A8: Visualizing Flight Delays

Manthan Thakar

November 7, 2017

Objective

Visualize the mean delay of the five most active airlines and for the five most active airports in the country from given historical data.

Data Processing

In order to obtain delay data for airports and airline, we first gather the data by discarding invalid records. We run a **single** MapReduce job for that purpose.

Mapper: The map phase of the mapreduce job is responsible for - Validating each record by performing sanity checks - Emitting valid records with `airline`, `airport`, `year` and `month` as a key and `delay` as value

Reducer: In the reduce phase all the same flights are aggregated and we calculate the mean delay for each flight. To reduce the shuffling of data over the network, we use the same reducer as **combiner** as well.

Since, there could be many similar flights in the same month of the same year, using combiner in our design significantly reduces the amount of data that is shuffled for the reduce phase.

Performance

AWS Cluster Configurations

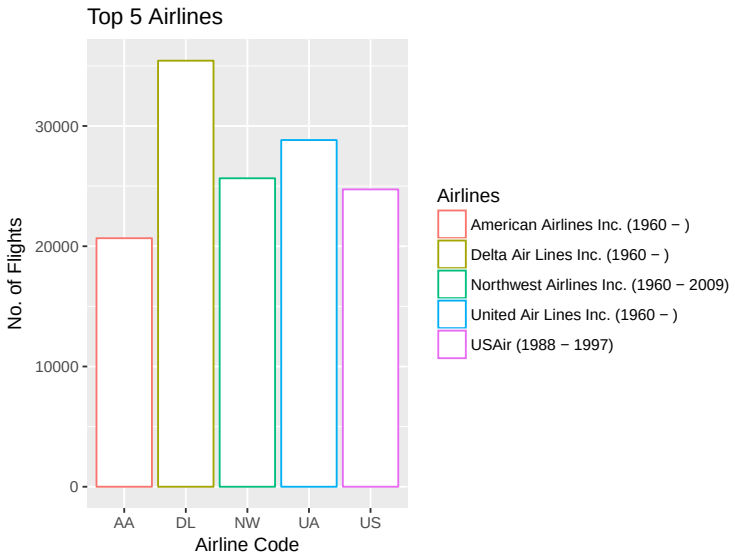
The Amazon Map Reduce cluster was setup using Amazon EMR.

config	value
Instance Type	m3.xlarge
Hadoop Distribution	Amazon 2.7.3 (EMR 5.8.0)
Memory	15GB
Storage	2 x 40GB SSD
vCPU	4
No. Nodes	4

On a 4-node m3.xlarge cluster, it takes about **13 minutes** to run our job. Note that this is a noticable improvement over previous submission where it took **19 minutes** to run jobs. This is because 3 mapreduce jobs were employed in that approach.

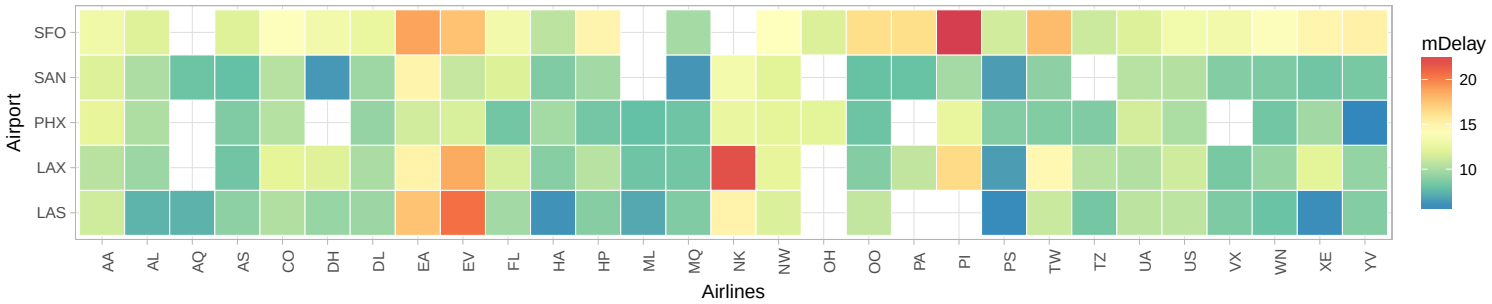
Most Active Airports and Airlines

The following graphs show the top 5 most active airlines and airports extracted from the data.



Delays

Fig 1. Mean delays from all airlines to top 5 airports across all years



As we can see in the figure above, airlineID 19790 Delta Airlines has more delays in flights when going to Chicago Airport (13930). Moreover, Delta Airlines is also one of the most active top 5 airlines.

Mean delays for top 5 destinations across months

Mean delays for top 5 destinations are displayed across months for all years in the graph below.

Here, we can see that airport ID `13930` (chicago airport), has more delays lying above 0.3 mean delay. This airport also happens to be the most active airport in the dataset.

Mean delays for destinations across years

The graph below shows mean delows for destinations across years. Interestingly, the delays for top 5 most active airports have increased after the year 2006. Moreover, in 2015 there they mean delay has increased significantly for `13930` which seems to have the highest mean delay in all other years.

Interestingly, the two datapoints for year 2013 are sort of the outliers. This could be because of sequestration-related furloughs being in effect during that year in Chicago.