

TASK:

Imagine that you have been hired as a data analyst for a company that plans to disrupt the airline industry by building an underground high-speed passenger rail tunnel. The company needs your help to decide which two major United States airports this tunnel should connect. The distance between the airports must be within a specified range, and the airports must have a large volume of air travellers flying between them in both directions. The company believes that these air travellers can be persuaded to switch to high-speed rail because of frustratingly long flight delays.

QUERY:

Identified the route by running the following SELECT statement using Impala on the VM:

```
SELECT
    origin as Origin,
    dest as Destination,
    avg(distance) as Average_Distance,
    count(flight)/10 as AVG_ANNUAL_NO_OF_FLIGHTS,
    sum(seats)/10 as AVG_ANNUAL_NO_OF_SEATS,
    avg(arr_delay) as AVF_DELAY

from flights f
LEFT OUTER JOIN planes p
ON f.tailnum = p.tailnum

WHERE f.distance BETWEEN 300 AND 400

GROUP BY Origin, Destination
HAVING AVG_ANNUAL_NO_OF_FLIGHTS > 5000
ORDER BY AVG_ANNUAL_NO_OF_SEATS DESC
LIMIT 10
```

QUERY OUTPUT:

	origin	destination	average_distance	avg_annual_no_of_flights	avg_annual_no_of_seats	avf_delay
1	SFO	LAX	337	14711.6	1996597	10.32287293126454
2	LAX	SFO	337	14539.9	1981058.5	13.762905484742598
3	PHX	LAX	370	8662.1000000000004	1219234.8999999999	5.8187046323245779
4	LAX	PHX	370	8650	1210173.1000000001	5.9509371554575523
5	PHX	SAN	304	6200.3999999999996	1067278.3999999999	4.8899259210912769
6	SAN	PHX	304	6215.8000000000002	1060204.1000000001	3.77428823011447
7	SLC	DEN	390.62256309128577	8012.1999999999998	920918.5	4.2241251305540528
8	DEN	SLC	390.62220888981636	7667.1999999999998	893437.40000000002	6.2365325728404146
9	BOS	DCA	399	8484.3999999999996	867688.30000000005	1.3520530960542629
10	DCA	BOS	399	8492.5	864009.09999999998	4.0304374240583236

Recommendation 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	14711	14539
Average annual passenger capacity	1996597	1981058
Average arrival delay in minutes	10	13

I recommended the above route because of the following factors: -

- 1) The average of seats per year is greater than any other routes.
- 2) There were ofcourse, other routes that were having more distance between them, but were falling short on seats.
- 3) Also, the average delay was considered before recommending the routes.


FLY DATABASE:

Tables






Table Name
airlines
airports
flights
planes

Column Details for all the Tables

fly.airlines

	Name	Type
1	 carrier	string
2	 name	string

fly.airports

	Name	Type
1	 faa	char(3)
2	 name	string
3	 lat	double
4	 lon	double
5	 alt	smallint

fly.flights

		Name	Type
1	i	year	smallint
2	i	month	tinyint
3	i	day	tinyint
4	i	dep_time	smallint
5	i	sched_dep_time	smallint
6	i	dep_delay	smallint
7	i	arr_time	smallint
8	i	sched_arr_time	smallint
9	i	arr_delay	smallint
10	i	carrier	string
11	i	flight	smallint
12	i	tailnum	string
13	i	origin	string
14	i	dest	string
15	i	air_time	smallint
16	i	distance	smallint

fly.planes

		Name	Type
1	i	tailnum	string
2	i	year	int
3	i	type	string
4	i	manufacturer	string
5	i	model	string
6	i	engines	int
7	i	seats	int
8	i	engine	string