

Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Ceyda Çaylak
Date: 09.07.2021

(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

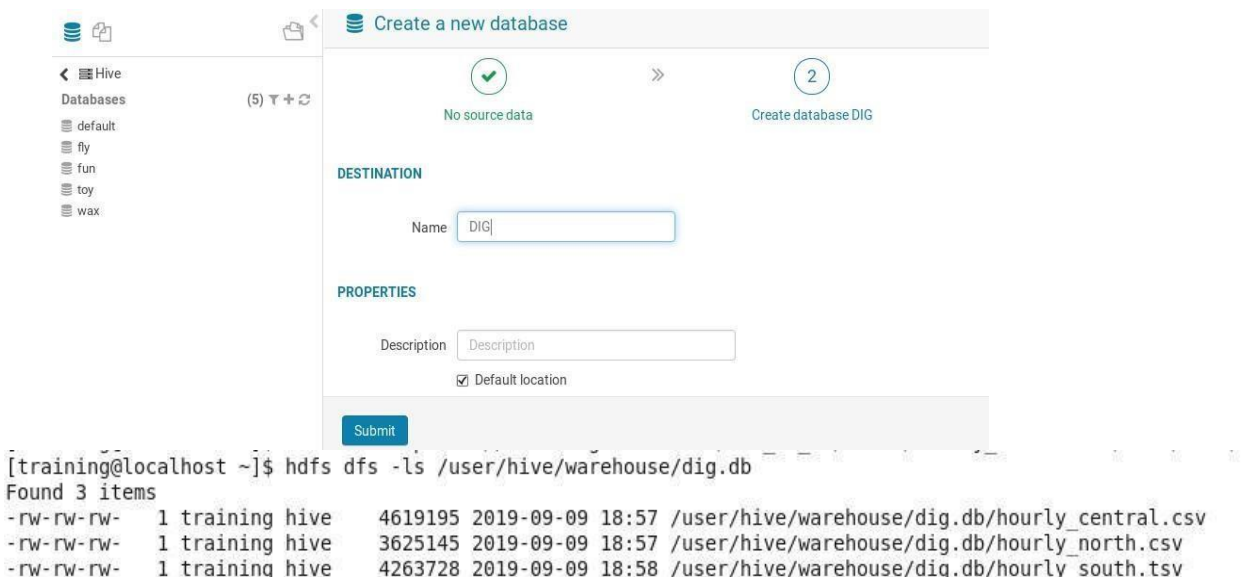
I performed the following steps to complete this task:

1. I need to load the 3 files from s3 to local directory:

```
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. Then import them into 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
```

FIELDS

Name	<input type="text" value="tbm"/>	Type	<input type="text" value="string"/>	<input type="text" value="Shai-Hulud"/>	<input type="text" value="Shai-Hulud"/>
Name	<input type="text" value="year"/>	Type	<input type="text" value="smallint"/>	<input type="text" value="2020"/>	<input type="text" value="2020"/>
Name	<input type="text" value="month"/>	Type	<input type="text" value="tinyint"/>	<input type="text" value="01"/>	<input type="text" value="01"/>
Name	<input type="text" value="day"/>	Type	<input type="text" value="tinyint"/>	<input type="text" value="02"/>	<input type="text" value="02"/>
Name	<input type="text" value="hour"/>	Type	<input type="text" value="tinyint"/>	<input type="text" value="09"/>	<input type="text" value="10"/>
Name	<input type="text" value="dist"/>	Type	<input type="text" value="decimal"/>	<input type="text" value="8"/>	<input type="text" value="2"/>
	<input type="text" value="0.00"/>			<input type="text" value="4.90"/>	
Name	<input type="text" value="lon"/>	Type	<input type="text" value="decimal"/>	<input type="text" value="9"/>	<input type="text" value="6"/>
	<input type="text" value="-121.345467"/>			<input type="text" value="999999"/>	
Name	<input type="text" value="lat"/>	Type	<input type="text" value="decimal"/>	<input type="text" value="9"/>	<input type="text" value="6"/>
	<input type="text" value="37.599819"/>			<input type="text" value="999999"/>	

[Back](#)

[Submit](#)

3.

(#I put all in one table here.)

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

(#Grouped by tbm and counted the number of rows)

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

(#for Hue optimization)

```
DESCRIBE dig
```

```
(#handling missing data)
ALTER TABLE dig.tbm
SET
TBLPROPERTIES("serialization.null.format"="99
999");
```

Results

```
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)

