## INFO 7275 Advance Database Management Systems

**Airline On-Time statistics and arrival data analysis**

**Introduction:**

According to the U.S Bureau of Transportation Statistics, more than 1 million flights were delayed between June 2015 to June 2016. This translates to nearly 64 million minutes’ worth of delay.  
Considering all the delays and its types, it becomes necessary to consider the efficient way of travel.

The aim of this project will be to use various Map Reduce design patterns like Summarization design pattern, Data Organization Patterns, Join Patterns and Filtering Patterns to graphically represent the following questions

* Best time to fly to minimize delays
* Main factors of delays
* Best airlines to avoid delays
* Top N busiest airports

**Dataset:**

The data set consists of on-time arrival for domestic flights by major air carriers and provides additional items as departure and arrival delays, origin and destination airports, flight numbers, schedule and actual departure and arrival times, cancelled and diverted flights, taxi-out and taxi-in times.

Other supplement data:

carriers.csv   
airports.csv  
plane-data.csv

Links to the data set:

<https://www.transtats.bts.gov/Fields.asp?Table_ID=236>

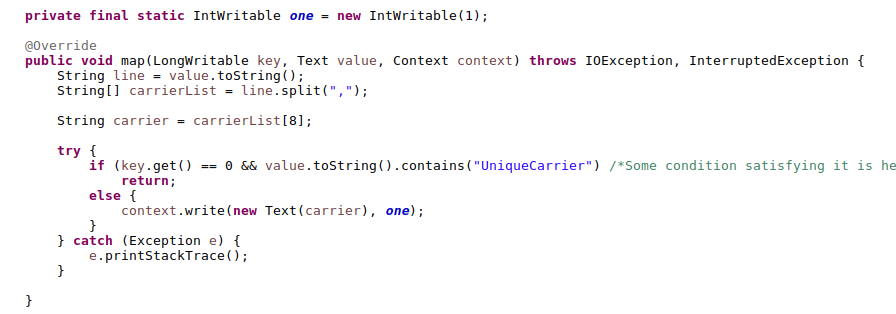
[http://stat-computing.org/dataexpo/2009/the-data.html](https://www.transtats.bts.gov/Fields.asp?Table_ID=236)

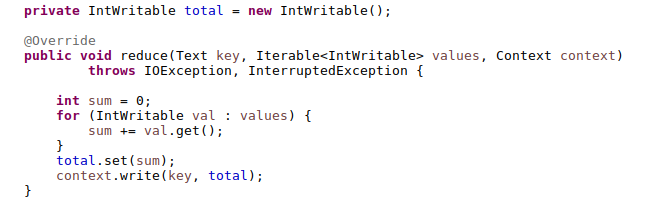
[https://www.transtats.bts.gov/TableInfo.asp](https://www.transtats.bts.gov/Fields.asp?Table_ID=236)

**Analysis performed:**

1. Total number of flights in a year

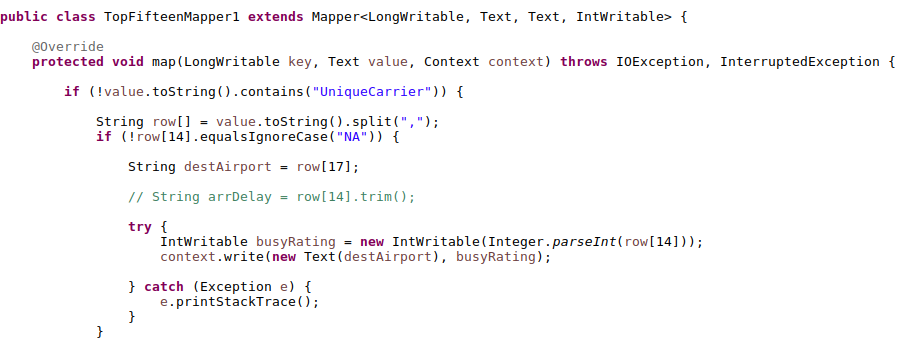
Using Summarization Pattern and calculating the total occurrence of each Unique Carrier, we can find out the total number of flights in year.

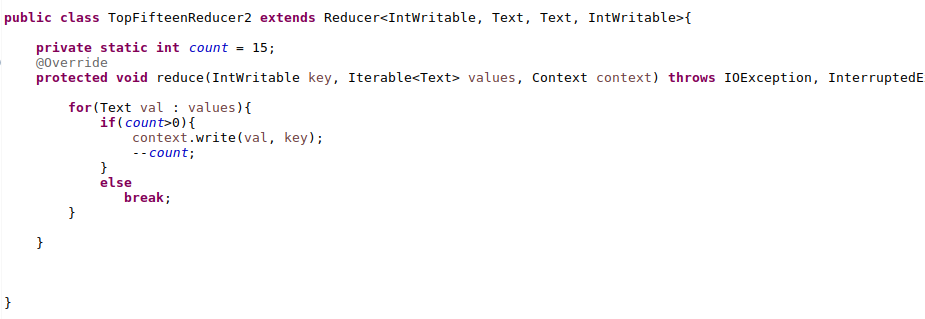


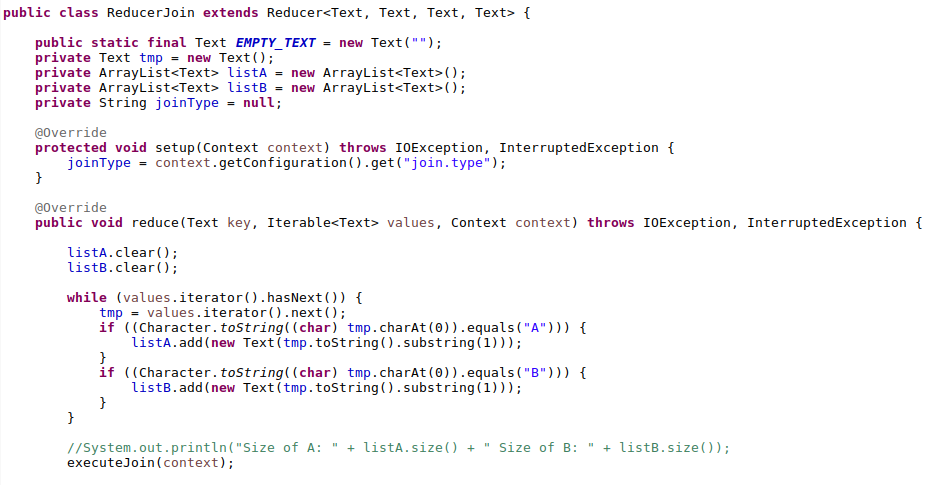


1. Top K busiest airports with respect to delay

Using Chaining of MR jobs, calculating the top K airports with the arrival delay of each airport and summarizing it.  
Using Join Pattern with the data dictionary of airports displaying the top airports with their name and address

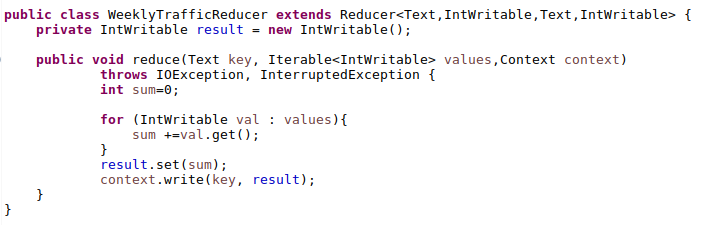






1. Weekly Analysis for the best time to fly

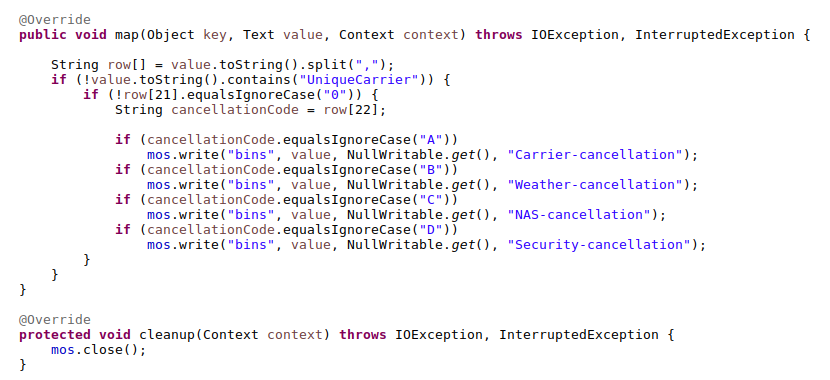
Using Partitioning Data Organization Pattern, partitioned the data based on the day of the week and calculating the weekly traffic for each day



1. Cancellation Reason

Using Binning Data Organization Pattern, binning the data for each of the four-cancellation reason and separating them out.

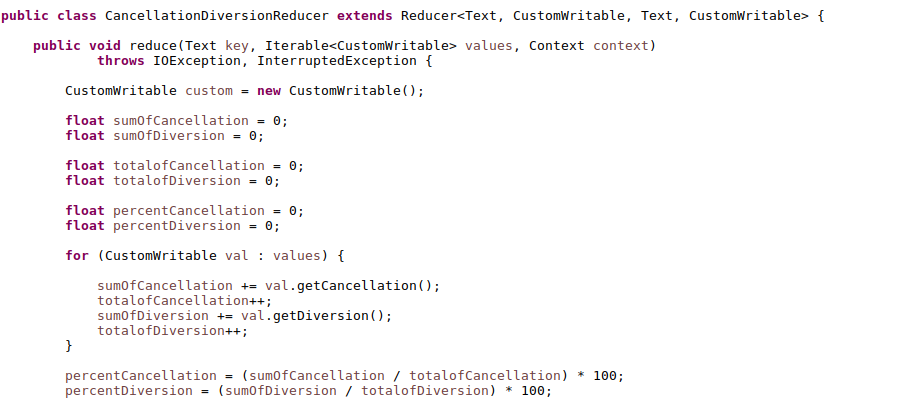
Reason for cancellation (A = carrier, B = weather, C = NAS, D = security)



1. Best Airlines to avoid cancellation and diversion

Using a custom writable class and summarization pattern, calculating the cancellation and diversion rate for each unique carrier, we found out the best airlines to avoid arrival/departure delays

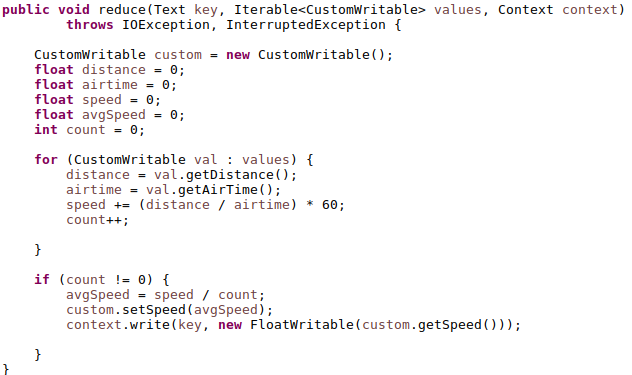




|  |  |
| --- | --- |
|  |  |

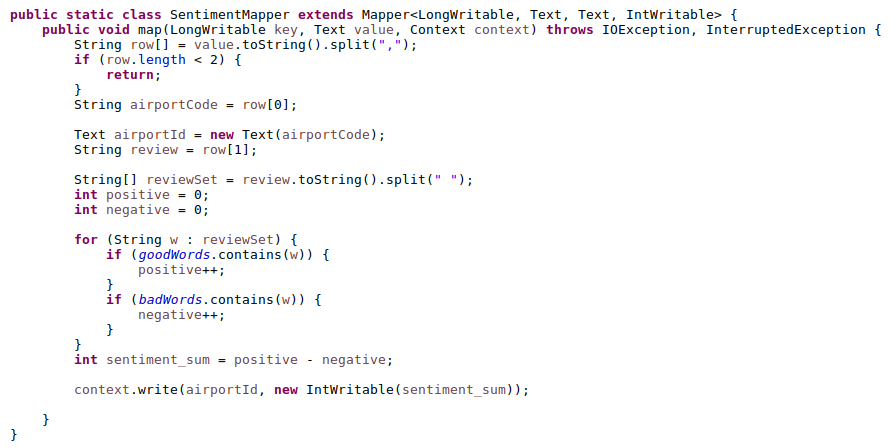
1. Average speed for the airlines

Using Summarization pattern along with custom writable class, calculating the mean speed of each airlines and comparing

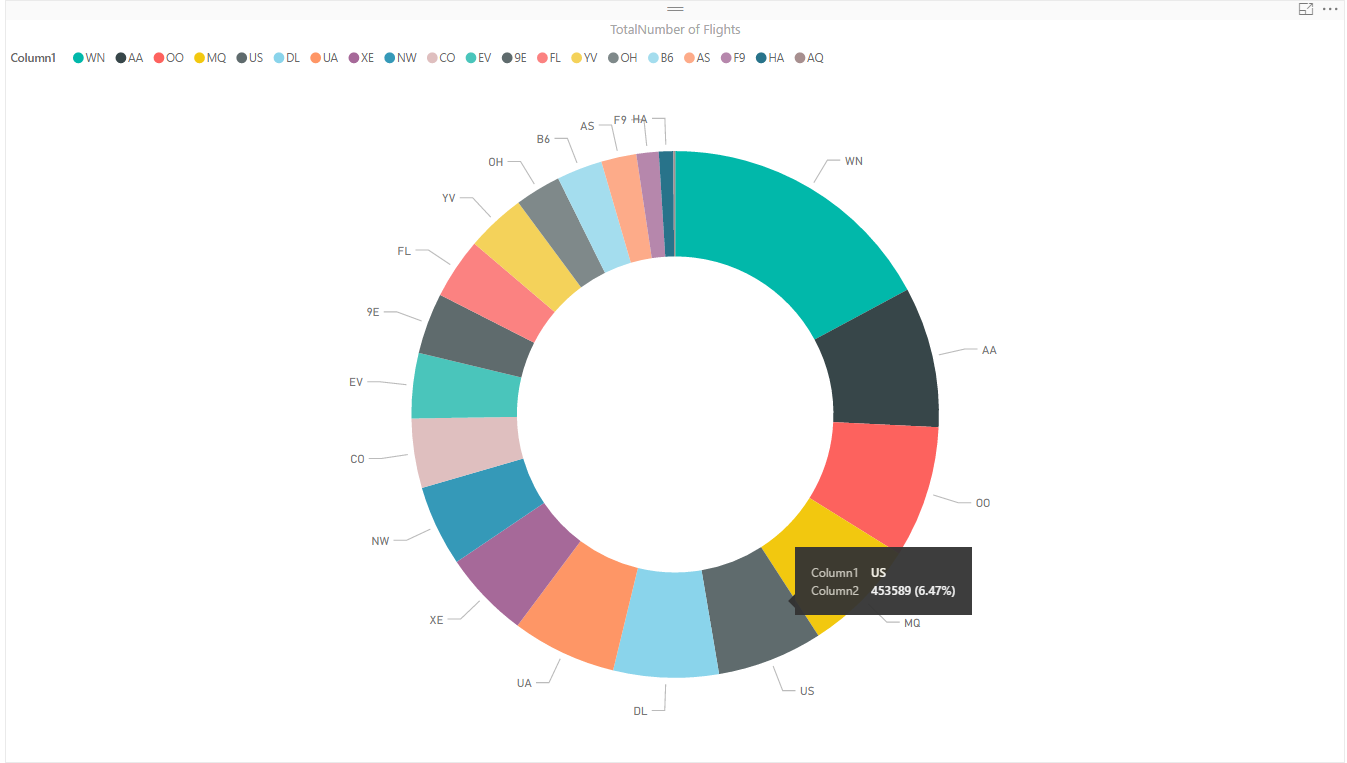


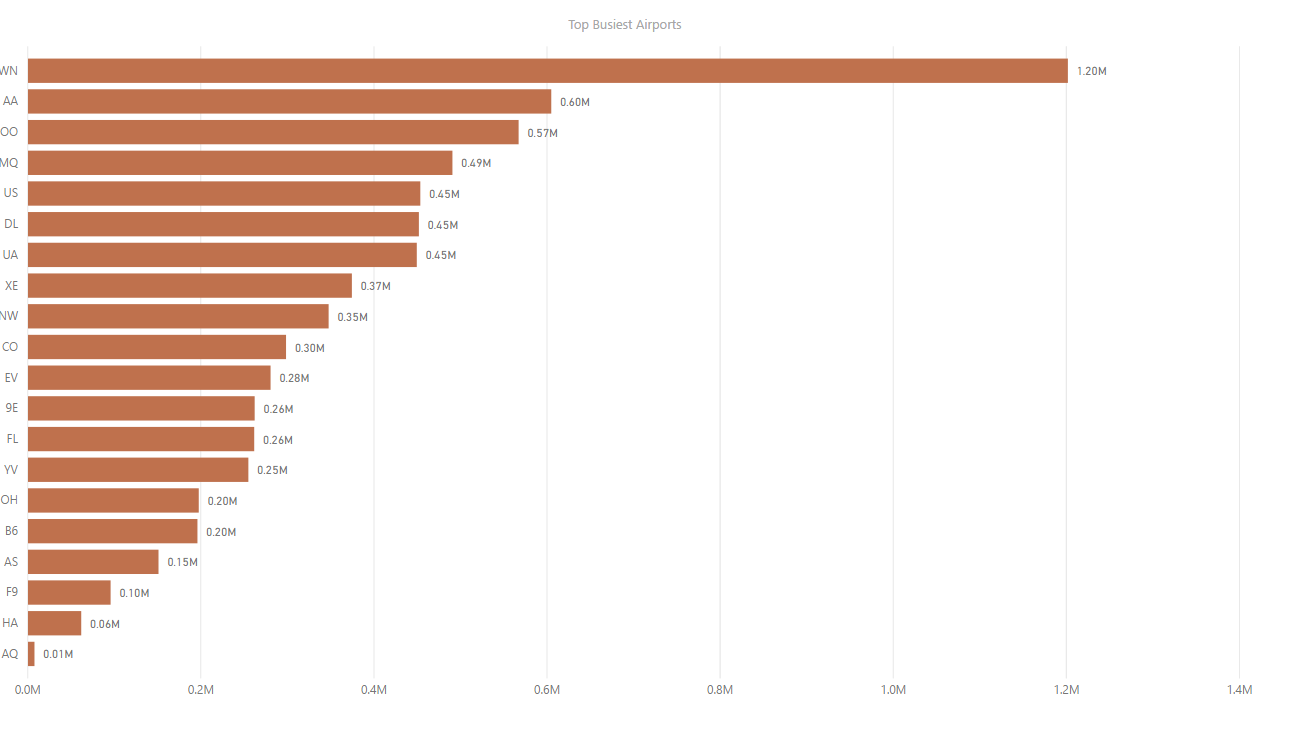
1. Airport Reviews

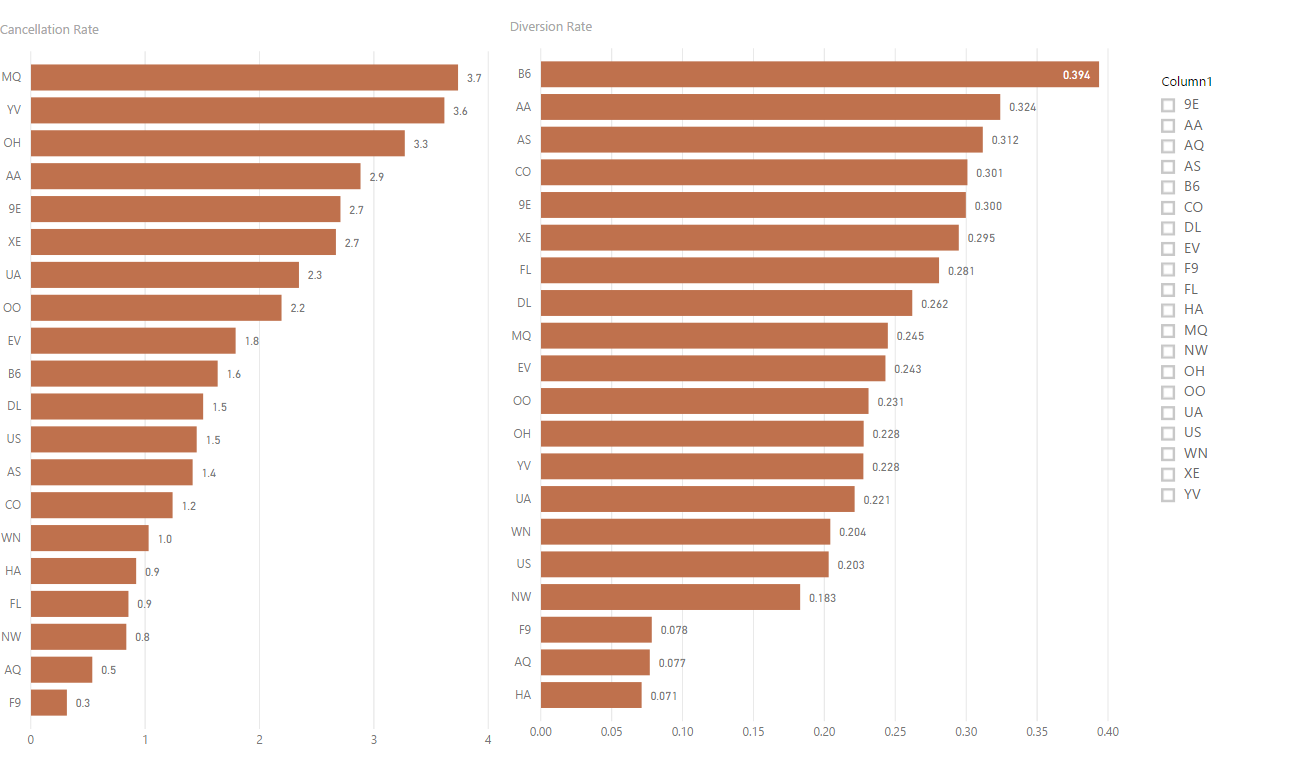
Using Sentiment Analysis, finding the type of review for each airport

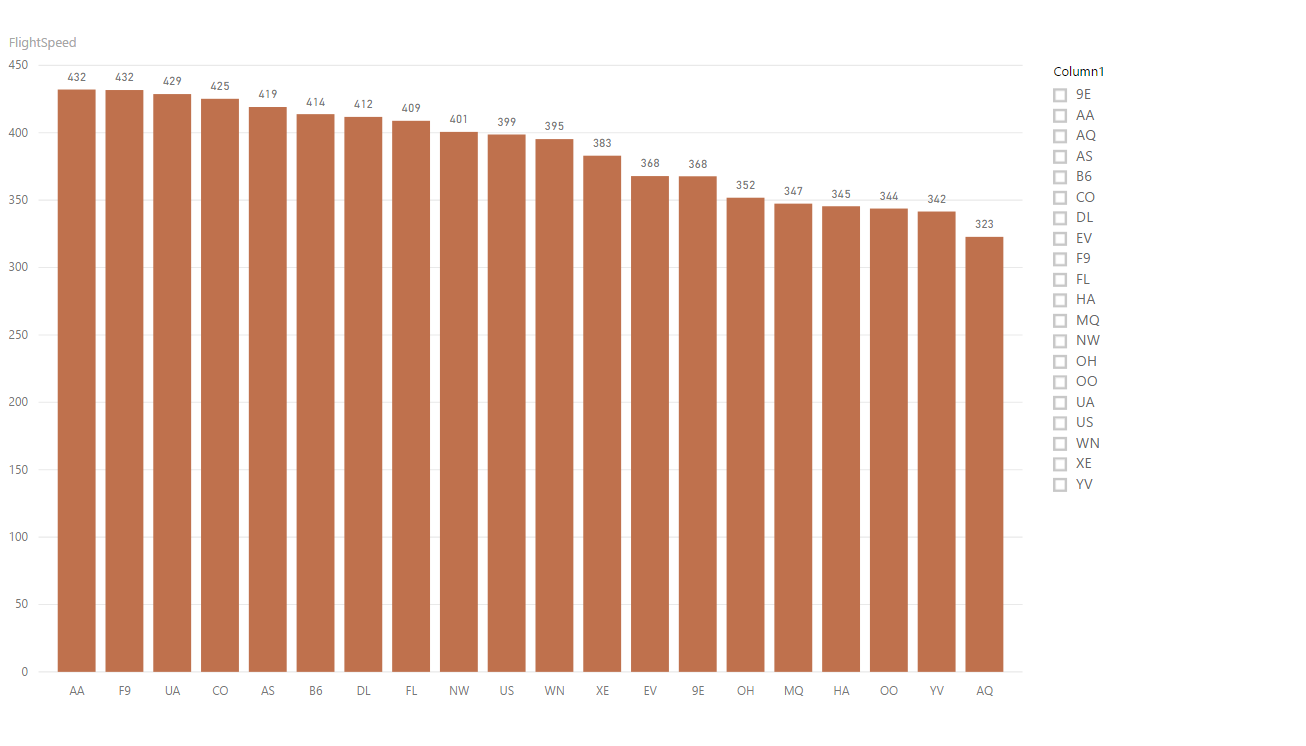


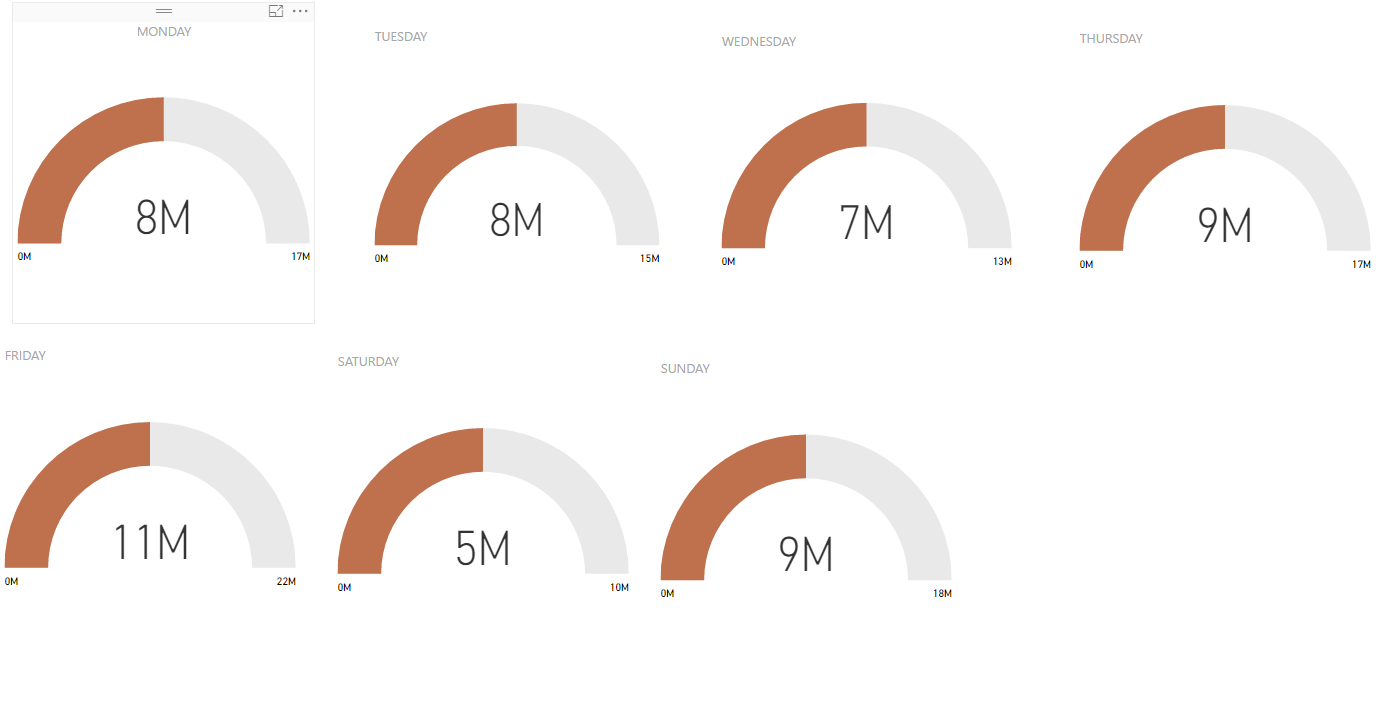
Dashboards for Visualization







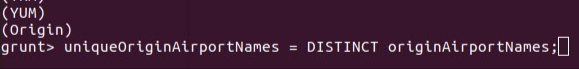




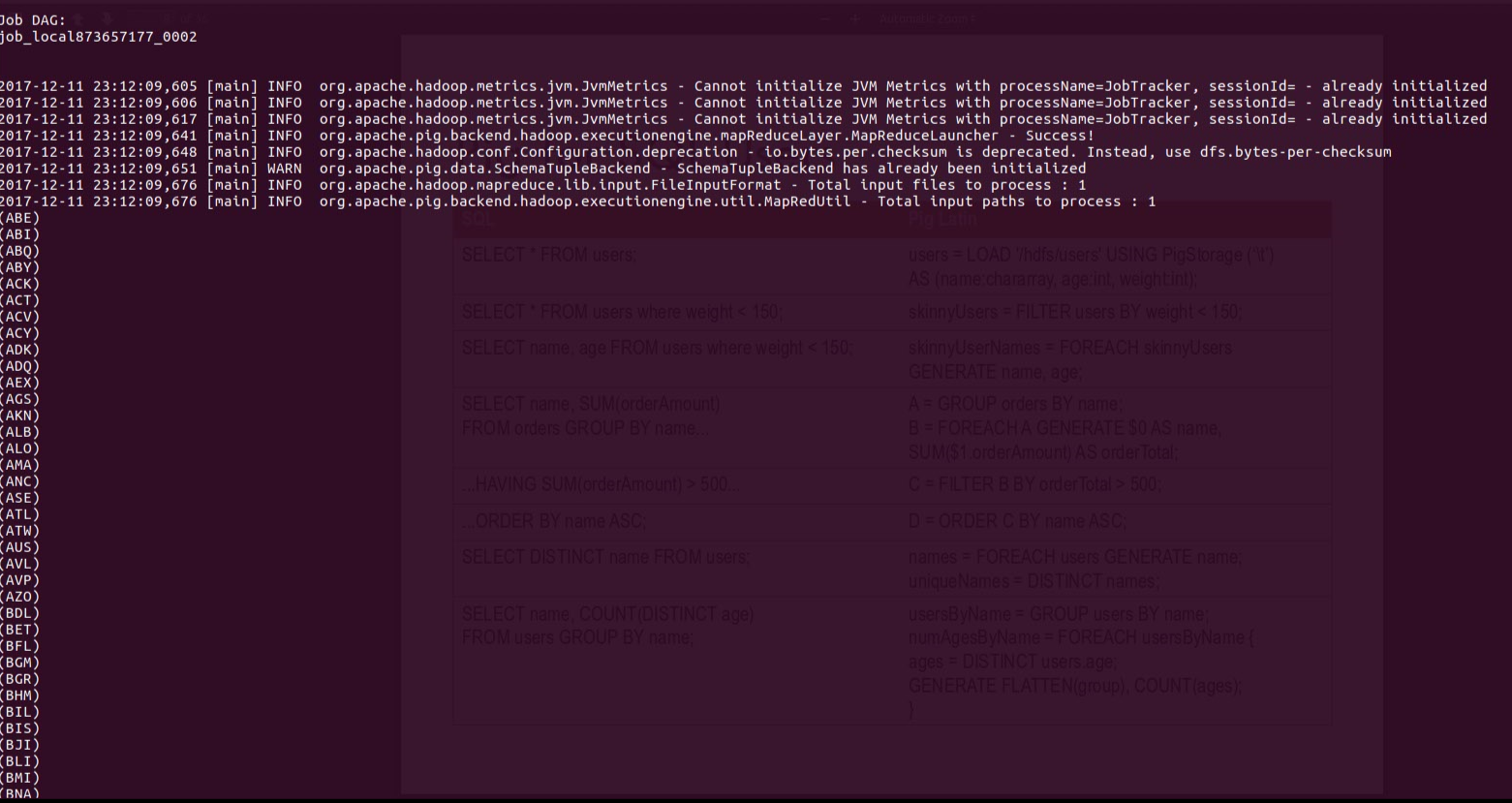
NoSQL Databases

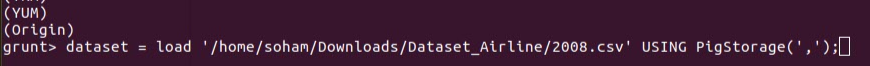
Pig

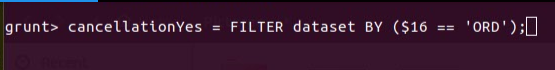


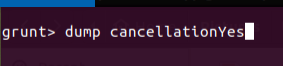


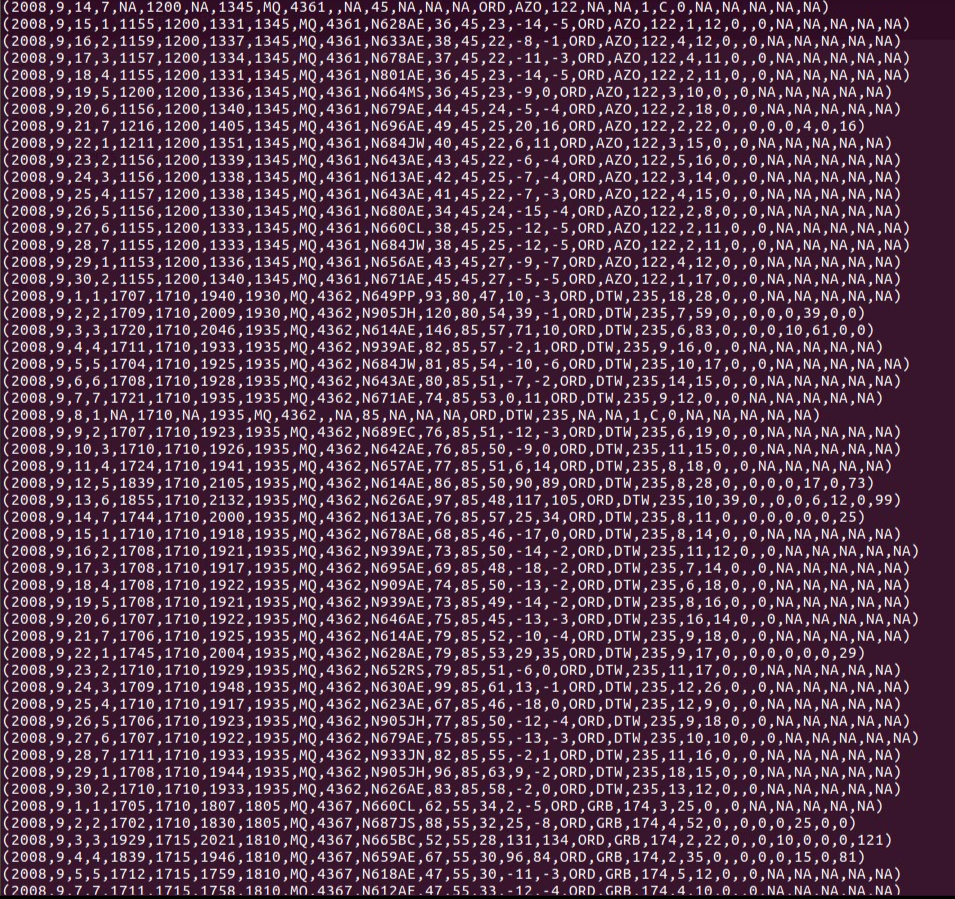












HIVE

