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	14712	14540
Average annual passenger capacity	1996597	1981059
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(avg(f.arr_delay)) as avg_arr_delay,
      round(count(*)/10)  as avg_flights,
      round(sum(p.seats)/10) as avg_capacity
from flights f
   left outer join planes as 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_capacity desc
```

Notes

Results are rounded so we can see the stats clearer