**Assignment 5.2**

**Problem Statement**: To implement the use case present in below blog link:

<https://acadgild.com/blog/aviation-data-analysis-using-apache-pig/>

**Required files**:

1. Delayed\_Flights.csv
2. Airports.csv
3. piggybank-0.15.0.jar

**Problem Statement 1**:

Find out the top 5 most visited destinations.

**Steps**:

REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';

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

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

C = filter B by dest is not null;

D = group C by dest;

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

F = order E by $1 DESC;

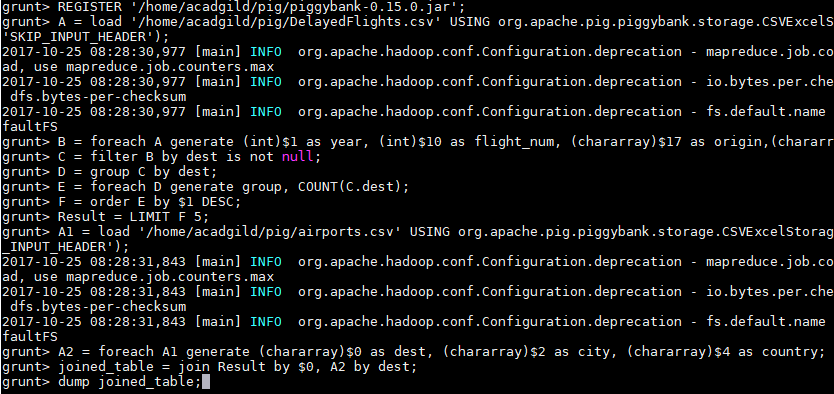
Result = LIMIT F 5;

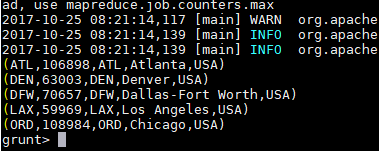
A1 = load '/home/acadgild/pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

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

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

dump joined\_table;





**Problem Statement 2**:

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

**Steps**:

REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';

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

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

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

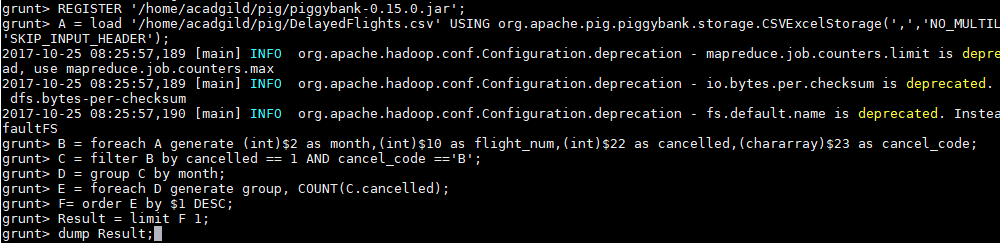
D = group C by month;

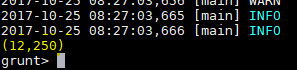
E = foreach D generate group, COUNT(C.cancelled);

F= order E by $1 DESC;

Result = limit F 1;

dump Result;





**Problem Statement 3**:

Top ten origins with the highest AVG departure delay.

**Steps**:

REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';

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

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

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

D1 = group C1 by origin;

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

Result = order E1 by $1 DESC;

Top\_ten = limit Result 10;

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

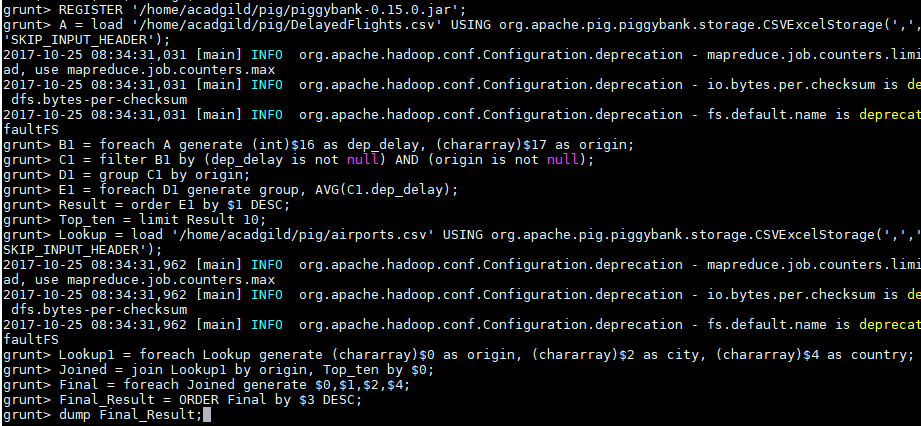
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;

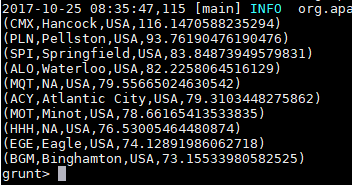
Joined = join Lookup1 by origin, Top\_ten by $0;

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

Final\_Result = ORDER Final by $3 DESC;

dump Final\_Result;





**Problem Statement 4**:

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

**Steps:**

REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';

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

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

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

D = GROUP C by (origin,dest);

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

F = ORDER E BY $1 DESC;

Result = limit F 10;

dump Result;

