Assignment

Construction has begun on a tunnel connecting **San Francisco** and **Los Angeles**. The tunnel will be dug over a period of ten years. It will be dug in three different sections by three tunnel boring machines (TBMs) named **Bertha II**, **Shai-Hulud**, and **Diggy McDigface**.

Solution

I performed the following steps to complete this task:

1. Examine the file directory and run following commends in console to see data
   hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv | head
   hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv | head
   hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv | head

   All of the files have 8 columns with the same data types hourly_central.csv has a header line and uses 999999 for missing values hourly_south.tsv uses tab as separator

2. Download three folders from S3 to local (training folder) using commend below
   hdfs dfs -get s3a://training-coursera2/tbm_sf_la/*

3. Create a Database named dig and create tables by uploading downloaded files via HUE

4. Import data to tables and specify relevant column names, data types, field separators in HUE

5. Table details are listed below:

   a. **dig.hourly_central**: tbm STRING, year SMALLINT, month TINYINT, day TINYINT, hour TINYINT, dist DECIMAL(8,2), lon DOUBLE, lat DOUBLE
   b. **dig.hourly_north**: tbm STRING, year SMALLINT, month TINYINT, day TINYINT, hour TINYINT, dist DECIMAL(8,2), lon DOUBLE, lat DOUBLE
   c. **dig.hourly_south**: tbm STRING, year SMALLINT, month TINYINT, day TINYINT, hour TINYINT, dist DECIMAL(8,2), lon DOUBLE, lat DOUBLE

6. Union all three tables using a SQL query below:

```
CREATE TABLE tbm_sf_la AS
        SELECT * FROM hourly_central
UNION ALL
        SELECT * FROM hourly_north
UNION ALL
        SELECT * FROM hourly_south
```

Result

After performing the steps 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)
Ion	DOUBLE
lat	DOUBLE