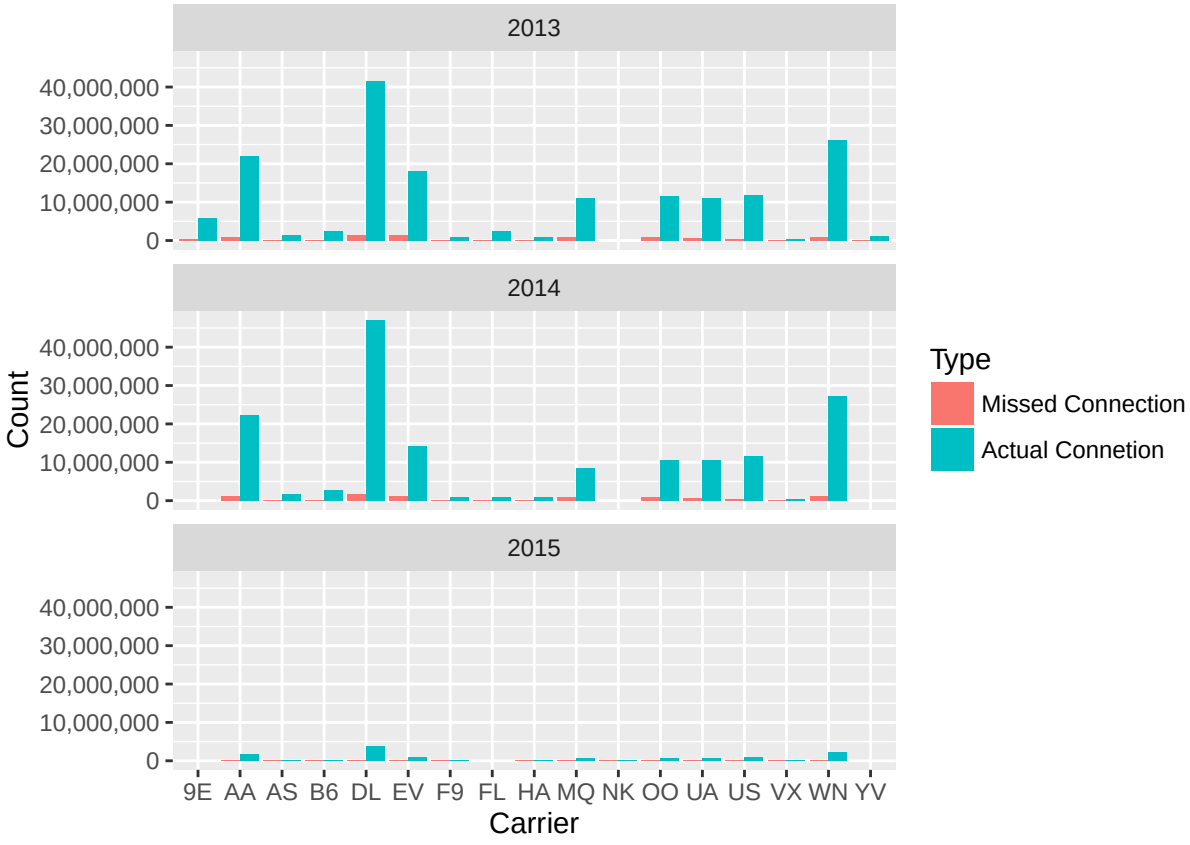
Assignment5_report

Sharma, Abhijeet and Khan, Afan Ahmad

February 20, 2016

Results

Following graph shows the plot displaying the count of Missed and Actual connections per Carrier per Year.



Timings

```
tabl <- "# Below are the listed times on various datasets
| Data        |    Pseudo(in mins)  |   Cloud *(in mins) |
| ----------- |:-------------------:|-------------------:|
| 2-Documents |      .16            |    5               |
| 25-Documents|      196            |    15              |
"
```

*(10 c.medium machines)

Implementing Map Reduce Job

This program consists of one map-reduce job and one RMarkdown script. ## FlightCount – Main class `FlightCount` configures the job 'flight count' , sets the Mapper class to 'FlightMapper', sets the Reducer class to 'FlightReducer', and sets the input and output paths from the command line arguments.

FlightMapper – Mapper Class

This class maps the carriercode , year, origin/destination and month as keys and Scheduled Arrival/Departure Time, Actual Arrival/Departure Time and Cancelled Status as values. This class will output the intermediate set of key-value pairs. Key -> custom combination of flight code and year separated by tab separation. Example, `AA\t2014\tALB\t1` where "AA" is the carrier code, "2014" is the flight year, "ALB" is the origin/destination, "1" is the Month. Value -> " 'arriving/departing'$\widehat{1}$2000011$\widehat{2}$000400$\widehat{0}$" where "arriving/departing" states the type of flight, "1200001" is the "CRS_ArrivalTime/CRS_DepartureTime"(depending on Type) in milliseconds, "ArrivalTime/DepartureTime" in milliseconds and Cancelled Status.

FlightReducer – Reducer Class

This class iterates though the iterable array "Iterable" for all keys. The reduce method then computes the missed and actual connections between flights and outputs to file.

SYSTEM SPECIFICATION:

1. Java 1.7.0_79
2. Ubuntu 14.04 64-bit/ MAC 10.11 64-bit
3. 8GB RAM
4. Pandoc (https://github.com/jgm/pandoc/releases/tag/1.16.0.2)
5. pdflatex (sudo apt-get install texlive-latex-base and sudo apt-get install texlive-fonts-recommended) (Only if running from KnitR)
6. R Packages: 6.1 Rcpp 6.2 R.utils 6.3 ggplot2 6.4 rmarkdown 6.5 plyr 6.6 reshape2

Implementation

1. We have implemented a Map Reduce job where the Mapper calls the map method. Each record is checked for sanity, their timestamps calculated and, depending on their cancelled status, are sent to Reducer(Once as origin location as part of key and again as destination as part of key).This is similar to an Equi-Join such that arrival.DEST == departing.ORIGIN. The Intermediate Key, Value Pairs are sent to the Reducer.
2. The Reducer calls the reduce method passing an iterable of values. The values for departing and arriving flights are identified and stored in separate Array Lists. The flights are then compared for connections and the variables "missed" and "connections" are updated accordingly.
3. The R Markdown document is then run which shows the plot between Price and Year for N=1 and N=200 and generates the report.
4. Plot is drawn with ggplot2 package.
5. Make file is created to automate all the above steps.

Assumptions

1. We have taken Carrier code, Year, Location and Month as Keys. So, we are only calculating for flights within a Year and Month. We are calculating rolling-overs in days in a month but we are not considering the edge case between Months(eg: connections between 31st of Month and 1st of next Month).
2. Since we have also taken Year as keys, we are also not assuming rolling overs between years, eg: connections between 31st of December in a year and 1st of January of next year.
3. Finally, we are not considering flights whose difference between **scheduled timings** is more than 6 hrs or less than 30 mins as connections.