Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Abrar Hyder Mohammed
Date: 19/09/2022

(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.

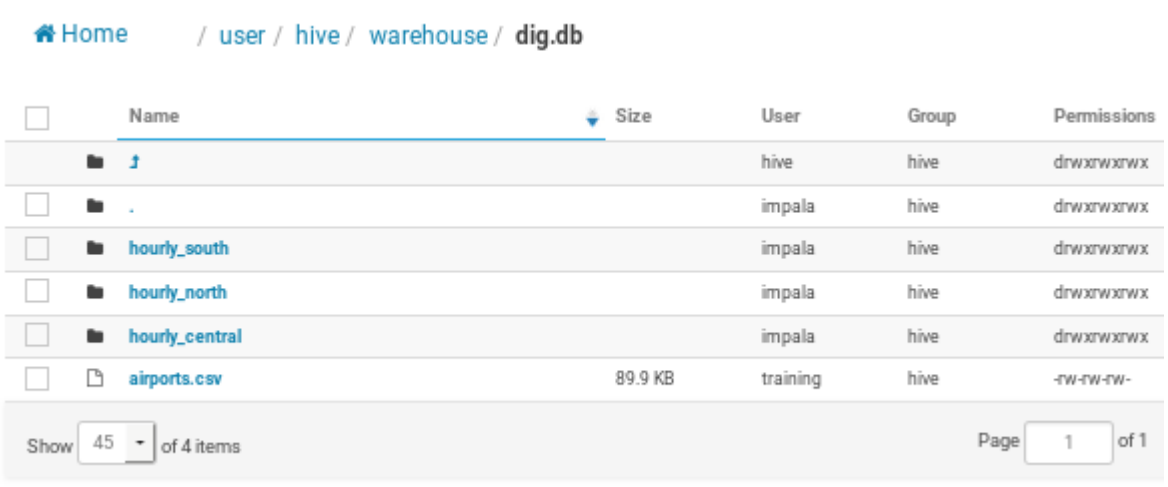
Solution

I performed the following steps to complete this task:

1. I download all the three files from tbm_sf_la directory located in s3 to local directory by using following commands

    - hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_central.csv .
    - hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv .
    - hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_south.csv .

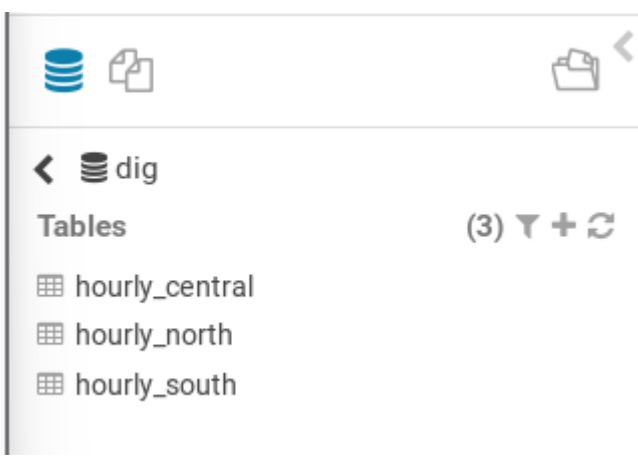
2. Then with the help of hue file browser I copied these three files from local directory to dig database .



Using the hdfs dfs -cat command printed the content to the screen to find the structure(contains header or not ,datatype) of each fie.

```
training@localhost:~
File  Edit  View  Search  Terminal  Help
Shai-Hulud,2030,10,02,18,370742.86,999999,999999
Shai-Hulud,2030,10,02,19,370749.14,999999,999999
Shai-Hulud,2030,10,02,20,370755.43,999999,999999
Shai-Hulud,2030,10,02,21,370761.71,999999,999999
Shai-Hulud,2030,10,02,22,370768.00,-118.934074,34.950340
[training@localhost ~]$ hdfs dfs -cat '/user/hive/warehouse/dig.db/hourly_
tbm,year,month,day,hour,dist,lon,lat
Shai-Hulud,2020,01,02,09,0.00,-121.345467,37.599819
Shai-Hulud,2020,01,02,10,4.90,999999,999999
Shai-Hulud,2020,01,02,11,9.79,999999,999999
Shai-Hulud,2020,01,02,12,14.69,999999,999999
Shai-Hulud,2020,01,02,13,19.59,999999,999999
Shai-Hulud,2020,01,02,14,24.48,999999,999999
Shai-Hulud,2020,01,02,15,29.38,999999,999999
Shai-Hulud,2020,01,02,16,34.28,999999,999999
Shai-Hulud,2020,01,02,17,39.17,999999,999999
cat: Unable to write to output stream.
[training@localhost ~]$ hdfs dfs -cat '/user/hive/warehouse/dig.db/hourly_
Bertha II,2020,01,02,09,0.00,-121.345947,37.600201
Bertha II,2020,01,02,10,5.00,\N,\N
Bertha II,2020,01,02,11,10.00,\N,\N
Bertha II,2020,01,02,12,15.00,\N,\N
Bertha II,2020,01,02,13,20.00,-121.346107,37.600319
Bertha II,2020,01,02,14,25.33,\N,\N
```

3. Finally I created three tables each for each file using the left assist panel and then queried it using the following SQL commands to get the results.



1.61s dig text ?

```
SELECT tbm, COUNT(*) AS num_rows FROM new
GROUP BY new.tbm
ORDER BY new.tbm;
```

Query History Saved Queries

Results (3)

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

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	Bigint
Month	Bigint
Day	Bigint
Hour	Bigint

Dist	Decimal
Lon	Decimal
lat	Decimal