

Peer-Graded Assignment: Analyzing Big Data with SQL

Name: HAKAN KALAYCI

Date: 06 SEP 2019

(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	PHX	LAX
Three-letter airport code for destination	LAX	PHX
Average flight distance in miles	370	370
Average number of flights per year	8662	8662
Average annual passenger capacity	1219235	1210173
Average arrival delay in minutes	6	6

(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:

Finding Airport origin and destination code along with average number of flights per year.

```
SELECT
    origin AS Origin,
    dest AS Destination,
    AVG(distance) AS avg_distance,
    ROUND(COUNT(flight)/10) AS avg_annual_num_of_flights,
    ROUND(SUM(seats)/10) AS avg_annual_seat_capacity,
    ROUND(AVG(arr_delay)) AS avg_delay
FROM flights f
LEFT OUTER JOIN planes p
```

```
ON f.tailnum = p.tailnum  
WHERE 300 <= f.distance AND f.distance <= 400
```

```
GROUP BY Origin, Destination  
HAVING avg_annual_num_of_flights > 5000  
ORDER BY avg_annual_seat_capacity DESC  
LIMIT 10;
```