

## Peer-Graded Assignment: Analyzing Big Data with SQL

Name: BatEl Yaish

Date: 12/25/2020

### 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	13141	12969
Average annual passenger capacity	1996597	1981058.5
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 'Origin',
    dest 'Dest',
    AVG(distance) 'avg_Flight_Distance',
    ROUND(COUNT(*)/10) 'avg_Num_Flights_Yearly',
    ROUND(AVG(arr_delay)) 'avg_Arrival_Delay',
    SUM(seats)/10 'avg_Annual_Passenger_Capacity'
FROM flights f LEFT JOIN planes p
ON f.tailnum = p.tailnum
WHERE distance BETWEEN 300 AND 400
GROUP BY origin,dest
HAVING avg_Num_Flights_Yearly > 5000
ORDER BY avg_Annual_Passenger_Capacity DESC
LIMIT 2;
```