Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: HAKAN KALAYCI
Date: 10.09.2019

(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.

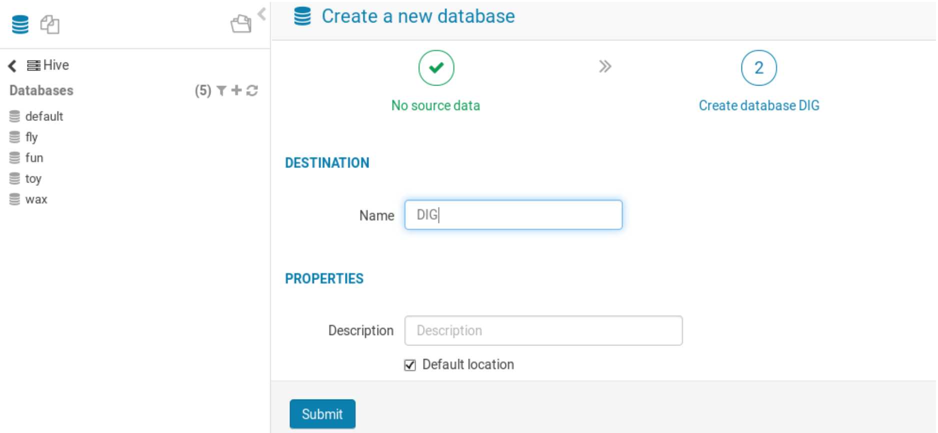
Solution

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
```



Hive
Databases (5)
default
fly
fun
toy
wax

Create a new database

No source data » 2 Create database DIG

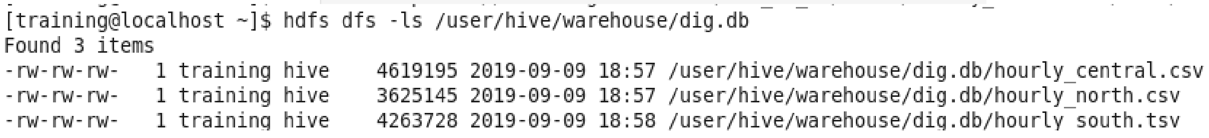
DESTINATION

Name: DIG

PROPERTIES

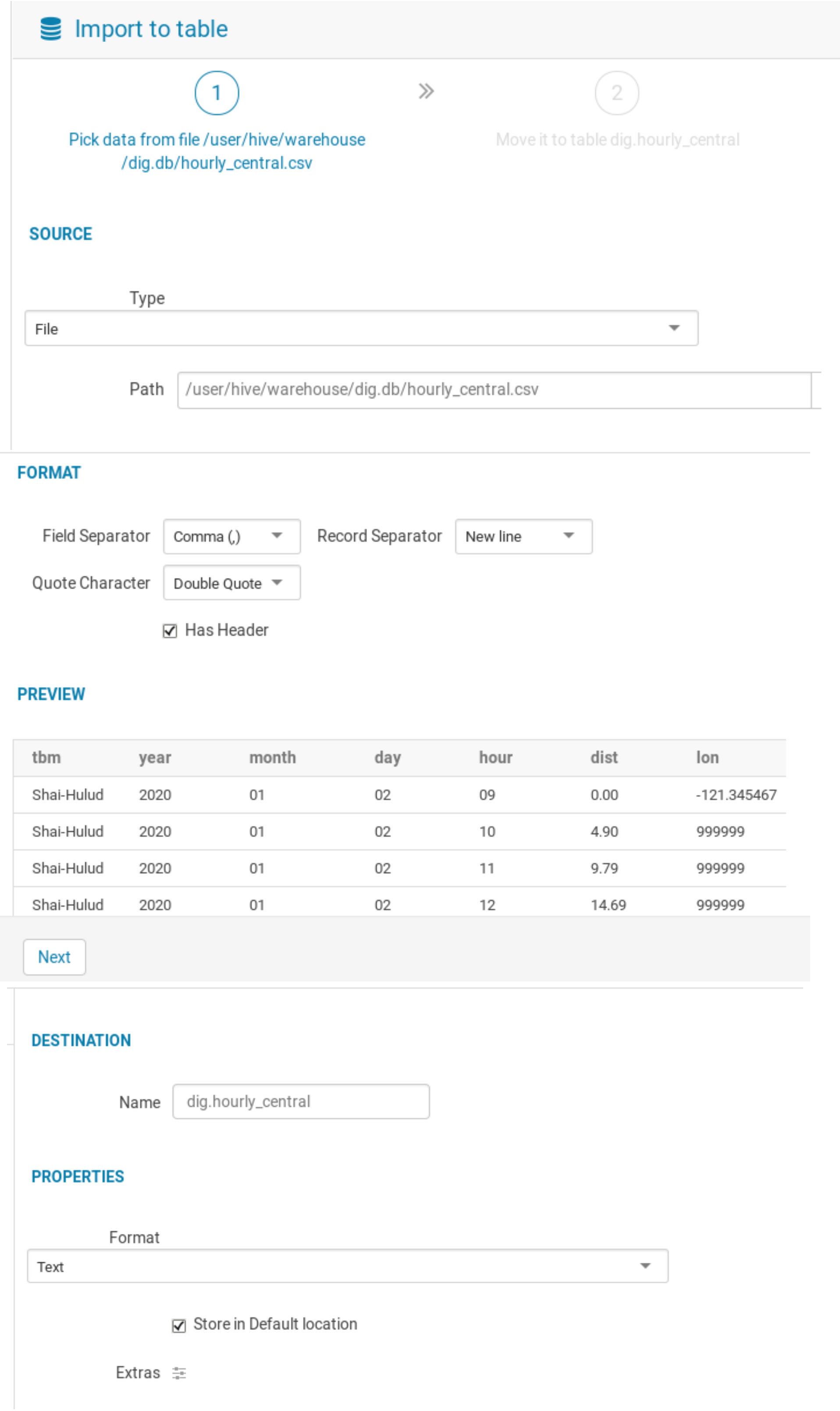
Description: Description

☑ Default location



```
[training@localhost ~]$ hdfs dfs -ls /user/hive/warehouse/dig.db
Found 3 items
-rw-rw-rw-   1 training hive    4619195 2019-09-09 18:57 /user/hive/warehouse/dig.db/hourly_central.csv
-rw-rw-rw-   1 training hive    3625145 2019-09-09 18:57 /user/hive/warehouse/dig.db/hourly_north.csv
-rw-rw-rw-   1 training hive    4263728 2019-09-09 18:58 /user/hive/warehouse/dig.db/hourly_south.tsv
```

- I executed below operation each csv files



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

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

FORMAT

Field Separator: Comma (,)

Record Separator: New line

Quote Character: Double Quote

☑ Has Header

PREVIEW

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

DESTINATION

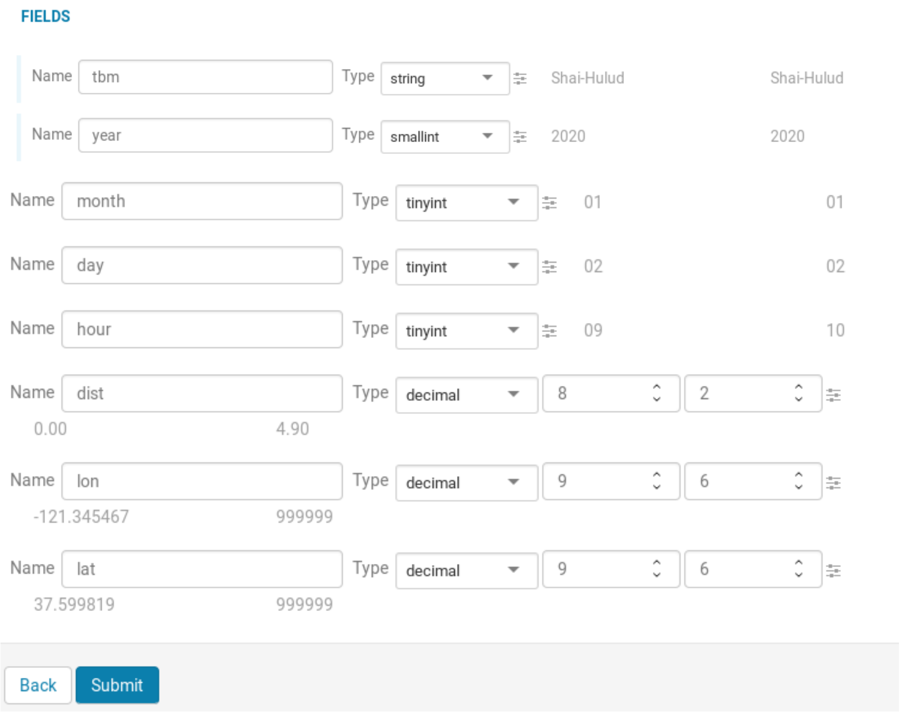
Name: dig.hourly_central

PROPERTIES

Format: Text

☑ Store in Default location

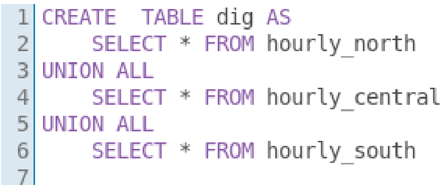
Extras



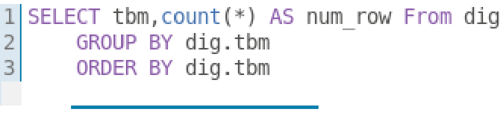
FIELDS

Name	Type				
tbm	string	Shai-Hulud	Shai-Hulud		
year	smallint	2020	2020		
month	tinyint	01	01		
day	tinyint	02	02		
hour	tinyint	09	10		
dist	decimal	8	2	0.00	4.90
lon	decimal	9	6	-121.345467	999999
lat	decimal	9	6	37.599819	999999

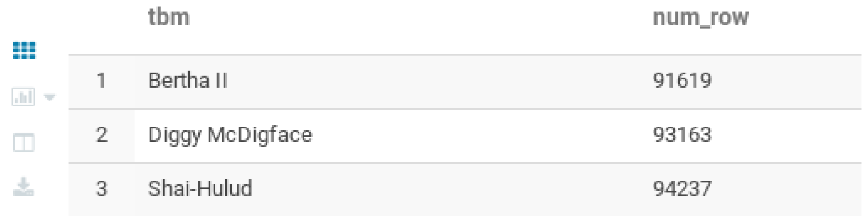
3.



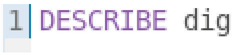
```
CREATE  TABLE dig AS
    SELECT * FROM hourly_north
UNION ALL
    SELECT * FROM hourly_central
UNION ALL
    SELECT * FROM hourly_south
```



```
SELECT tbm,count(*) AS num_row From dig
    GROUP BY dig.tbm
    ORDER BY dig.tbm
```

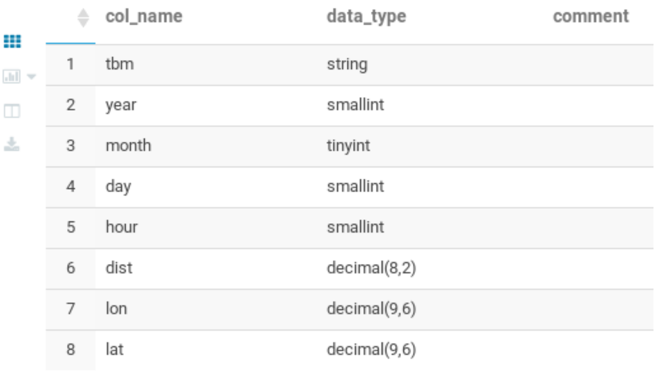


	tbm	num_row
1	Bertha II	91619
2	Diggy McDigface	93163
3	Shai-Hulud	94237



```
DESCRIBE dig
```

*hue optimized data type



	col_name	data_type	comment
1	tbm	string	
2	year	smallint	
3	month	tinyint	
4	day	smallint	
5	hour	smallint	
6	dist	decimal(8,2)	
7	lon	decimal(9,6)	
8	lat	decimal(9,6)	

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