



Session 5: Exploring Pig

# Assignment 2 Question

### Problem Statement

Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.

NOTE: You must submit a word file containing steps and screenshots.

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

Downloaded data from below given link:

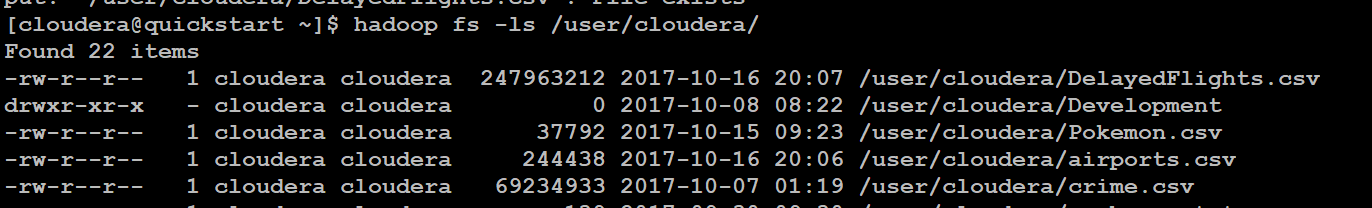
<https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVFOdzlJNWM/view>

<https://drive.google.com/file/d/0B_Qjau8wv1KocDR3djk1Qm96Mmc/view>

[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/Downloads/airports.csv /user/cloudera

[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/Downloads/DelayedFlights.csv /user/cloudera





## **Problem Statement 1**

*Find out the top 5 most visited destinations*.

grunt> REGISTER '/home/cloudera/Downloads/piggybank.jar';

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

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

grunt> C = filter B by dest is not null;

grunt> D = group C by dest;

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

grunt> F = order E by $1 DESC;

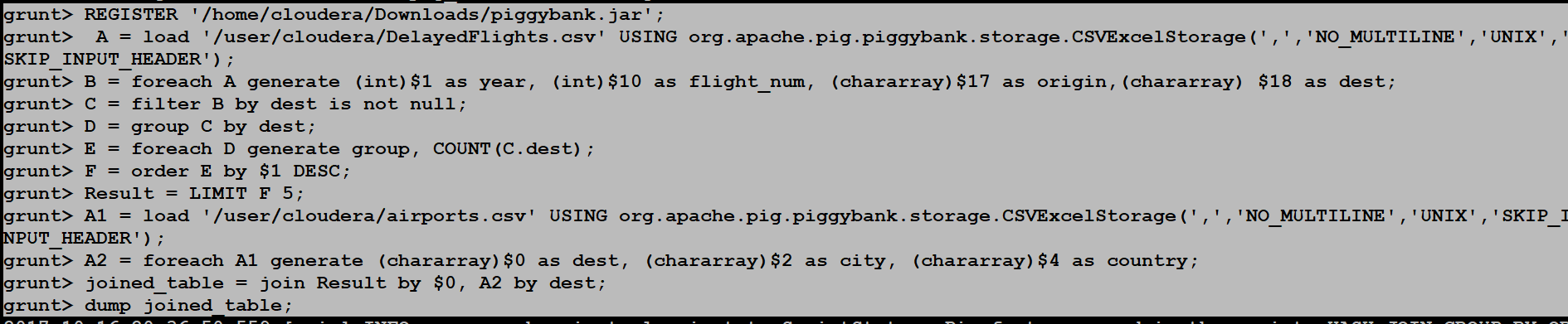
grunt> Result = LIMIT F 5;

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

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

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

grunt> dump joined\_table;



Output:-

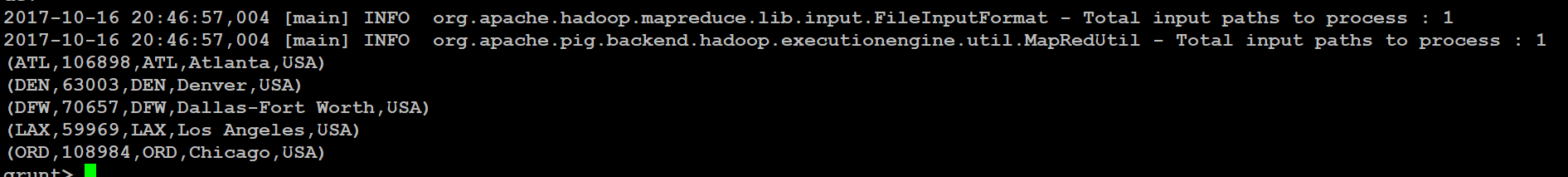
(ATL,106898,ATL,Atlanta,USA)

(DEN,63003,DEN,Denver,USA)

(DFW,70657,DFW,Dallas-Fort Worth,USA)

(LAX,59969,LAX,Los Angeles,USA)

(ORD,108984,ORD,Chicago,USA)



## **Problem Statement 2**

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

grunt> REGISTER '/home/cloudera/Downloads/piggybank.jar';

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

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

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

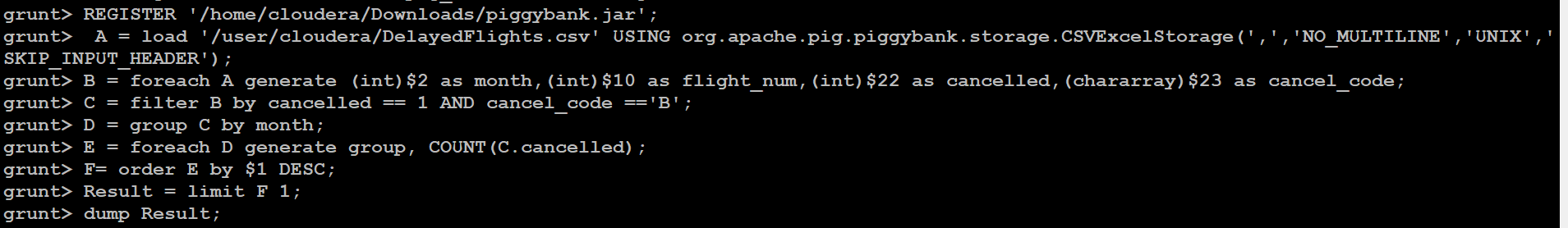
grunt> D = group C by month;

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

grunt> F= order E by $1 DESC;

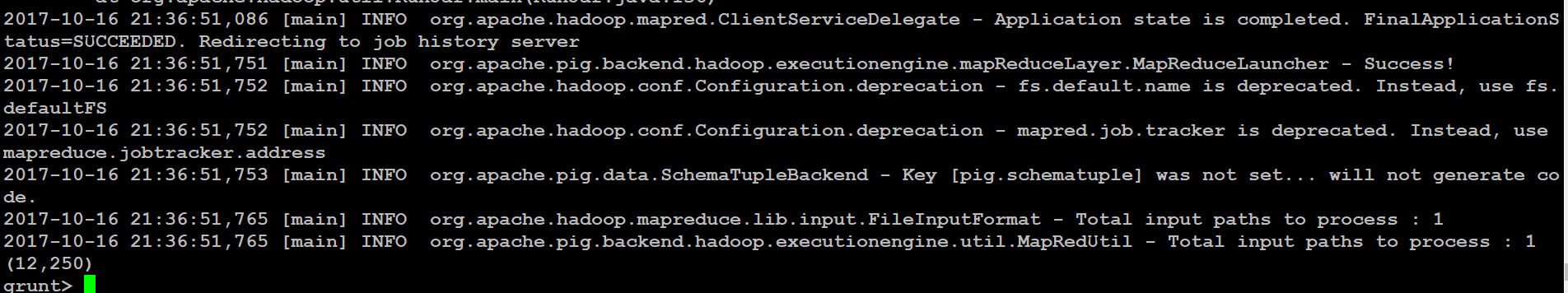
grunt> Result = limit F 1;

grunt> dump Result;



Output:-

(12,250)



## **Problem Statement 3**

*Top ten origins with the highest AVG departure delay*

grunt> REGISTER '/home/cloudera/Downloads/piggybank.jar';

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

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

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

grunt> D1 = group C1 by origin;

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

grunt> Result = order E1 by $1 DESC;

grunt> Top\_ten = limit Result 10;

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

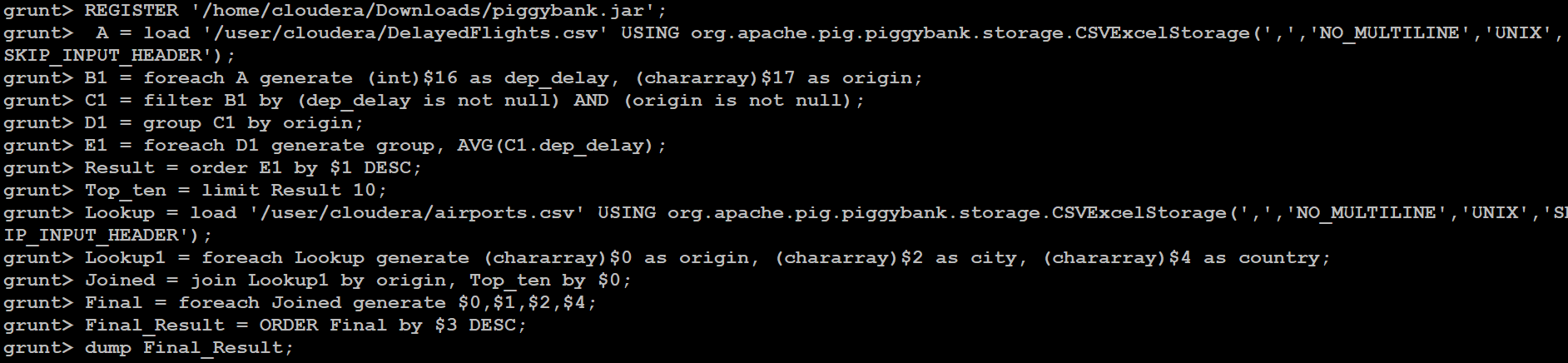
grunt> Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;

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

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

grunt> Final\_Result = ORDER Final by $3 DESC;

grunt> dump Final\_Result;



Output:-

(CMX,Hancock,USA,116.1470588235294)

(PLN,Pellston,USA,93.76190476190476)

(SPI,Springfield,USA,83.84873949579831)

(ALO,Waterloo,USA,82.2258064516129)

(MQT,NA,USA,79.55665024630542)

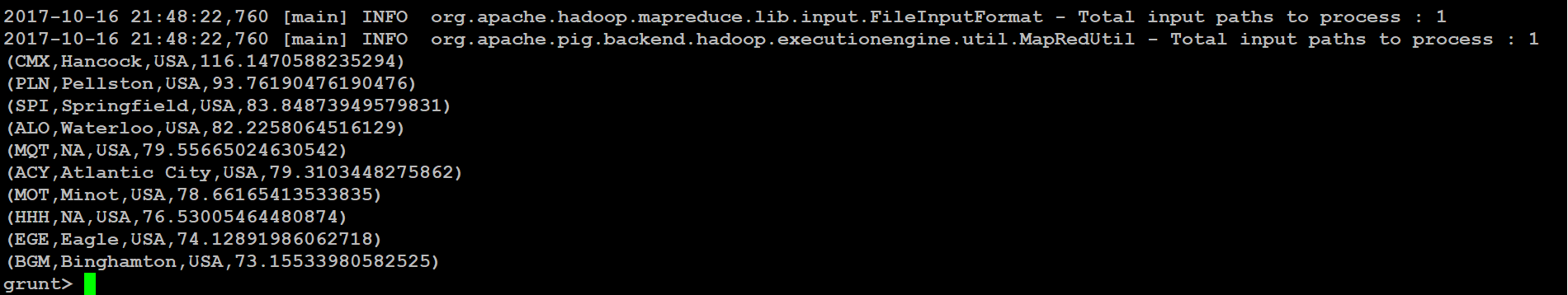
(ACY,Atlantic City,USA,79.3103448275862)

(MOT,Minot,USA,78.66165413533835)

(HHH,NA,USA,76.53005464480874)

(EGE,Eagle,USA,74.12891986062718)

(BGM,Binghamton,USA,73.15533980582525)



## **Problem Statement 4**

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

grunt> REGISTER '/home/cloudera/Downloads/piggybank.jar';

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

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

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

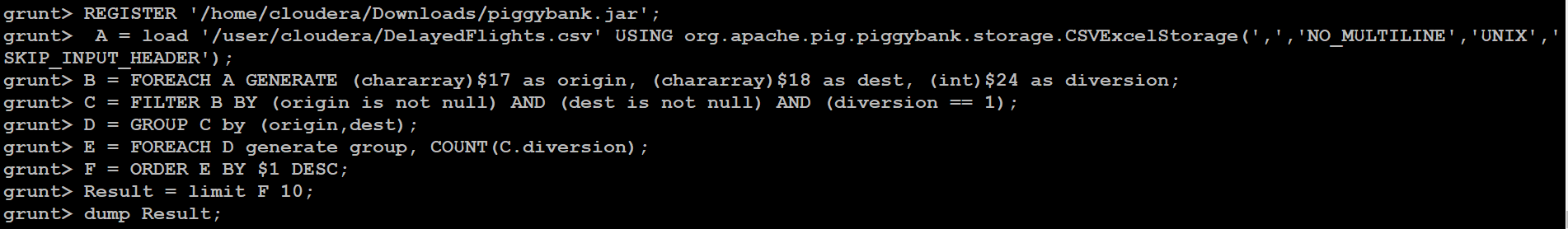
grunt> D = GROUP C by (origin,dest);

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

grunt> F = ORDER E BY $1 DESC;

grunt> Result = limit F 10;

grunt> dump Result;



Output:-

((ORD,LGA),39)

((DAL,HOU),35)

((DFW,LGA),33)

((ATL,LGA),32)

((ORD,SNA),31)

((SLC,SUN),31)

((MIA,LGA),31)

((BUR,JFK),29)

((HRL,HOU),28)

((BUR,DFW),25)

