Peer-Gradedw Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Hoang Trung Nghia
Date: 7/10/2021

(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. Get 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. Import 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
   ```

```
CREATE TABLE dig.tbm_sf_la ROW FORMAT DELIMITED FIELDS TERMINATED BY
‘|’ SOTRED AS csv AS
SELECT *
FROM hourly_central
UNION ALL
SELECT *
FROM hourly_north
UNION ALL
SELECT *
FROM hourly_south TBLPROPERTIES(“serialization.mull.format”=”999999”)
```

(Describe all the steps you performed. Include the commands or SQL statements you ran.)

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)

Notes

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