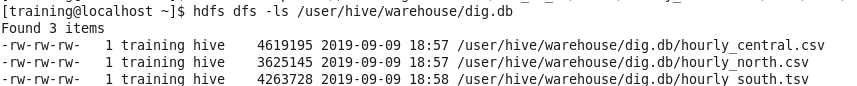
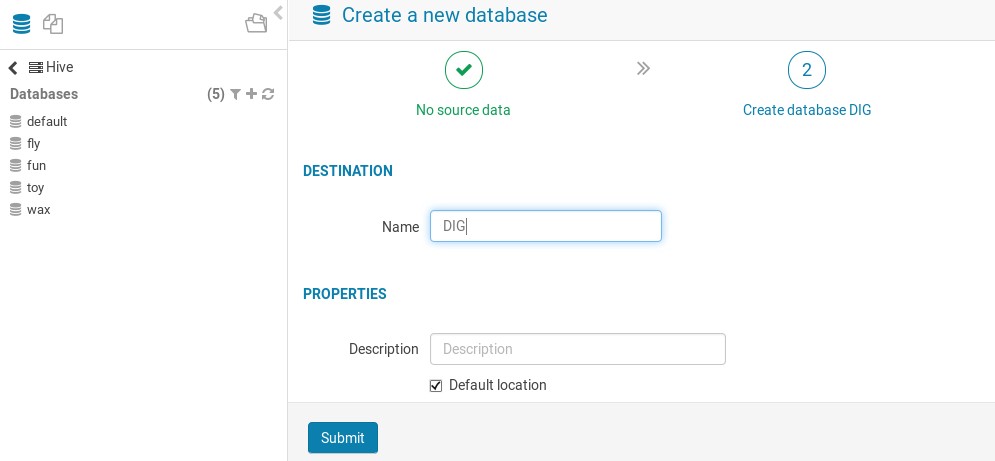
**Peer-Graded Assignment:** Data Management



**Course:** Managing Big Data in Clusters and Cloud Storage

**Name:** Sayan Chandra

**Date:** 28/06/2021

*(Include your name and today’s date above.)*

# Assignment

Create a table named **tbm\_sf\_la** in the database named **dig** to store the data from three tunnel boring machines (TBMs), which is currently stored in S3 in three separate subdirectories under a directory named **tbm\_sf\_la** in the bucket named **training-coursera2**. In this document, describe the steps taken to complete this task.

# Solution

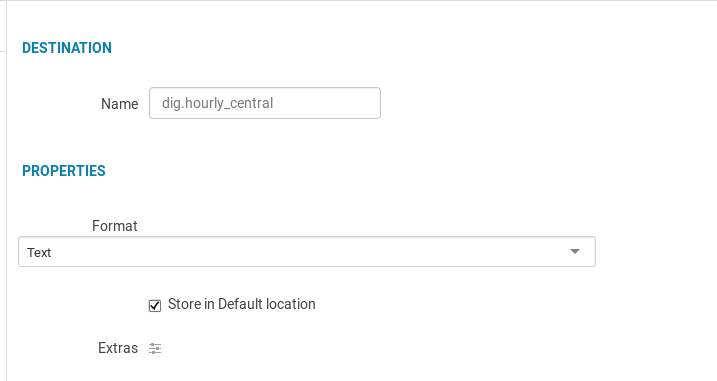
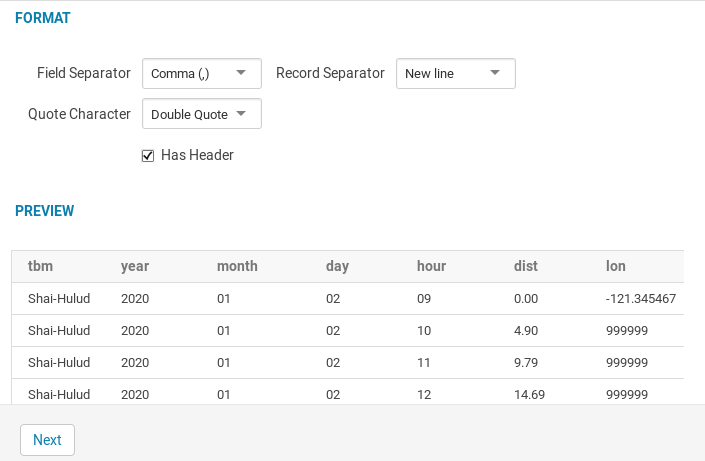
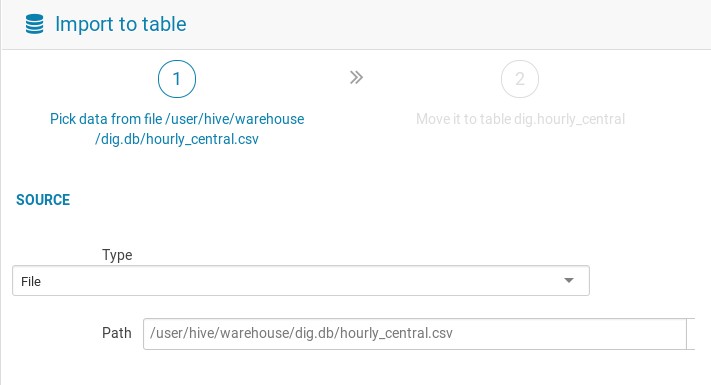
I performed the following steps to complete this task:

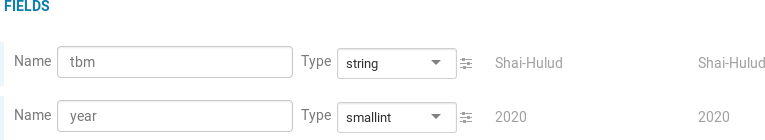
1. I got below three files from s3 to local directory via terminal
   * “hdfs dfs – get s3a://training-coursera2/tbm\_sf\_la/south/hourly\_south.tsv .”
   * “hdfs dfs – get s3a://training-coursera2/tbm\_sf\_la/north/hourly\_north.csv .”
   * “hdfs dfs – get s3a://training-coursera2/tbm\_sf\_la/central/hourly\_central.csv .”
2. I imported Local directory to Hue Browser

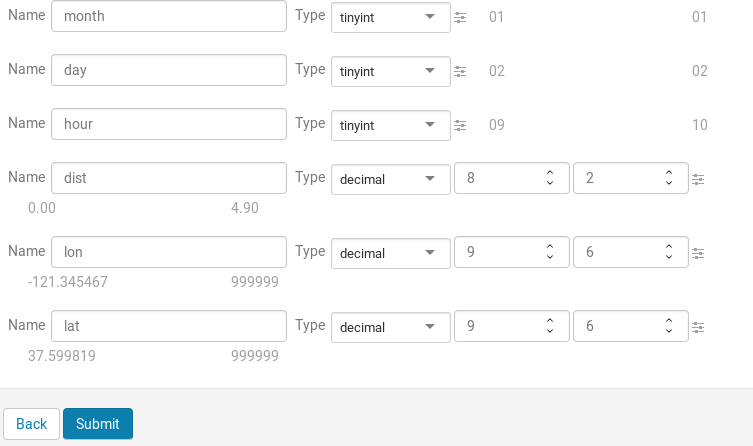
hdfs dfs –mkdir /user/hive/warehouse/dig.db

hdfs dfs -cp s3a://training-coursera2/tbm\_sf\_la/central/hourly\_central.csv /user/hive/warehouse/dig.db hdfs dfs -cp s3a://training-coursera2/tbm\_sf\_la/north/hourly\_north.csv /user/hive/warehouse/dig.db hdfs dfs -cp s3a://training-coursera2/tbm\_sf\_la/south/hourly\_south.tsv /user/hive/warehouse/dig.db

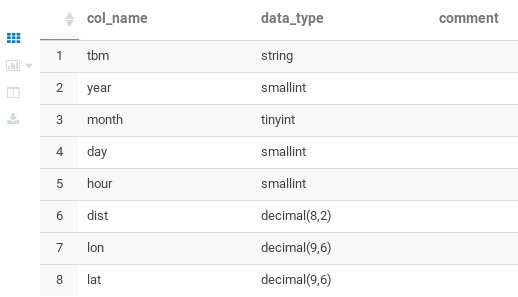
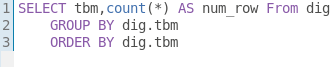
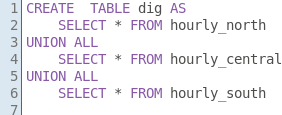
* I executed below operation each csv files







3.



\*hue optimized data type

# Result

After performing the steps described above, I ran the following queries and they produced the following result sets:

**SELECT tbm, COUNT(\*) AS num\_rows FROM dig.tbm\_sf\_la GROUP BY tbm ORDER BY tbm;**

|  |  |
| --- | --- |
| **tbm** | **num\_rows** |
| Bertha II | 91619 |
| Diggy McDigface | 93163 |
| Shai-Hulud | 94237 |

**DESCRIBE dig.tbm\_sf\_la;**

|  |  |
| --- | --- |
| **name** | **type** |
| tbm | string |
| year | smallint |
| Month | tinyint |
| Day | smallint |
| Hour | smallint |
| dist | decimal (8,2) |
| lon | decimal (9,6) |
| lat | decimal (9,6) |

# Notes

Same operation will executed in terminal