**Peer-Graded Assignment:** Analyzing Big Data with SQL
**Name:** Alfred Chan
**Date:** 17/9/2020

*(Include your name and today's date above.)*

## 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** | 13,140.8 | 12,969.4 |
| **Average annual passenger capacity** | 1,996,597 | 1,981,058.5 |
| **Average arrival delay in minutes** | 10.5 | 13.9 |

*(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
   f.origin,
   f.dest,
   count(*)/10 AS flight_count_per_year,
   avg(f.distance) AS avg_distance,
   avg(f.arr_delay) AS avg_arrival_delay,
   sum(p.seats)/10 AS avg_annual_capacity

FROM fly.flights AS f LEFT OUTER JOIN fly.planes AS p
   ON f.tailnum = p.tailnum

WHERE distance BETWEEN 300 AND 400
```

```
GROUP BY origin, dest

HAVING flight_count_per_year >= 5000

ORDER BY
   flight_count_per_year DESC,
   avg_annual_capacity DESC,
   avg_arrival_delay DESC,
   avg_distance DESC;
```

*(Fill in the blank to indicate whether you used Hive or Impala, and fill in the SQL query.)*

## Notes

*(This section is optional. You may use it to describe your process, add details or caveats, explain your interpretations, or describe any further analysis that you performed.)*

- *ORDER  BY indicates the priority of choosing the airports*
    1. *flight count*
    2. *annual capacity*
    3. *arrival delay*
    4. *distance*

- *As high usage of tunnel should be top priority -> hence number of flights + potential number of people can rail can transit should be priortised*

- *SFO & LAX airport are top in number 1, 2 and 3 -> hence the choice of SFO & LAX*