Departamento de Ciencias de la Computación
UNIVERSIDAD DE CHILE

ANALYSING THE AIRLINE DATASET

Massive Data Processing (CC5212)

Date: August 10th, 2016
Members: Francisco Peters
Diego Salazar
Professor: Aidan Hogan

I. Goal

The goal of the project was to analyze a large dataset about flights in a variety of airports in the USA during the last few years using hadoop. Using techniques learned from the labs and also learn a few new tricks using distributed computing in the DCC server cluster. All map and reduce processes were run remotely using an ssh connection.

The dataset included a wide variety of attributes so we decided to use them to find a correlation between flights delays and some other variables. Some of the results are showed using visualizations to get a more clear understanding.

II. Data

The source of the data is: http://stat-computing.org/dataexpo/2009.
This dataset contains flight arrival and departure details for all commercial flights within the USA during 2008. We decided to work on this dataset because we thought it was both interesting and useful to work on it, because one can get very valuable data out of it, and even get parameters to decide which one is the best airline, or when it is convenient to travel, and more. We tried to find chilean airlines dataset, but sadly we couldn't do so, that's why we chose to work on the available USA airlines dataset.

The dataset comes in the format .csv.bz2 (although it is not necessary to uncompress it) and it weights 657Mb. This dataset contains 7.009.728 lines, featuring the following variables: Name, Year, Month, DayofMonth, DayOfWeek, DepTime, CRSDepTime, **ArrTime**, **CRSArrTime**, UniqueCarrier, FlightNum, TailNum, ActualElapsedTime, CRSElapsedTime, **AirTime**, ArrDelay, DepDelay, Origin, Dest, Distance, TaxiIn, TaxiOut, Cancelled, CancellationCode, Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay.
The variables in bold are the ones we worked with.

III. Methods

We first ran a Map-Reduce method using Hadoop and a simple function to calculate the number of total flights, just using a sum and getting all the flights related to one airline.

The second method was to get the number of delayed flights, this was calculated by comparing the arrival time (in minutes) and the official arrival time (in minutes) for each flight, when the arrival time was bigger than the official time it was counted as a delay, after this we could get a relation between the two and show it as a little quality measure for airlines going from 0.0 to 1.0.

Last function was to take all the airtime for each flight and make a sum to relate them to the airline. There are detailed explanations in the code even though it's pretty easy to understand.

Classes:

AirlineDelayByStartTime.java: Main class, it sets the mapper and reducer classes, sets the number of reducers tasks, the input and output and runs the jobs.

AirlineDelayByStartTimeMapper.java: Mapper class, it takes info from the dataset and translate it into variables.

AirlineDelayByStartTimeReducer.java: Reducer class, it's in charge of making all the calculations and writing to the output.

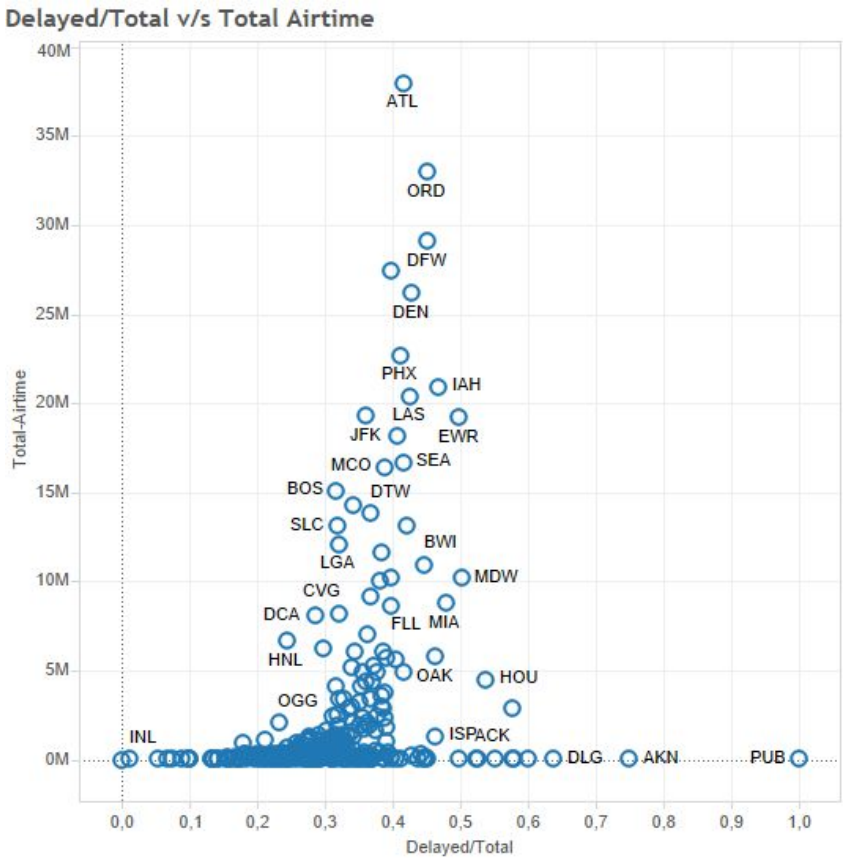
IV. Results

Here are some visualizations using the data obtained by the map/reduce jobs:

Graph 1: Total Flights vs Delayed Flights.

Here in the Graph 1 in the X axis we have the total flights and in the Y axis the delayed flights, each point represents an airline.

We observe that the number of flights per airline has a direct correlation with the number of delayed flights (duh), it's surprising though that this is pretty much a perfect straight line.



Graph 2: (Delayed Flights)/(Total Flights) vs Total Airtime

In the Graph 2 we can see in the X axis the relation between Delayed Flights and Total Flights and in the Y axis we have the total Airtime, there's an accumulation of points in the center and something that resembles a Gaussian bell, we believe these results may be related to some level of expertise the airlines adquire after having longer flights, may be related to the fact that longer flights means the need to be synchronous with other airlines from other countries. Also note that a bigger number of flights also correlates with a bigger total airtime, so more flights may be related to some level of expertise too. The left half of the graph seems pretty obvious since bigger time may cause more delays in the journey.

V. Conclusion

I think the biggest achievement from this project was to apply techniques and lessons from the lab into real world datasets, it wasn't just a matter of getting some data from a database and then showing it, using distributed computing techniques helped us to have a better understanding of Hadoop and map/reducing in general, as a personal note I was fascinated by uploading a .jar file into the cluster and then run it remotely and view the results almost instantly (even though this same thing was done in the lab, it didn't had the same impact since the cluster server was in the same building).

VI. Apendix

To read some information about the dataset:
http://stat-computing.org/dataexpo/2009/the-data.html

It has detailed information about the columns and what do they mean.