Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: BatEl Yaish
Date: 12/27/2020

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. I Examined the Data by running the following commands in terminal:
   - hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv | head
   - hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv | head
   - hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv| head

2. I created external tables for each of the files located in S3:

```
/* Central file, handling the header row */
CREATE EXTERNAL TABLE dig.central (
    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT,
    Hour SMALLINT,
    Dist DECIMAL(8,2),
    Lon DECIMAL(10,6),
    Lat DECIMAL(10,6)
)
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
    STORED AS TEXTFILE
    LOCATION ' s3a://training-coursera2/tbm_sf_la/central/'
    TBLPROPERTIES('skip.header.line.count' = '1');
```

```
/* Centeral file, 999999 to represent missing values was handled during create */
CREATE EXTERNAL TABLE dig.north (
    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT,
```

```
    Hour SMALLINT,
    Dist DECIMAL(8,2),
    Lon DECIMAL(10,6),
    Lat DECIMAL(10,6)
)
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
    STORED AS TEXTFILE
    LOCATION ' s3a://training-coursera2/tbm_sf_la/north/';
```

```
/* South file, handling the tab delimiter */
CREATE EXTERNAL TABLE dig.south (
    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT,
    Hour SMALLINT,
    Dist DECIMAL(8,2),
    Lon DECIMAL(10,6),
    Lat DECIMAL(10,6)
)
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t'
    STORED AS TEXTFILE
    LOCATION ' s3a://training-coursera2/tbm_sf_la/south/';
```

3. I used CTA to create the table and simultaneously loaded the data of tables north, central and south.

```
CREATE TABLE dig.tbm_sf_la
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    AS
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.north
UNION
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.south
UNION
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.central;
```

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(10,6)
lat	decimal(10,6)

Notes

We can create the table not STORED AS a text file (the default), but as a PARQUET file like so:

```
CREATE TABLE dig.tbm_sf_la
    STORED AS PARQUET
    AS
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.north
UNION
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.south
UNION
SELECT tbm,year,month,day,hour,dist, lon , lat
 FROM dig.central;
```

And then examine the PARQUET file in terminal:
hdfs dfs -cat /user/hive/warehouse/dig.db/tbm_sf_la/* | head