# A4 - On-Time Performance Data

We will explore the world of airline on-time performance data analysis. Your task will be to implement a pipeline that runs a set of pseudo distributed map reduce tasks to plot the mean delay of the five most active airlines and for the five most active airports in the country. Specifically, your code should take a set of data files as input and produce visualizations of delays. You are free to choose how to visualize delays, how many graphs you want to produce, etc. You will be asked to defend your choice.

## Expected elements of a solution:

- A Makefile that builds and runs the pipeline (including graph generation);
- One or more map reduce tasks that read and clean the data, any corrupt row can be deleted, and output data that can be used to generate a graph
- A R Markdown file that includes code to generate graphs from the output of the MR task and explanation of the results.
- Make sure to include all JAR files needed to build your code and to only use local paths.
- It should be possible to run the pseudo distributed version of your code with a single local data file easily

## Fine print

- Data for one month is here (https://drive.google.com/open?id=0B2Yl023GVEqHZnVBbWV3TGNibmM)
- The full data set will be made available on AWS
- Some helpful code will be shared via piazza
- A description of the fields in the data is here (https://www.bts.gov) (Look for online time performance of airlines)
- The sanity test is for corrupt data (field names may vary slightly)

```
All rows should be of the same length
All columns should hold one type of values
CRSArrTime and CRSDepTime should not be zero
timeZone = CRSArrTime - CRSDepTime - CRSElapsedTime;
timeZone % 60 should be 0
AirportID,  AirportSeqID, CityMarketID, StateFips, Wac should be larger than 0
Origin, Destination,  CityName, State, StateName should not be empty
For flights that are not Cancelled:
ArrTime -  DepTime - ActualElapsedTime - timeZone should be zero
if ArrDelay > 0 then ArrDelay should equal to ArrDelayMinutes
if ArrDelay < 0 then ArrDelayMinutes should be zero
if ArrDelayMinutes >= 15 then ArrDel15 should be true
```

- Aggregate flights per month (i.e. add up all flight delays in a month for each airline and destination)
- Use the destination airport for each flight
- If there are several airlines or airports with the same number of flights, use lexicographical sort
- A cancelled flight counts as being delay 4 times the CSRElapsedTime (i.e. its scheduled length)
- Rather than using delay in minutes, normalize the delay with respect to the scheduled length of each flight
- In your visualization show the mean delay per flight, month, airline and destination