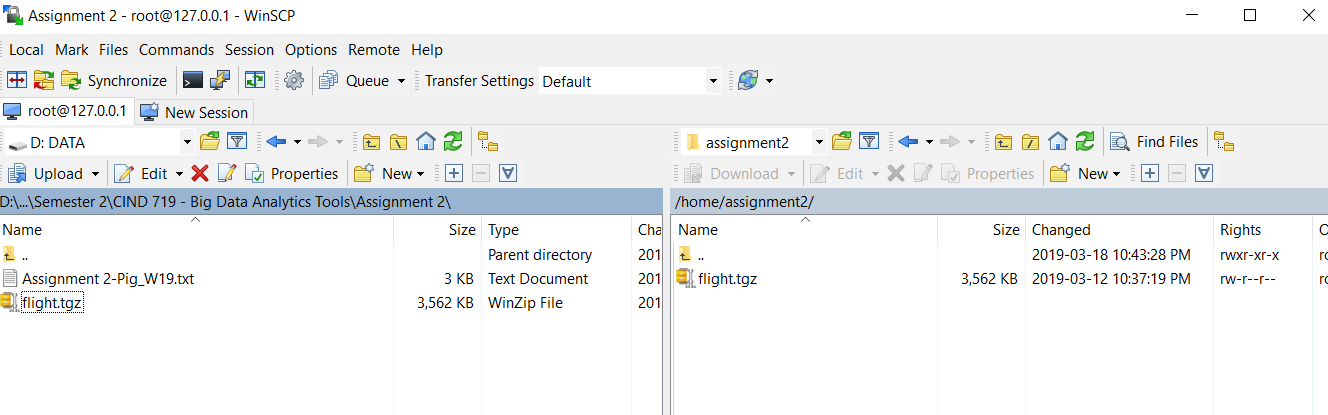
**CIND 719 Assignment 2 Chantal Sylvestre**

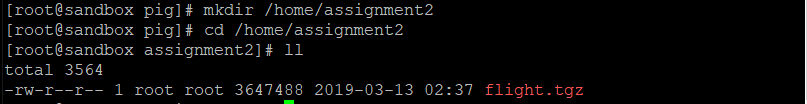
--Made assignment2 directory on linux machine

mkdir /home/assignment2

-- Used filezilla to move tgz file to linux machine then extracted files

tar xvf /home/lab/flight.tgz -C /home/assignment2/

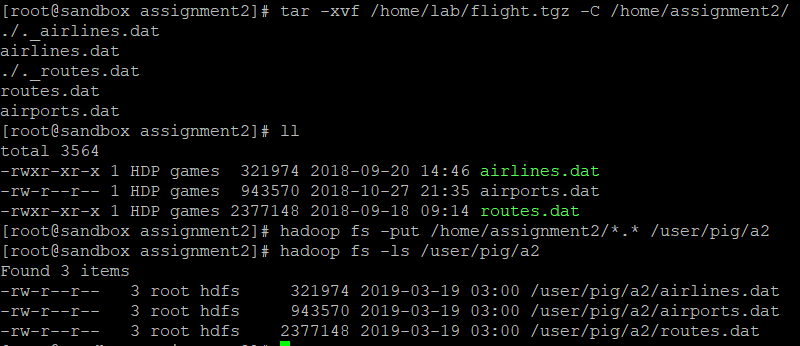




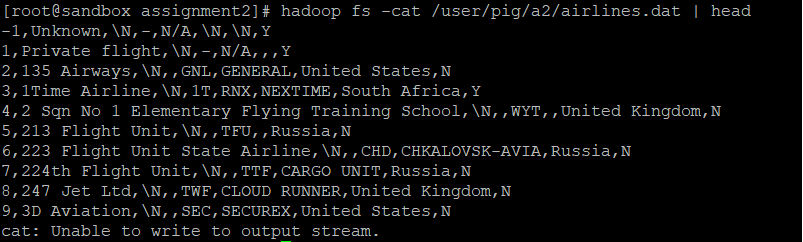
--Made a directory on HDFS

hadoop fs -mkdir /user/pig/a2

--Put extracted files onto HDFS

hadoop fs -put /home/assignment2/\*.\* /user/pig/a2

--The first few lines of the airlines dataset:



-------------------------

----QUESTION 1----

-------------------------

--Load airlines dataset into pig

airlines = LOAD '/user/pig/a2/airlines.dat' using PigStorage(',') AS (airline\_id:chararray, airline\_name:chararray, airline\_alias:chararray, airline\_iata:chararray, airline\_icao:chararray, airline\_call:chararray, airline\_country:chararray, airline\_active:chararray);

--Filtered for air canada, case insensitive with or without a space using Regex

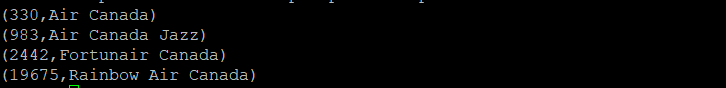
air\_canada = FILTER airlines BY airline\_name MATCHES '(?i).\*air.?canada.\*';

--Generate only the necessary fields

air\_canada2 = foreach air\_canada GENERATE airline\_id, airline\_name;

--Output

dump air\_canada2;



------------------------

----QUESTION 2----

------------------------

--load in the airports dataset

airports = LOAD '/user/pig/a2/airports.dat' using PigStorage(',') AS (airport\_id:int, airport\_name:chararray, airport\_city:chararray, airport\_country:chararray, airport\_iata:chararray, airport\_icao:chararray, airport\_lat:float, airport\_lon:float, airport\_alt:chararray, airport\_tz:chararray, airport\_dst:chararray, airport\_timezone:chararray, airport\_source:chararray);

-- grouped by country

airport\_country = GROUP airports BY airport\_country;

--count the number of airports per country

airport\_count = FOREACH airport\_country GENERATE group AS country, COUNT(airports) AS count;

--order data descending

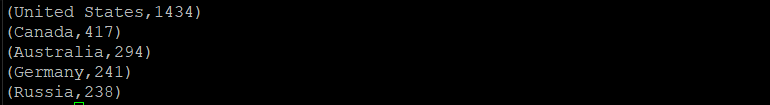
airport\_order = ORDER airport\_count by count DESC;

--limit to the top 5 highest number of airports per country

airport\_top = limit airport\_order 5;

--output for question 2

dump airport\_top;



------------------------

----QUESTION 3----

------------------------

--load routes dataset

routes = LOAD '/user/pig/a2/routes.dat' using PigStorage(',') AS (r\_airline:chararray, r\_airline\_id:int, r\_source:chararray, r\_source\_id:int, r\_dst:chararray, r\_dst\_id:int, r\_code:chararray, r\_stops:chararray, r\_eq:chararray);

--selecting the information I want displayed

routes2 = foreach routes generate r\_source, r\_source\_id, r\_dst, r\_dst\_id;

-- only selecting the distinct rows so there are no duplicate routes in the output

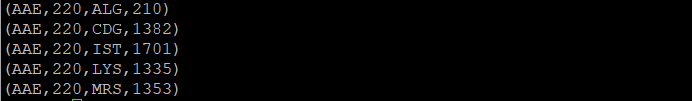
routes3 = distinct routes2;

--limiting by 5 to display first 5 rows

routes4 = limit routes3 5;

--question 3 output

dump routes4;



-------------------------

----QUESTION 4----

-------------------------

--load in necessary data from airports dataset

lat\_lon = foreach airports generate airport\_iata, airport\_id, airport\_name, airport\_lat, airport\_lon;

--join the airport data to the route source data to link the lat and lon and airport name

src\_loc = join lat\_lon by airport\_iata, routes3 by r\_source;

--rename the lat and lon and airport name and id to indicate that these are linked to the source airport for the route

src\_loc2 = foreach src\_loc generate lat\_lon::airport\_iata, lat\_lon::airport\_id as s\_id, lat\_lon::airport\_name as s\_name, lat\_lon::airport\_lat as lat1, lat\_lon::airport\_lon as lon1, routes3::r\_dst, routes3::r\_dst\_id;

--join the airport data to the route destination data to link the lat and lon and airport name

dst\_loc = join lat\_lon by airport\_iata, src\_loc2 by r\_dst;

--generated the required data for both the source and destination data into a variable

dst\_src = foreach dst\_loc generate lat\_lon::airport\_name as dst\_name, lat\_lon::airport\_id as dst\_id, lat\_lon::airport\_lat as lat2, lat\_lon::airport\_lon as lon2, src\_loc2::s\_name, src\_loc2::s\_id as s\_id, src\_loc2::lat1, src\_loc2::lon1;

--calculated the distance and then generated the desired fields using a nested foreach

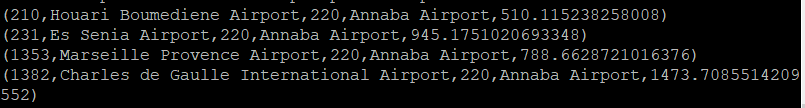
d = foreach dst\_src {

distance = SQRT((lat2 - lat1) \* (lat2 - lat1) + (lon2 - lon1) \* (lon2 - lon1)) \* 111;

generate s\_id, s\_name, dst\_id, dst\_name, distance;};

--limit 5 to output the first 5 rows

e = limit d 5;



--stored output as a directory on the Hadoop fs

store d into '/user/pig/a2/routes\_with\_distances';

