Peer-graded Assignment: Data Management

i

It looks like this is your first peer-graded assignment. Learn more

Submit your assignment soon

Even though your assignment is due on Oct 25, 1:59 PM +07, try to submit it 1 or 2 days early if you can. Submitting early gives you a better chance of getting the peer reviews you need in time.

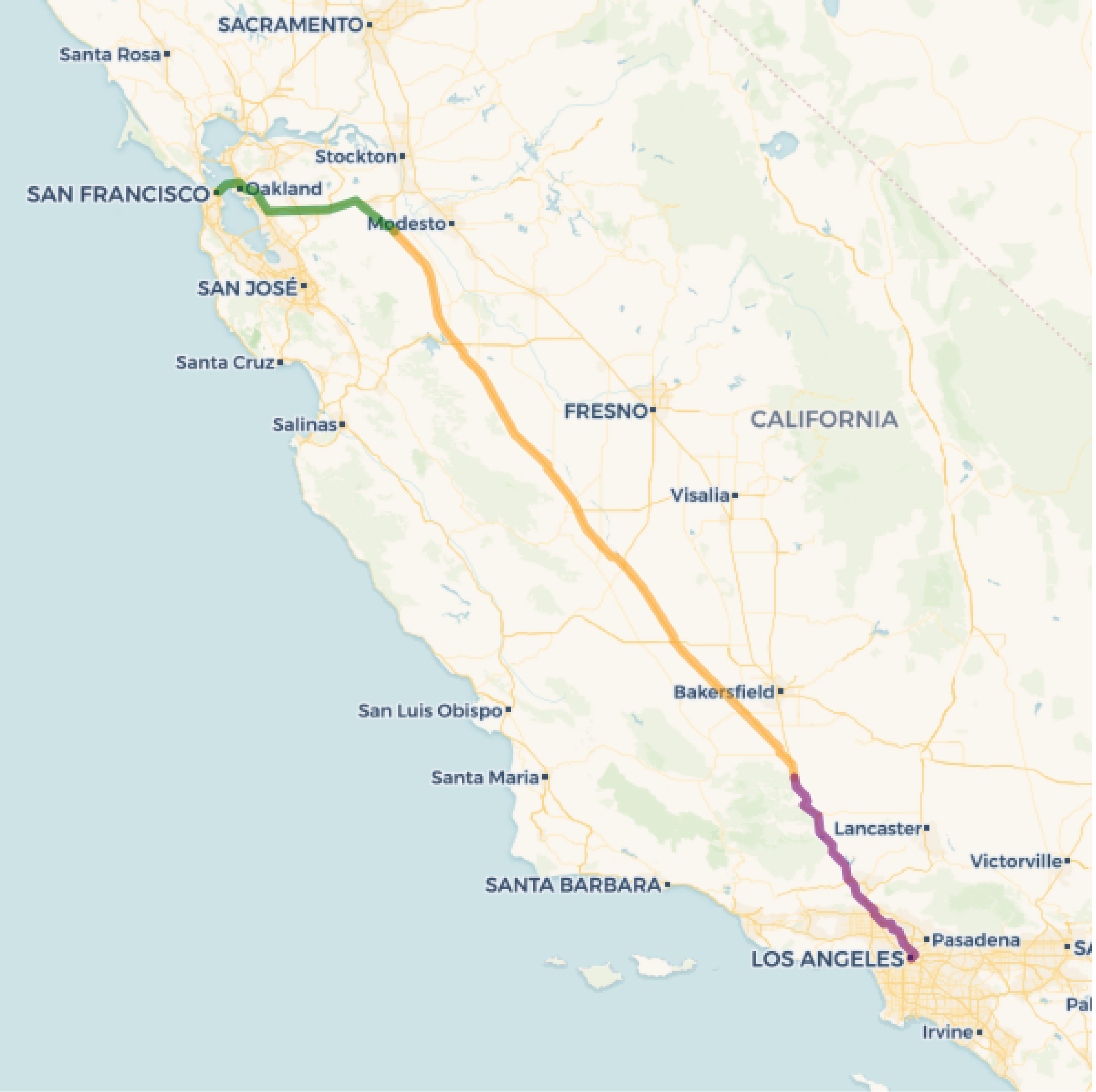
Instructions
My submission
Discussions

For this assignment, you will create a table with data describing an underground tunneling project.

If you took the second course in this specialization (*Analyzing Big Data with SQL*), recall that the peer-reviewed assignment asked you to analyze flights data to select a profitable route for an underground high-speed rail tunnel. Based on your analysis and on other factors, construction has begun on a tunnel connecting San Francisco and Los Angeles. The tunnel will be dug over a period of ten years. It will be dug in three different sections by three tunnel boring machines (TBMs) named Bertha II, Shai-Hulud, and Diggy McDigface.

Each of these TBMs will generate a large volume of data as it operates. Each TBM will generate the data slightly differently. Simulated versions of the three TBM-generated datasets are provided. You must create a table on the VM and load these datasets into it. Then you must create and upload a document describing the steps you performed to complete this task.

Note: Please review the Coursera Honor Code if you are unfamiliar with it. Specifically, be sure that your work is your own, and do not share it with others (including in the forums here or in outside locations including, but not limited to, LinkedIn and GitHub).


SACRAMENTO
Santa Rosa
Stockton
SAN FRANCISCO
Oakland
Modesto
SAN JOSÉ
Santa Cruz
Salinas
FRESNO
CALIFORNIA
Visalia
Bakersfield
San Luis Obispo
Santa Maria
Lancaster
Victorville
SANTA BARBARA
Pasadena
LOS ANGELES
Irvine

Review criteria

less

Your submission will be graded based on the following criteria:

- How carefully you followed the instructions provided here (for example: are the table and columns named exactly as the instructions specify?)
- How well you selected data types (for example: did you use the DECIMAL type with properly chosen scale and precision for representing exact decimal numbers?)
- The accuracy of the results (for example: do the queries of dig.tbm_sf_la return the expected results?)
- Whether you handled the differences between the three data files for the three different TBMs
- Whether the steps you describe produce the results you describe
- How clearly the document explains your process
- How readable the commands and SQL statements in the document are (for example: did you use line breaks and indentation in your SQL statements?)

Example Submissions

less

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

document_templateDOC File

Download file

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 Instructions

less

Prepare

1. Review the Coursera Honor Code if you are unfamiliar with it. Specifically, be sure that your work is your own, and do not share it with others (including in the forums here or in outside locations including, but not limited to, LinkedIn and GitHub).
2. Download and open the template document provided above
3. Start the course VM and ensure that it is connected to the internet

Examine the Data

Use the commands you learned about in this course to list and examine the three files containing the tunnel boring machine data. They are delimited text files, each containing tens of thousands of lines. They are stored in Amazon S3 in subdirectories under a directory named tbm_sf_la in the S3 bucket named training-coursera2. You have read access to this bucket.

Hint: List these files and view their contents by running commands in the terminal. Use chaining to apply the head command to display only the first few rows of each file.

Notice what these three files have in common:

- Each file contains eight columns representing the same eight fields
- The data types of the eight columns are the same in all three files
- The rows of the table represent hourly time intervals

Notice the differences between these three files:

- They use different delimiters

- One of the files uses the string 999999 to represent missing values
- One of the files has a header line

Create the Table

Create a table named tbm_sf_la in the database named dig to store the data from all three of the TBMs. Use what you learned by examining the data to decide how best to do this.

When creating this table, you must:

- Use the exact column names shown in the header line in one of the files
- Specify appropriate data types for the columns, based on what you observed when examining the data and based on what you have learned about data types in this course
- Ensure that the table can be queried by both Hive and Impala

Remember that all the files in one table's storage directory should be uniformly formatted; they should all use the same delimiter, they should all use the same strings to represent missing values, and they should either all have a header line or none of them should. However, the files in S3 are *not* uniformly formatted, so you cannot simply copy the three files to the new table's directory.

Hint: You might decide to create three separate tables (one for each TBM) as an intermediate step before creating one table to store the data for all three TBMs.

Keep in mind that there are many different ways to successfully complete this task. Think about everything you have learned in this course and consider alternative approaches.

Load the Data into the Table

Load the data for all three TBMs into the table. That this might be a one-step process or a multi-step process—or it might not be necessary at all—depending on what approach you decided to use. Keep track of all the commands and statements you run, so that you can include them all in the document.

Run some SELECT queries on the resulting table and check that they return the expected results.

Describe Ways to Optimize the Table

Think about possible ways to optimize the table you created. Consider whether the approach you used would enable analysts to run queries quickly and efficiently even if the data was much larger (for example, if each row represented one second instead of one hour). What could you do differently? Describe this in your document.

Complete Your Document

Starting with the provided template, finish the document describing the steps you performed to complete this task.

Your document must include:

- A complete description of all the steps you performed.
- All commands and SQL statements you ran to complete the task.

- Descriptions of how you handled the different delimiters, the missing values represented as 999999 in one file, and the header line in one file
- The result of the SQL statement:

```
SELECT tbm, COUNT(*) AS num_rows FROM dig.tbm_sf_la GROUP BY tbm ORDER BY tbm;
```

- The result of the SQL statement:

```
DESCRIBE dig.tbm_sf_la;
```

Try to make the document clear and concise. Format the commands and SQL statements in the document so they 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:

Why isn't my table (or the data in it) showing up in Impala?

Did you remember to invalidate or refresh Impala's metadata cache?

What tools and techniques may I use to complete this assignment?

You should use the tools and techniques that are taught in this course. Although there are other tools and techniques that you could use to complete this assignment, your peers might not be able to review your work if you use them.

Is it OK if dig.tbm_sf_la is a *view* instead of a table?

Yes. This is not the only way to complete the problem, but it is one way. Views are covered in Week 5 (the honors week) of this course; you may want to skip ahead to view the content about views before completing this assignment. However, it's not necessary to use views, so if you are not planning to complete the honors week, you do not need to skip ahead.

Is it OK if I create the table dig.tbm_sf_la and load the data into it in a single step?

Yes. For example, if you use a CTAS statement, then the table creation and data loading will happen in a single step.

What should I do to omit the header line that is in one of the files from the table's data?

There are several different ways to do this. The best way depends on which approach you decide to use to complete the task. *Hint:* One way is by setting the table property skip.header.line.count for the table that has the data file containing the header line.

What should I do with the missing values that are represented by 999999 in one of the files?

Make these values appear as NULL in query results. There are several different ways do do this. *Hint:* You could achieve this by setting the table property serialization.null.format.

In the file where missing values are represented by 999999, are there any places where 999999 does *not* represent a missing value?

No. In that file, the string 999999 always represents a missing value.

How do you specify multiple table properties in one TBLPROPERTIES clause?

Separate the 'key'='value' pairs with commas. For example: TBLPROPERTIES ('key1'='value1', 'key2'='value2')