Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Bhavin Patel
Date: 10th July 2020

Assignment

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 largest total number of seats on the planes that flew between them. Your 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	14712	14540
Average annual passenger capacity	1996597	1981059
Average arrival delay in minutes	10	14

(Replace AAA and BBB with the actual airport codes, and fill in all the cells of the table.)

Method

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

```
SELECT flights.origin, flights.dest,
ROUND(COUNT(flights.flight)/10) AS avg_flights_year,
ROUND(SUM(planes.seats)/10) AS avg_seats_cap,
ROUND(AVG(flights.distance)) AS avg_flights_dist,
ROUND(AVG(flights.arr_delay)) AS avg_arr_delay
FROM flights
LEFT OUTER JOIN planes
ON flights.tailnum = planes.tailnum
WHERE flights.distance BETWEEN 300 AND 400
GROUP BY flights.dest, flights.origin
HAVING avg_flights_year >= 5000
ORDER BY avg_seats_cap DESC NULLS LAST
LIMIT 10;
```