Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Ross McKenna
Date: 16/11/2020

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	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 d.origin, d.dest, ROUND(AVG(d.distance)) AS flight_distance,
        round((COUNT(*))/10) AS flights_per_year,
        round(AVG(d.arr_delay)) AS avg_arrival_delay,
        round(sum(p.seats)/10) AS annual_seat_cap
  FROM  fly.flights AS d LEFT OUTER JOIN fly.planes AS p
    ON d.tailnum = p.tailnum
  WHERE d.distance BETWEEN 300 AND 400
  GROUP BY d.origin, d.dest
  HAVING annual_seat_cap > 5000
  ORDER BY annual_seat_cap DESC
  LIMIT 10;
```

Notes

SFO	LAX	337	13141	10	1996597
LAX	SFO	337	12969	14	1981059
PHX	LAX	370	8397	6	1219235
LAX	PHX	370	8376	6	1210173
PHX	SAN	304	6072	5	1067278
SAN	PHX	304	6025	4	1060204
SLC	DEN	391	7990	4	920919
DEN	SLC	391	7643	6	893437
BOS	DCA	399	7839	1	867688
DCA	BOS	399	7830	4	864009