**Assignment 5.2**

**Aviation Data Analysis Using Apache Pig**

**INPUT:**

[Delayed\_Flights.csv](https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVFOdzlJNWM/view?usp=sharing)

[Airports.csv](https://drive.google.com/file/d/0B_Qjau8wv1KocDR3djk1Qm96Mmc/view?usp=sharing)

These are 2 different datasets, i.e., Delayed\_Flights.csv and Airports.csv. Let us understand one at a time.

**Delayed\_Flights.csv Datasets**

There are 29 columns in this dataset. Some of them have been mentioned below:

* Year: 1987 – 2008
* Month: 1 – 12
* FlightNum: Flight number
* Canceled: Was the flight canceled?
* CancelleationCode: The reason for cancellation.

**Airports.csv Datasets**

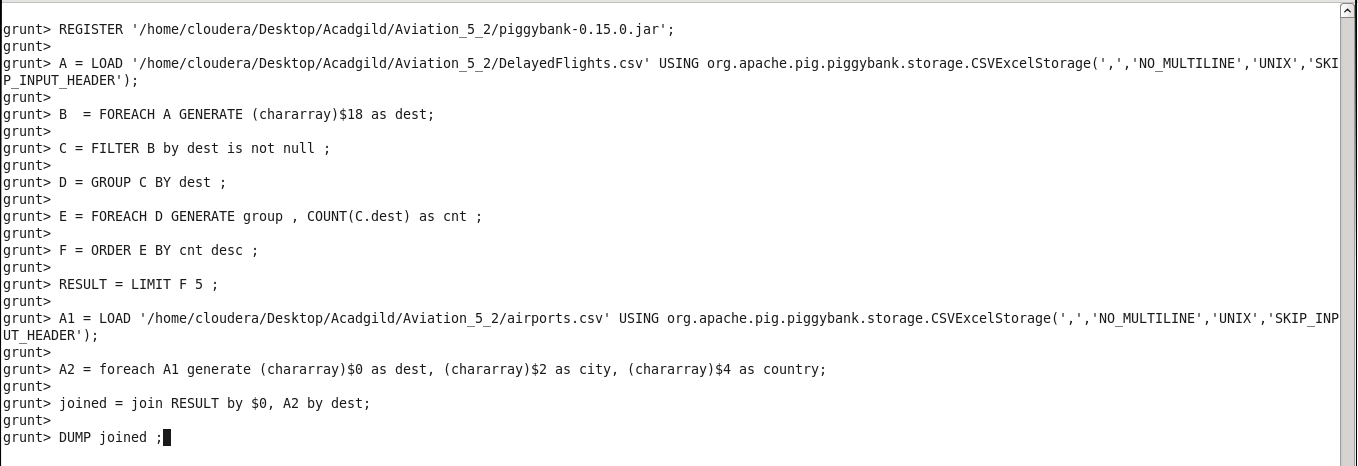
* iata: the international airport abbreviation code
* name of the airport
* city and country in which airport is located.
* lat and long: the latitude and longitude of the airport

Register Piggybank jar

## Problem Statement 1

Find out the top 5 most visited destinations.

**CODE:**



**Explanation:** In this code ,

* 1. Register the piggybank jar
  2. Loading the data in the pig
  3. Transforming the data
  4. Filtering the Null destination
  5. Counting the Destination visit
  6. Ordering to find the top 5
  7. Limit the result to 5

We have use join also to print the city and country as well

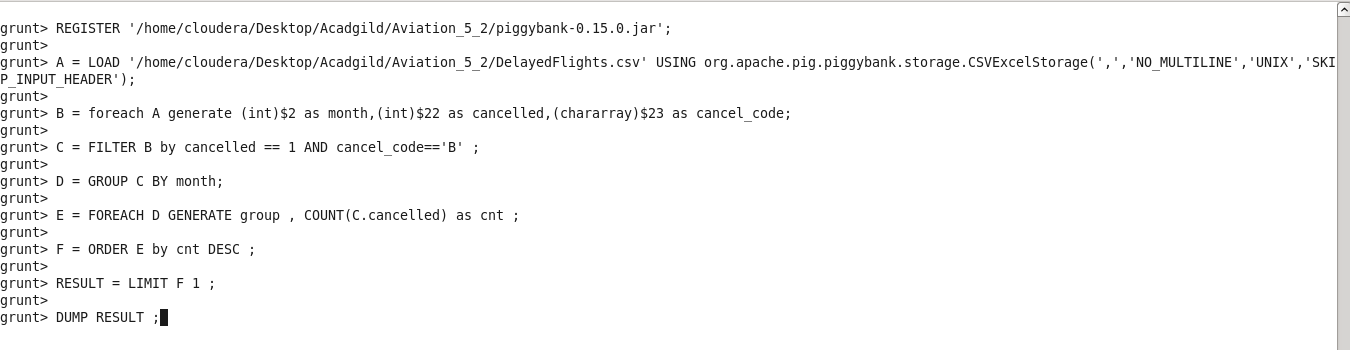
* 1. Loading the data
  2. Transforming the data
  3. Joining both the result
  4. Printing the final output

**OUTPUT:**



## **Problem Statement 2**

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



**EXPLANATION:**

1. Registering the Piggybank jar file
2. Loading the input data
3. Transforming the data
4. Filter the data on basis of criteria
5. Grouping the data on month
6. Counting the cancellations on the each month
7. Order the data by which we will get the month with most cancellations
8. Selecting the top month
9. Dumping the data

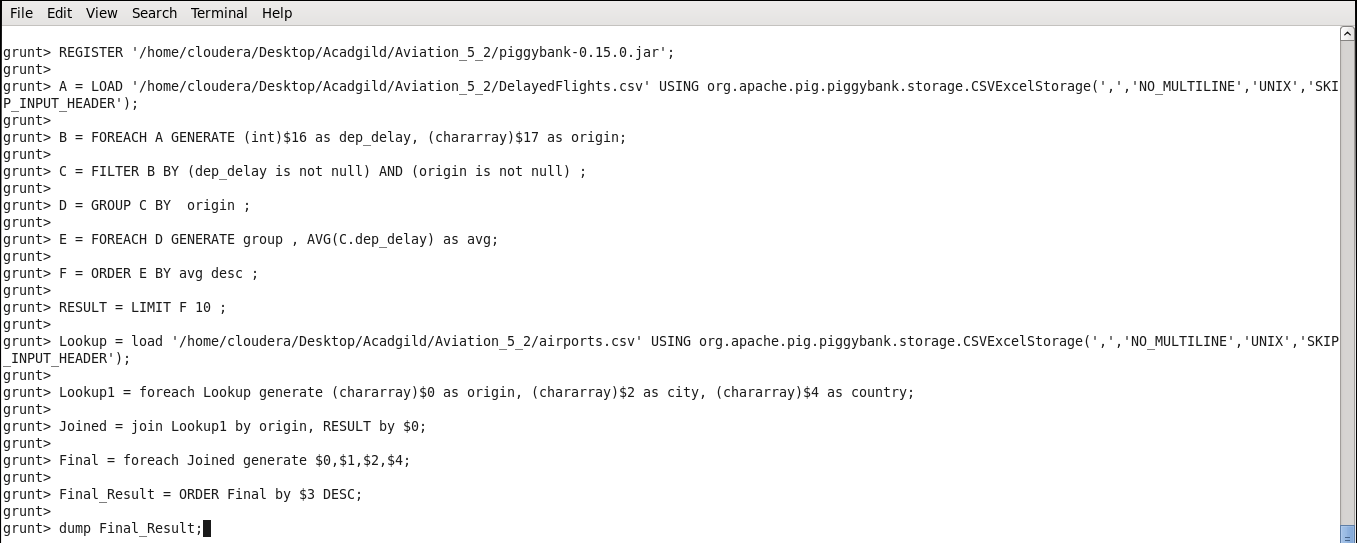
**OUTPUT:**



## **Problem Statement 3**

**Top ten origins with the highest AVG departure delay**

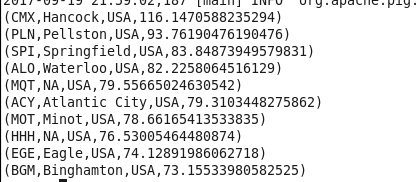
**CODE:**



**EXPLAINATION:**

1. First we register the Piggybank jar
2. Loading the input
3. Transforming the data
4. Filtering the data
5. Group by origin
6. Finding avg delay
7. Ordering the data by delay
8. In Result only top 10
9. Loading the look up table for other information
10. Transforming the data
11. Joining the previous data with this lookup
12. Print the final output

**OUTPUT :**



## **Problem Statement 4**

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

**CODE:**



**EXPLANATION :**

1. Register Piggybank jar
2. Loading the data in pig
3. Transforming the data
4. Filtering the data
5. Group by origin and destination
6. Counting the route which has maximum diversion
7. Ordering the data
8. Printing the top data
9. Printing the result



**NOTE : THE GITHUB repository will contain the Code and Document regarding Assignment**