Course: Managing Big Data in Clusters and Cloud Storage

Name: Bhagyesh

Date: 19/12/23

Assignment: Create Table for Tunnel Boring Machine Data

Solution:

1. **List Files in S3 Bucket:**

```bash
hdfs dfs -ls s3a://training-coursera2/tbm_sf_la/central
```

2. **Display Head of CSV/TSV Files:**

```bash
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
```

3. **Data Exploration:**

hourly_central.csv

- Header: tbm, year, month, day, hour, dist, lon, lat
- Delimiter: Comma
- Null Values: "999999"

hourly_north.csv

- No Header
- Delimiter: Comma
- Null Values: "\N"

hourly_south.tsv

- No Header
- Delimiter: Tab
- Null Values: "\N"

Column Types:

```sql
tbm VARCHAR(100),
year SMALLINT,
month SMALLINT,
day SMALLINT,
hour SMALLINT,
dist DECIMAL(7, 2),
lon DECIMAL(10, 6),
lat DECIMAL(10, 6)
```

4. **Create Database and External Tables:**

```sql
CREATE DATABASE IF NOT EXISTS dig;
USE dig;

CREATE EXTERNAL TABLE hourly_central (
```

```
    tbm VARCHAR(100),
    year SMALLINT,
    month SMALLINT,
    day SMALLINT,
    hour SMALLINT,
    dist DECIMAL(7, 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');

CREATE EXTERNAL TABLE hourly_north (
    tbm VARCHAR(100),
    year SMALLINT,
    month SMALLINT,
    day SMALLINT,
    hour SMALLINT,
    dist DECIMAL(7, 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';
```

```
CREATE EXTERNAL TABLE hourly_south (
    tbm VARCHAR(100),
    year SMALLINT,
    month SMALLINT,
    day SMALLINT,
    hour SMALLINT,
    dist DECIMAL(7, 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';
```

5. **Create Combined Table:**

```sql
CREATE TABLE tbm_sf_la LIKE hourly_central;

INSERT INTO tbm_sf_la
SELECT * FROM hourly_central
UNION ALL
SELECT * FROM hourly_north
UNION ALL
SELECT * FROM hourly_south;
```

6. **Query and Results:**

```sql
-- Query 1
SELECT tbm, COUNT(*) AS num_rows FROM dig.tbm_sf_la GROUP BY tbm ORDER BY tbm;

-- Query 2
DESCRIBE dig.tbm_sf_la;
```

Results:

tbm	num_rows
Bertha II	91619
Diggy McDigface	93163
Shai-Hulud	94237

tbm	num_rows
Bertha II	91619
Diggy McDigface	93163
Shai-Hulud	94237

```sql
name type
tbm varchar(100)
year smallint
month smallint
day smallint
hour smallint
```

dist decimal(7,2)

lon decimal(10,6)

lat decimal(10,6)

```

7. **Conclusion and Notes:**

- Attempted local file input but found the S3 approach more elegant.

- Consider further optimizing by shortening the `VARCHAR` length for the `tbm` column.
```