**Session 5: EXPLORIN PIG**

Assignment 5.2

Student Name: Abarajithan SA

Course: Big Data Hadoop & Spark Training

Start Date:  2017-09-09

End Date:  2017-11-26

**Assignment 5.2**–

Implement the use case present in below blog link and share the complete steps along with

Screenshot from your end.

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

Contents

[Input Data Sets: 2](#_Toc496444425)

[Problem Statement 1 2](#_Toc496444426)

[Find out the top 5 most visited destinations 2](#_Toc496444427)

[Output: the top 5 most visited destinations 3](#_Toc496444428)

[Problem Statement 2 4](#_Toc496444429)

[Which month has seen the most number of cancellations due to bad weather? 4](#_Toc496444430)

[Output: month has seen the most number of cancellations due to bad weather? 4](#_Toc496444431)

[Problem Statement 3 5](#_Toc496444432)

[Top ten origins with the highest AVG departure delay 5](#_Toc496444433)

[Output Top ten origins with the highest AVG departure delay 5](#_Toc496444434)

[Problem Statement 4 7](#_Toc496444435)

[Which route (origin & destination) has seen the maximum diversion? 7](#_Toc496444436)

[Outputwhich route (origin & destination) has seen the maximum diversion? 7](#_Toc496444437)

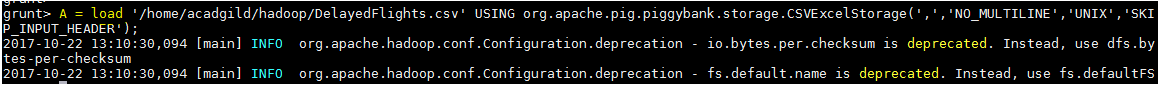
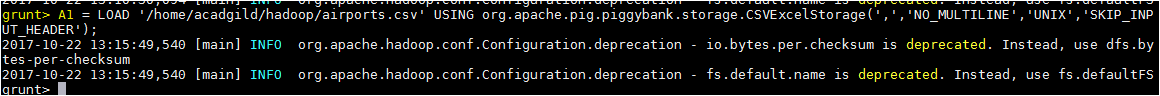
# Input Data Sets:

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

# Problem Statement 1

## Find out the top 5 most visited destinations

Codes:

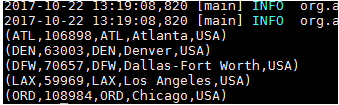
1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
3. **B = FOREACH A GENERATE (int)$1 as year, (int)$10 as flight\_num, (chararray)$17 as origin,(chararray) $18 as dest;**
4. **C = FILTER B by dest is NOT Null;**
5. **D = GROUP C BY dest;**
6. **E = FOREACH D GENERATE group, COUNT(C.dest);**
7. **F = ORDER E by $1 DESC;**
8. **Result = LIMIT F 5;**
9. **A1 = LOAD '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
10. **A2 = FOREACH A1 GENERATE (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;**
11. **joined\_table = JOIN Result by $0, A2 by dest;**
12. **DUMP joined\_table;**
13. 
14. 
15. 
16. 
17. 
18. 
19. 
20. 
21. 
22. 
23. 
24. 

## Output: the top 5 most visited destinations

Original Output:



Executed Output:

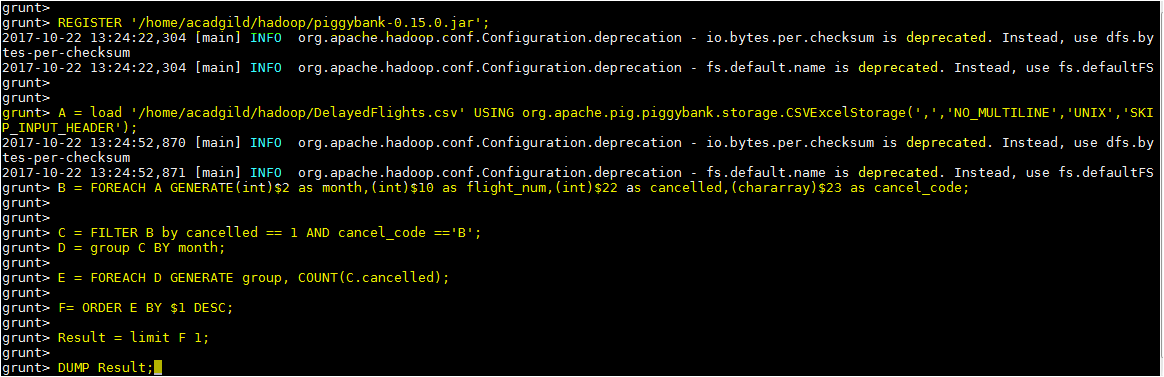


# Problem Statement 2

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

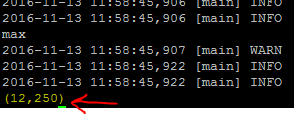
Codes:

1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
3. **B = FOREACH A GENERATE(int)$2 as month,(int)$10 as flight\_num,(int)$22 as cancelled,(chararray)$23 as cancel\_code;**
4. **C = FILTER B by cancelled == 1 AND cancel\_code =='B';**
5. **D = group C BY month;**
6. **E = FOREACH D GENERATE group, COUNT(C.cancelled);**
7. **F= ORDER E BY $1 DESC;**
8. **Result = limit F 1;**
9. **DUMP Result;**

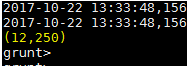


## Output: month has seen the most number of cancellations due to bad weather?

Original Output:



Executed Output:

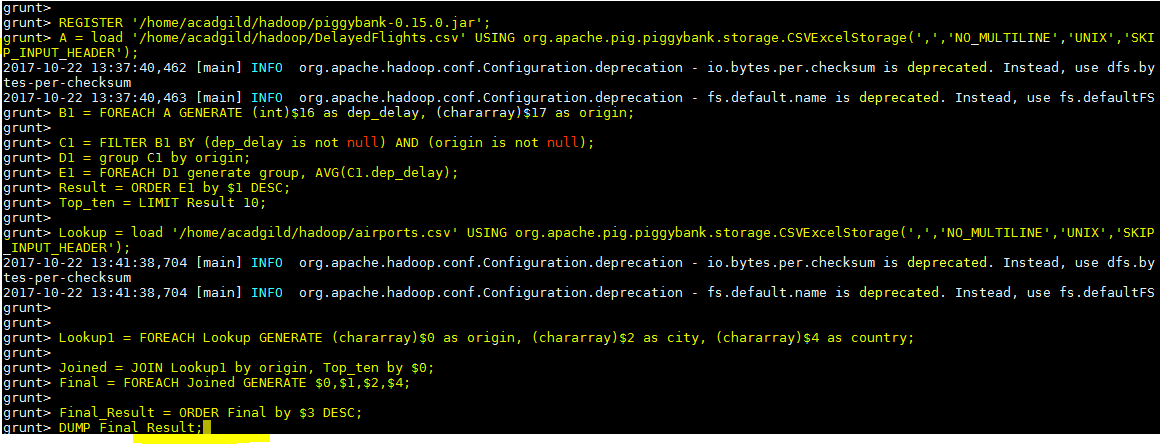


# Problem Statement 3

## Top ten origins with the highest AVG departure delay

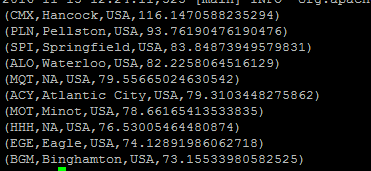
Codes:

1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
3. **B1 = FOREACH A GENERATE (int)$16 as dep\_delay, (chararray)$17 as origin;**
4. **C1 = FILTER B1 BY (dep\_delay is not null) AND (origin is not null);**
5. **D1 = group C1 by origin;**
6. **E1 = FOREACH D1 generate group, AVG(C1.dep\_delay);**
7. **Result = ORDER E1 by $1 DESC;**
8. **Top\_ten = LIMIT Result 10;**
9. **Lookup = load '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
10. **Lookup1 = FOREACH Lookup GENERATE (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;**
11. **Joined = JOIN Lookup1 by origin, Top\_ten by $0;**
12. **Final = FOREACH Joined GENERATE $0,$1,$2,$4;**
13. **Final\_Result = ORDER Final by $3 DESC;**
14. **DUMP Final\_Result;**

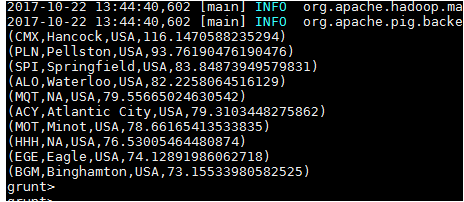


## Output Top ten origins with the highest AVG departure delay

Original Output



Executed Output:

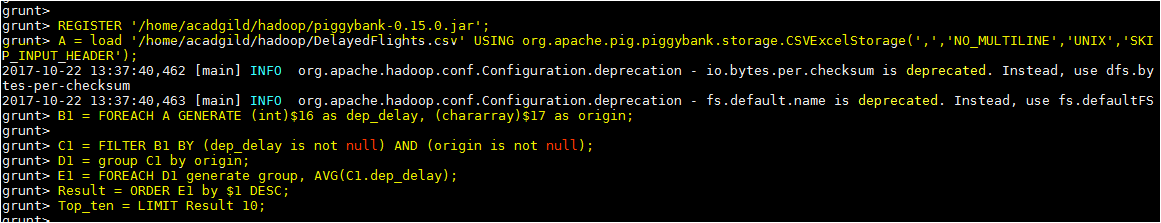


# Problem Statement 4

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

Codes:

1. **REGISTER '/home/acadgild/airline\_usecase/piggybank.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');**
3. **B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;**
4. **C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);**
5. **D = GROUP C by (origin,dest);**
6. **E = FOREACH D generate group, COUNT(C.diversion);**
7. **F = ORDER E BY $1 DESC;**
8. **Result = LIMIT F 10;**
9. **DUMP Result;**



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

Original output:



Executed output:

