**Project:** A duplet of airports!

**Name:** Sombik Roy

**Date:** 29-June-2021

**Objective**

We have to recommend which pair of United States airports should be connected with a high-speed passenger rail tunnel. To do this, we should 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. Then we must 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.

**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:

Finding airport origin and destination code along with average number of flights per year.

**SELECT**

**origin AS Origin,**

**dest AS Destination,**

**Round(AVG(distance)) AS avg\_distance,**

**ROUND(COUNT(flight)/10) AS avg\_annual\_num\_of\_flights,**

**ROUND(SUM(seats)/10) AS avg\_annual\_seat\_capacity,**

**ROUND(AVG(arr\_delay)) AS avg\_delay**

**FROM flights f**

**LEFT OUTER JOIN planes p**

**ON f.tailnum = p.tailnum**

**WHERE 300 <= f.distance AND f.distance <= 400**

**GROUP BY Origin, Destination**

**HAVING avg\_annual\_num\_of\_flights > 5000**

**ORDER BY avg\_annual\_seat\_capacity DESC**

**LIMIT 10**

**Notes**

*Database Used: fly*

*Query Interface : Hue*

*Query Engine : Impala. Impala gives better response times than Hive.*

*Tables used : flights, planes*

*Flights table gives details about flights between each route, departure, arrival times, distance between airports, origin and destination airports, flight number, flight tail num that has flown, year, day, month of the flight.*

*Planes table has details regarding the manufacturer of the plane, manufacturer year of the plane, type of the plane, tailnum of the plane, engines etc.*

*Below table gives data for pair of airports whose average flights per year > 5000.Considering average flights per year and highest average annual capacity and average arrival delay of 10 and 14 minutes, the route between* ***SFO (San Francisco)*** *and* ***LAX (Los Angeles)*** *is considered for recommendation for high-speed rail connectivity*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **origin** | **dest** | **avg\_distance** | **avg\_flights\_year** | **avg\_annual\_passenger\_capacity** | **avg\_arrival\_delay** |
| SFO | LAX | 337 | 14712 | 1996597 | 10 |
| LAX | SFO | 337 | 14540 | 1981059 | 14 |
| PHX | LAX | 370 | 8397 | 1219235 | 6 |
| LAX | PHX | 370 | 8376 | 1210173 | 6 |
| PHX | SAN | 304 | 6072 | 1067278 | 5 |
| SAN | PHX | 304 | 6025 | 1060204 | 4 |
| SLC | DEN | 390.623628 | 7990 | 920919 | 4 |
| DEN | SLC | 390.62344 | 7643 | 893437 | 6 |
| BOS | DCA | 399 | 7839 | 867688 | 1 |
| DCA | BOS | 399 | 7830 | 864009 | 4 |
| ORD | MSP | 334 | 5687 | 834917 | 8 |
| OAK | LAX | 337 | 6088 | 814200 | 8 |
| LAX | OAK | 337 | 6089 | 813435 | 7 |
| MSP | ORD | 334 | 5509 | 811846 | 8 |
| ATL | RDU | 356 | 5276 | 776510 | 2 |
| RDU | ATL | 356 | 5273 | 774922 | 4 |
| SJC | LAX | 308 | 5911 | 696069 | 6 |
| LAX | SJC | 308 | 5900 | 694074 | 6 |
| SMF | LAX | 373 | 5411 | 582367 | 6 |
| LAX | SMF | 373 | 5406 | 580252 | 7 |
| BOS | BWI | 369.355502 | 5615 | 546348 | 6 |
| BWI | BOS | 369.355721 | 5612 | 545351 | 8 |