Problem Statement: The price of a ticket depends, in part, on the amount of fuel consumed in the particular itinerary (a factor of distance and winds). Furthermore, prices have a tendency to increase over time. To understand airline pricing, computer a simple linear regression that models the cost of tickets for different airlines. Give a new ranking of airlines with respect to price.

Group assignment, two students Create a pipeline of jobs to run on EMR (map-reduce) and locally (R, maybe Bash scripts). Determine the airlines active in 2015. Work with the flights by those airlines in 2010-2014. For each of those airlines compute linear regressions for {distance traveled, flight time} to price from 2010-2014. Automatically generate graphs (with R) showing a linear fit of the variable to price for each airline. Include your graphs in your report, and conclude which airline is cheapest. Is distance traveled or flight time a better variable? Why?

Airlines active in 2015

HA, EV, MQ, OO, US, B6, WN, UA, DL, NK, VX, AS, F9, AA (Of these, NK does not have any data in 2010-2014)

Graphs showing a linear fit of the variable to price for each airline

Cheapest Airline

Is distance traveled or flight time a better variable?

Why?

ANALYSIS

Conclusion: