**Peer-Graded Assignment:** Data Management
**Course:** Managing Big Data in clusters and Cloud Storage
Name: Monika Shailesh
Date: 11/07/2021

# ASSIGNMENT

For this assignment, you will create a table with data describing an underground tunneling project.

If you took the second course in this specialization (*Analyzing Big Data with SQL*), recall that the peer-reviewed assignment asked you to analyze flights data to select a profitable route for an underground high-speed rail tunnel. Based on your analysis and on other factors, 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**.

Each of these TBMs will generate a large volume of data as it operates. Each TBM will generate the data slightly differently. Simulated versions of the three TBM-generated datasets are provided. You must create a table on the VM and load these datasets into it. Then you must create and upload a document describing the steps you performed to complete this task.

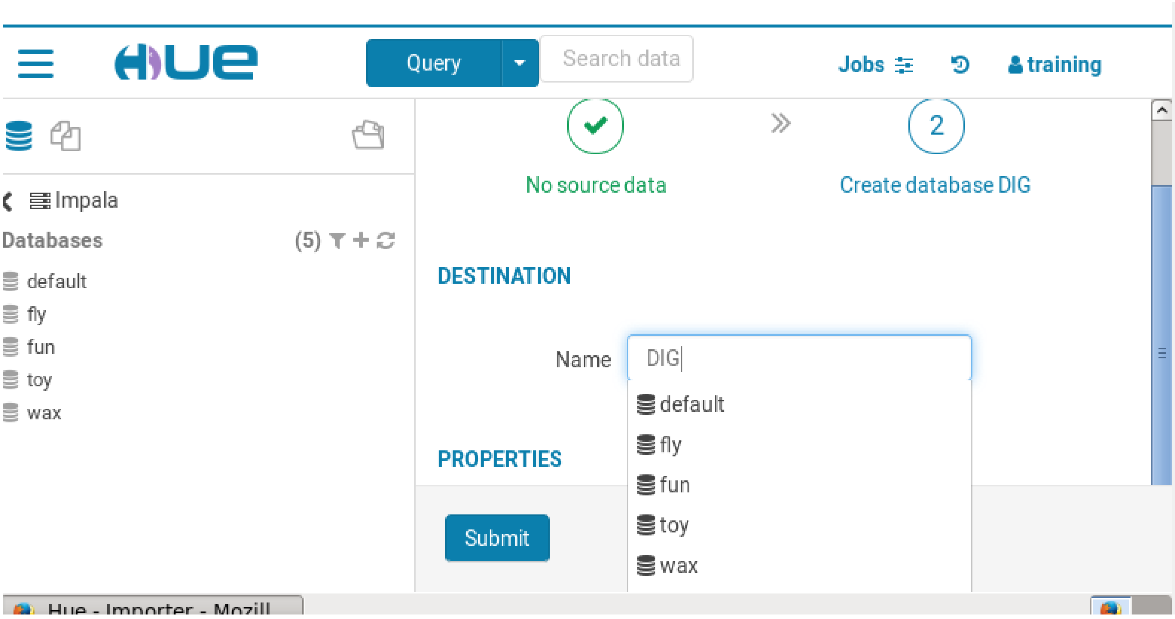
# SOLUTION

I performed the following steps to complete this task:

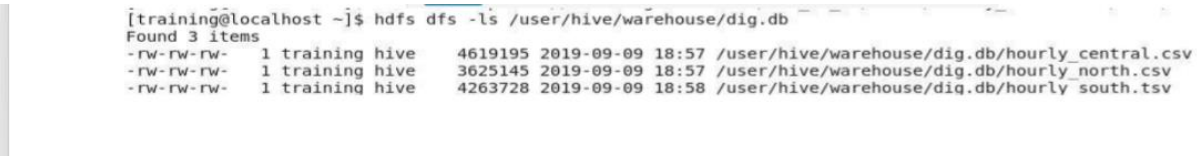
1.I mentioned below three files from s3 to local directory via terminal

- ‘’hdfs dfs -get s3a://training-coursera2/tbm_sf_la/south/hourly_south.csv.’’
- ‘’hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_south.csv.’’
- ’hdfs dfs -get s3a://training-coursera2/tbm_sf_la/central/hourly_south.csv.’’

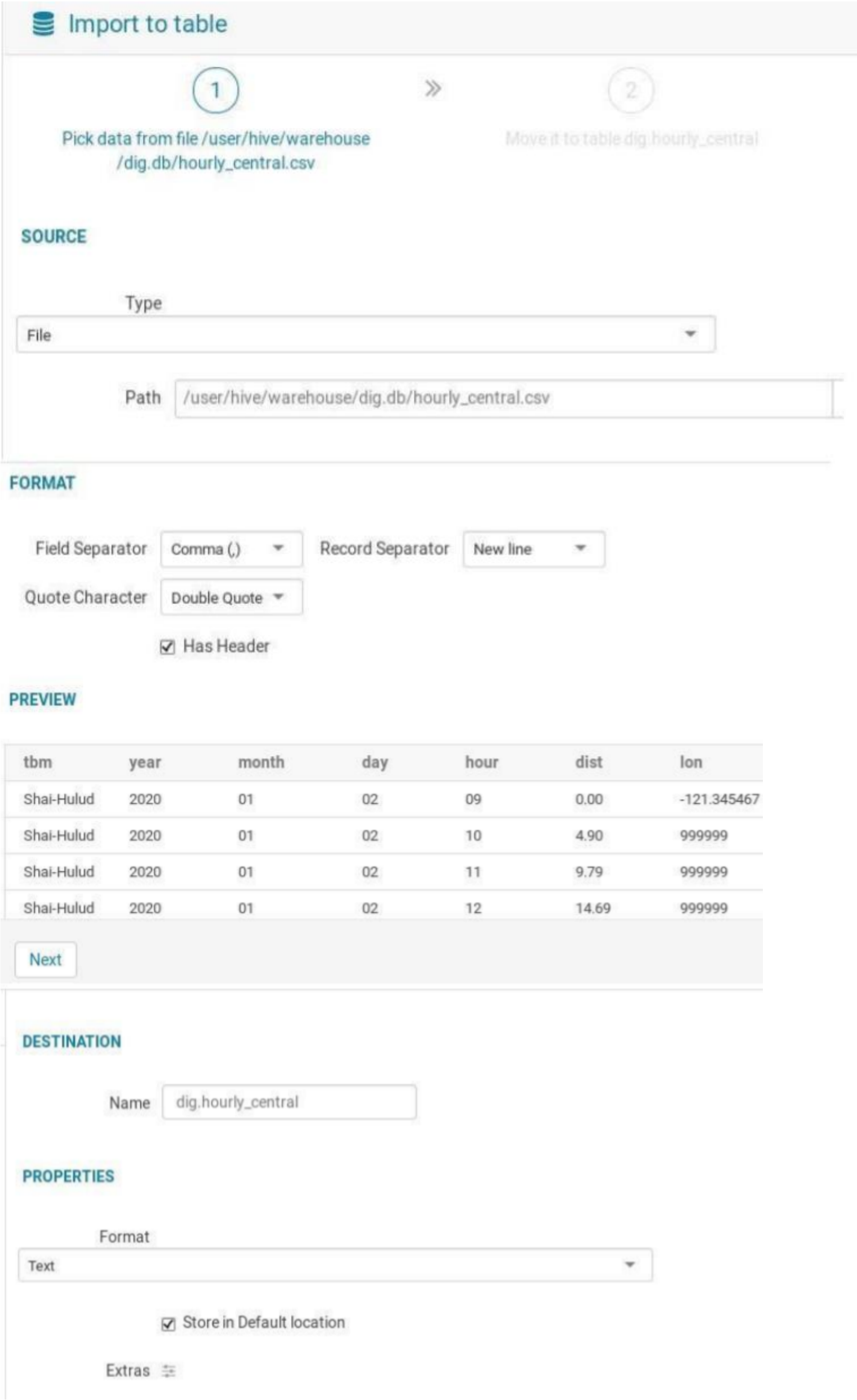
2.



Name DIG



```
[training@localhost ~]$ hdfs dfs -ls /user/hive/warehouse/dig.db
Found 3 items
-rw-rw-rw-   1 training hive    4619195 2019-09-09 18:57 /user/hive/warehouse/dig.db/hourly_central.csv
-rw-rw-rw-   1 training hive    3625145 2019-09-09 18:57 /user/hive/warehouse/dig.db/hourly_north.csv
-rw-rw-rw-   1 training hive    4263728 2019-09-09 18:58 /user/hive/warehouse/dig.db/hourly south.tsv
```



Import to table

1 Pick data from file /user/hive/warehouse/dig.db/hourly_central.csv

2 Move it to table dig.hourly_central

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 |
|---|---|---|---|---|---|---|
| Shai-Hulud | 2020 | 01 | 02 | 09 | 0.00 | -121.345467 |
| Shai-Hulud | 2020 | 01 | 02 | 10 | 4.90 | 999999 |
| Shai-Hulud | 2020 | 01 | 02 | 11 | 9.79 | 999999 |
| Shai-Hulud | 2020 | 01 | 02 | 12 | 14.69 | 999999 |

DESTINATION

Name: dig.hourly_central

PROPERTIES

Format: Text

☑ Store in Default location

3. For putting it on one table named “dig.tbm_sf_la” I ran this query

```
CREATE TABLE dig.tbm_sf_la AS
 SELECT* FROM hourly_central
 UNION ALL
```

```
SELECT*FROM hourly_north
UNION ALL
SELECT * FROM hourly_south
```

4. ALTER TABLE dig.tbm_sf_la SET TBLPROPERTIES (“serialization.null.format”= “99999”);

## Result

After performing the steps described mentioned above, I querried the following code and then got the following result set:

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

| Tbm | num_rows |
|---|---|
| Bertha ll | 91619 |
| Diggy McDigface | 93163 |
| Shai-Hulud | 94237 |

```
DESCRIBE dig.tbm_sf_la;
```

| name | type |
|---|---|
| tbm | string |
| Year | smallint |
| Month | tinyint |
| Day | smallint |
| Hour | ssmallint |
| dist | Decimal (8,2) |
| lon | Decimal (8,2) |
| lat | Decimal (8,2) |