

# A5 Paths Report

*Pankaj Tripathi, Kartik Mahaley*

*March 09, 2016*

This is a report generated from R and it has the required results of our assignment A5.

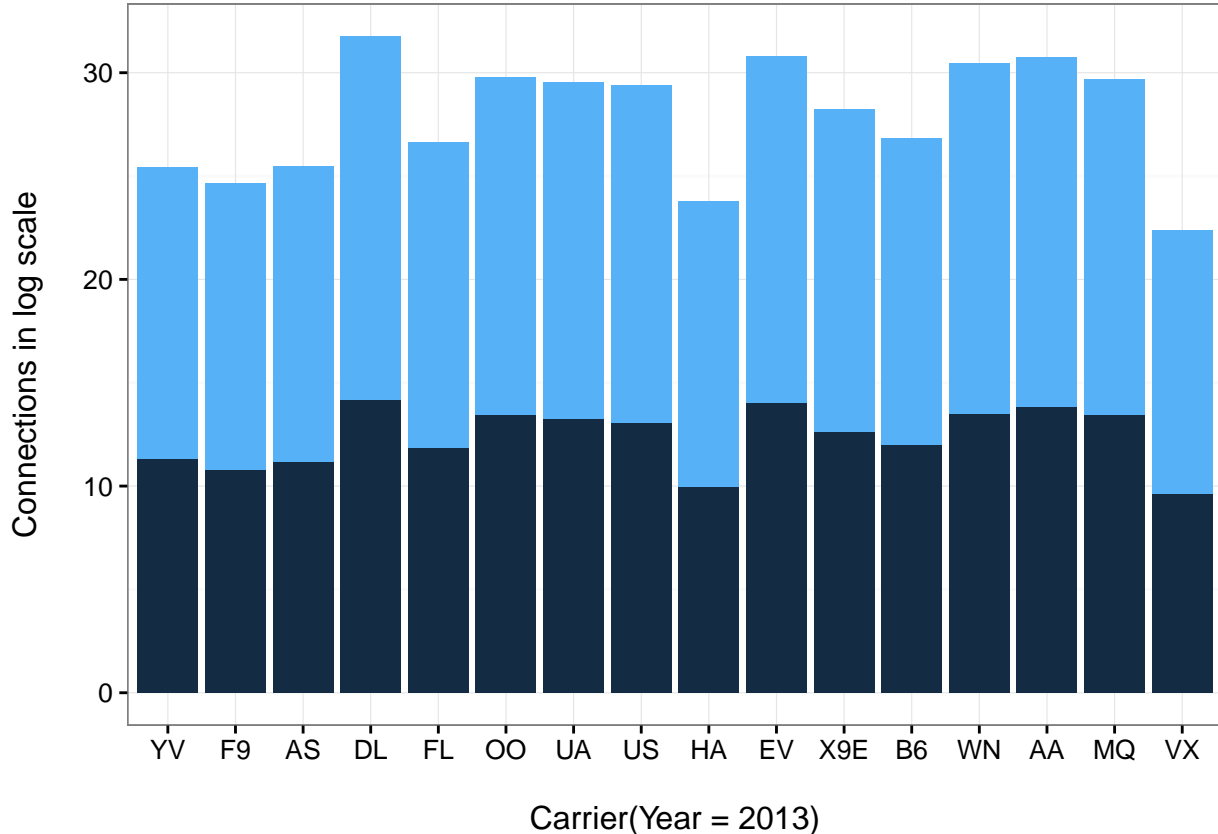
**INTRODUCTION:** This assignment is to display number of connections and missed connection between flights of same airline for same year. We have written mapreduce chain job to accomplish the same.

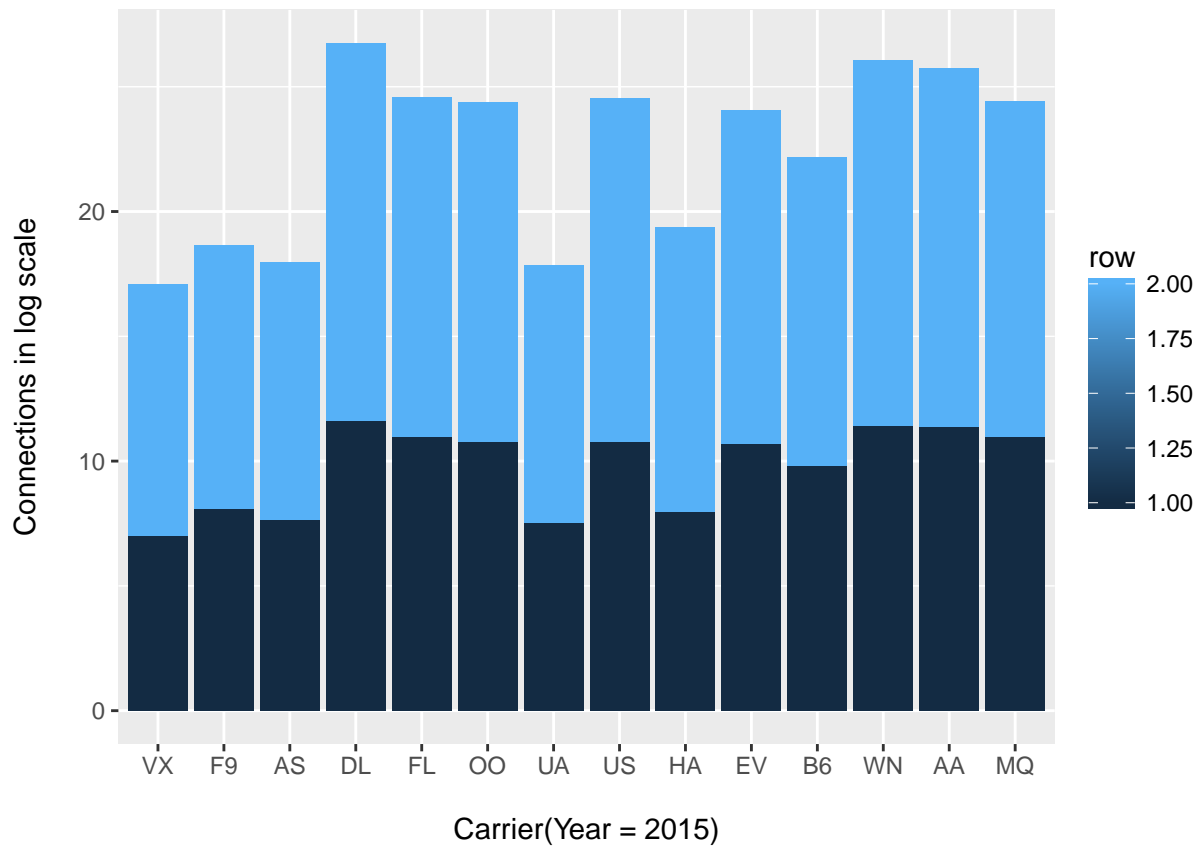
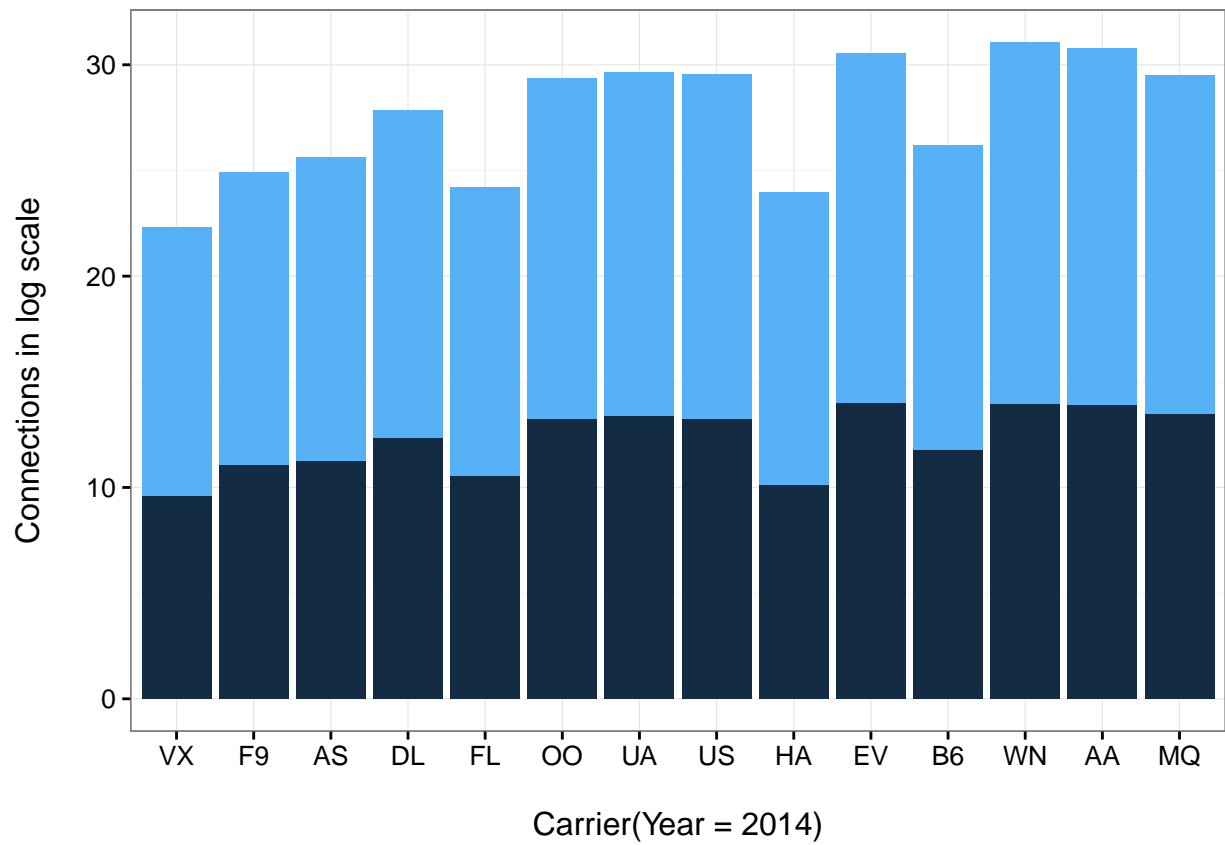
**DESIGN:** We have taken each row and divided it into two records for departing and arrival flights. Our Mapper takes key as carrier year and destination/origin airport. If key contains origin airport then corresponding value is false with departure details. If key contains destination airport then corresponding value is true and arrival details.

**IMPLEMENTATION:** Our code includes chain mapreduce with map1-reduce1-map2-reduce2. 1st mapreduce job creates intermediate files with key as carrier, year, airport(destination/origin). This intermediate files are then taken by job2 to combine all the files generated by job1 and produces final output for airline, year and connections/missedconnections. The conditions checked for connection flight and missed flight is given below. 1. Any pair of flight F and G of the same carrier such as F.Destination = G.Origin. 2. The scheduled departure of G is  $\leq 6$  hours and  $\geq 30$  minutes after the scheduled arrival of F. 3. A connection is missed when the actual arrival of F  $< 30$  minutes before the actual departure of G.

**PERFORMANCE:** As per our analysis for 653 MB gzipped data(25 files) on i5-8gb Mac machine and 653MB on i5-8gb Linux machine pseudo mode takes 150 mins and on cluster it takes 30 mins with 1:10 m3 xlarge machine.

**GRPAHS:**





## NULL

