**Delayed\_Flights.csv Datasets**

Name Description

1 Year 1987-2008

2 Month 1-12

3 DayofMonth 1-31

4 DayOfWeek 1 (Monday) - 7 (Sunday)

5 DepTime actual departure time (local, hhmm)

6 CRSDepTime scheduled departure time (local, hhmm)

7 ArrTime actual arrival time (local, hhmm)

8 CRSArrTime scheduled arrival time (local, hhmm)

9 UniqueCarrier unique carrier code

10 FlightNum flight number

11 TailNum plane tail number

12 ActualElapsedTime in minutes

13 CRSElapsedTime in minutes

14 AirTime in minutes

15 ArrDelay arrival delay, in minutes

16 DepDelay departure delay, in minutes

17 Origin origin IATA airport code

18 Dest destination IATA airport code

19 Distance in miles

20 TaxiIn taxi in time, in minutes

21 TaxiOut taxi out time in minutes

22 Cancelled was the flight cancelled?

23 CancellationCode reason for cancellation (A = carrier, B = weather, C = NAS, D = security)

24 Diverted 1 = yes, 0 = no

25 CarrierDelay in minutes

26 WeatherDelay in minutes

27 NASDelay in minutes

28 SecurityDelay in minutes

29 LateAircraftDelay in minutes

**Airports.csv Datasets**

"iata",

"airport",

"city",

"state",

"country",

"lat",

"long"

<https://github.com/Rahulagn/Assignment-5.1>

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

==============================================================

pig -x local

REGISTER '/home/cloudera/lib/piggybank.jar';

A = load '/home/cloudera/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE');

===============================================

P1:

===============================================

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/cloudera/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE');

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;

P2:

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;

P3:

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/cloudera/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE');

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;

P4:

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;

**Problem Statement 1**

Find out the top 5 most visited destinations.

Image :1

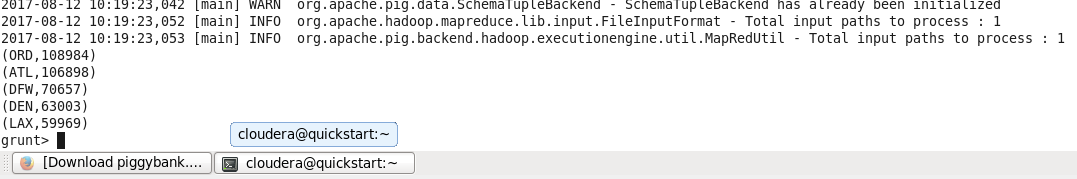


Image 2:



## Problem Statement 2

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

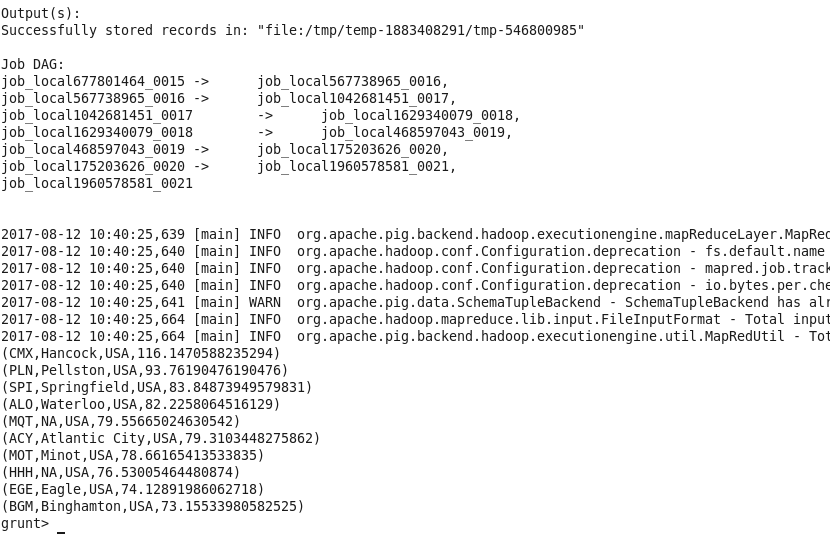
Image : 1



## Problem Statement 3

Top ten origins with the highest AVG departure delay

Image 1:



## Problem Statement 4

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

Image 1:

