**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. | KOSURI DIVESH | 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 analyse 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

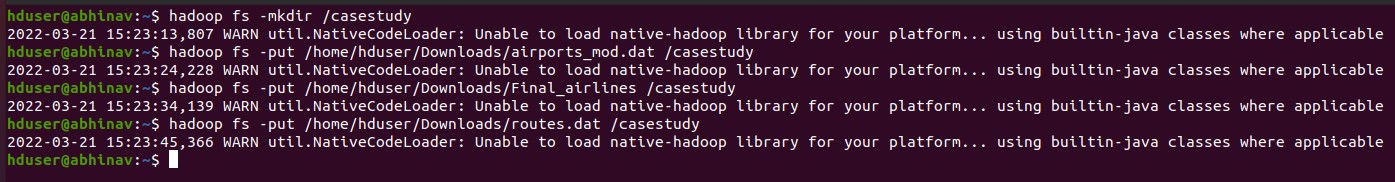
**Outcome:**

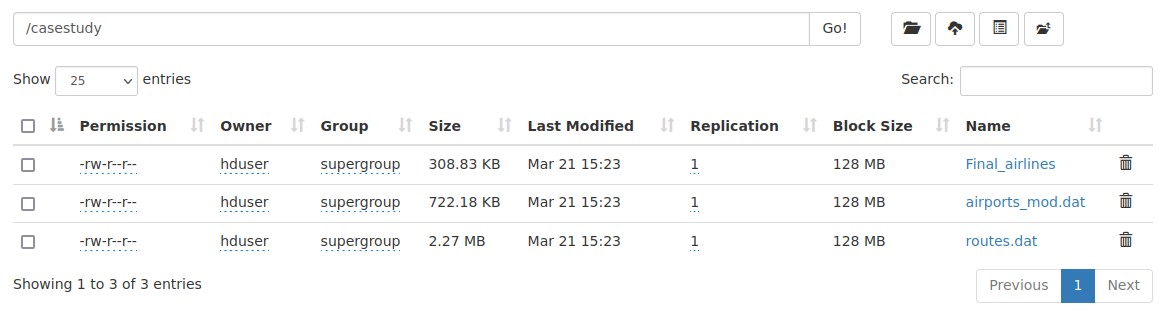
We tried to explore detailed analysis on airline datasets such as listing airports operations, list of airlines having no halts etc., Here we mainly focussed on the processing of big datasets using hive component of Hadoop ecosystem in distributed environment.

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

**Loading the Dataset:**

**Hive:**

****

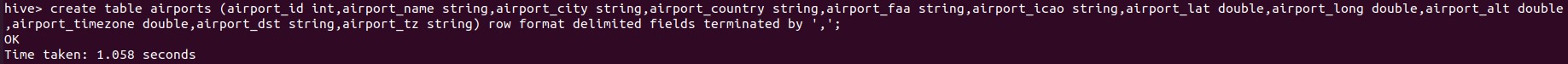
****

**Queries:**

**Hive:**

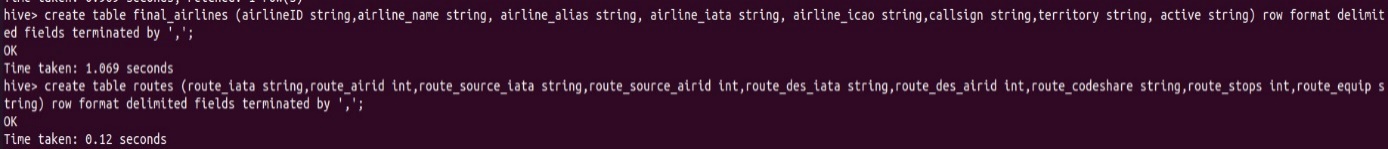
1. **Creating table airport for airports\_mod dataset:**

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 ',';

****

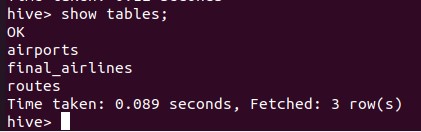
1. **Creating table final airlines for Final\_airlines :**

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 ',';



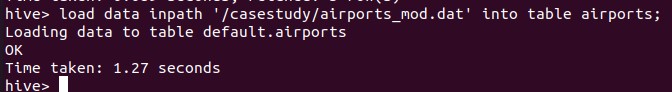
1. **Creating table route for routes.dat**:

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 ',';



1. **loading data into airport table**

load data inpath '/airports\_mod.dat' into table airports;

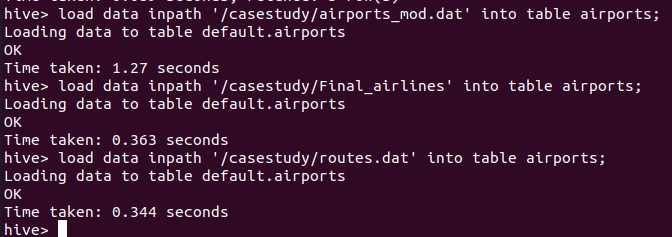


1. **loading data into final airlines table**

load data inpath '/Final\_airlines' into table final\_airlines;

1. **loading data into route table**

load data inpath '/routes.dat' into table routes;



**RAVELLA ABHINAV – CB.EN.U4CSE19453**

# **Queries:**

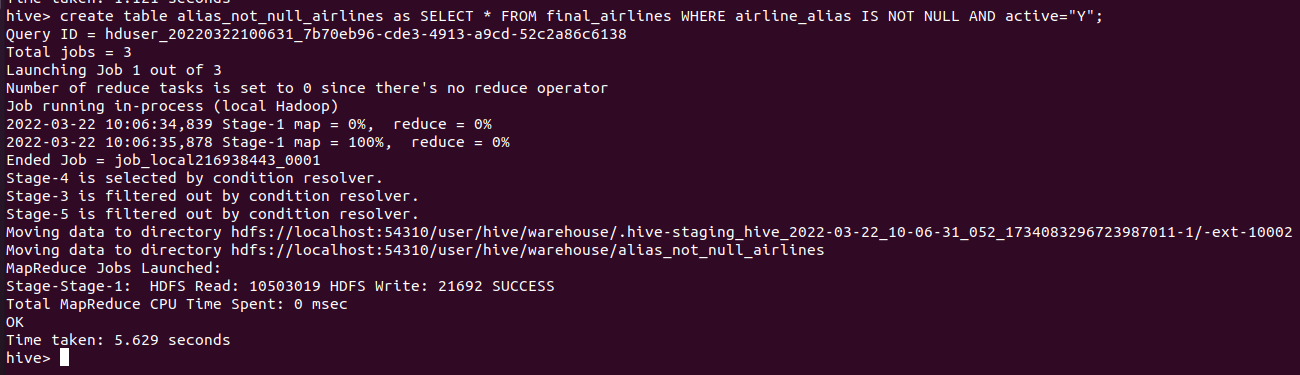
## **Hive:**

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 :**

****

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

**Query:** SELECT \* FROM alias\_not\_null\_airlines LIMIT 10;

****

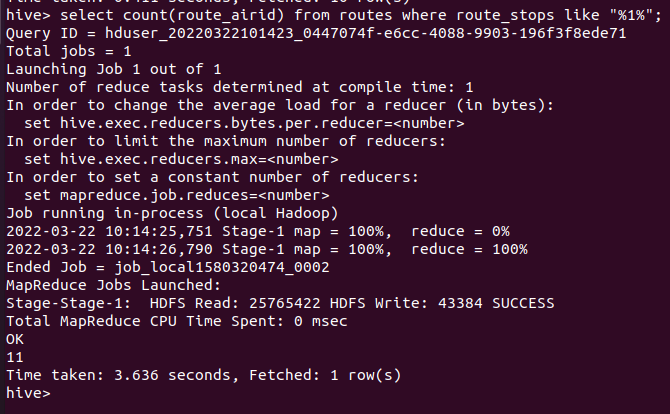
**Explaination:**

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.

1. 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:**

****

**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”.

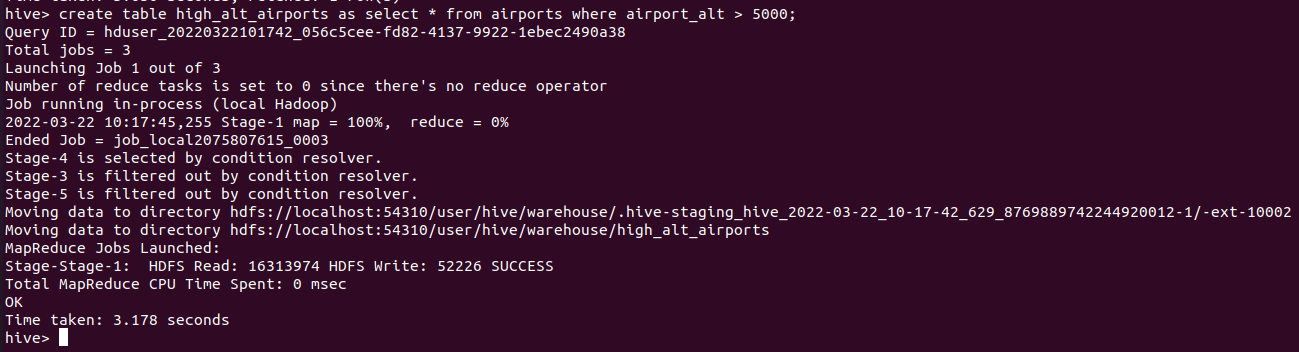
1. select all the route id’s which have their no of stops equal to 1

2. Add the aggregate function “count” to count the no of ids that are resulted as a result of first query.

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

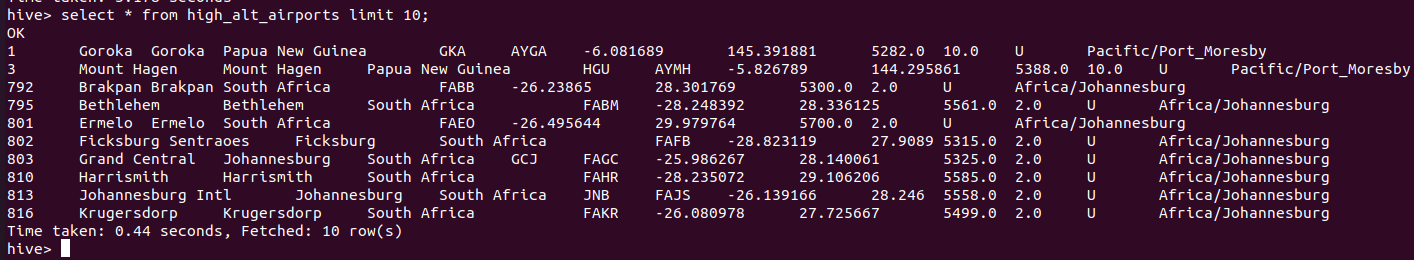
**Query:** create table high\_alt\_airports as select \* from airports where airport\_alt > 5000;

**Result:**

****

**Subquery:**

select \* from high\_alt\_airports limit 10;

****

**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.

**PIG**

**Ravella Abhinav**

**CB.EN.U4CSE19453**

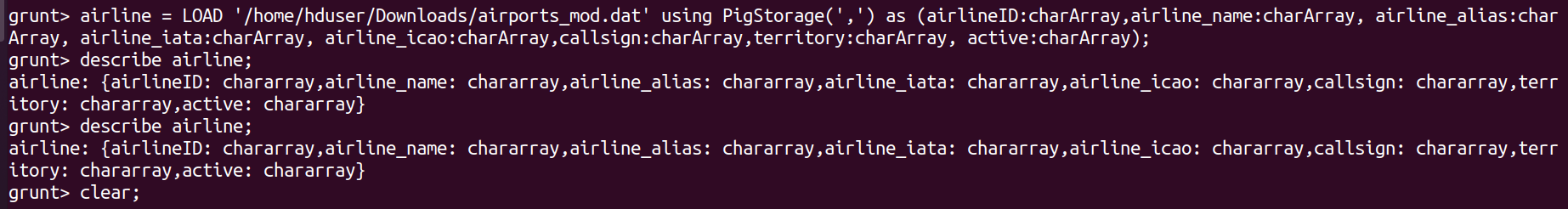
**Load datasets:**

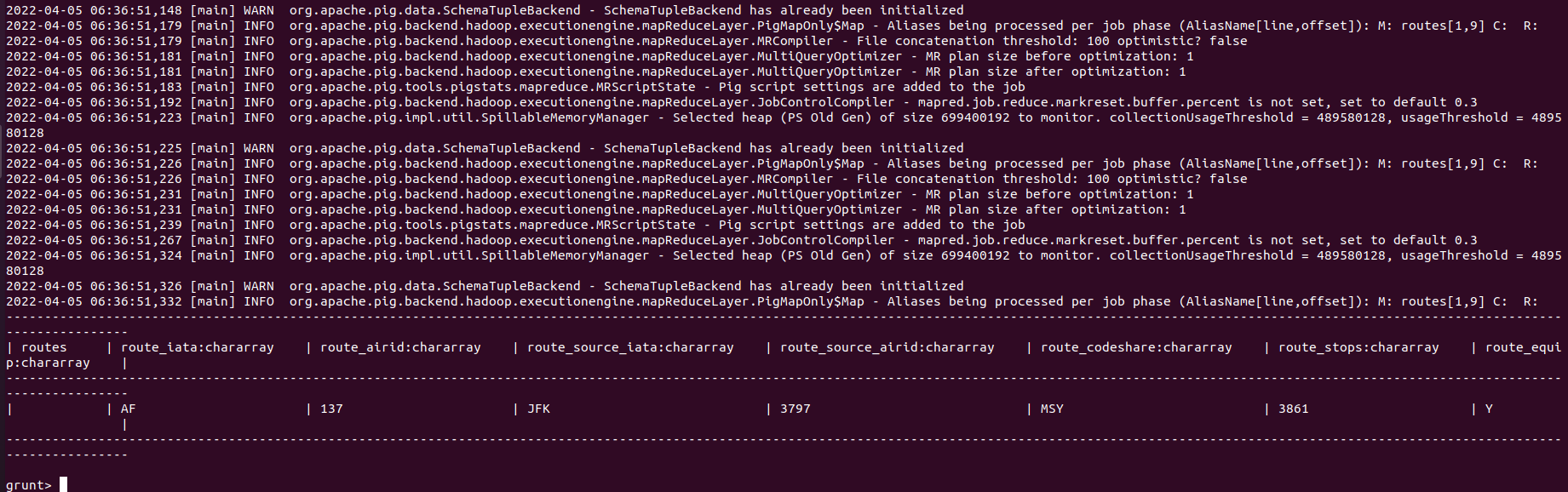
**Airlines:**

**Query:**

airline = LOAD '/home/hduser/Downloads/airports\_mod.dat' using PigStorage(',') as (airlineID:charArray,airline\_name:charArray, airline\_alias:charArray, airline\_iata:charArray, airline\_icao:charArray,callsign:charArray,territory:charArray, active:charArray);

**Output:**



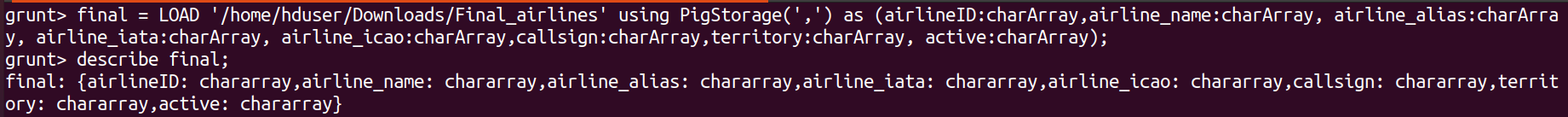


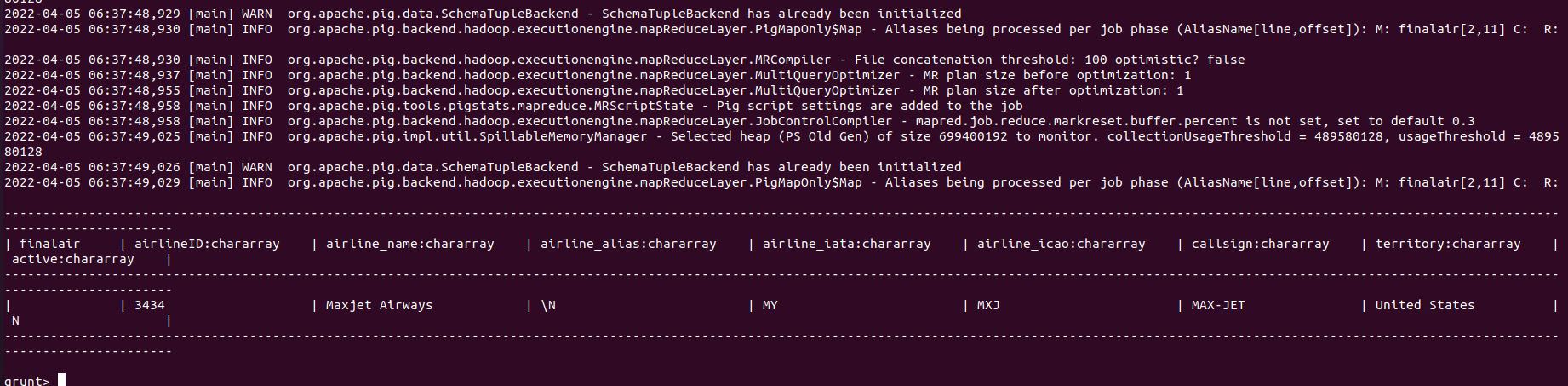
**Final Airlines:**

**Query:**

final = 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);

**Output:**

****

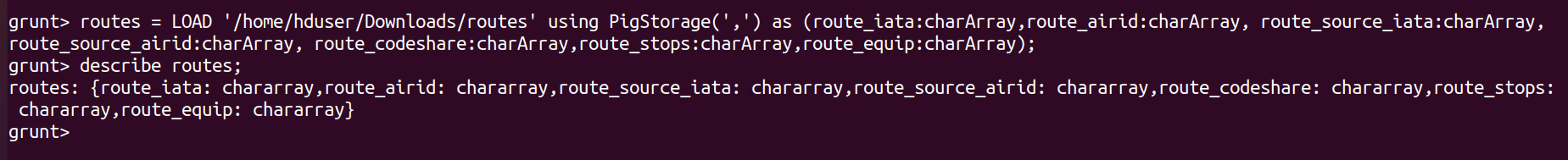
****

**Routes:**

**Query:**

routes = LOAD '/home/hduser/Downloads/routes' using PigStorage(',') as (route\_iata:charArray, route\_airid:charArray, route\_source\_iata:charArray, route\_source\_airid:charArray, route\_codeshare:charArray, route\_stops:charArray, route\_equip:charArray);

**Output:**

****

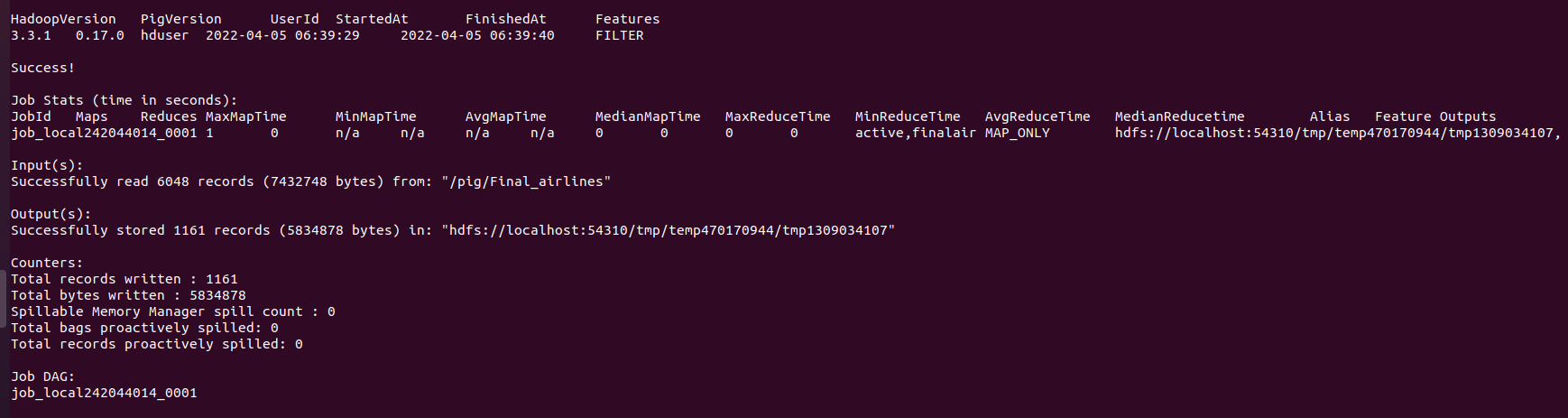
****

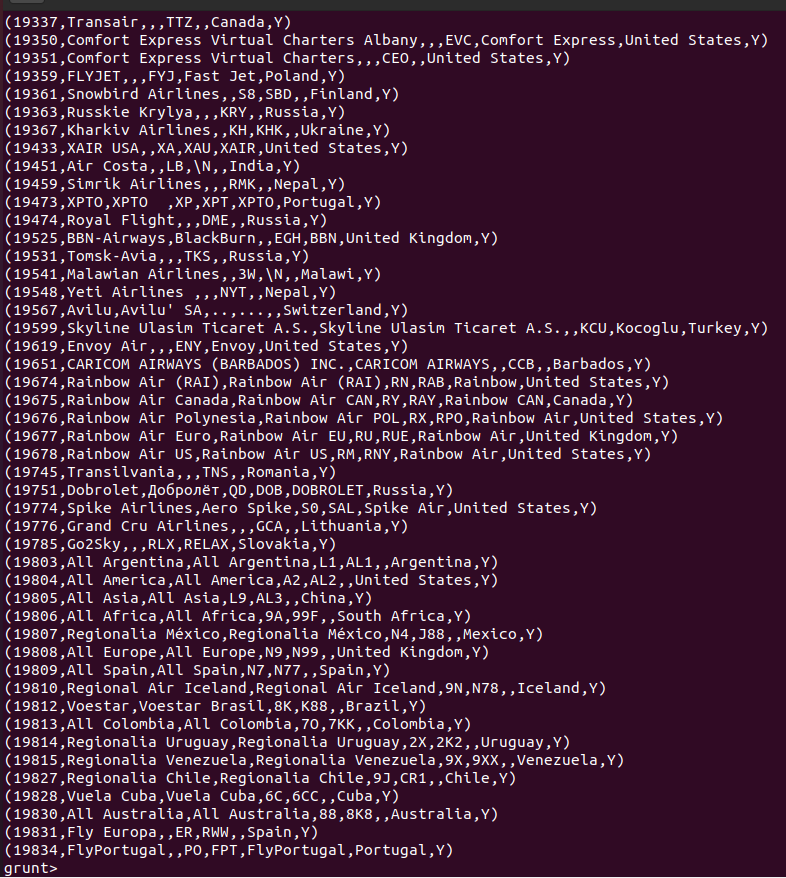
**Queries:**

1. **Find the airlines that are active:**

* active = filter final by Active == 'Y';
* limit = LIMIT active 50;
* dump active;
* dump limit;

**Output:**

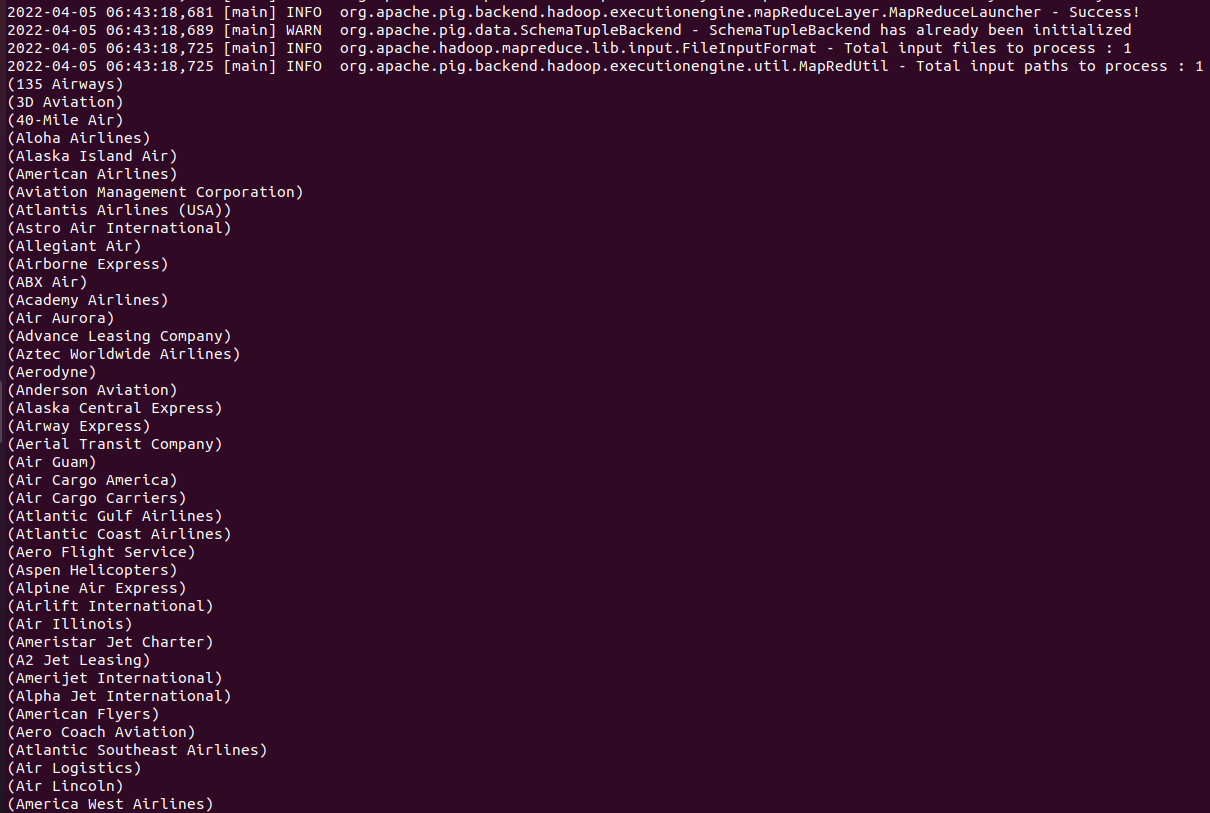


****

1. **Find the names of active airlines in territory ‘Unites States’?**

Query:

* active\_airlines = filter filterair by active == 'Y';
* active\_airlines\_usa = filter active\_airlines by territory == 'United States';
* find\_active\_airline\_names\_in\_usa = foreach active\_airlines\_usa generate airline\_name;
* dump find\_active\_airline\_names\_in\_usa;

****

**Explanation:**

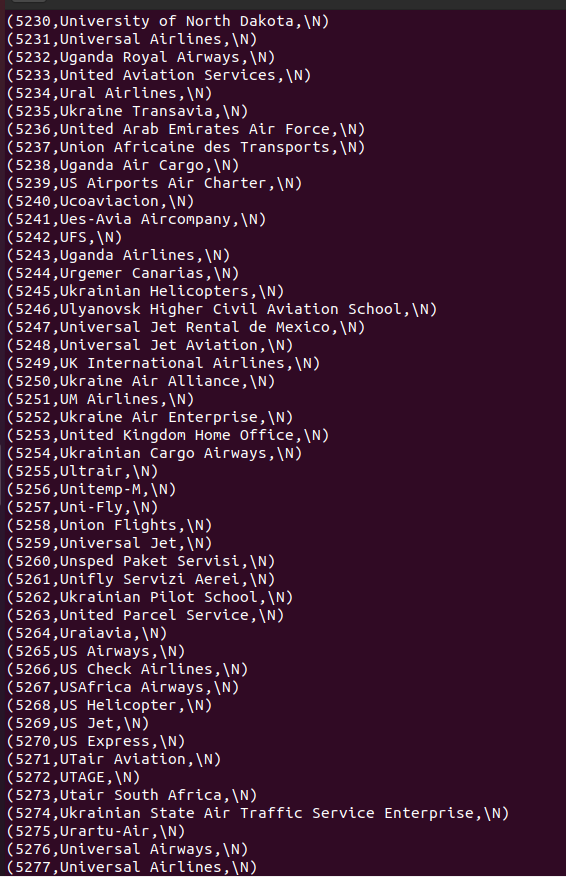
Job is to find all the airlines that are active in the United States which are currently active. So first filtering all the airlines that are active using filter and by keywords and saving them in a table and then querying that table to check for the United States territory using the same keywords and printing the airline names only using foreach and generate keywords.

1. **Get all the alias names of all airlines with airlineID**

**Query:**

* foreach\_data = FOREACH finalair GENERATE airlineID,airline\_name,airline\_alias;
* dump foreach\_data;

**Output:**

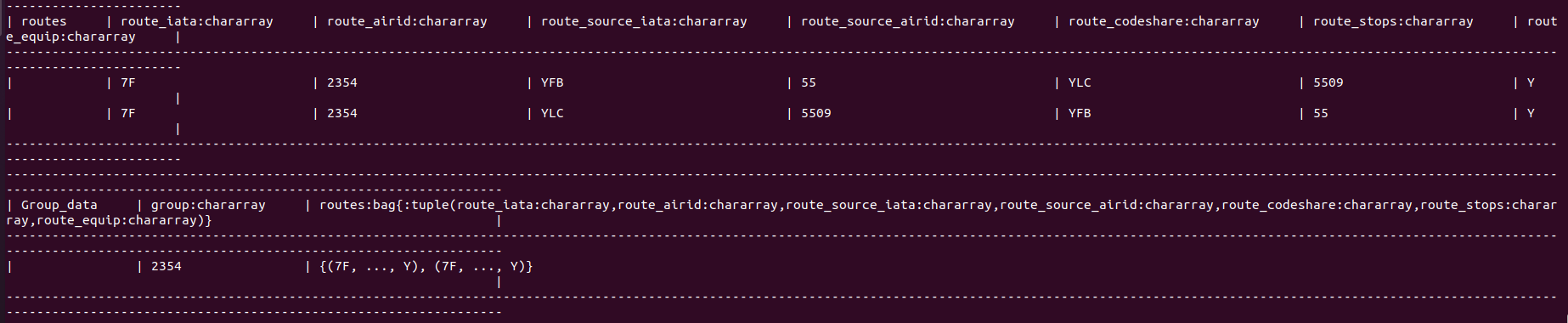
****

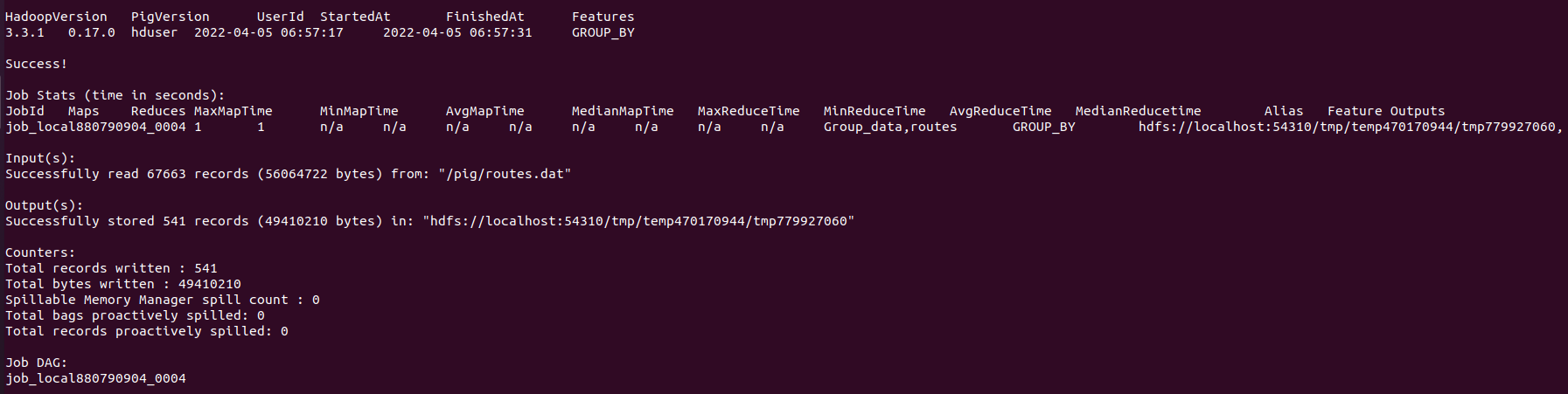
1. **Exploring group command:**

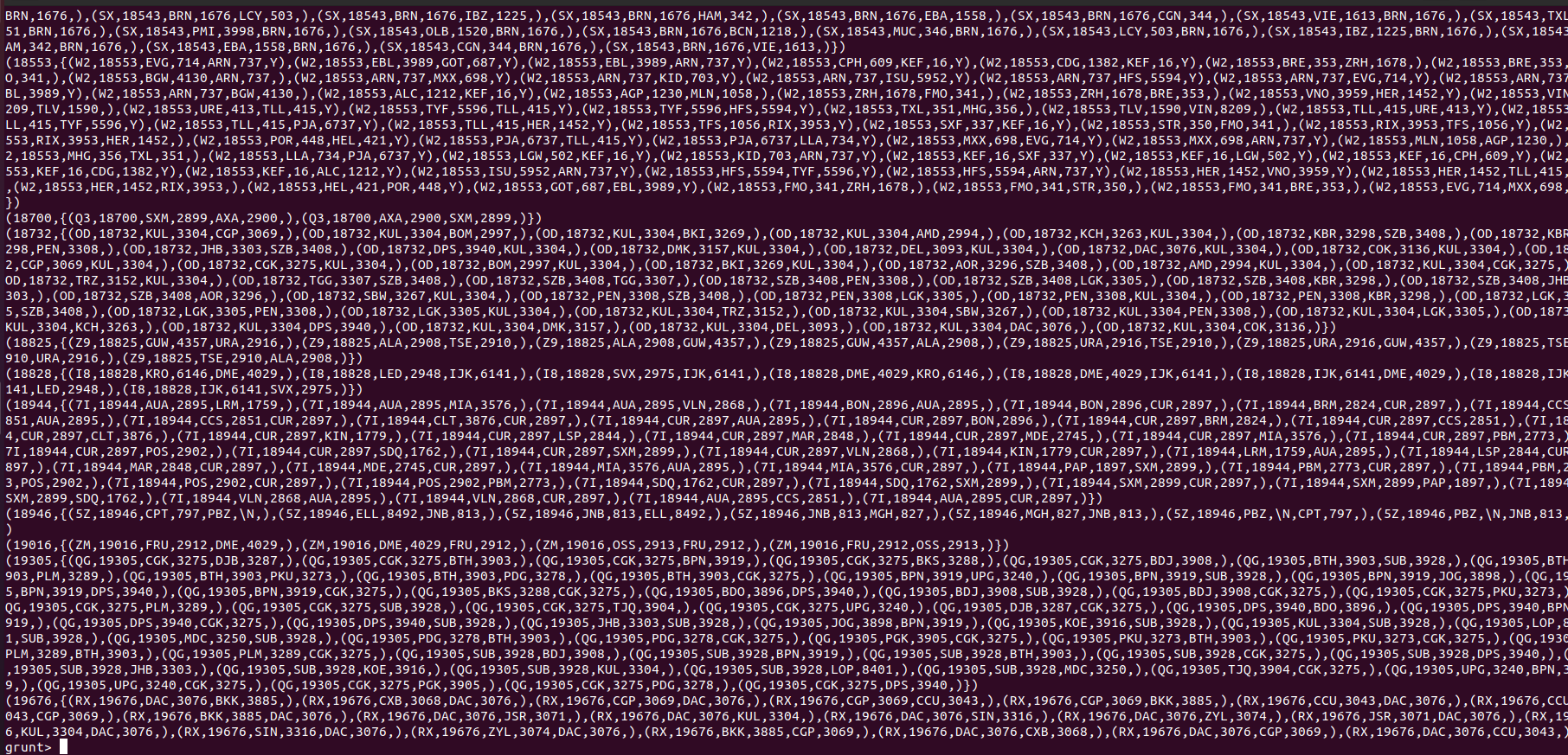
**Query:**

* Group\_data = GROUP routes BY (route\_iata,route\_airid);
* Limit\_group = LIMIT Group\_data 10;
* Dump Limit\_group;
* group\_all = GROUP routes All;

**Output:**

****

****

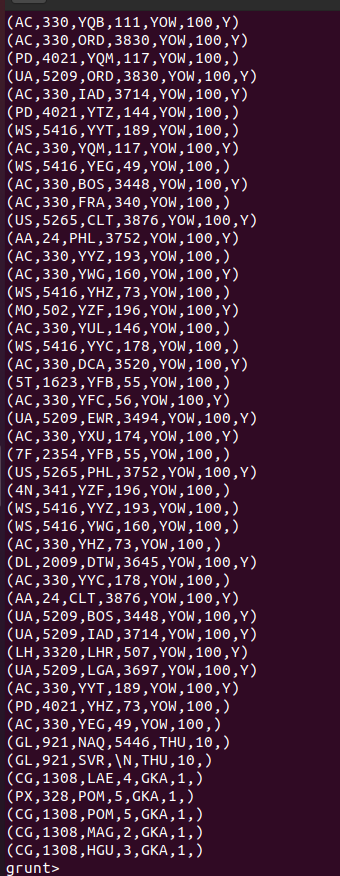
****

1. **Find the flights that has most number of stops?**

**Query:**

* stops = ORDER routes by route\_stops DESC;
* dump stops;

**Output:**

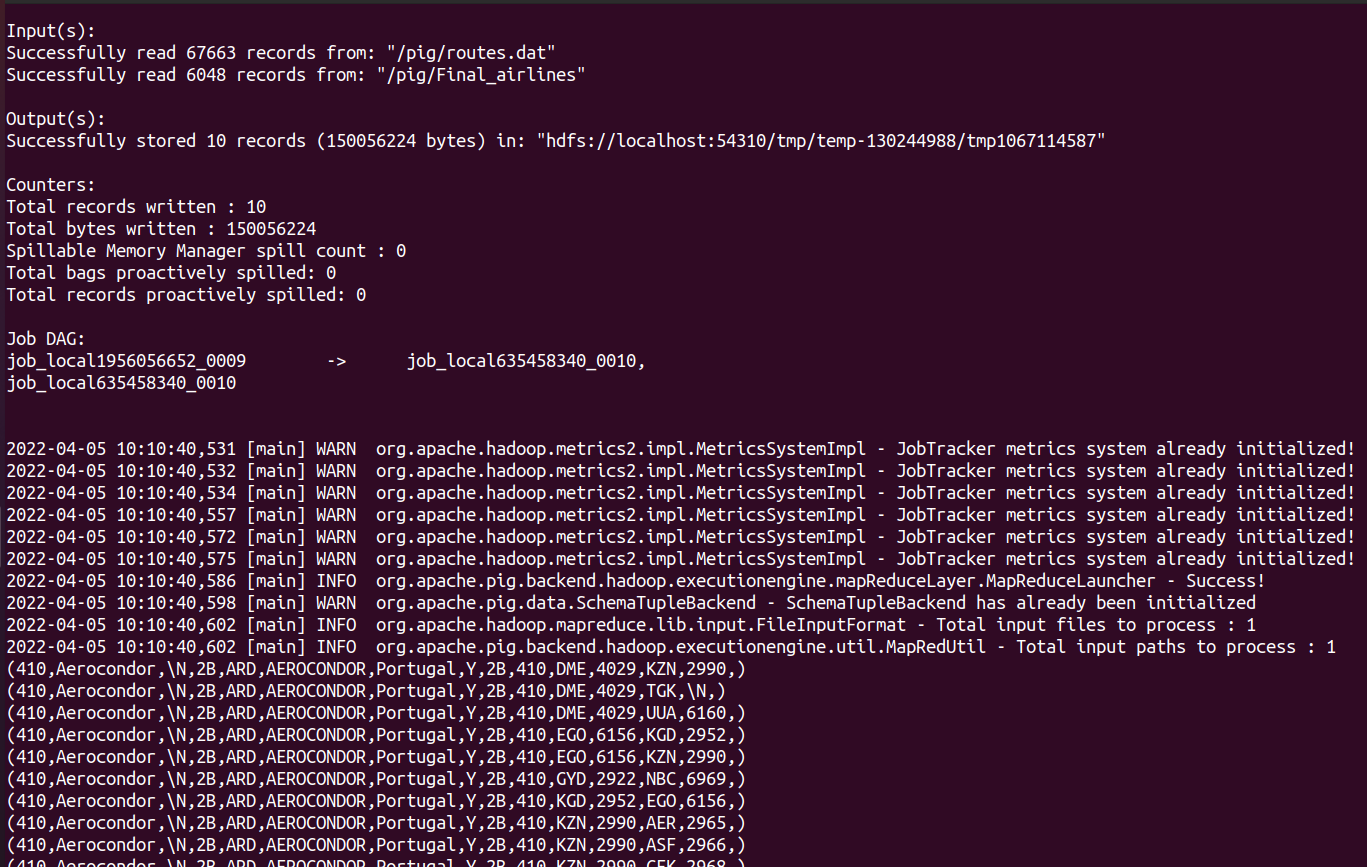
****

1. **JOIN Command:**

**Query:**

* right\_join = JOIN finalair BY airline\_iata RIGHT OUTER, routes BY route\_iata;
* LIM = LIMIT right\_join 10;
* dump LIM;

**Output:**

****

**Query:**

**Output:**