Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Sohail Makhani
Date: December 26th, 2023

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. In Impala query editor, I created a database **dig** by this command

   CREATE DATABASE dig. By default, it is stored in hive warehouse directory.

2. Using shell, I copied the files from S3 bucket to hive metastore by these commands

   ```
   hdfs dfs -cp s3a://training-coursera/tbm_sf_la/central/hourly_central.csv
   hdfs:///user/hive/warehouse/dig.db
   hdfs dfs -cp s3a://training-coursera/tbm_sf_la/north/hourly_north.csv
   hdfs:///user/hive/warehouse/dig.db
   hdfs dfs -cp s3a://training-coursera/tbm_sf_la/south/hourly_south.csv
   hdfs:///user/hive/warehouse/dig.db
   ```

3. Using Impala query editor, I created 3 tables named **central, north**, and **south** defining the delimeters and skipping the header column in hourly_central.csv files. After viewing these files in Hue files browser, the 3 tables are created as follows:

   For **hourly_central.csv** file

   ```
   create table dig.central(
   tbm string,
   year int,
   month int,
   day int,
   hour int,
   dist decimal(9,2),
   lon decimal(9,6),
   lat decimal(9.6)
   ```

```
)
row format delimited
fields terminated by ‘,’
tblproperties(‘skip.header.line.count’=’1’)
```

For **hourly_north.csv** file

```
create table dig.north(
tbm string,
year int,
month int,
day int,
hour int,
dist decimal(9,2),
lon decimal(9,6),
lat decimal(9.6)
)
row format delimited
fields terminated by ‘,’
```

For **hourly_south.tsv** file

```
create table dig.south(
tbm string,
year int,
month int,
day int,
hour int,
dist decimal(9,2),
lon decimal(9,6),
lat decimal(9.6)
)
row format delimited
fields terminated by ‘\t’
```

4. After creating these tables, I loaded files into these tables using the following command

   ```
   load data inpath ‘/user/hive/warehouse/dig.db/hourly_central.csv’ into table central;
   load data inpath ‘/user/hive/warehouse/dig.db/hourly_north.csv’ into table north;
   load data inpath ‘/user/hive/warehouse/dig.db/hourly_south.tsv’ into table south;
   ```

5. Then I created my final table **tbm_sf_la**

   ```
   create table dig.tbm_sf_la(
   tbm string,
   year int,
   ```

```
month int,
day int,
hour int,
dist decimal(9,2),
lon decimal(9,6),
lat decimal(9.6)
)
```

6. Then I inserted the data in the final table from those 3 tables as follows:

```
insert into table dig.tbm_sf_la
select * from dig.central
union all
Select * from dig.north
union all
select * from dig.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	Int
Month	Int
Day	Int
Hour	Int
Dist	Decimal(9,2)
Lon	Decimal(9,6)
Lat	Decimal(9.6)