Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Chaitanya DA
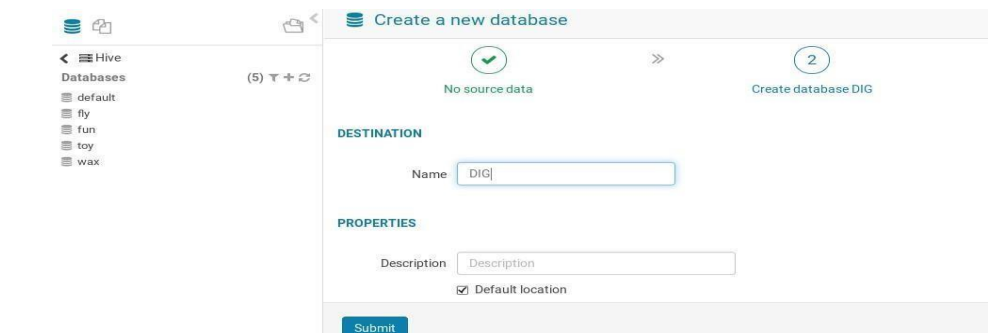**Date:** 18/05/2021

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. 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. 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


```
[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

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

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

Next

DESTINATION

Name | dig.hourly_central

PROPERTIES

Format

Text ▾

☑ Store in Default location

Extras ⇔

3. For Putting it in one table named "**dig.tbm_sf_la**" I ran this query

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

4. 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
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 (8,2)
lat	Decimal (8,2)