For this assignment, you will query and analyze flights data to inform a business decision.

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 travelers flying between them in both directions. The company believes that these air travelers can be persuaded to switch to high-speed rail because of frustratingly long flight delays.

You must write a SQL statement and analyze the result to recommend which two airports this rail tunnel should connect. Then you must create and upload a document describing the SQL statement you ran and the tunnel route you recommend.

Review criteria

Your submission will be graded based on the following criteria:

- How carefully you followed the instructions provided (for example: did you include all requested details in the document?)
- Whether your recommendation adheres to all the requirements described (for example: was the tunnel route you recommended within the range of allowed distances?)
- Whether the results of your SQL statement accurately answer the questions that are asked
- Whether your peers can reproduce your results using the SQL statement you provide
- Whether your SQL statement uses the appropriate clauses and an appropriate type of join
- How readable is the SQL statement that you provide (for example: did you use line breaks and indentation in the clauses of the statement?)

Example Submissions

Use the template provided here to create the document required for this assignment.

document_template.doc

This file can be opened using word processing applications such as Microsoft Word or LibreOffice, or it can be imported into Google Drive and edited using Google Docs.

Step-By-Step Assignment Instructionsless

Prepare

1. Download and open the template document provided above.
2. Start the course VM, open Hue, go to the Hive or Impala Query Editor, and select **fly** as the current database.

Understand the Task

Your job is to recommend which pair of United States airports should be connected with a high-speed passenger rail tunnel. The company you work for has given you the following strict requirements:

These two airports must:

- Be between **300** and **400** miles apart
- Average at least **5,000** (five thousand) flights per year between them, *in each direction*

Among the pairs of airports that meet these requirements, you must identify the one pair that has the **largest total number of seats on the planes that flew between them**.

The company is also interested to know the **average arrival delay** for flights between these two airports, because they believe that routes with a history of delayed arrivals will make it easier to persuade air travelers to switch to high-speed rail.

For the pair of airports you recommend, you must provide the following details:

- The three-letter codes identifying both airports
- The average flight distance in miles for flights between the airports, in each direction
- The average number of flights per year between the airports, in each direction
- The average annual passenger capacity (average yearly total number of seats on the planes) for flights between the airports, in each direction
- The average arrival delay for flights between the airports, in each direction

Write the SQL Query

You must write a SELECT statement to identify the pair of airports that fulfills all the requirements listed above. This SELECT statement must also return all the required details listed above.

The following hints might help you:

- The **flights** table has a column named **distance** which gives the distance in miles of each flight. Use the values in this column as the distances between airports.
- The **planes** table contains ten years of flights data, so to get per-year (annual) average totals, **divide the full-table totals by ten**.
- The **planes** table has a column named **seats** which gives the number of seats on each plane.

- The first two rows in the result of your query should show your recommended tunnel route. These top two rows should both show the same pair of airports, but with the **origin** and **dest** switched.

See the **Frequently Asked Questions** section below if you get stuck and need additional hints.

Complete Your Document

Starting with the provided template, finish the document describing your recommended tunnel route and the query you ran to find it.

Your document must include:

- All the required details of your recommended route
- The SQL query that you wrote and ran to identify this route

Try to make the document clear and concise. Format the SQL query so it can be easily read and understood by your peers.

Include your name and the date at the top of your document.

Submit Your Document

Save your document as a PDF file, then upload it in the **My submission** tab.

Review Your Peers

You must review and grade ***three (3)*** of your peers' assignments, and get ***three (3)*** reviews of your own. If, after some time, you still need some reviews of your assignment, you can add a thread to the discussion forum and post a link to your submission.

Frequently Asked Questions less

Here are some frequently asked questions about the assignment:

What SQL engine should I use?

You must use Hive or Impala on the course VM.

Can I use two or more SELECT statements for this assignment?

No, you must run just one single SELECT statement to complete this assignment.

What clauses should my SELECT statement include?

Your SELECT statement should include:

- A SELECT clause
- A FROM clause
- A WHERE clause
- A GROUP BY clause
- A HAVING clause
- An ORDER BY clause

Should my SELECT statement use a join?

Yes, your SELECT statement should use a join. You should decide which type of join is most appropriate.

Should my SELECT statement use a subquery?

No, subqueries were not covered in this course. It is not necessary to use them for this assignment. While *you* might understand subqueries well enough to help with this assignment, you can't be sure that your peers grading your assignment will understand them, and that could cost you points.

How should I calculate the average number of flights per year and the average aircraft passenger capacity per year?

The data in the **flights** table contains **ten full years** of data, representing flights from January 1, 2008 through December 31, 2017. To find the average number of flights *per year*, **calculate the total number** of flights in the table and **divide that number by ten**. Similarly, to find the average aircraft passenger capacity per year, calculate the total aircraft passenger capacity for all ten years and divide that number by ten. Do not assume that you must use the aggregate function **AVG** when you are asked to provide an average!

Which tables should I use?

You must use the **flights** and **planes** tables in the **fly** database on the VM. *Hint:* These tables can be joined using the column that is named **tailnum** in both tables as the join key column.

Are there some rows in the **flights** table with **tailnum** values that are not in the **planes** table?

Yes. Because of this, you should be careful about what type of join you pick, or else you might undercount the numbers of flights.

Also note that because of this, the total numbers of seats on the flights are likely higher in reality than in the numbers your queries will return—but there is nothing you can do to resolve that.

Can I use other data besides what is available in the **fly** database on the VM?

No, do not use any data except the tables in the **fly** database on the VM. If you would like, in the **Notes** section of the document that you submit, you may optionally present further work you did that used other data, but this is entirely optional and you will not be graded on this.

Can I recommend a tunnel connecting three or more airports?

No. For this assignment, you must recommend which **two** airports to connect.

Can I combine airports for cities like New York that have more than one airport?

For this assignment, do *not* attempt to combine nearby airports together. Analyze the data as if each airport stands alone, and report your recommendation using only two airports. If you would like, in the **Notes** section of the document that you submit, you may optionally present further recommendations based on combining nearby airports together, but this is entirely optional and you will not be graded on this.