**5 Days FDP on Big Data Analytics**

**Day 3-Lecture 12 Demo**

**“Flight Data Set Analysis using Apache Hive”**

**Problem Statement:** An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it’s, they have ability to analyze limited data from databases. This demo intension is to for the airline data to provide platform for new analytics using **Apache Hive** based on the following queries

**In this use case, there are 3 data sets:**

1. airports\_mod.dat
2. Final\_airlines,
3. routes.dat,

**1) Creating Table and Loading data for airports\_mod.dat**

|  |
| --- |
| hive>create table if not exists day3airportsTable (AirportID bigint, Name string, City string, Country string, IATA varchar(3), ICAO varchar(4), Latitude double, Longitude double, Altitude bigint, Timezone double, DST varchar(1), Tz string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile; |

**Crosscheck whether table is created or not;**

hive> show tables;

|  |
| --- |
| hive> load data local inpath '/home/cloudera/Downloads/airports\_mod.dat' into table day3airportsTable; |

**To check the schema of table**

Hive> describe day3airportsTable;

Print the data

Hive> select \* from day3airportsTable;

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

|  |
| --- |
| hive> select airportID, Name, City from day3airportsTable where country="India"; |

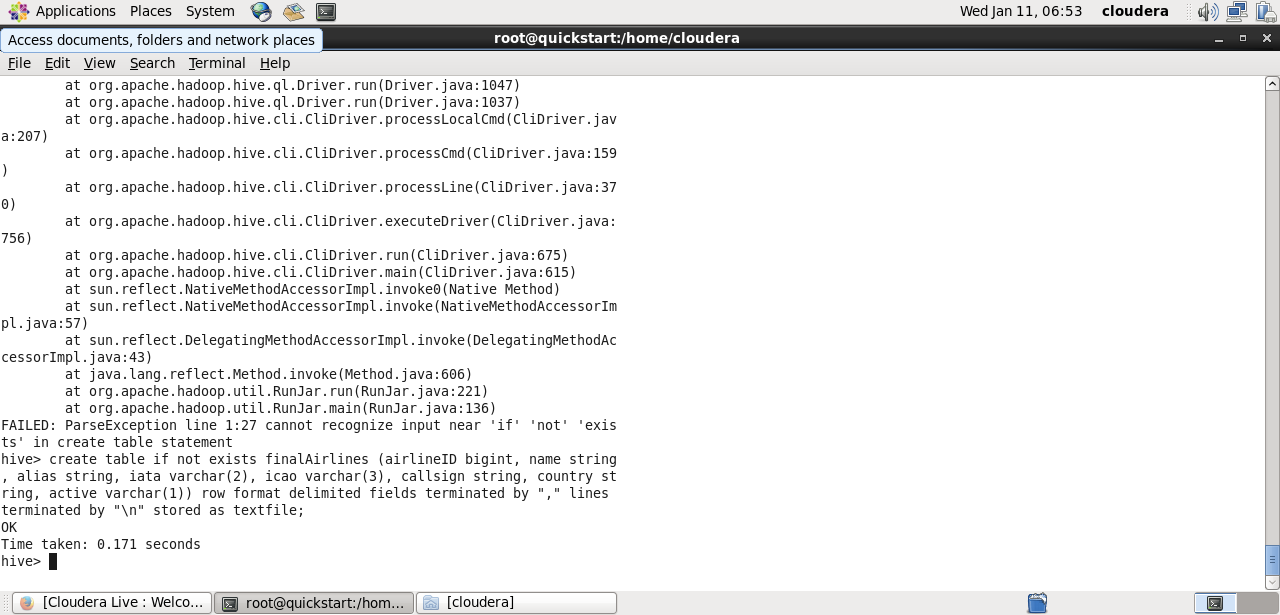
Graphical user interface, text, application

Description automatically generated

**Dataset 2: Final\_airlines**

**1) Creating and Loading data for Final\_airlines**

|  |
| --- |
| hive> create table if not exists finalAirlines (airlineID bigint, name string, alias string, iata varchar(2), icao varchar(3), callsign string, country string, active varchar(1)) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile; |



|  |
| --- |
| hive> load data local inpath '/home/cloudera/Downloads/Final\_airlines' into table finalAirlines; |

**Crosscheck data is loaded successfully or not**

**hive>** select \* from finalAirlines;

**2. Find the list of Active Airlines in United state**

|  |
| --- |
| hive> select airlineid, name from finalairlines where country="United States" AND active="Y”; |

**3. Which country (or) territory having highest Airports**

|  |
| --- |
| hive> select country, count(\*) as A from airportstable group by country order by A desc limit 1; |

Graphical user interface, text, application, email

Description automatically generated

Answer: US

**Dataset 3:** routes.dat

**1) Creating Table and Loading data for routes.dat**

|  |
| --- |
| hive> create table if not exists routes(airline varchar(3), airlineID bigint, source varchar(4), sourceID bigint, destination varchar(4), destinationID bigint, codeshare varchar(1), stops int, equipment varchar(3)) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile; |

|  |
| --- |
| hive> load data local inpath '/home/cloudera/Downloads/routes.dat’ into table routes; |

**Crosscheck data is loaded successfully or not**

**hive>** select \* from routes;

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

**Hint: Use Join operation over routes.dat and Final airlines**

|  |
| --- |
| hive>select distinct(a.airlineid),a.name from finalairlines a join routes on  a.airlineid=routes.airlineid where stops=0; |

Output screenshot

