Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Jessica Chen
Date: SEP 26 2020

(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. Download 3 data files from S3 bucket to my local HDFS directory

```
hdfs dfs –get s3a://training-coursera2/tbm_sf_la/betha/hourlydata.csv        /user/training/tbm/beth_hourlydata.csv
hdfs dfs –get s3a://training-coursera2/tbm_sf_la/shaihulud/hourlydata.csv    /user/training/tbm/shaihulud_hourlydata.csv
hdfs dfs –get s3a://training-coursera2/tbm_sf_la/diggy/hourlydata.csv        /user/training/tbm/diggy_hourlydata.csv
```

2. Create external hive table based on those cvs files copied to local HDFS in step1

```
create table dig.tbm_sf_la_text(
tbm                         string,
year                        smallint,
month                       tinyint,
day                         smallint,
hour                        smallint,
dist                        decimal(8,2),
lon                         decimal(9,6),
lat                         decimal(9,6)
)
row format delimited fields terminated by ‘,’
location ‘hdfs://user/training/tbm’
tblproperies (‘skip.header.line.count’=’1’, ‘serialization. null. format'='')
stored as textfile;
```

3. Create internal hive table store as ORC to improve storage/performance

```
Create table dig.tbm_sf_la (
tbm                         string,
year                        smallint,
month                       tinyint,
day                         smallint,
hour                        smallint,
dist                        decimal(8,2),
lon                         decimal(9,6),
lat                         decimal(9,6)
)
Stored as ORC;
```

4. Copy data from external hive table to internal hive table using ITAS

```
Insert into table dig.tbm_sf_la select * from dig.tbm_sf_la_text;
```

(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
list	decimal(8,2)
lon	decimal(9,6)
lat	decimal(9,6)

(Fill in the above tables.)

Notes

(In this section, describe ways that you could further optimize the table. You may also describe other methods you considered or attempted.)
I can also try use HUE file browser to import cvs into hive table. But I choose command line so I can repeat this process by automatic script in case there is any data files added in the S3 bucket.