

Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Balasubramanian PG
Date: 24/12/2023

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	8650
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:

```
SELECT fl.origin, fl.dest,  
round(avg(fl.distance)) AS avg_dist,  
round(count(fl.flight/10)) AS avg_flights_per_year,  
round(sum(pl.seats/10)) AS seat_cap,  
round(avg(fl.arr_delay)) as avg_delay  
FROM fly.flights AS  
fLEFT OUTER JOIN  
fly.planes as pl
```

```
ON fl.tailnum = pl.tailnum
WHERE fl.distance >= 300 AND fl.distance <= 400
GROUP BY fl.origin, fl.dest
HAVING avg_flights_per_year >
5000ORDER BY seat_cap DESC
```

LIMIT 10