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 historical airline on-time performance data.

Data Processing

In order to obtain delay data for airports and airline, we gather the delay data for each airline and airport per year and per month by discarding invalid records. A **single** Map Reduce job (`DelayJobDriver.java`) is run for that purpose.

The output types of the job are described below:

Output Key (Text): The output key of the `DelayJob` is a comma-separated `Text` object containing fields, `airline`, `airport`, `year` and `month`. This provides the flexibility to extract delay data per airline, per airport, per year and per month without running separate jobs.

Output Value (DelayWritable): DelayWritable is a custom writable that consists of two fields: **Delay** in minutes and **Count** which is the count of flights for the given key. It's important to emit count along with the delay so that it can be used to calculate correct mean delays as well as to find the most active airports and airlines.

The Map and Reduce phase of the job are described below:

Mapper: The map phase of the mapreduce job is responsible for - Validating each record by performing sanity checks - Emitting valid records as key-value pairs

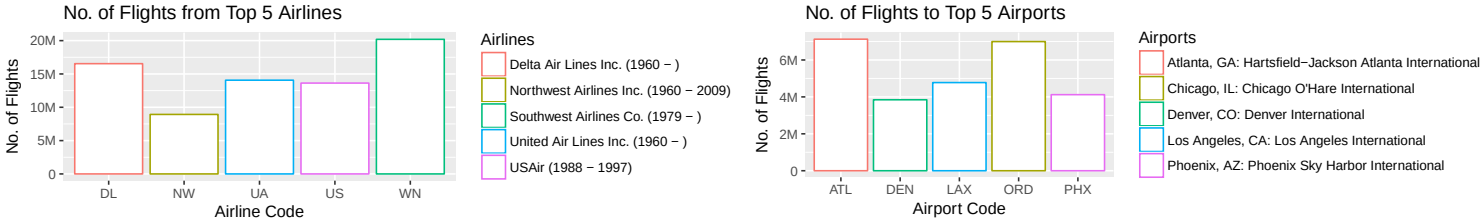
Reducer: In the reduce phase all the flights having the same output key (same airline, airport, year and month) are combined by aggregating the delay and flight count in `DelayWritable`. Since, there could be many flights with same key in one file, we apply the same reducer as **combiner** which significantly reduced the amount of data shuffled for the reduce phase.

Performance

AWS EMR Cluster Configuration	
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 noticeable improvement over previous submission where it took **19 minutes** to run jobs. This is because 3 mapreduce jobs were employed in that approach.

Figure 1: Top 5 Most active airports & airlines



Most Active Airports and Airlines

Figure 1 shows top 5 most active airports and airlines along with the number of flights as the measure of activity. *South West Airlines* is the most active airline and *Atlanta's Hartsfield-Jackson* is the most active airport, while *Chicago's O'Hare* closely comes at second place.

Mean delay per year

In this section, mean delays per year for top 5 airlines and airports are analyzed. Figure 2 shows mean delay per year for top airlines and airports and apply loess smoothing to obtain a trend line with confidence intervals. Note that, outliers are clipped in the graphs shown below in favor of readable plots.

Among the top 5 airlines in fig 2.1, North West airlines (NW) seems to be having very high variations in mean delays across years. Mean delays for NW seem to be decreasing starting from 1988 to 1993. They increase again starting from 1994.

It can also be observed that United Airlines (UA) has wider error bounds, indicating highly varying mean delays across years, making the pattern less predictable. Moreover, South West airlines (WN) has been seeing steady increase in mean delays.

Fig 2.1. Mean delays for top 5 airlines per year

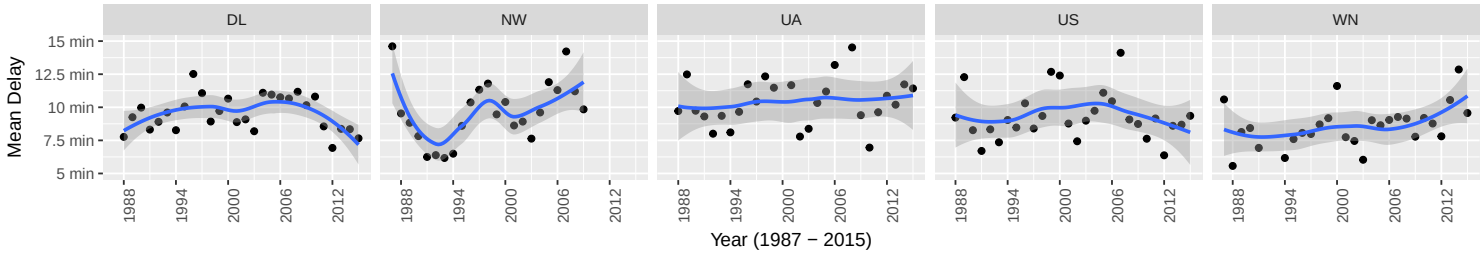


Fig 2.2. Mean delays for top 5 airports per year

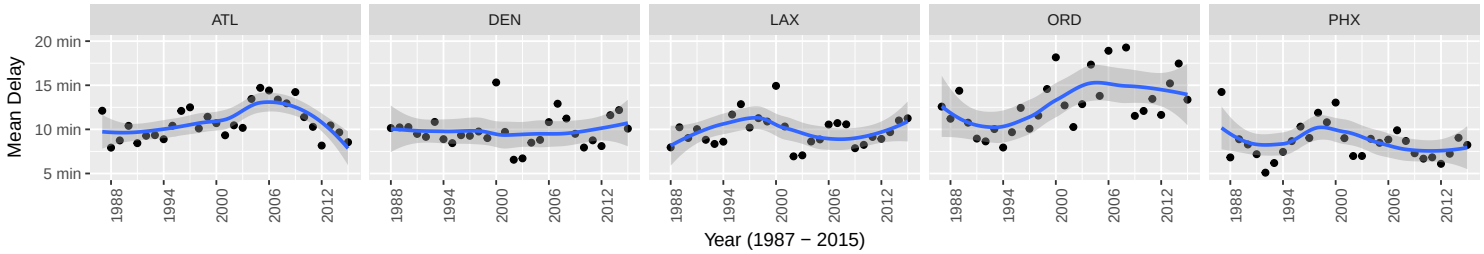


Figure 2: Mean delay per year for top 5 airlines & airports

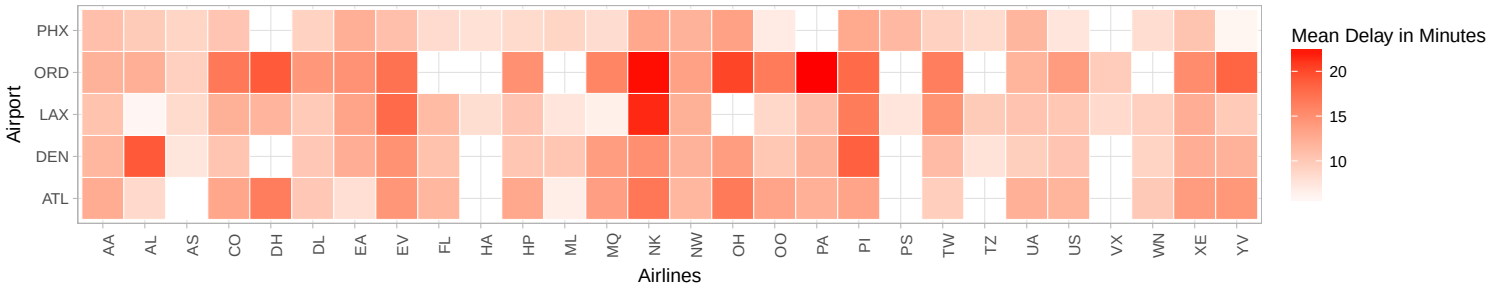


Figure 3: Mean delays from all airlines to top 5 airports across all years

Similarly in fig 2.2, we can observe that the mean delays for ATL airport increases up to 2006 and then steeply decreases. Interestingly, LAX and ORD seem to be having exactly the opposite trends for mean delays. While the delay trend in ORD across years roughly resembles letter "S", the trend line in LAX looks like an upside-down version of it. The Denver airport (DEN) seems to be having the least variation in delays with trend line staying around 10 minutes mark.

Mean delay across all years

Figure 3 shows a heat map of mean delays across all years from all airlines to top 5 airports. Each square block in the heat map shows the mean delay from an airline to an airport and the delay is proportional to the intensity of the color in that block. It's important to note that the white blocks with lines inside show NA values, meaning there's no flight data available from an airline to a particular destination.

Among top 5 airports, *ORD* contains many dark blocks across all airlines, suggesting mean delays above 15 minutes while *PHX* has most of the mean delays close to 10 minutes. As a new airline, you wouldn't want your first flight to go to ORD.

Out of all airlines, NK - Spirit Airlines, seems to be having highest mean delays while traveling to top 5 airports. The delays increase especially for LAX and ORD. *AS* seems to be having least delays going to top 5 airports. As a traveler going to Los Angeles, from the historical data, you'd be better off traveling in *MQ* than *NK*.

All of the top 5 airlines relatively have lesser delays while going to top 5 airports, despite of having higher volume of flights.

Mean delays per month

Fig 4.1. Mean delays for top 5 airlines per month

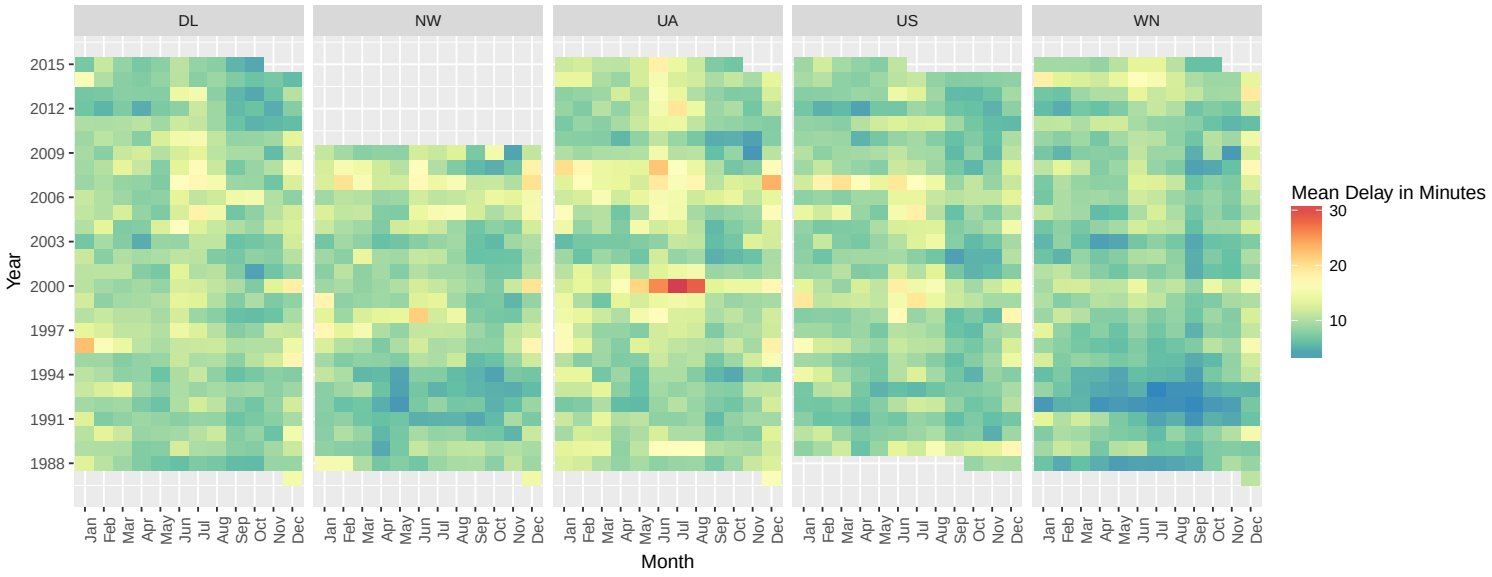


Fig 4.2. Mean delays for top 5 airports per month

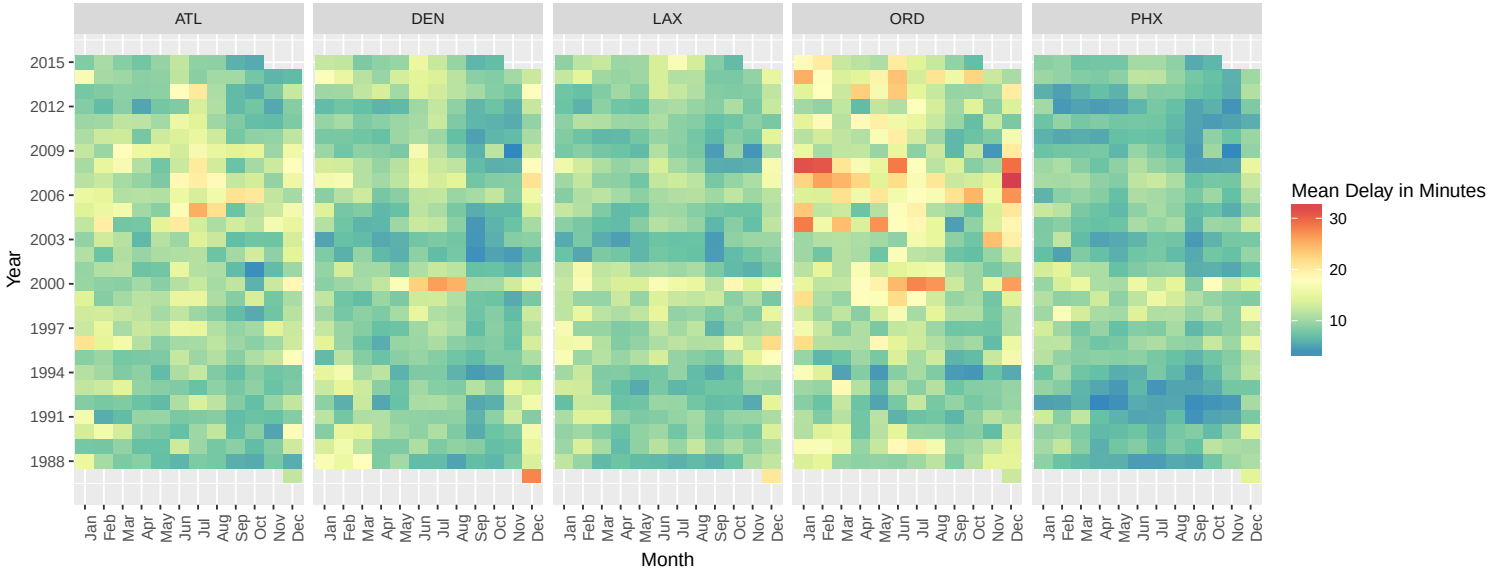


Figure 4: Mean delays per month for top 5 airlines & airports

Figure 4 shows mean delays for top 5 airlines and airports across all months of all years, as a heat map. We decided to explore the idea of visualizing mean delays per month per year, because it wouldn't be useful to aggregate delays for a month across all years. As Figure 3, the darker the block higher the delay. But here, different delay categories have different colors. In general, blue blocks indicate smaller delays and red blocks indicate higher delays. Blocks with shades of green, yellow and orange generally fall into medium range.

Figure 4.1 shows mean delays per month for top 5 airlines for all years. For WN, which is also the most active airline, September seems to be the month having least delays across all years. In fact, for all the top 5 airlines, in terms of delays September seems to be the best month across all years. Conversely, December is the worst month for delays for all of the top 5 airlines.

Figure 4.2 similarly shows mean delays per month for top 5 airports for all years. This graph seems to be agreeing with the previous observation that the delays are smaller in September across years. One interesting observation in figure 4.2 is that all of the top 5 airports seems to be having higher delays in the year of 2000. Moreover, for year 2000 airports DEN and ORD seems to be having similar delay patterns for months Jun, July and August and December to some extent.

This visualization makes it easy to point out particularly bad months in a year. Moreover, we can identify airports that have higher seasonal delays or had higher seasonal delay in any particular year. For example, ORD in general has a lot of orange/red blocks. But one pattern that stands out starts around December 2006 and continues till December 2008. This pattern shows that there very high delays starting from December continuing till February of next year. This could be attributed to stormy winters which would generally occur around those months.

Conclusion

With the new approach to Map Reduce job described in the first section, it was easier to get all of the data in one job which is a better approach than previously employed approach. For the visualization, instead of using network graphs (as done in `old.pdf`), we explore heat maps. We show how it was better to recognize certain patterns with those visualizations. The idea of visualizing mean delays per year as well as per month is also explored which was missing previously. From both of the visualizations, some interesting patterns emerge, which in spite of having data hand, were not presented previously.