diff: Changes made to A4

Manthan Thakar

11/8/2017

Number of Jobs

- The previous submission used 3 mapreduce jobs - two separate jobs for getting airport and airline delays and one job for getting activity across airlines and airports.
- In A8, we only use one Mapreduce job which gets mean delays for each flight. This significantly reduces the amount of data that we need to process in R. So we calculate top active airlines and airports in R. This helps us discard the activity job from the previous approach.

Sanity Checks

- The previous submission had an erroneous condition for validating `timezone` values, which resulted in dropping most of the records. In this submission, that is fixed by using `time` subtractions instead of `float` subtractions for Arrival times and departure times.
- With the previous erroneous approach for file `323.csv` we observed 65K valid records. This number increases to 240K+ with the new (correct) approach.

The airlineID and airportID hell

- Previously, the numeric IDs for both airlines and airports were used. This resulted in using an extra lookup while plotting graphs.
- In this approach we use `CARRIER` and `DEST` fields in the mapreduce jobs.