Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Priyanshu Saxena
Date: 21 July, 2021

Assignment

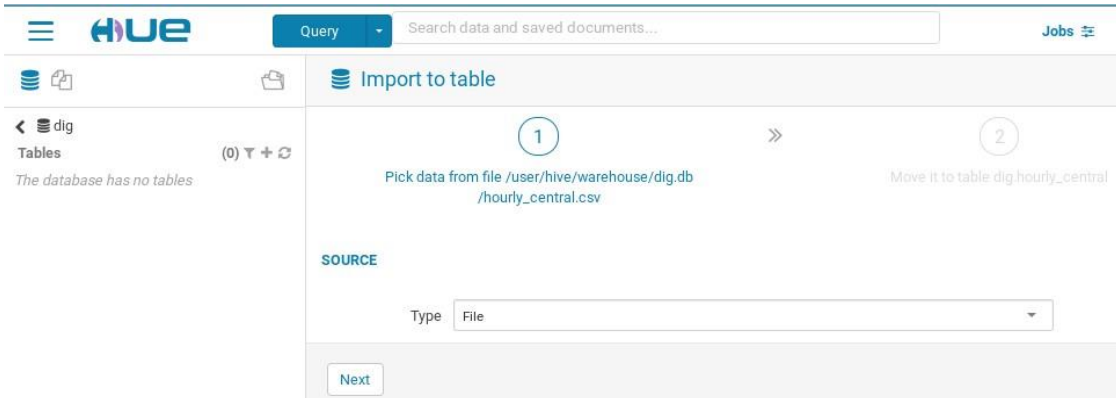
Create a table named **tbm_sf_la** in the database named **dig** to store the data from three tunnel boring machines (TBMs), which is currently stored in S3 in three separate subdirectories under a directory named **tbm_sf_la** in the bucket named **training-coursera2**. In this document, describe the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

Solution

I performed the following steps to complete this task:

1. Following are the steps which were ran on the terminal to download the files from bucket on the local system:
   “hdfs dfs -get s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv”
   “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”

2. Imported the data from local file system to the Hue Browser. Following are the screenshots for a csv (comma separated) file, the process remains the same for tsv (tab separated) file:



The database has no tables

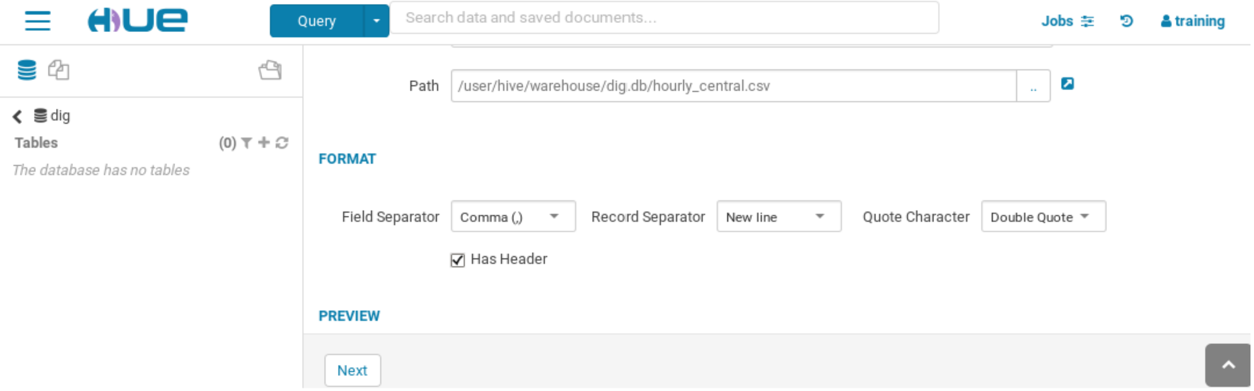
Import to table

1 Pick data from file /user/hive/warehouse/dig.db/hourly_central.csv

2 Move it to table dig.hourly_central

SOURCE

Type: File



The database has no tables

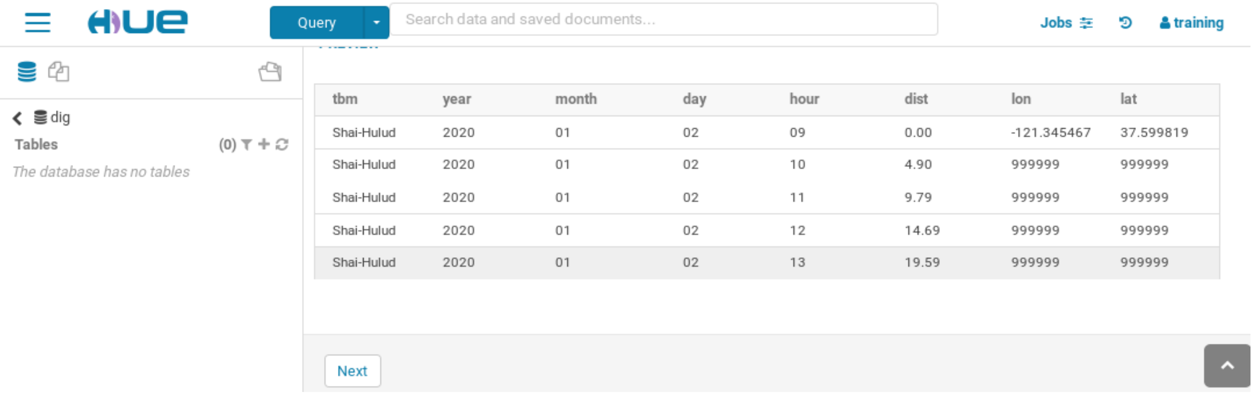
Path: /user/hive/warehouse/dig.db/hourly_central.csv

FORMAT

Field Separator: Comma (,) Record Separator: New line Quote Character: Double Quote

☑ Has Header

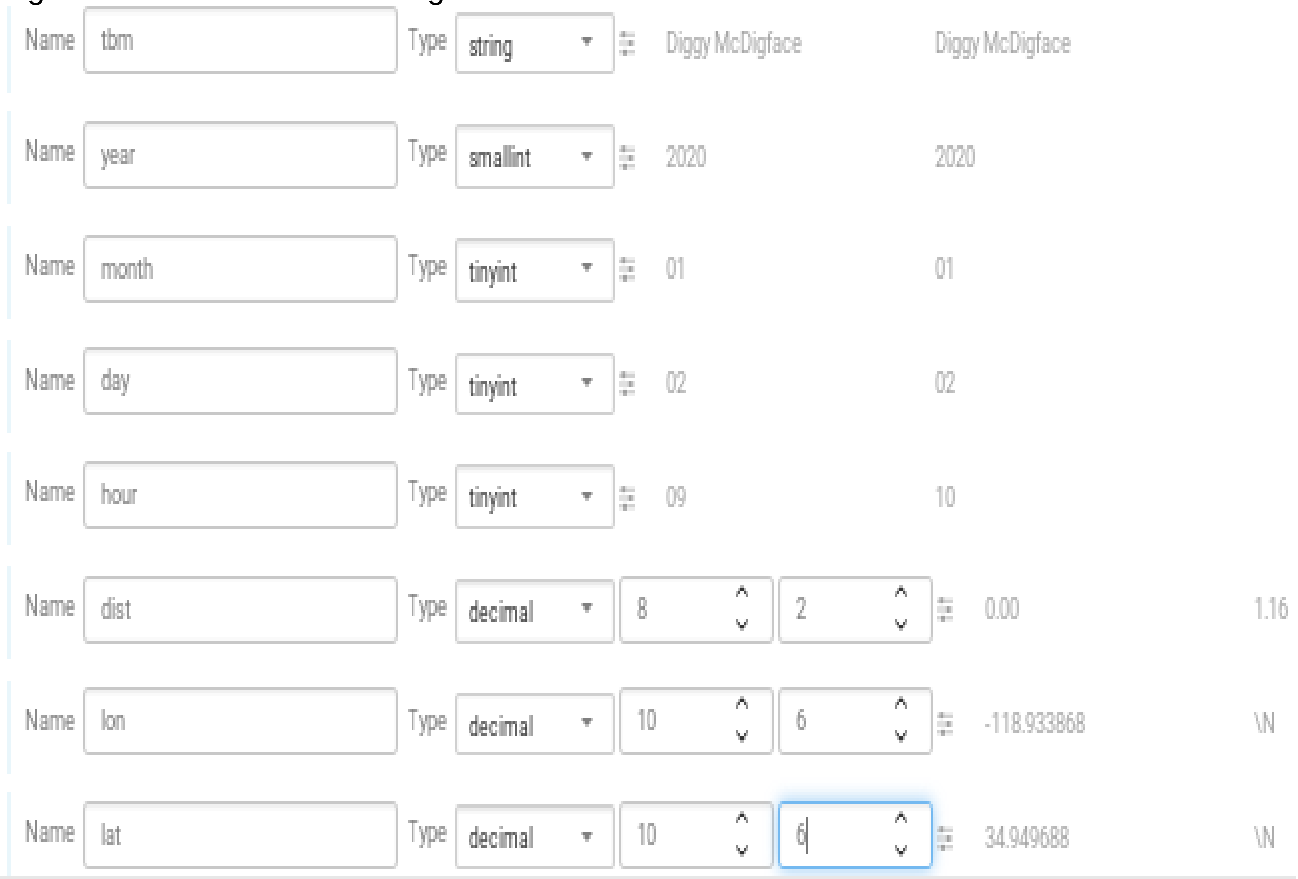
PREVIEW



The database has no tables

tbm	year	month	day	hour	dist	lon	lat
Shai-Hulud	2020	01	02	09	0.00	-121.345467	37.599819
Shai-Hulud	2020	01	02	10	4.90	999999	999999
Shai-Hulud	2020	01	02	11	9.79	999999	999999
Shai-Hulud	2020	01	02	12	14.69	999999	999999
Shai-Hulud	2020	01	02	13	19.59	999999	999999

3. Following Screenshots are for moving data to the table:



Name	tbm	Type	string			Diggy McDigface	Diggy McDigface
Name	year	Type	smallint			2020	2020
Name	month	Type	tinyint			01	01
Name	day	Type	tinyint			02	02
Name	hour	Type	tinyint			09	10
Name	dist	Type	decimal	8	2	0.00	1.16
Name	lon	Type	decimal	10	6	-118.933868	\N
Name	lat	Type	decimal	10	6	34.949688	\N

4. I ran the following commands on the HUE Browser Query:

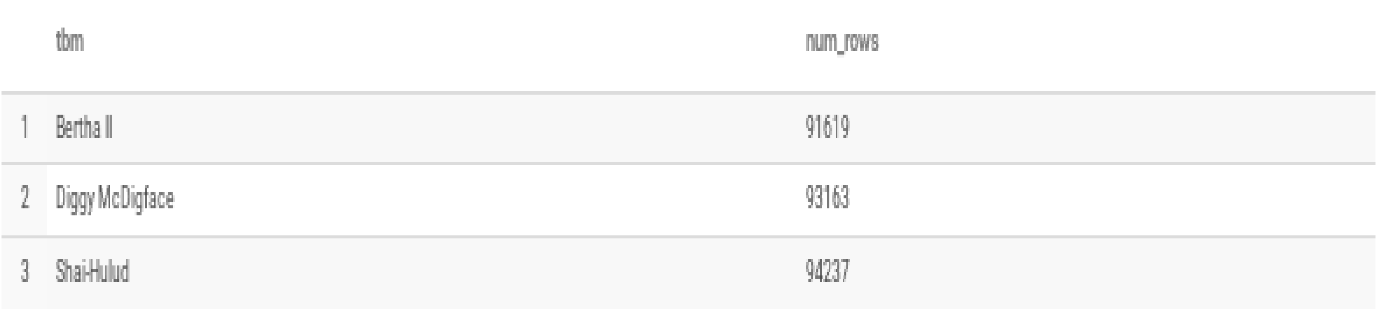
a. CREATE TABLE **dig.tbm_sf_la** AS SELECT *
FROM **hourly_central** UNION ALL
SELECT * FROM **hourly_north**
union all
SELECT * FROM **hourly_south**;

b. ALTER TABLE **dig.tbm_sf_la**
SET TBLPROPERTIES("serialization.null.format" = "**99999**");

Result

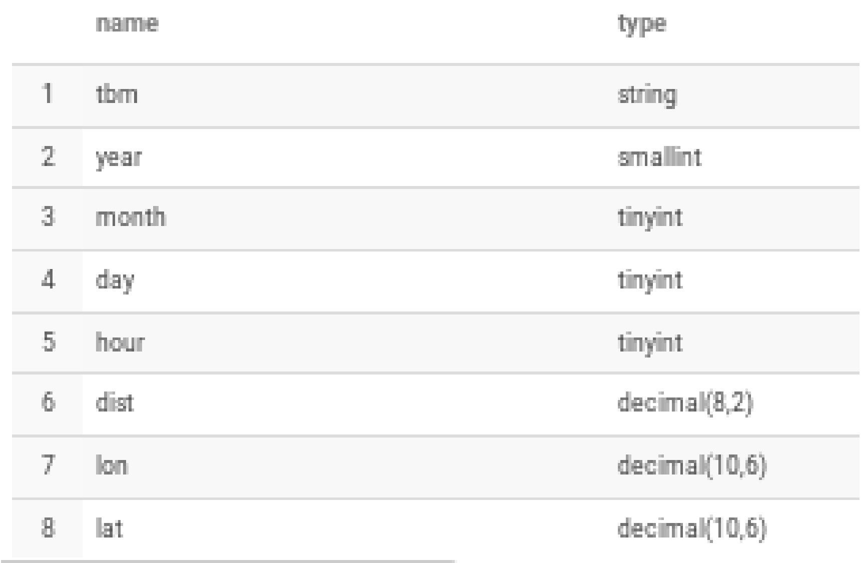
After performing the steps described above, I ran the following queries and they produced the following result sets:

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



	tbm	num_rows
1	Bertha II	91619
2	Diggy McDigface	93163
3	Shai-Hulud	94237

DESCRIBE dig.tbm_sf_la;



	name	type
1	tbm	string
2	year	smallint
3	month	tinyint
4	day	tinyint
5	hour	tinyint
6	dist	decimal(8,2)
7	lon	decimal(10,6)
8	lat	decimal(10,6)

Notes

(In this section, describe ways that you could further optimize the table. You may also describe other methods you considered or attempted.)