**Aviation Data Analysis Using Apache Pig**

These are 2 different datasets, i.e., Delayed\_Flights.csv and Airports.csv.

**1. Delayed\_Flights.csv Datasets**

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

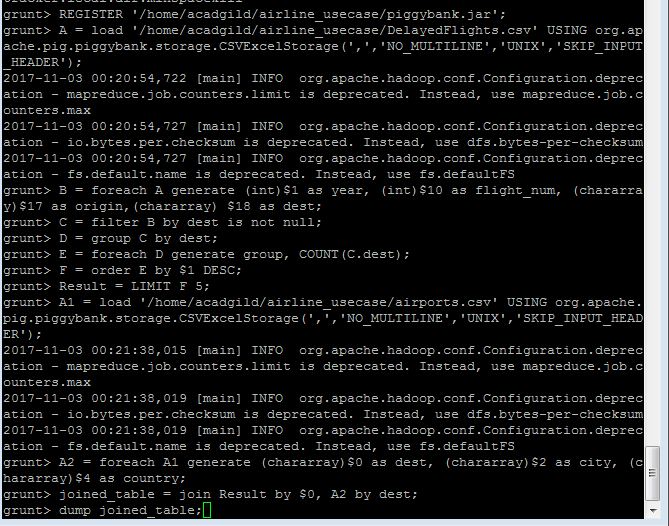
* Year: 1987 – 2008
* Month: 1 – 12
* FlightNum: Flight number
* Canceled: Was the flight canceled?
* CancelleationCode: The reason for cancellation.

**2. Airports.csv Datasets**

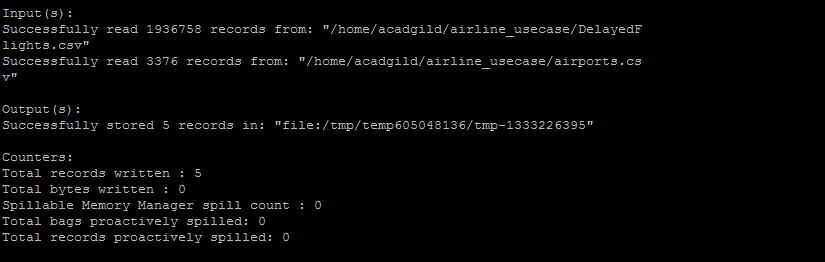
* iata: the international airport abbreviation code
* name of the airport
* city and country in which airport is located.
* lat and long: the latitude and longitude of the airport

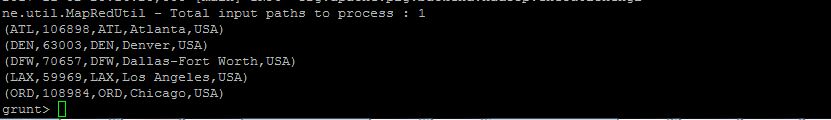
**Problem - 1**

* Find out the top 5 most visited destinations.



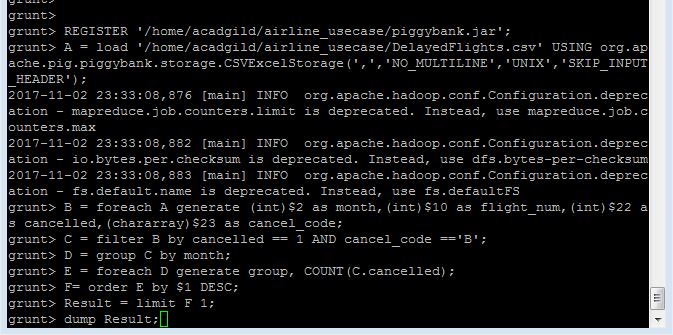
1. **In Line 1**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
2. In relation **A**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
3. In relation **B**, we are generating the columns that are required for processing and explicitly typecasting each of them.
4. In relation **C**, we are filtering the null values from the “dest” column.
5. In relation **D**, we are grouping relation C by “dest.”
6. In relation **E**, we are generating the grouped column and the count of each.
7. Relation **F** and **Result** is used to order and limit the result to top 5.
8. These are the steps to find the top 5 most visited destinations. However, adding few more steps in this process, we will be using another table to find the city name and country as well.
9. In relation **A1**, we are loading another table to which we will look-up and find the city as well as the country.
10. In relation **A2**, we are generating dest, city, and country from the previous relation.
11. In relation **joined\_table**, we are joining Result and A2 based on a common column, i.e., “dest”
12. Finally, using dump, we are printing the result.



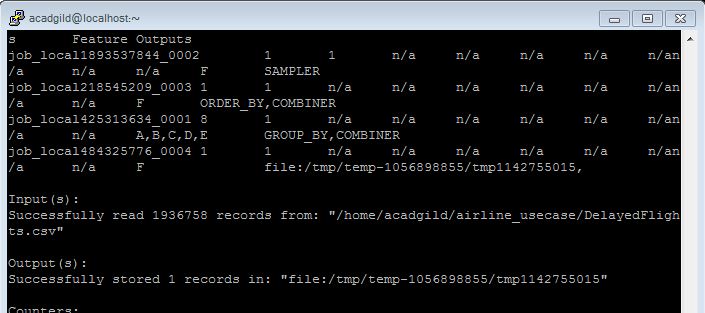


## Problem - 2

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



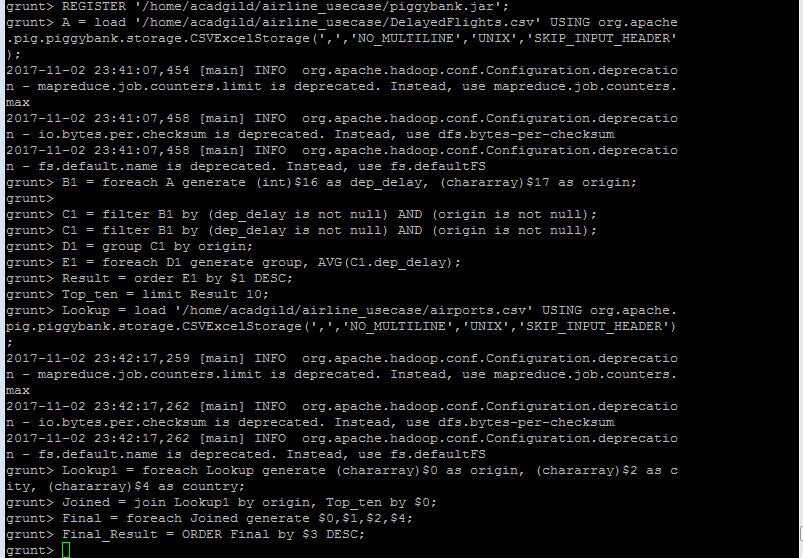
1. **In Line 1**: We are registering *piggybank* jar in order to use the CSVExcelStorage class.
2. In relation **A**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and header.
3. In relation **B**, we are generating the columns which are required for processing and explicitly typecasting each of them.
4. In relation **C**, we are filtering the data based on cancellation and cancellation code, i.e., canceled = 1 means flight have been canceled and cancel\_code = ‘B’ means the reason for cancellation is “weather.” So relation C will point to the data which consists of canceled flights due to bad weather.
5. In relation**D**, we are grouping the relation C based on every month.
6. In relation **E**, we are finding the count of canceled flights every month.
7. Relation **F** and **Result** is for ordering and finding the top month based on cancellation.



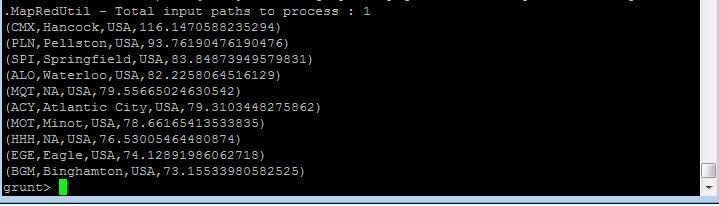


## Problem - 3

* Top ten origins with the highest AVG departure delay

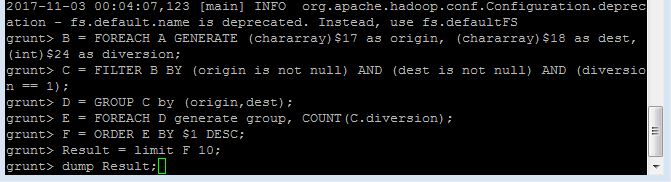


1. **In Line 1**: We are registering *piggybank* jar in order to use the CSVExcelStorage class.
2. In relation **A**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and header.
3. In relation **B1**, we are generating the columns which are required for processing and explicitly typecasting each of them.
4. In relation**C1**, we are removing the null values fields present if any.
5. In relation **D1**, we are grouping the data based on column “origin.”
6. In relation **E1**, we are finding average delay from each unique origin.
7. Relations named **Result** and **Top\_ten** are ordering the results in descending order and printing the top ten values.
8. These steps are good enough to find the top ten origins with the highest average departure delay.
9. However, rather than generating just the code of origin, we will be following a few more steps to find some more details like country and city.
10. In the relation **Lookup**, we are loading another table to which we will look up and find the city as well as the country.
11. In the relation L**ookup1,** we are generating the destination, city, and country from the previous relation.
12. In the relation **Joined**, we are joining relation Top\_ten and Lookup1 based on common a column, i.e., “origin.”
13. In the relation**Final,**we are generating required columns from the Joined table.
14. Finally, we are ordering and printing the results



## Problem - 4

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



1. **In Line 1**: We are registering *piggybank* jar in order to use CSVExcelStorage class.
2. In relation **A**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
3. In relation **B**, we are generating the columns which are required for processing and explicitly type-casting each of them.
4. In relation **C**, we are filtering the data based on “not null” and diversion =1. This will remove the null records, if any, and give the data corresponding to the diversion taken.
5. In relation **D**, we are grouping the data based on origin and destination.
6. Relation **D** finds the count of diversion taken per unique origin and destination.
7. Relations **F** and **Result** orders the result and produces top 10 results.

