Assignment 5.2 ( Exploring PIG)

Task1: Top 5 most visited destinations.

Step1: Register the piggybank-0.14.0.jar in order to use the CSVExcelStorage class.

REGISTER '/home/acadgild/assignment\_5.2/piggybank-0.14.0.jar';

Step2: Load DelayedFlights.csv dataset using CSVExcelStorage into relation A

A = load '/home/acadgild/assignment\_5.2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step3: Selecting columns year, flight\_num, origin, dest and form relation B

B = foreach A generate (int)$1 as year, (int)$10 as flight\_num, (chararray)$17 as origin,(chararray) $18 as dest;

Step3: Filter B for dest which are not null and form relation C

C = filter B by dest is not null;

Step4: Group C dest and form relation D

D = group C by dest;

Step5: Generate the grouped column and the count of each.

E = foreach D generate group, COUNT(C.dest);

Step6: Sort desceding by grouped column

F = order E by $1 DESC;

Step6: Sort desceding by grouped column

Step7: Take only 5 records for F and form relation Result

Result = LIMIT F 5;

Step7: Load airports.csv dataset using CSVExcelStorage into relation A1

A1 = load '/home/acadgild/assignment\_5.2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step8: Select columns dest, city, country from A1 and form relation A2

A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;

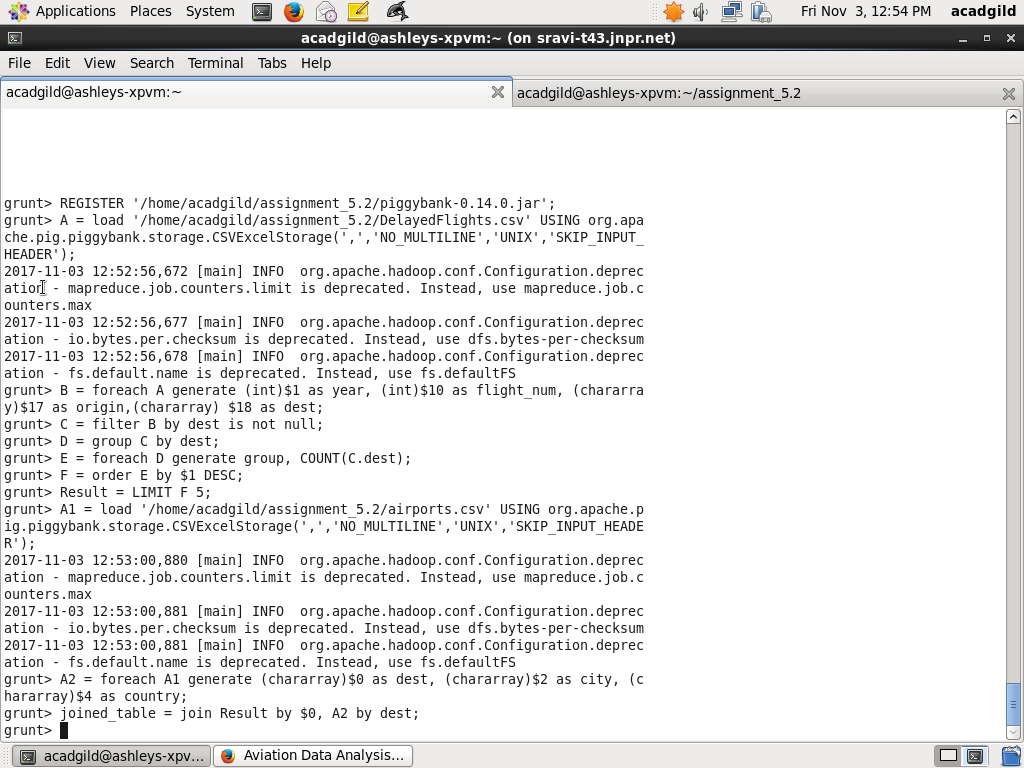
Step8: Join Result and A2 and form relation joined\_table

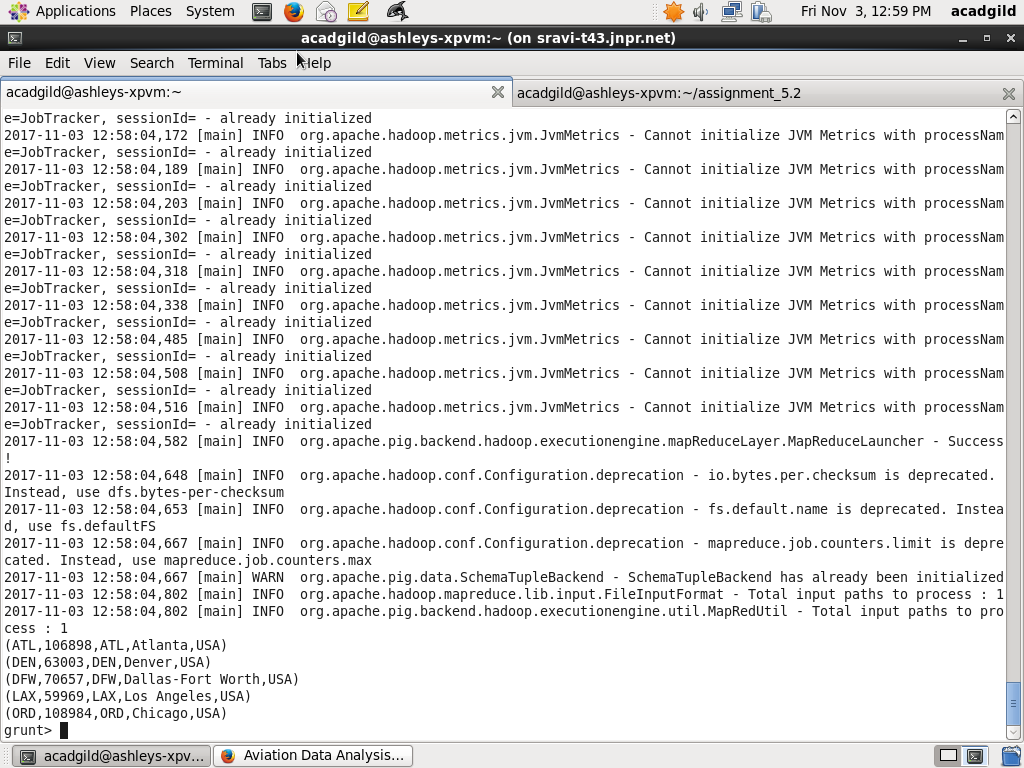
joined\_table = join Result by $0, A2 by dest;

Step9: dump joined\_table

dump joined\_table;

Screenshot of query with output is as below:





Task2: Which month has seen the most number of cancellations due to bad weather

Step1: Register the piggybank-0.14.0.jar in order to use the CSVExcelStorage class.

REGISTER '/home/acadgild/assignment\_5.2/piggybank-0.14.0.jar';

Step2: Load DelayedFlights.csv dataset using CSVExcelStorage into relation A

A = load '/home/acadgild/assignment\_5.2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step3: Select columns month, Flight\_num, cancelled, cancel\_code and form relation B

B = foreach A generate (int)$2 as month,(int)$10 as flight\_num,(int)$22 as cancelled,(chararray)$23 as cancel\_code;

Step3: Filter relation B by cancelled and cancellation code and form relation C

C = filter B by cancelled == 1 AND cancel\_code =='B';

Step3: Group relation C by every month and form relation D

D = group C by month;

Step4: Count of canceled flights every month and form relation E

E = foreach D generate group, COUNT(C.cancelled);

Step5: Order E and find the top month

F= order E by $1 DESC;

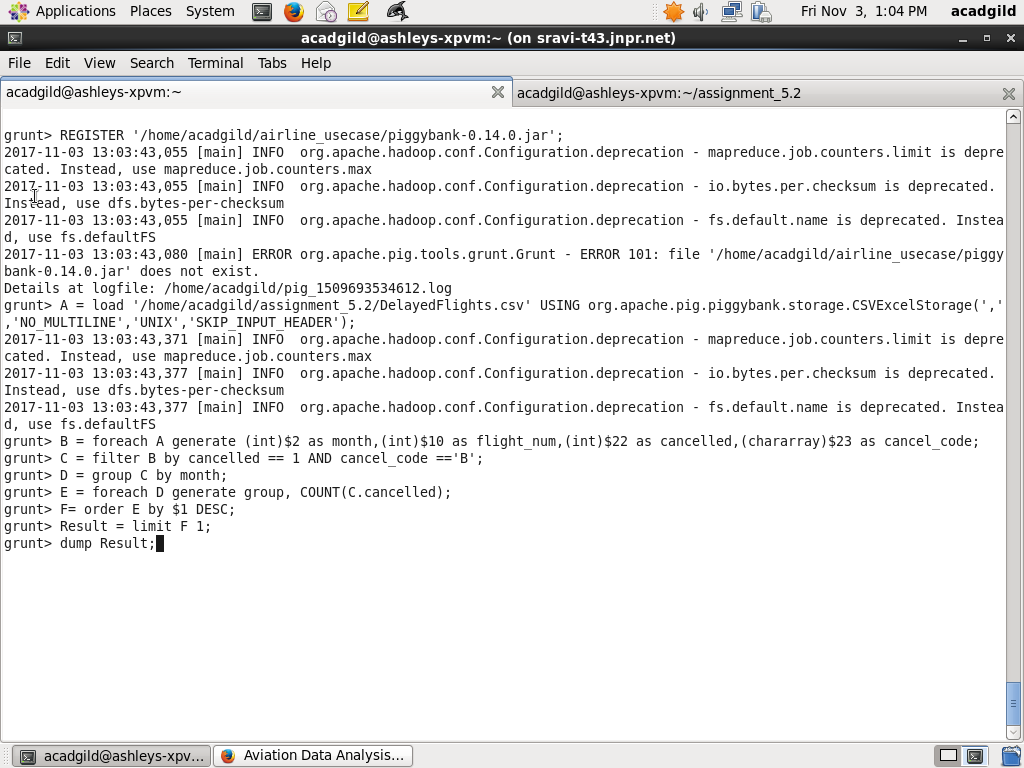
Step5: Select first tuple and form relation Result

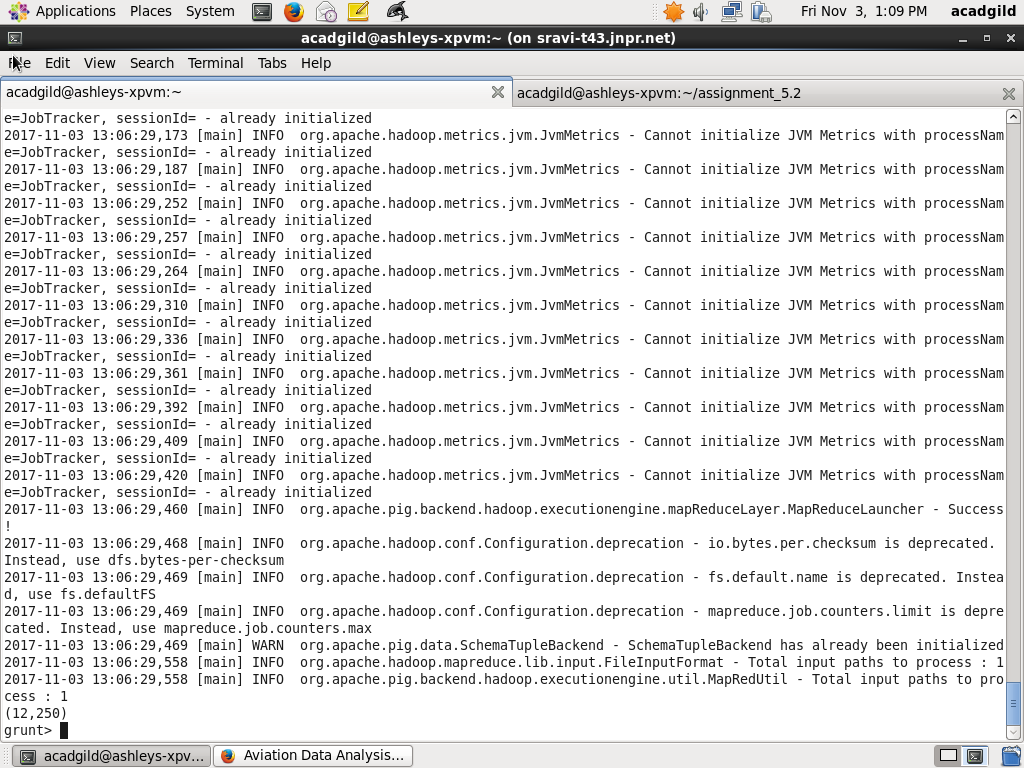
Result = limit F 1;

Step6: dump Result

dump Result

The screenshot of query and result is as below:





Task3:

Top ten origins with the highest AVG departure delay

Step1: Register the piggybank-0.14.0.jar in order to use the CSVExcelStorage class.

REGISTER '/home/acadgild/assignment\_5.2/piggybank-0.14.0.jar';

Step2: Load DelayedFlights.csv dataset using CSVExcelStorage into relation A

A = load '/home/acadgild/assignment\_5.2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step3: Select fields departure delay and origin from A and form relation B1

B1 = foreach A generate (int)$16 as dep\_delay, (chararray)$17 as origin;

Step4: From relation B1 filter those tuples which has departure delay and origin is not null

C1 = filter B1 by (dep\_delay is not null) AND (origin is not null);

Step5: Group C1 by origin and form relation D1

D1 = group C1 by origin;

Step6: Find average departure delay for each origin

E1 = foreach D1 generate group, AVG(C1.dep\_delay);

Step6: Order E1 descending by origin and form relation Result

Result = order E1 by $1 DESC;

Step7: Get top 10 from Result

Top\_ten = limit Result 10;

Step8: Load airports.csv dataset using CSVExcelStorage into Lookup

Lookup = load '/home/acadgild/assignment\_5.2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step8: From relation Lookup select origin, city, country from Lookup

Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;

Step8: Join relations Lookup1 an Top\_ten by origin and form relation Joined

Joined = join Lookup1 by origin, Top\_ten by $0;

Step9: Get the field State, City, Country, Average Delay and form relation Final

Final = foreach Joined generate $0,$1,$2,$4;

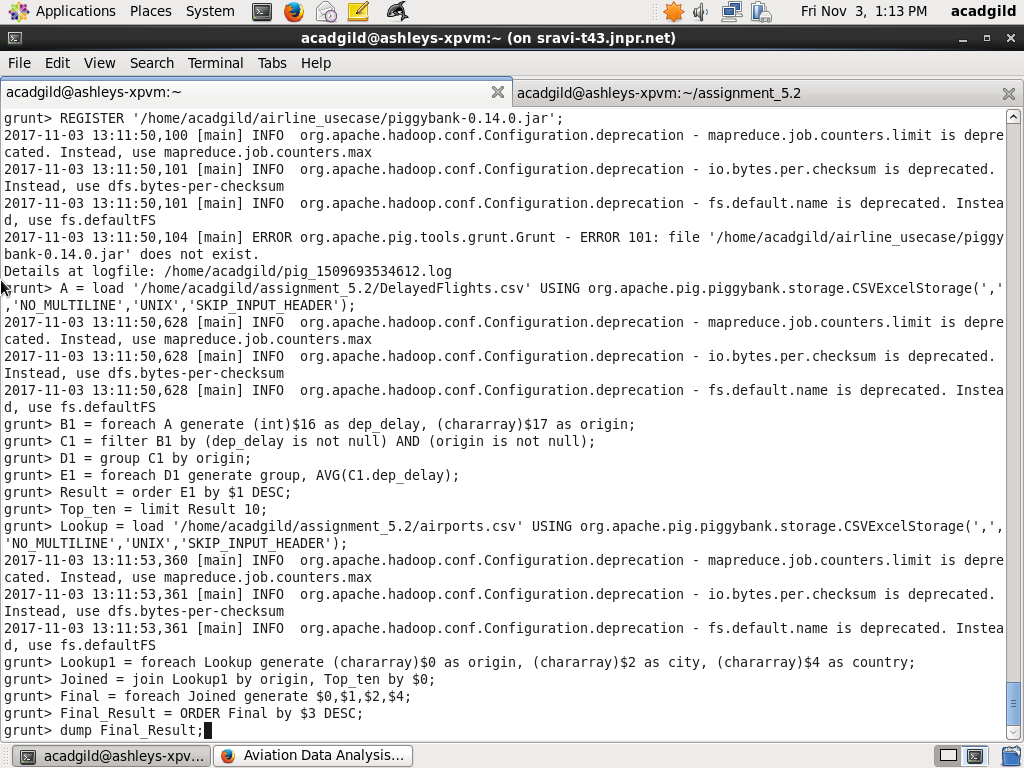
Step10: Sort relation by average delay descending and form relation Final\_Result

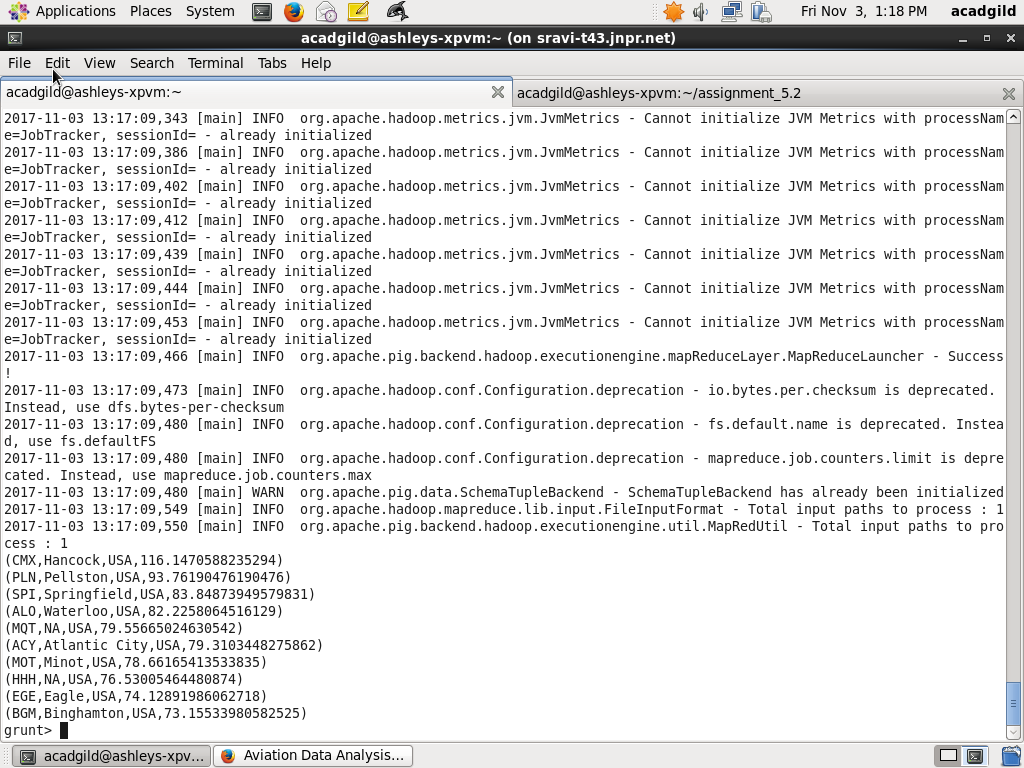
Final\_Result = ORDER Final by $3 DESC;

Step11: dump Final\_Result

dump Final\_Result;

Screenshot of query and output is as below:





Task 4: Which route (origin & destination) has seen the maximum diversion?

Step1: Register the piggybank-0.14.0.jar in order to use the CSVExcelStorage class.

REGISTER '/home/acadgild/assignment\_5.2/piggybank-0.14.0.jar';

Step2: Load DelayedFlights.csv dataset using CSVExcelStorage into relation A

A = load '/home/acadgild/assignment\_5.2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Step3: Select fields origin, destination, diversion from A and form relation B

B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;

Step4: Filter B for tuples for which origin is not null, destination is not null and diversion has occurred and form relation C

C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);

Step5: Group C by origin and destination

D = GROUP C by (origin,dest);

Step6: For each group count number of diversions

E = FOREACH D generate group, COUNT(C.diversion);

Step6: Sort E by origin

F = ORDER E BY $1 DESC;

Step7: Take only first 10 tuples and form relation Result

Result = limit F 10;

Step8: dump the Result

dump Result;

