

Peer-Graded Assignment: Data Management

Course: Managing Big Data in Clusters and Cloud Storage

Name: Bhavin Patel

Date: 11th July 2020

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. Examine and Copy the Data

```
hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/hdfs:///user/hive/warehouse/dig.db/
```

2. Create the Table

```
create table tbm_sf_la_central (tbm string, year smallint, month tinyint, day tinyint, hour  
tinyint, dist decimal (8,2), lon decimal (9,6), lat decimal (9,6))  
row format delimited fields terminated by ","  
tblproperties('skip.header.line.count'='1','serialization.null.format'='')
```

```
create table tbm_sf_la_north (tbm string, `year` smallint, `month` tinyint, `day` tinyint,  
`hour` tinyint, dist decimal (8,2), lon decimal(9,6), lat decimal(9,6)) row format delimited  
fields terminated by ','
```

```
create table tbm_sf_la_south (tbm string, `year` smallint, `month` tinyint, `day` tinyint,  
`hour` tinyint, dist decimal (8,2), lon decimal(9,6), lat decimal(9,6)) row format delimited  
fields terminated by '\t'
```

3. Load the Data into the Table

```
LOAD DATA INPATH
```

```
 '/user/hive/warehouse/dig.db/tbm_sf_la/central/hourly_central.csv' INTO TABLE  
dig.tbm_sf_la_central;
```

```
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la/north/hourly_north.csv'  
INTO TABLE dig.tbm_sf_la_north;
```

```
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la/south/hourly_south.tsv'  
INTO TABLE dig.tbm_sf_la_south;
```

4. Combine all the Data into Single Table

```
create table tbm_sf_la as select * from dig.tbm_sf_la_central union all  
select * from dig.tbm_sf_la_north union all  
select * from dig.tbm_sf_la_south;
```

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	tinyint
hour	tinyint
dist	decimal(8,2)
lon	decimal(9,6)
lat	decima(9,6)