Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: RAMASRAVANI TALARI
Date: 8/1/2020

(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:

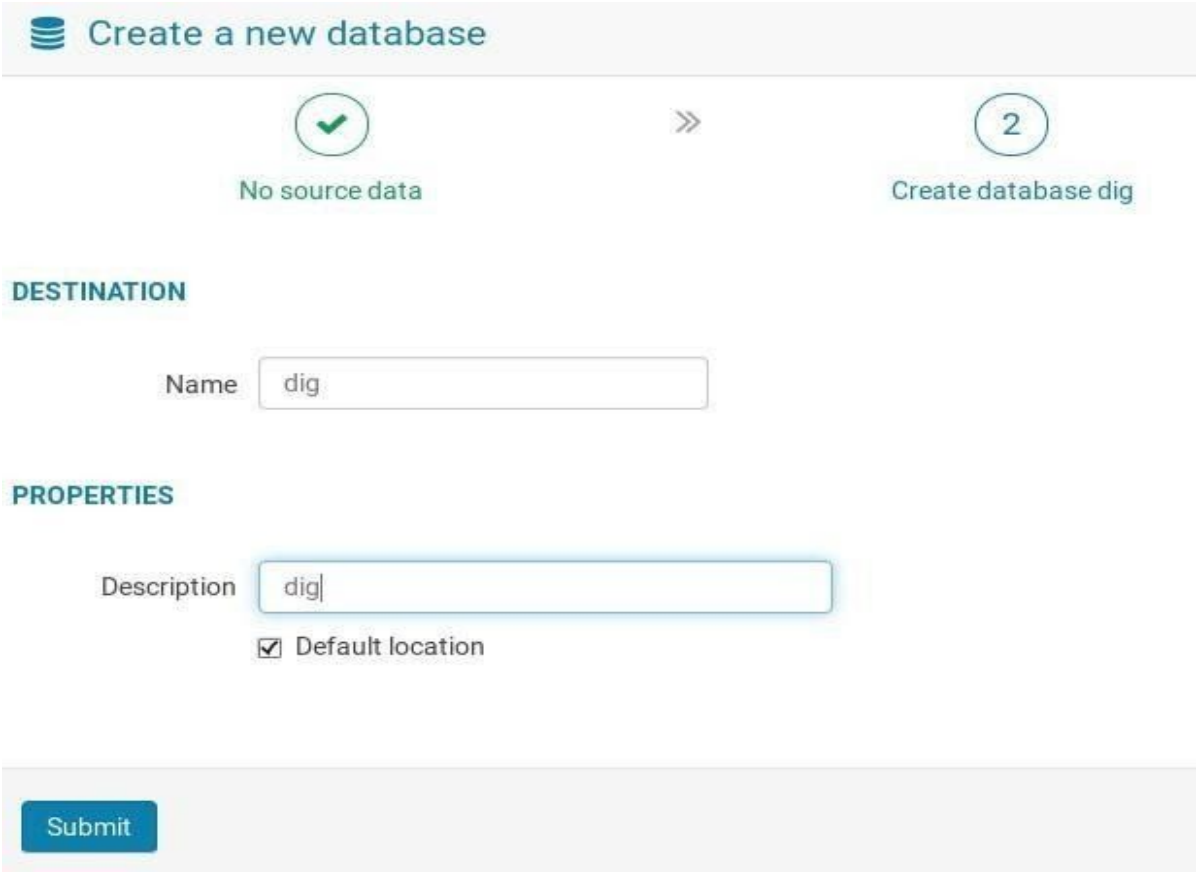
1. **I Have copied the below three files from s3 to hdfs via terminal to the training directory in hdfs**

   hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv /user/training
   hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv /user/training
   hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv /user/training

```
[training@localhost ~]$ hdfs dfs -ls /user/training
Found 7 items
drwxrwxrwx   - training supergroup          0 2020-07-27 00:37 /user/training/2015_11_18
drwxrwxrwx   - training supergroup          0 2020-07-27 00:37 /user/training/2015_11_19
drwxrwxrwx   - training supergroup          0 2020-07-27 00:37 /user/training/2015_11_20
drwxrwxrwx   - training supergroup          0 2020-07-27 00:37 /user/training/2015_11_21
-rw-rw-rw-   1 training supergroup    4619195 2020-07-27 06:17 /user/training/hourly_central.csv
-rw-rw-rw-   1 training supergroup    3625145 2020-07-27 06:15 /user/training/hourly_north.csv
-rw-rw-rw-   1 training supergroup    4263728 2020-07-27 03:48 /user/training/hourly_south.tsv
```

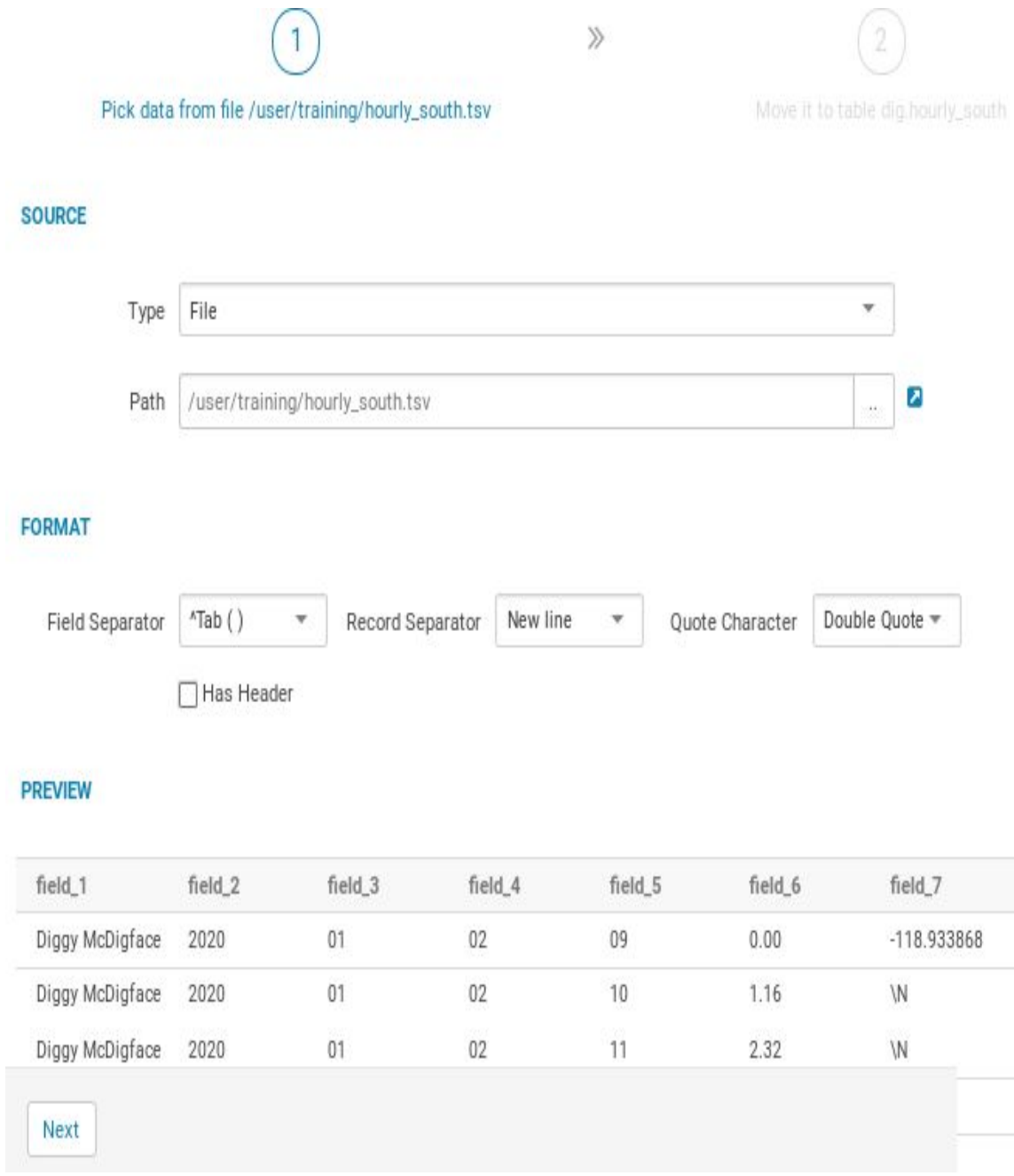
2. **Created a database and tables and loaded data into the tables.**

2.1. Create database name "**dig**"



2.2. Create 3 different tables for each file imported from S3 storage, having the same number of columns, data types and same format for handling missing values.

2.2.1. Create table '**hourly_south'** using Hue table creator with Field Separator in **dig** database



FIELDS

Name	Type			
tbm	string		Diggy McDigface	Diggy McDigface
year	smallint		2020	2020
month	tinyint		01	01
day	tinyint		02	02
hour	tinyint		09	10
dist	decimal	8	2	
0.00	1.16			
lon	decimal	9	6	
-118.933868	\N			
lat	decimal	9	6	
34.949688	\N			

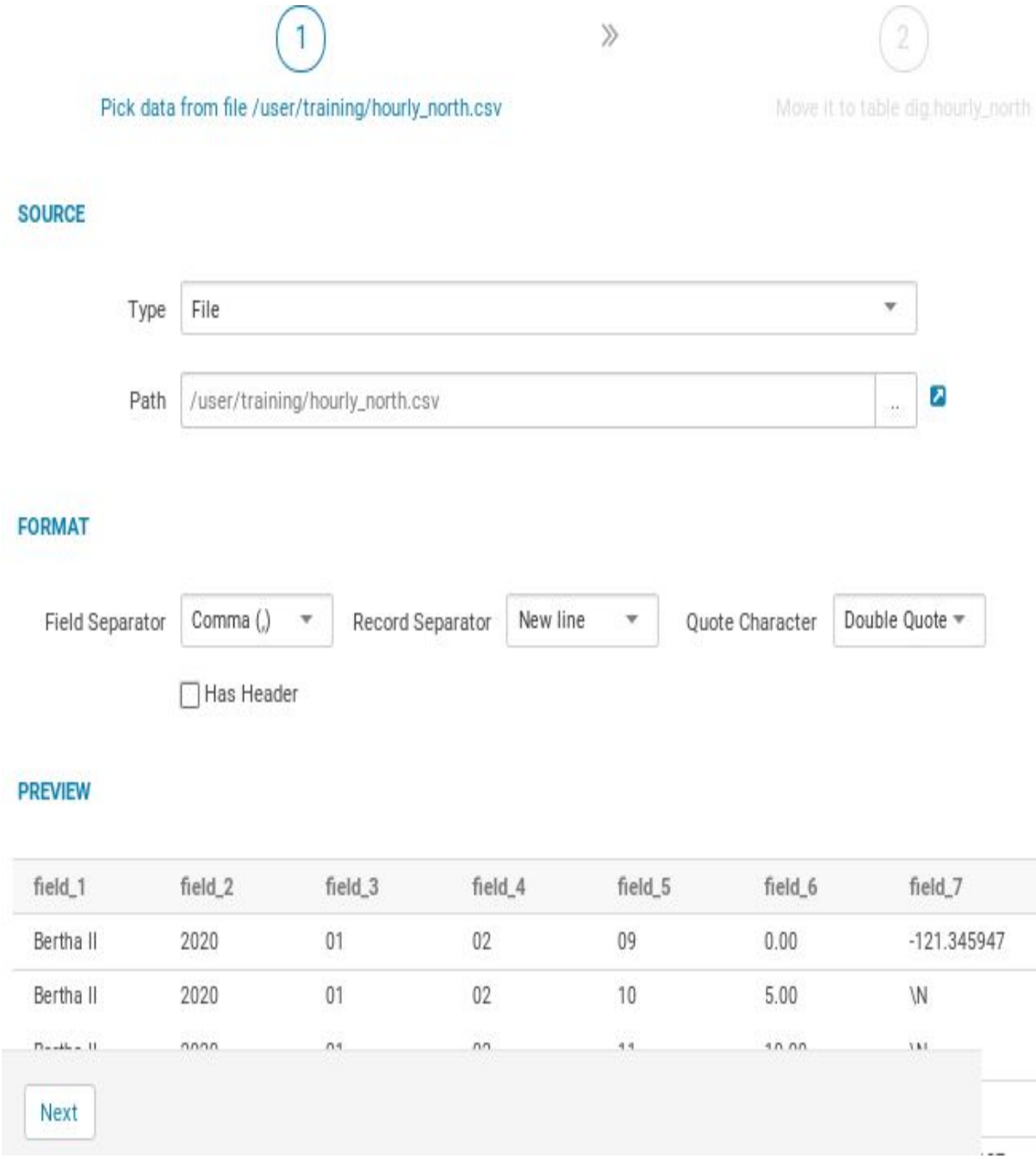
Databases > dig > hourly_south

No description available

Overview | Columns (8) | Sample | Details

	tbm	year	month	day	hour	dist	lon	lat
1	Diggy McDigface	2020	1	2	9	0.00	-118.933868	34.949688
2	Diggy McDigface	2020	1	2	10	1.16	NULL	NULL
3	Diggy McDigface	2020	1	2	11	2.32	NULL	NULL
4	Diggy McDigface	2020	1	2	12	3.49	NULL	NULL
5	Diggy McDigface	2020	1	2	13	4.65	NULL	NULL
6	Diggy McDigface	2020	1	2	14	5.81	NULL	NULL

2.2.2. Create table '**hourly_north'** using Hue table creator in **dig** database



FIELDS

Name	Type			
tbm	string		Bertha II	Bertha II
year	smallint		2020	2020
month	tinyint		01	01
day	tinyint		02	02
hour	tinyint		09	10
dist	decimal	8	2	0.00 5.00
lon	decimal	9	6	-121.345947 \N
lat	decimal	9	6	37.600201 \N

Table Browser

Databases > dig > hourly_north

No description available

Overview | Columns (8) | Sample | Details

	tbm	year	month	day	hour	dist	lon	lat
1	Bertha II	2020	1	2	9	0.00	-121.345947	37.600201
2	Bertha II	2020	1	2	10	5.00	NULL	NULL
3	Bertha II	2020	1	2	11	10.00	NULL	NULL
4	Bertha II	2020	1	2	12	15.00	NULL	NULL
5	Bertha II	2020	1	2	13	20.00	-121.346107	37.600319
6	Bertha II	2020	1	2	14	25.33	NULL	NULL
7	Bertha II	2020	1	2	15	30.67	NULL	NULL

2.2.3. Create table '**hourly_central**' using Hive Query in **dig** database

```
CREATE TABLE hourly_central (
   tbm  string,
   year  smallint,
   month  tinyint,
   day  tinyint,
   hour  tinyint,
   dist  DECIMAL(8,2),
   lon DECIMAL(9,6),
   lat DECIMAL(9,6)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1','serialization.null.format' = '999999');
```

6.20s dig text ?

```
LOAD DATA INPATH '/user/training/hourly_central.csv' INTO TABLE hourly_central;
```

Table Browser

Databases > dig > hourly_central

No description available

Overview Columns (8) Sample Details

	tbm	year	month	day	hour	dist	lon	lat
1	Shai-Hulud	2020	1	2	9	0.00	-121.345467	37.599819
2	Shai-Hulud	2020	1	2	10	4.90	NULL	NULL
3	Shai-Hulud	2020	1	2	11	9.79	NULL	NULL
4	Shai-Hulud	2020	1	2	12	14.69	NULL	NULL
5	Shai-Hulud	2020	1	2	13	19.59	NULL	NULL
6	Shai-Hulud	2020	1	2	14	24.48	NULL	NULL
7	Shai-Hulud	2020	1	2	15	29.38	NULL	NULL
8	Shai-Hulud	2020	1	2	16	34.28	NULL	NULL

3. **Union all tables created above and Create new table tbm_sf_la. Using below query**

27m, 36s dig text ?

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

```

	name	type
1	tbm	string
2	year	smallint
3	month	tinyint
4	day	tinyint
5	hour	tinyint
6	dist	decimal(8,2)
7	lon	decimal(9,6)
8	lat	decimal(9,6)

5.57s dig text ?

```
1 SELECT tbm, COUNT(*) AS num_rows
2 FROM dig.tbm_sf_la
3 GROUP BY tbm
4 ORDER BY tbm
```

Query History Saved Queries Results (3)

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

(Describe all the steps you performed. Include the commands or SQL statements you ran.)

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 (9,6)
lat	decimal (9,6)

Notes

(In this section, describe ways that you could further optimize the table. You may also describe other methods you considered or attempted.)