Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Mireya Lyons
Date: 26 August 2022

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. **To begin, I discovered the names for the three subdirectories stored in S3 via terminal using:**

   [training@localhost ~]$ hdfs dfs -ls s3a://training-coursera2/tbm_sf_la/

   I then got the resulting files from S3 to the Local Directory using:

   [training@localhost ~]$ hdfs dfs -get s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv .

   [training@localhost ~]$ hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv .

   [training@localhost ~]$ hdfs dfs -get s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv .

2. **After obtaining the files, I imported the newly obtained files from Local Directory into the Hue Browser**

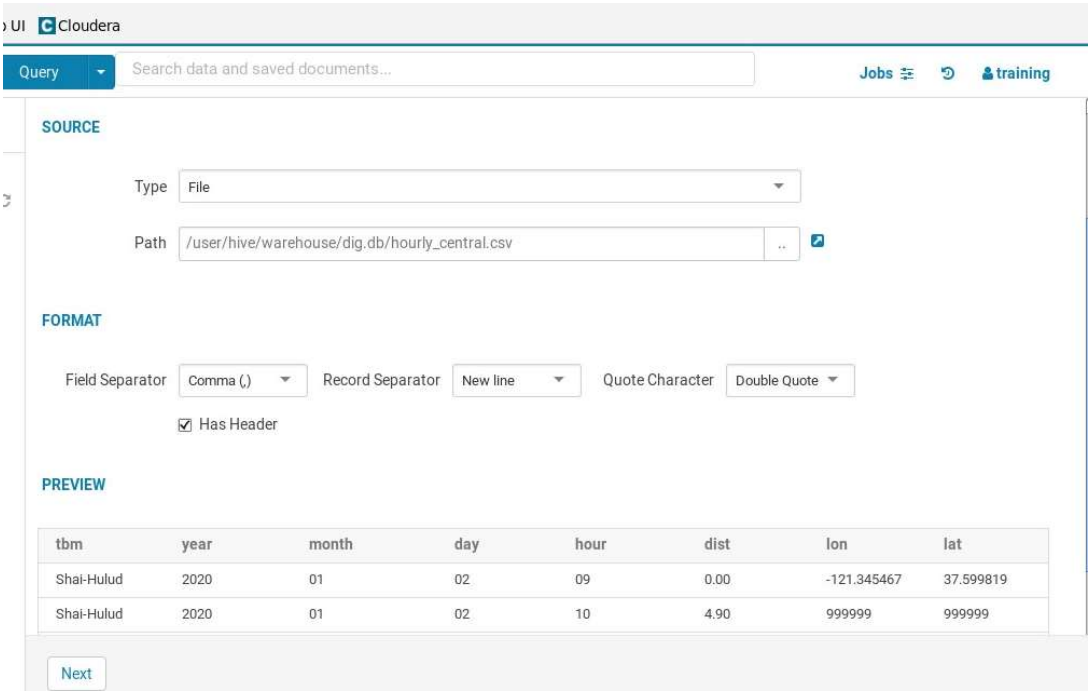
   [training@localhost ~]$ hdfs dfs -mkdir /user/hive/warehouse/dig.db

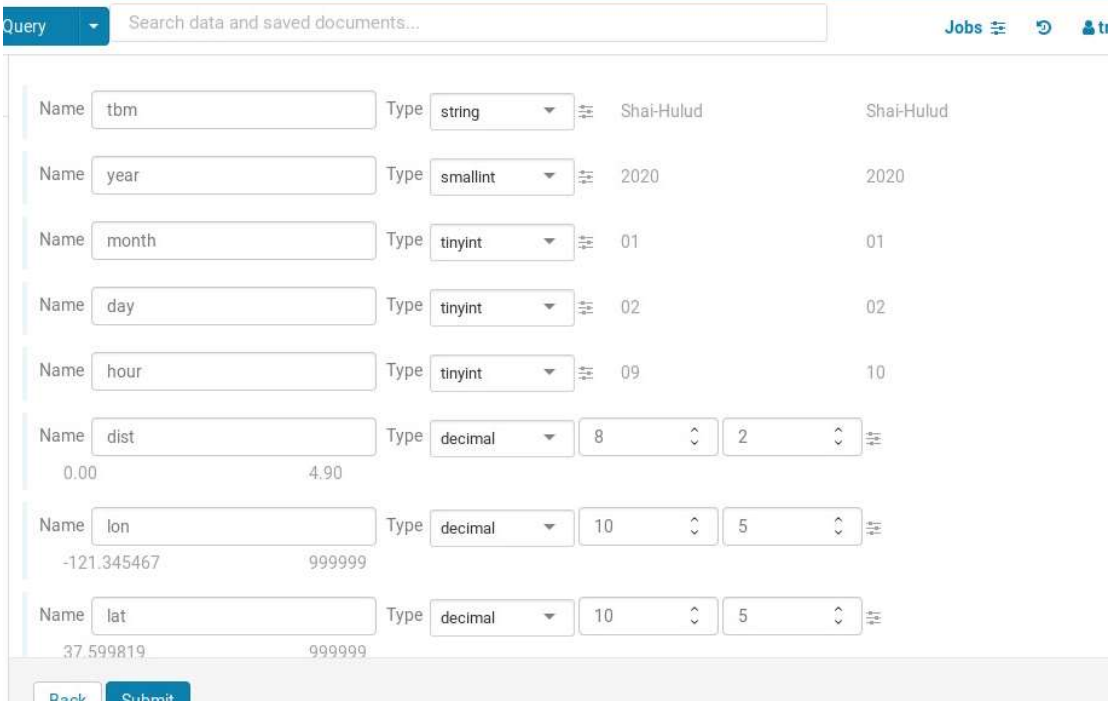
   [training@localhost ~]$ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv /user/hive/warehouse/dig.db

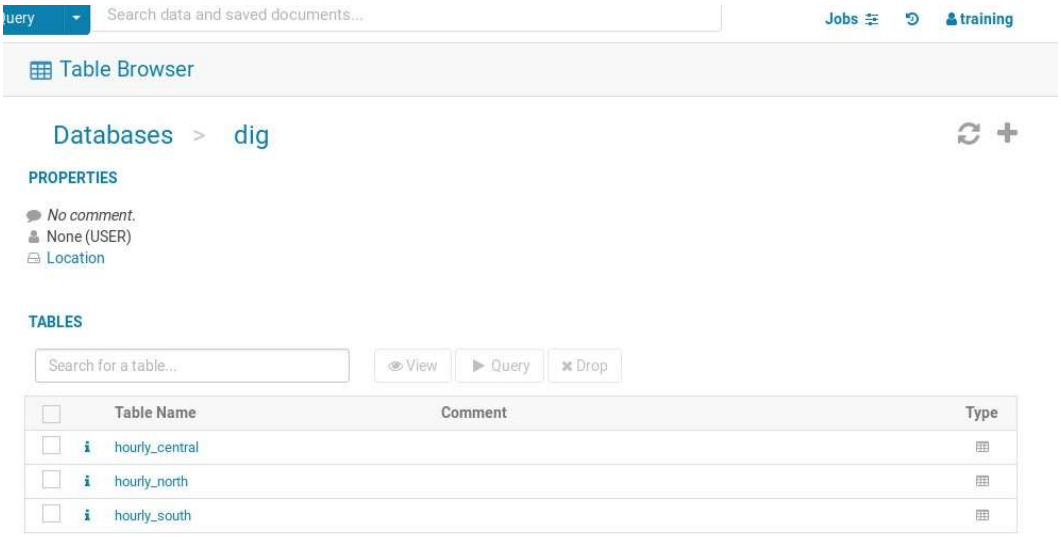
   [training@localhost ~]$ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv /user/hive/warehouse/dig.db

   [training@localhost ~]$ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv /user/hive/warehouse/dig.db

   Using the Hue Browser I created the database, Dig, and imported the files into the database using the same field names and types


UI
Cloudera
Query
Search data and saved documents...
Jobs
training
SOURCE
Type
File
Path
/user/hive/warehouse/dig.db/hourly_central.csv
FORMAT
Field Separator
Comma (,)
Record Separator
New line
Quote Character
Double Quote
Has Header
PREVIEW
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
Next


Query
Search data and saved documents...
Jobs
Name tbm Type string Shai-Hulud Shai-Hulud
Name year Type smallint 2020 2020
Name month Type tinyint 01 01
Name day Type tinyint 02 02
Name hour Type tinyint 09 10
Name dist Type decimal 8 2
0.00 4.90
Name lon Type decimal 10 5
-121.345467 999999
Name lat Type decimal 10 5
37.599819 999999
Back
Submit



3. **I combined the data from the three tables into a new table named tbm_sf_la in the dig database**

   I ran this query to combine the data and create the new table in the in Hue Browser:

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

   To work around the null conversion error in Impala I used this query:

```
ALTER TABLE tbm_sf_la SET
TBLPROPERTIES("serialization.null.format"="99999")
```

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	tinyint
Hour	tinyint
Dist	decimal (8,2)
Lon	decimal(10,5)
Lat	decimal(10,5)

Notes

If I were to do this process again, I would adjust the decimal data types for the fields slightly.