Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Chandrashekhar Kalnad
Date: 10/19/2019

(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	BET	BOS
Three-letter airport code for destination	ANC	DCA
Average flight distance in miles	399	399
Average number of flights per year	6278.19	5664.39
Average annual passenger capacity	21217.59	19176.29
Average arrival delay in minutes	8.3	97.7

(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 fly.flights.origin, fly.flights.dest, fly.flights.distance, sum(isnull((fly.flights.flight)/10, 0)) as
avg_flight, sum(isnull((fly.planes.seats)/10, 0)) as avg_seats, sum(isnull((fly.flights.arr_delay)/10, 0))
as avg_delay from fly.flights
RIGHT OUTER JOIN fly.planes ON fly.flights.tailnum = fly.planes.tailnum
where (fly.flights.distance> 300 AND  fly.flights.distance<400) GROUP BY fly.flights.tailnum,
fly.flights.origin, fly.flights.dest, fly.flights.distance
ORDER BY avg_seats DESC, avg_flight DESC, avg_delay DESC;
```

(Fill in the blank to indicate whether you used Hive or Impala, and fill in the SQL query.)

Notes

My preference order is for selection of flights is most seats -> most flights -> most arrival delay