Title: "Routing - Assignment 7 Report"

author: "Neelesh and Habiba"

output: pdf document

Work allocation

- Brain Storming: Habiba Shaik, Neelesh
- Program 1:
  - Mapper: Habiba Shaik
  - Reducer: Neelesh
- Program 2:
  - Mapper: Neelesh
  - Reducer: Habiba Shaik
- RScript: Habiba Shaik
- Bash Script:
  - Local: Habiba Shaik
  - EMR: Habiba Shaik
- Report : Habiba Shaik, Neelesh
- Performance calculation: Neelesh

Implementation details for the Map Reduce programs:

Program 1

This program calculates the average duration of each one-hop route using the 36 history files.

Mapper:

- **Key-** Airline-Month-Destination-Origin
- **Value-** Actual-elapsed-time
- **Notes-** The Mapper maps all the rows based on the above mentioned key combination.

Reducer:

- **Notes:** The reducer calculates the averge of durations of all one-hop routes.

How to invoke:

We created a shell script which invokes calls to 1. Hadoop program for MapReduce in EMR mode 2. EMR commandline call to create a cluster and run the step.

Program 2:

This program builds all possible connections from the iternaries in test file that qualify the layover conditions.

- **Key1:** Airline-Month-Day-City
- **Value1:** "A/D"-Destination/Origin-DepartureTime/Arrival-FlightDate-FlightNum

Mapper:

Each iternary in test file is sent as two keys one with origin and the other with departure. This facilitates the reducer to get all the flights arriving and departing from a connecting city.

Reducer:

A reducer get all the flights arriving and departing from the city given in its key. The A/D string at the beginning of the key helps in differentiating between arriving and departuring flights.

All possible connections with qualifying layover conditions are calculated.

Output from R script

Output files from the programs will be written as below to the same folder as Rscript and makefile/

```
"airline","day","month","origin","connection","destination","origin_flight","destin
ation_flight","Flight_Duration"
    "CO",1,1,"BOS","CLE","LAX",1431,1035,491
    "DL",1,1,"JFK","CVG","LAX",441,679,497
```

Implementation Details of R script

This Script reads from 4 files and creates 4 data frames. For each request files, it finds the possible connections and calculates the total duration for each route. The route with minimum duration is proposed as the best route for the given request route.

Conclusion

Accuracy:

There is no prediction in this program and hence there is no need to calculate accuracy.

Performance

- Pseudo Mode: ~19 mins
- EMR Mode: ~14 mins