Hadoop & Big Data Project

AIRLINE ANALYSIS

Step 1:

* Downloading the given dataset and using unrar command to unrar the rar files.
* Creating a directory(Airlines) and copying the data files into this new directory.

Syntax –

*Hadoop hdfs –copyFromLocal home/edureka/Downloads/Final\_airlines /Airlines*

Step 2:

* Using Pig script to execute the problem statements.

Question A

Find list of Airports operating in the Country India

A = load '/Airlines/airports\_mod.dat' using PigStorage(',') as (AirportID:int, Airportname:chararray, city:chararray, country:chararray, IATA:chararray, ICAO:chararray, latitude:float, longitude:float, altitude:int, timezone:double, dst:chararray, tz:chararray);

B = filter A by country == 'India';

C = foreach B generate Airportname;

store C into '/Airlines/ResultA';

Question B

B. Find the list of Airlines having zero stops

A = load '/Airlines/Final\_airlines' using PigStorage(',') as (airlineID:int, airlinename:chararray, alias:chararray, IATA:chararray, ICAO:chararray, callsign:chararray, Country:chararray, active:chararray);

B = load '/Airlines/routes.dat' using PigStorage(',') as (airline:chararray, airlineID:int, sourceairport:chararray, sourceID: int, destairport:chararray, destID:int, codeshare:chararray, stops: int,equipment:chararray);

C = filter B by stops == 0;

D = join C by airlineID, A by airlineID;

E = foreach D generate airlinename;

F = distinct E;

store F into '/Airlines/Result(B)';

Question C

List of Airlines operating with code share

A = load '/Airlines/Final\_airlines' using PigStorage(',') as (airlineID:int, airlinename:chararray, alias:chararray, IATA:chararray, ICAO:chararray, callsign:chararray, Country:chararray, active:chararray);

B = load '/Airlines/routes.dat' using PigStorage(',') as (airline:chararray, airlineID:int, sourceairport:chararray, sourceID: int, destairport:chararray, destID:int, codeshare:chararray, stops: int,equipment:chararray);

C = filter B by codeshare == 'Y';

D = join C by airlineID, A by airlineID;

E = foreach D generate airlinename;

F = distinct E;

store G into '/Airlines/Result-C';

Question D

Which country (or) territory having highest Airports

A = load '/Airlines/airports\_mod.dat' using PigStorage(',') as (AirportID:int, Airportname:chararray, city:chararray, country:chararray, IATA:chararray, ICAO:chararray, latitude:float, longitude:float, altitude:int, timezone:double, dst:chararray, tz:chararray);

B = foreach A generate country, Airportname;

C = group B by country;

D = foreach C generate flatten (group) as country, COUNT(B) as total;

E = order D by total DESC;

F = limit E 1;

store F into '/Airlines/RESULT-D';

Question E

Find the list of Active Airlines in United States.

A = load '/Airlines/Final\_airlines' using PigStorage(',') as

(airlineID:int, airlinename:chararray, alias:chararray,

IATA:chararray, ICAO:chararray, callsign:chararray,

Country:chararray, active:chararray);

B = load '/Airlines/routes.dat' using PigStorage(',') as

(airline:chararray, airlineID:int, sourceairport:chararray,

sourceID: int, destairport:chararray, destID:int,

codeshare:chararray, stops: int,equipment:chararray);

C = distinct B;

D = join C by airlineID, A by airlineID;

E = filter D by active == 'Y';

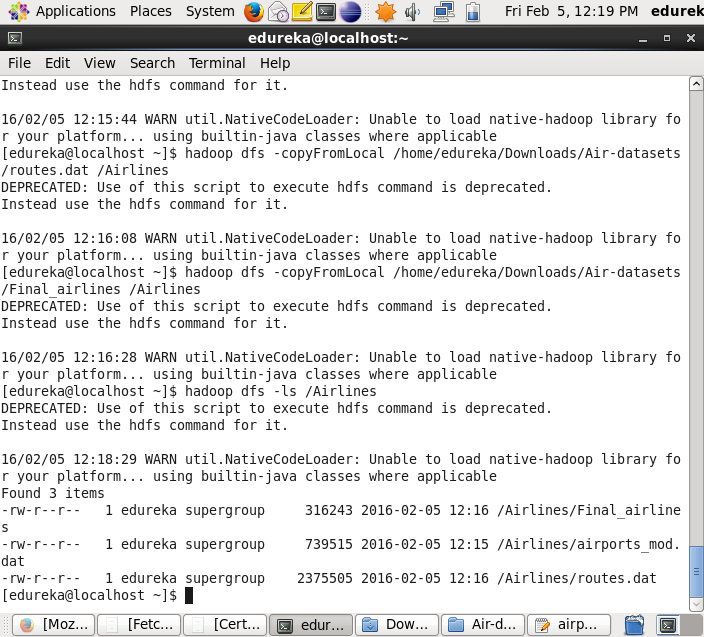
F = foreach E generate airlinename, Country, active;

G = filter F by Country == 'United States';

H = distinct G;

store H into '/Airline/Result-E';

The screenshots of results are shown below -

1. Copying files from local file system to hdfs.**
2. Success statement after executing the first pig script for each question.

