Project: Analyzing Big Data with SQL
Name: José Medardo Tapia Téllez
Date: 27/07/2021

Project

Recommend which pair of United States airports should be connected with a high-speed passenger rail tunnel. To do this, write and run a SELECT statement to return pairs of airports that are between **300** and **400** miles apart and that had at least **5,000** (five thousand) flights per year on average *in each direction* between them. Arrange the rows to identify which one of these pairs of airports has the largest total number of seats on the planes that flew between them. The SELECT statement must return all the information required to fill in the table below.

Recommendation

I recommend the following tunnel route:

	First Direction	**Second Direction**
Three-letter airport code for origin	SFO	LAX
Three-letter airport code for destination	LAX	SFO
Average flight distance in miles	337	337
Average number of flights per year	14,712	14,540
Average annual passenger capacity	1,996,597	1,981,059
Average arrival delay in minutes	10	14

Method

I identified this route by running the following SELECT statement using impala on the VM:

```
SELECT origin, dest,
        ROUND(COUNT(*)/10) AS number_of_flights_per_year,
        ROUND(AVG(distance)) AS avg_flight_distance_miles,
        ROUND(SUM(seats)/10) AS avg_annual_passenger_capacity,
        ROUND(AVG(arr_delay)) AS avg_arrival_delay_min

  FROM flights LEFT OUTER JOIN planes
       ON flights.tailnum = planes.tailnum

  WHERE distance >= 300 AND distance <= 400
  GROUP BY origin, dest
  HAVING number_of_flights_per_year > 5000
```

```
    ORDER BY avg_annual_passenger_capacity DESC;
```

Notes

The “SELECT” part of the code includes origin and dest, which are the unique pair of airport routes; this is followed by number of flights per year, average flight distance in miles, average annual passenger capacity and average arrival delay in minutes. We take this data “FROM” the flights tables using a “LEFT OUTER JOIN” in order to include all the flights even if they don’t match in tailnumber with the planes table. We utilize the “WHERE” clause to select airports which are at a distance between 300 and 400 miles. We “GROUP BY” origin and dest to obtain unique pairs of airports and to aggregate base on this. Since we are interested only in pairs of airports which have an average number of flights per year greater than 5,000 we use “HAVING”. We finally “ORDER BY” the average annual passenger capacity in descending order, as indicated in the assignment.

My interpretations are that building this route would be truly beneficial, since the distance is in the range solicited, the number of flights in this route is of great number since the next possible route doubles in size, the same idea goes through the average annual passenger capacity, it almost doubles in size from the next possible route. Finally, I do not consider that the average arrival delay in minutes is the best option to sell this route, I would instead consider to sell this route as a viable opportunity not to waste the two hours needed for checking into a plane.