Assignment 5.2

**Problem Statement 1**

Find out the top 5 most visited destinations.

### Register the piggybank jar in order to use the CSVExcelStorage class.

## 

## load the dataset using CSVExcelStorage

## 

## Generate the required columns, filter NULL values from ‘dest’ column, COUNT by grouped ‘dest’. LIMT result to 5

In relation **A1**, we are loading another table to which we will look-up and find the city as well as the country.

In relation **A2**, we are generating dest, city, and country from the previous relation.

## In relation **joined\_table**, we are joining Result and A2 based on a common column, i.e., “dest”

## 

## 

## Problem Statement 2

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

In relation **A**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.

In relation **B**, we are generating the columns that are required for processing and explicitly typecasting each of them.

In relation **C**, we are filtering the null values from the “dest” column.

In relation **D**, we are grouping relation C by “dest.”

In relation **E**, we are generating the grouped column and the count of each.

Relation **F** and **Result** is used to order and limit the result to top 5.

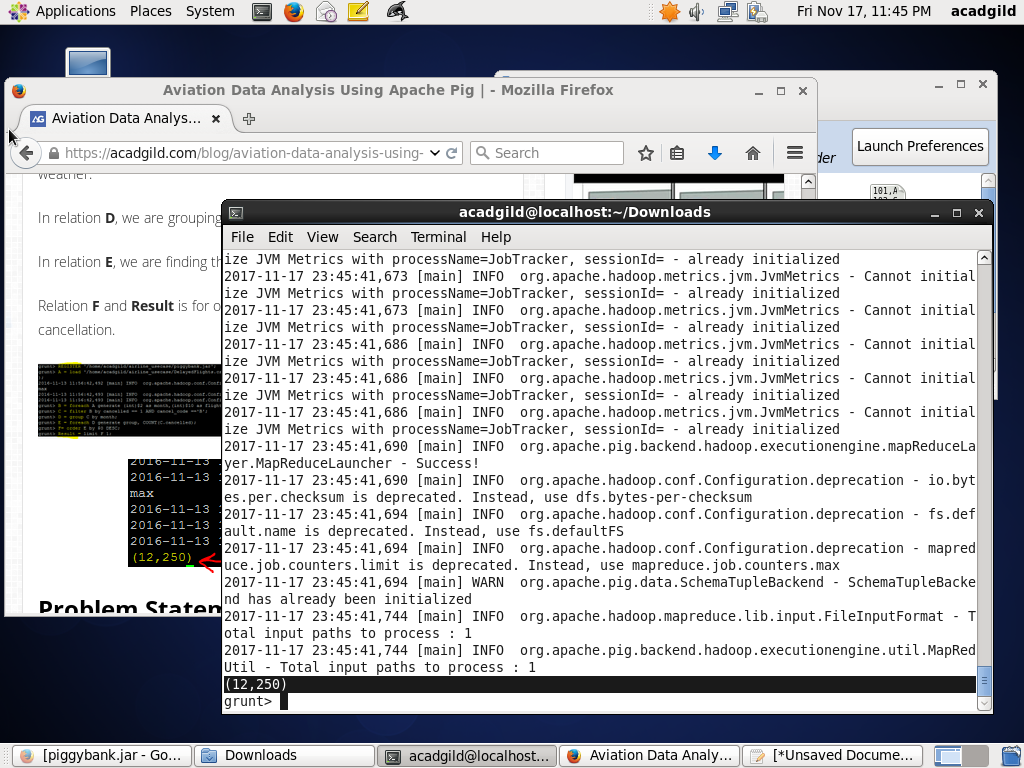
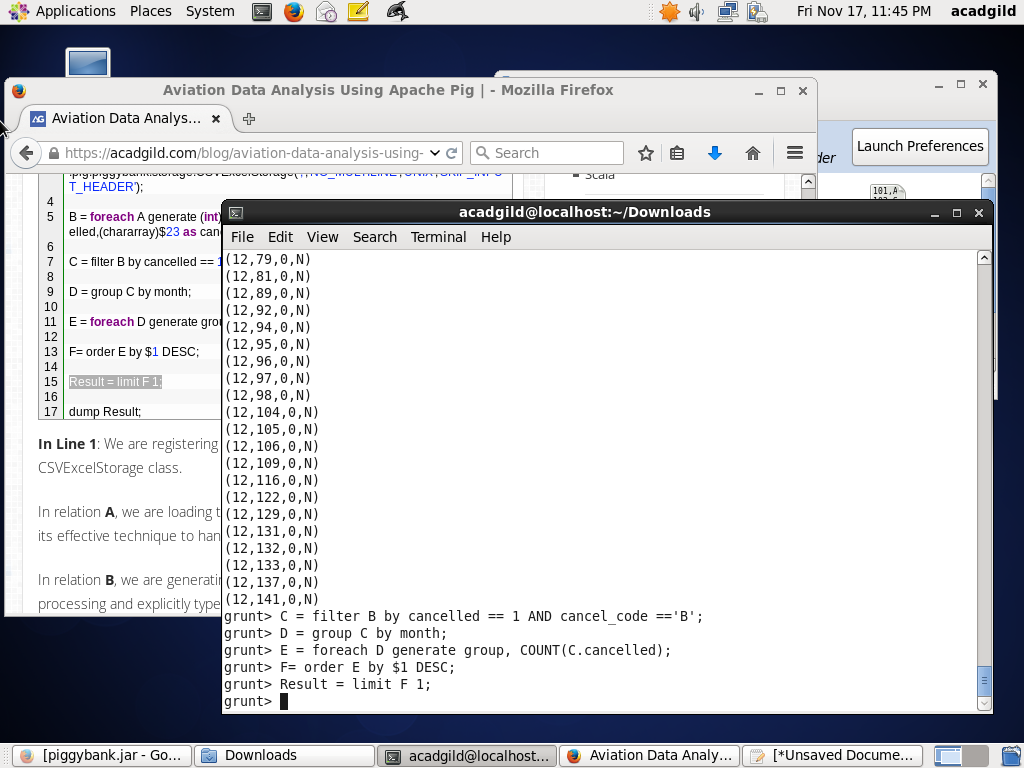
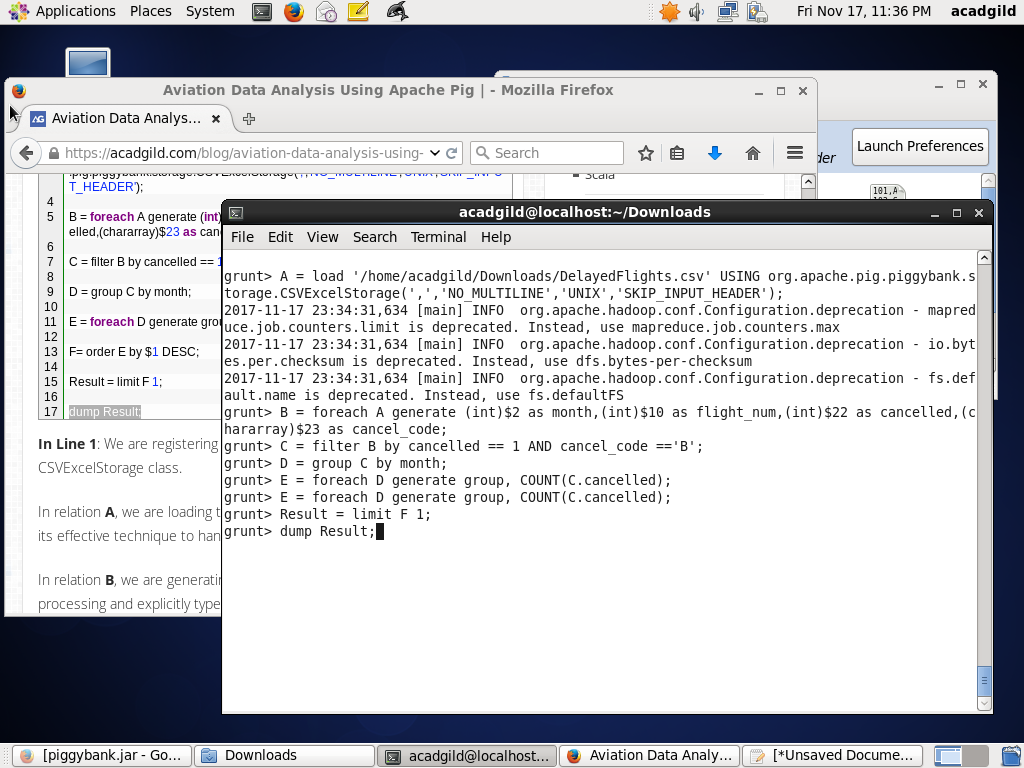
These are the steps to find the top 5 most visited destinations. However, adding few more steps in this process, we will be using another table to find the city name and country as well.

In relation **A1**, we are loading another table to which we will look-up and find the city as well as the country.

In relation **A2**, we are generating dest, city, and country from the previous relation.

In relation **joined\_table**, we are joining Result and A2 based on a common column, i.e., “dest”

Finally, using dump, we are printing the result.



## Problem Statement 3

Top ten origins with the highest AVG departure delay

In relation**C1**, we are removing the null values fields present if any.

In relation **D1**, we are grouping the data based on column “origin.”

In relation **E1**, we are finding average delay from each unique origin.

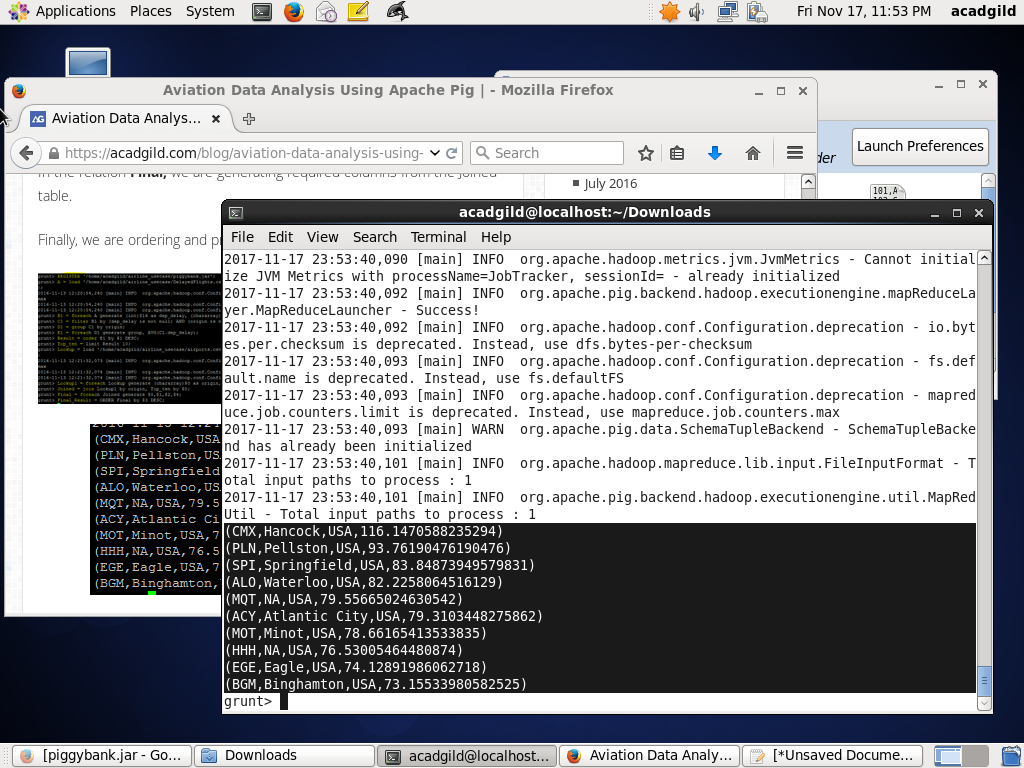
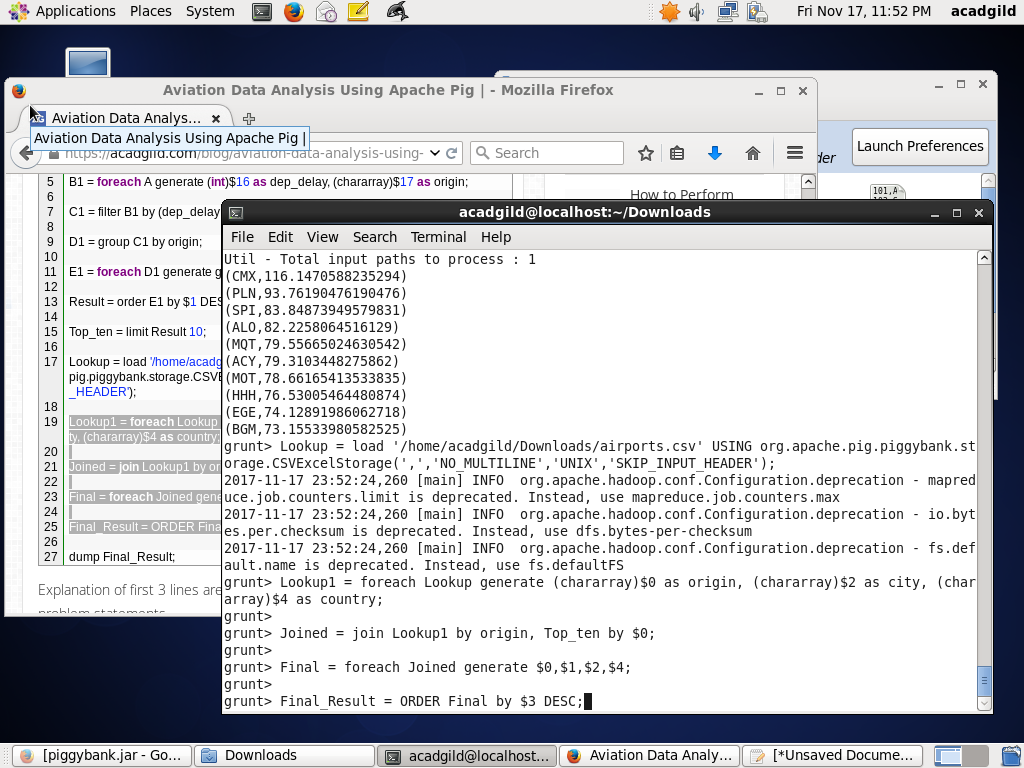
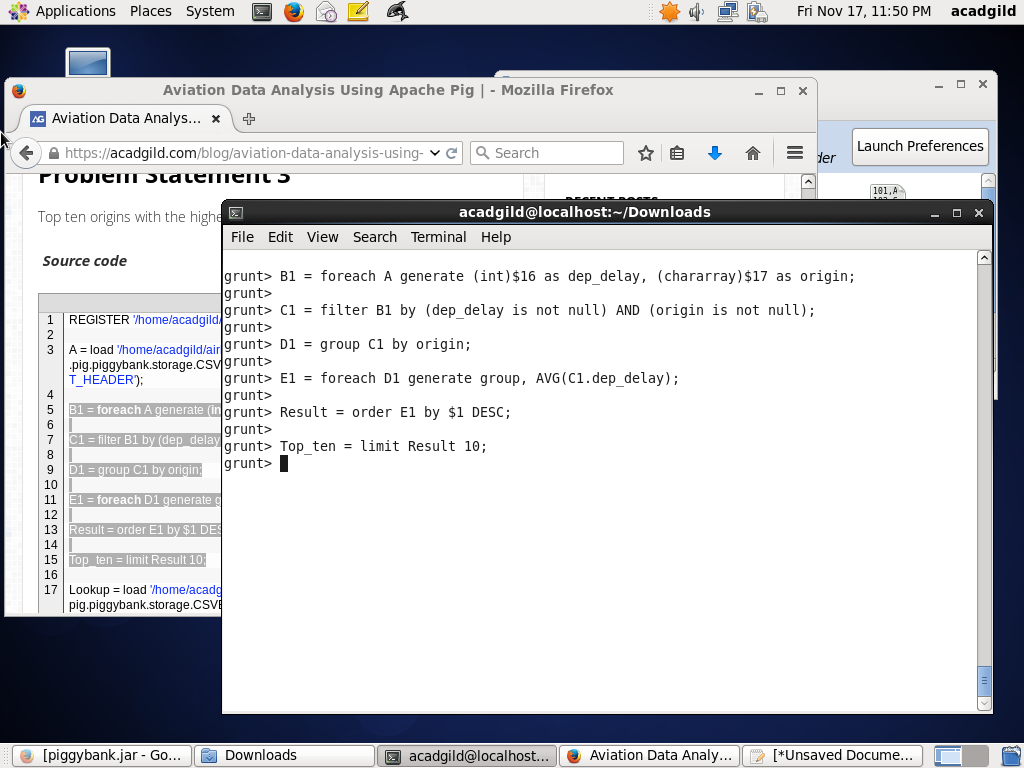
n the relation **Lookup**, we are loading another table to which we will look up and find the city as well as the country.

In the relation L**ookup1,** we are generating the destination, city, and country from the previous relation.

In the relation **Joined**, we are joining relation Top\_ten and Lookup1 based on common a column, i.e., “origin.”

In the relation**Final,**we are generating required columns from the Joined table.

Finally, we are ordering and printing the results.



## Problem Statement 4

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

**Solution**

In relation **B**, we are generating the columns which are required for processing and explicitly type-casting each of them.

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.

In relation **D**, we are grouping the data based on origin and destination.

Relation **D** finds the count of diversion taken per unique origin and destination.

Relations **F** and **Result** orders the result and produces top 10 results.

