Changes to original Assignment A4

Shreysa Sharma

11/9/2017

Assignment A4 was done as part of assignment A8. Below are the changes I made to the assignment:

- A4 was using 3 Map and 2 reduce jobs for calculating the mean delay for airlines and airports where the first mapper and reducer were just handling cleaning, then second mapper would map keys and value for airline then 3rd mapper will do the same for airports and finally the 2nd reducer will calculate the mean delay for the outputs from 2nd and 3rd mapper and output it in 2 seperate files. Although the report said that the 2nd and 3rd map reduce job ran in parallel but while working again on the implementation I noticed that they were chained and not parallel while A8 uses just 1 map reduce job that cleans, maps key value for both airline and airport and the reducer calculates the mean delay for airlines and airports and outputs a single output file.
- A4 implementation did not account for converting the arrTime, depTime, crsArrivalTime and crsDepTime into minutes due to which it was generating wrong set of airlines and airports while A8 converts the HHMM to total minutes and then does the checks and mean delay calculation.
- A4 report was only comparing results of 6 recent years while A8 does for all the 30 years.
- A4 took 33 minutes to run on EMR cluster while the current implementation takes 14.3 mins on AWS.