Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Suhaimi William Chan
Date: March 21, 2021

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:

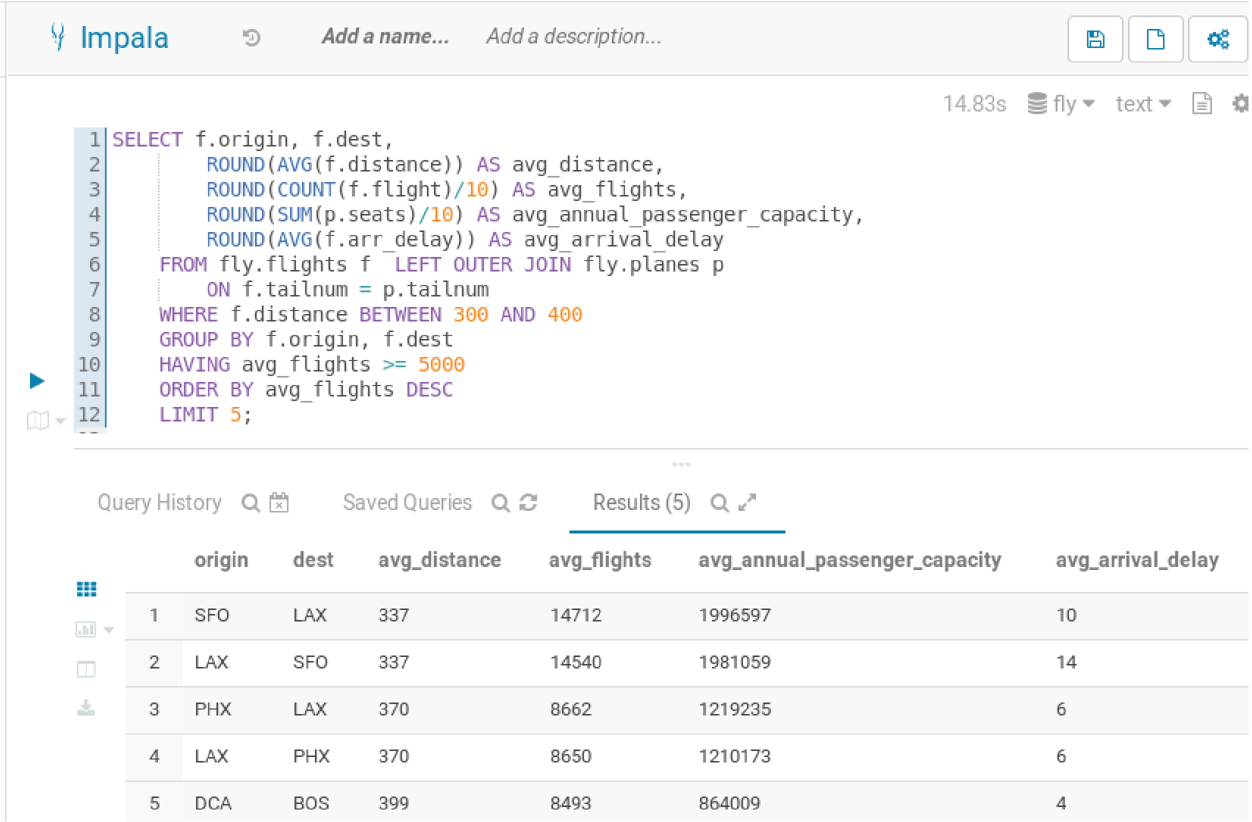
	First Direction	**Second Direction**
Three-letter airport code for origin	SFO	LAX
Three-letter airport code for destination	LAX	SFO
Average flight distance in miles	337	337
Average number of flights per year	14,712	14,540
Average annual passenger capacity	1,996,597	1,981,059
Average arrival delay in minutes	10	14

Method

I identified this route by running the following SELECT statement using Impala on the VM:

```
SELECT f.origin, f.dest,
    ROUND(AVG(f.distance)) AS avg_distance,
    ROUND(COUNT(f.flight)/10) AS avg_flights,
    ROUND(SUM(p.seats)/10) AS avg_annual_passenger_capacity,
    ROUND(AVG(f.arr_delay)) AS avg_arrival_delay
  FROM fly.flights f  LEFT OUTER JOIN fly.planes p
    ON f.tailnum = p.tailnum
  WHERE f.distance BETWEEN 300 AND 400
  GROUP BY f.origin, f.dest
  HAVING avg_flights >= 5000
  ORDER BY avg_flights DESC
  LIMIT 5;
```

Notes



```
SELECT f.origin, f.dest,
        ROUND(AVG(f.distance)) AS avg_distance,
        ROUND(COUNT(f.flight)/10) AS avg_flights,
        ROUND(SUM(p.seats)/10) AS avg_annual_passenger_capacity,
        ROUND(AVG(f.arr_delay)) AS avg_arrival_delay
    FROM fly.flights f  LEFT OUTER JOIN fly.planes p
        ON f.tailnum = p.tailnum
    WHERE f.distance BETWEEN 300 AND 400
    GROUP BY f.origin, f.dest
    HAVING avg_flights >= 5000
    ORDER BY avg_flights DESC
    LIMIT 5;
```

	origin	dest	avg_distance	avg_flights	avg_annual_passenger_capacity	avg_arrival_delay
1	SFO	LAX	337	14712	1996597	10
2	LAX	SFO	337	14540	1981059	14
3	PHX	LAX	370	8662	1219235	6
4	LAX	PHX	370	8650	1210173	6
5	DCA	BOS	399	8493	864009	4

(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.)