498
340 497
3484 492
3670 469
France 466
3797 456
580 453
China 438
Brazil 426
United Kingdom 418