Big Data Final Project

Team Members:

Praveen Mareedu

Ronish Medjo

Archna P

This Final Project contains multiple Analytic components. The analytics are as described below.

**Analytic 1**: Counted the number of Trip from source to destination and then used PIG Query to find most frequent trips from one source to destination after MAP REDUCE.

Things we have considered when counting number of trips. We have to know near by locations to group them as same source, we have achieved it by rounding off the latitude and longitude co-ordinates. Locations that are close are grouped together because most of the digits match in their coordinates.

ex : -73.776243,40.651281 and -73.776240,40.651280 are near by locations . They only differ in the last digit

**Analytic 2**: Suggesting restaurants based on the taxi trips based on the time intervals. Idea is to get a sense of places visited in these time ranges. We have used Pig to segregate into lunch, breakfast and dinner timings.

This Analytic uses Map Reduce to get results, Mapper takes key as the time rage and drop location as key . Most visited places from this analytic could be a good place to open a new restaurant.

**Analytic 3**: Ride Sharing from Airport. As a pre analytic we are using HIVE to get all the trips that are starting and ending at Airport location (JFK) in a buffer of 1 hour and 30 min.

Hive is used to get the trips that involves airport as destination or source. Then we ran Map Reduce to get trips that happened for each 30 min buffer on the output of above Hive query. Which gives the idea of how people travelling how many journeys are happening to near by locations to determine whether ride share is possible for these trips or not. These points could be a good place to advertise Ride Share apps as well.

**Analytic 4**: Run MapReduce on both the taxi and the flight data. In the taxi data we are filtering all the data based on the drop-off or pickup locations as the JFK airport for each cab. In the flight data we are grouping all the data in an hourly basis i.e. counting the number of flights for each hour, for every day.

Create two external tables using the above outputs and each table has the columns date and time. Using these common columns we can join the two tables and get the count of flights for the respective cab for the same time. This count gives the respective cab the number of flights that are going to land, thus recommending him to stay to pickup a passenger.