Peer-Graded Assignment: Data Management
Course: Managing Big Data in Clusters and Cloud Storage
Name: Süleyman Baver Özkeskin
Date: 25/03/2022

Assignment

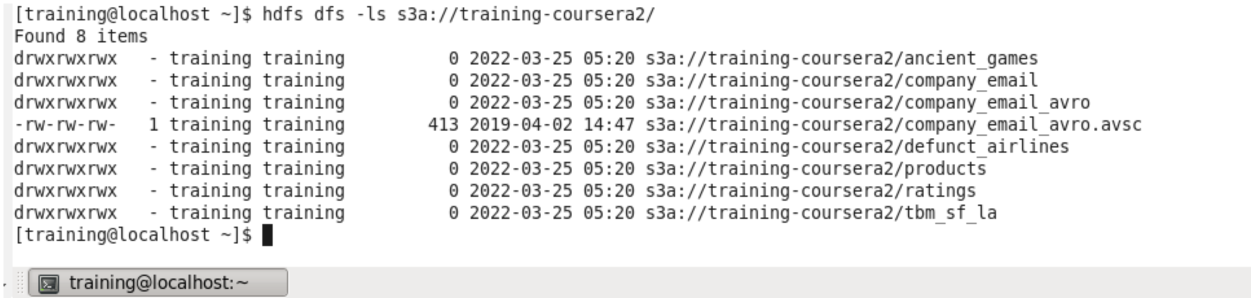
Create a table named **tbm_sf_la** in the database named **dig** to store the data from three tunnel boring machines (TBMs), which is currently stored in S3 in three separate subdirectories under a directory named **tbm_sf_la** in the bucket named **training-coursera2**. In this document, describe the steps taken to complete this task.

Solution

I performed the following steps to complete this task:

1. Following command gives the list the content of S3 bucket:

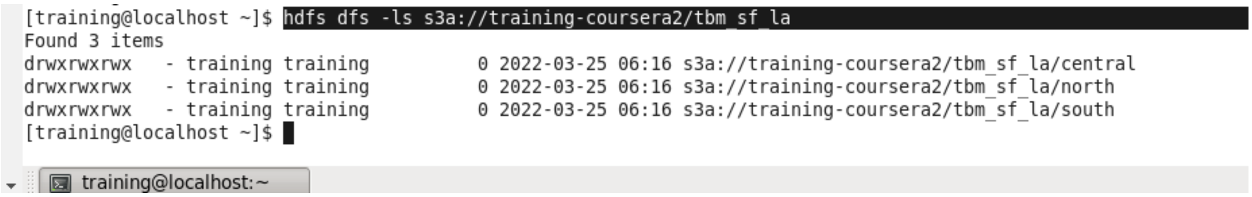
   hdfs dfs -ls s3a://training-coursera2/



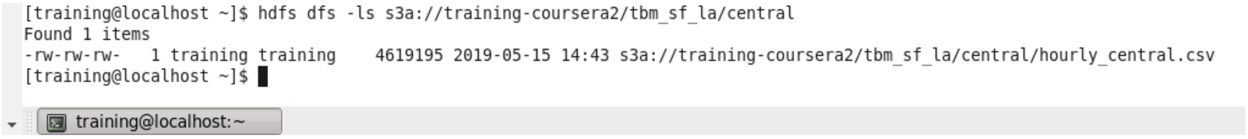
```
[training@localhost ~]$ hdfs dfs -ls s3a://training-coursera2/
Found 8 items
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/ancient_games
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/company_email
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/company_email_avro
-rw-rw-rw-   1 training training        413 2019-04-02 14:47 s3a://training-coursera2/company_email_avro.avsc
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/defunct_airlines
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/products
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/ratings
drwxrwxrwx   - training training          0 2022-03-25 05:20 s3a://training-coursera2/tbm_sf_la
[training@localhost ~]$ █
```

2. After ensuring the directory ,   I got the list of desired files and with the following commands

   hdfs dfs -ls s3a://training-coursera2/tbm_sf_la



```
[training@localhost ~]$ hdfs dfs -ls s3a://training-coursera2/tbm_sf_la
Found 3 items
drwxrwxrwx   - training training          0 2022-03-25 06:16 s3a://training-coursera2/tbm_sf_la/central
drwxrwxrwx   - training training          0 2022-03-25 06:16 s3a://training-coursera2/tbm_sf_la/north
drwxrwxrwx   - training training          0 2022-03-25 06:16 s3a://training-coursera2/tbm_sf_la/south
[training@localhost ~]$ █
```



```
[training@localhost ~]$ hdfs dfs -ls s3a://training-coursera2/tbm_sf_la/central
Found 1 items
-rw-rw-rw-   1 training training    4619195 2019-05-15 14:43 s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv
[training@localhost ~]$ 
```

3. After steps above , we have all the necessary directories with appropriate file types we can proceed to next step which is copying the files from S3 bucket to the local file system. For that we will use the “ hdfs dfs -get “ command as shown below.

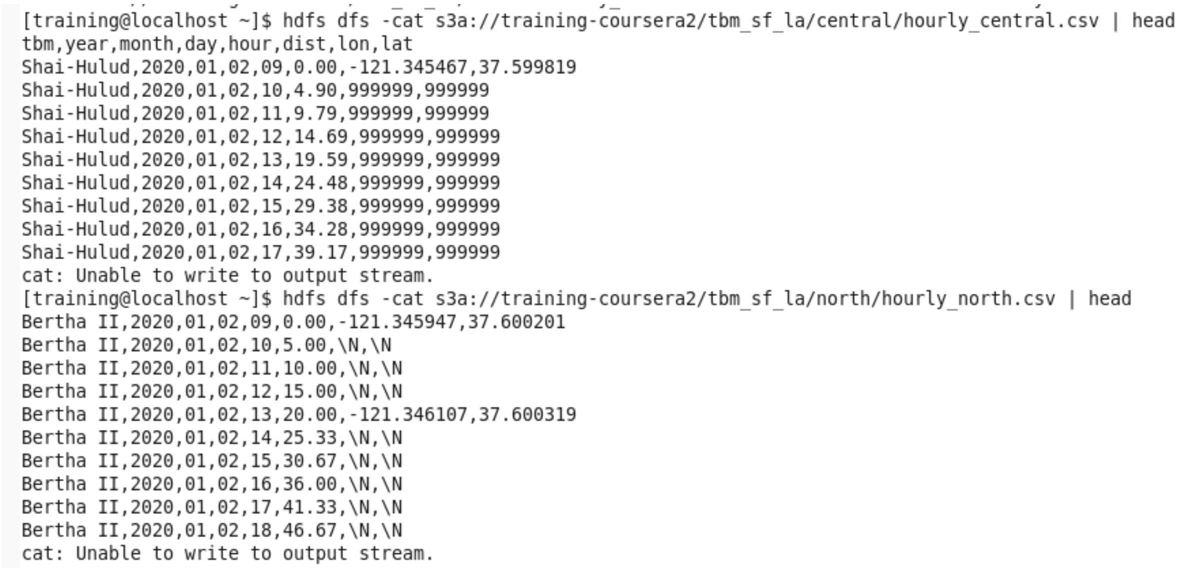
   hdfs dfs -get s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv .

   hdfs dfs -get s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv .

   hdfs dfs -get s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv .

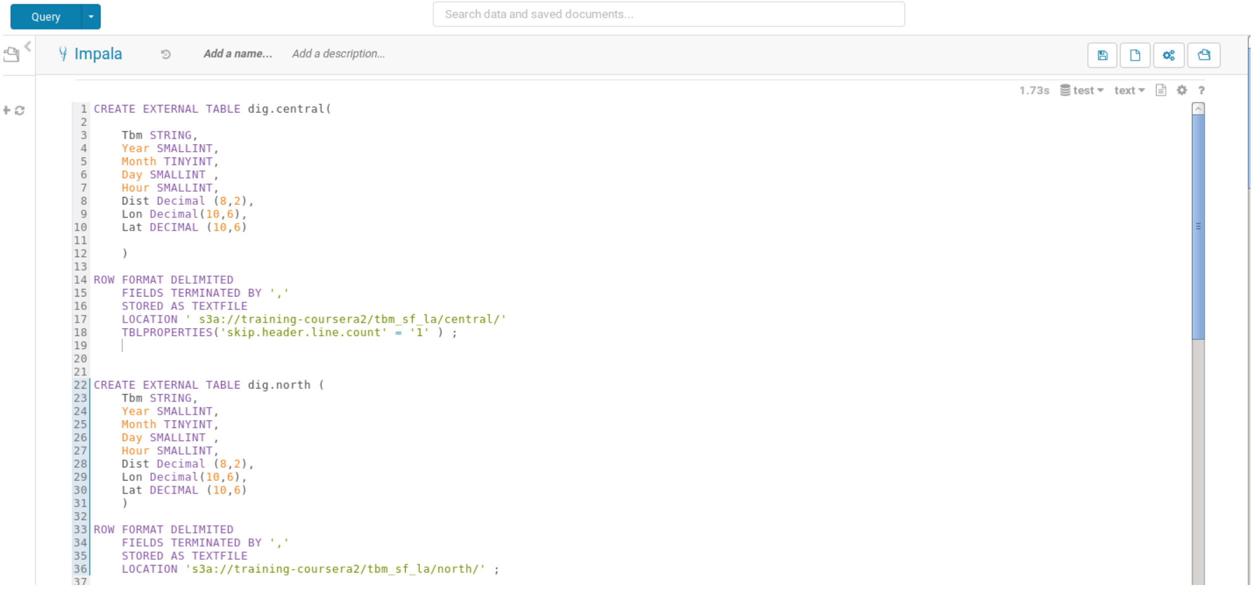
4. With the command below , we can examine a sample from the files

   hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv | head

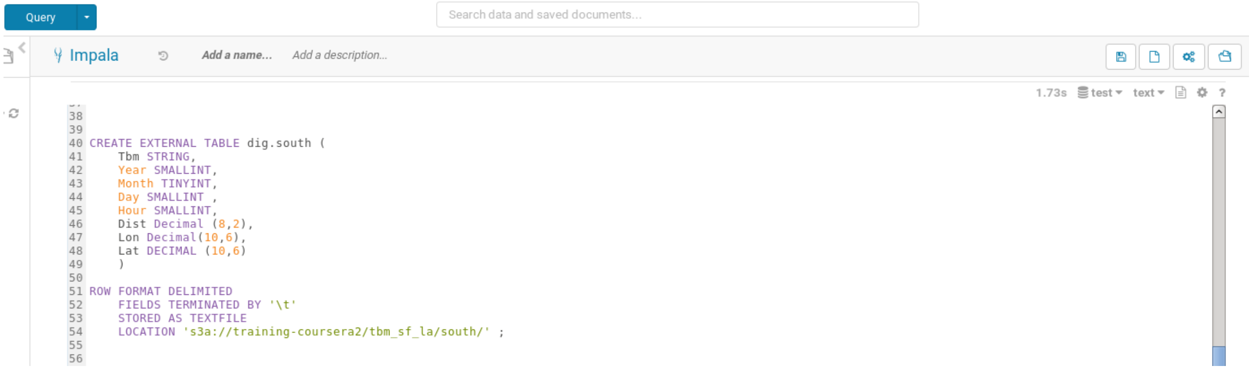


```
[training@localhost ~]$ hdfs dfs -cat s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv | head
tbm,year,month,day,hour,dist,lon,lat
Shai-Hulud,2020,01,02,09,0.00,-121.345467,37.599819
Shai-Hulud,2020,01,02,10,4.90,999999,999999
Shai-Hulud,2020,01,02,11,9.79,999999,999999
Shai-Hulud,2020,01,02,12,14.69,999999,999999
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 s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv | head
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
Bertha II,2020,01,02,15,30.67,\N,\N
Bertha II,2020,01,02,16,36.00,\N,\N
Bertha II,2020,01,02,17,41.33,\N,\N
Bertha II,2020,01,02,18,46.67,\N,\N
cat: Unable to write to output stream.
```

5. Here are the SQL queries that has been used for creating tables

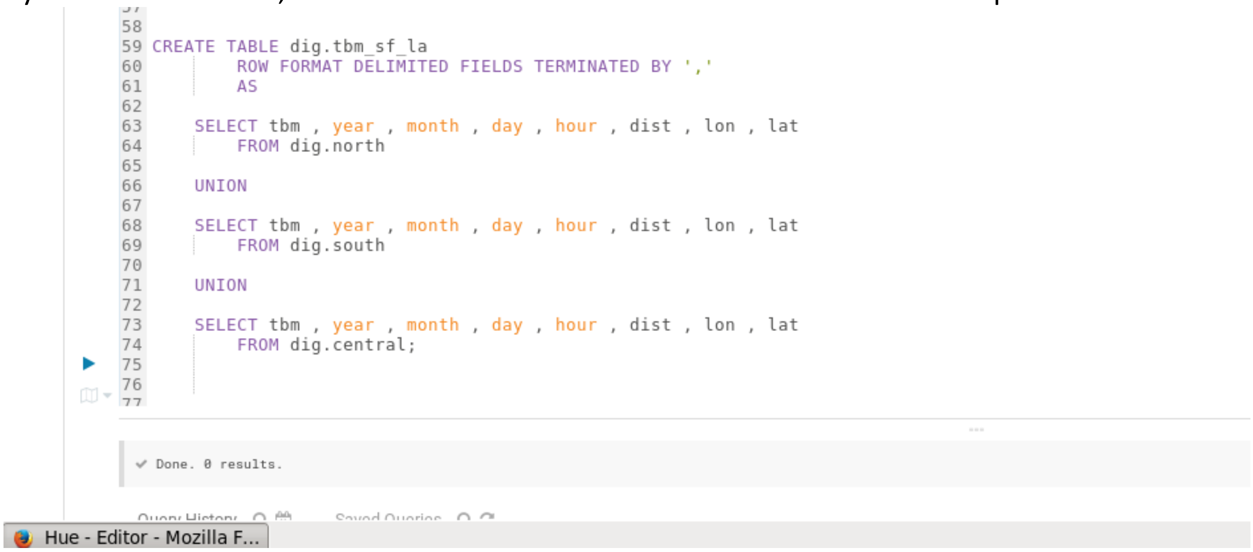


```
CREATE EXTERNAL TABLE dig.central(

    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT ,
    Hour SMALLINT,
    Dist Decimal (8,2),
    Lon Decimal(10,6),
    Lat DECIMAL (10,6)

    )

ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
    STORED AS TEXTFILE
    LOCATION ' s3a://training-coursera2/tbm_sf_la/central/'
    TBLPROPERTIES('skip.header.line.count' = '1' ) ;



CREATE EXTERNAL TABLE dig.north (
    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT ,
    Hour SMALLINT,
    Dist Decimal (8,2),
    Lon Decimal(10,6),
    Lat DECIMAL (10,6)
    )

ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
    STORED AS TEXTFILE
    LOCATION 's3a://training-coursera2/tbm_sf_la/north/' ;
```



```


CREATE EXTERNAL TABLE dig.south (
    Tbm STRING,
    Year SMALLINT,
    Month TINYINT,
    Day SMALLINT ,
    Hour SMALLINT,
    Dist Decimal (8,2),
    Lon Decimal(10,6),
    Lat DECIMAL (10,6)
    )

ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t'
    STORED AS TEXTFILE
    LOCATION 's3a://training-coursera2/tbm_sf_la/south/' ;
```

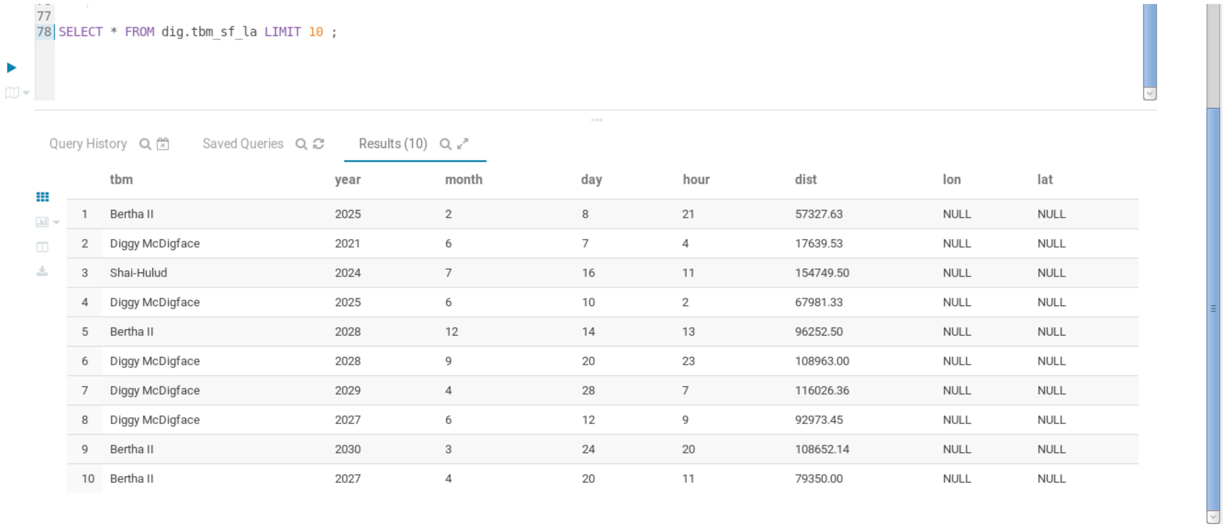
6. By UNION command , we will combine all 3 tables that we created in the step number 5



```

CREATE TABLE dig.tbm_sf_la
        ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
        AS

    SELECT tbm , year , month , day , hour , dist , lon , lat
        FROM dig.north

    UNION

    SELECT tbm , year , month , day , hour , dist , lon , lat
        FROM dig.south

    UNION

    SELECT tbm , year , month , day , hour , dist , lon , lat
        FROM dig.central;
```

```
✓ Done. 0 results.
```

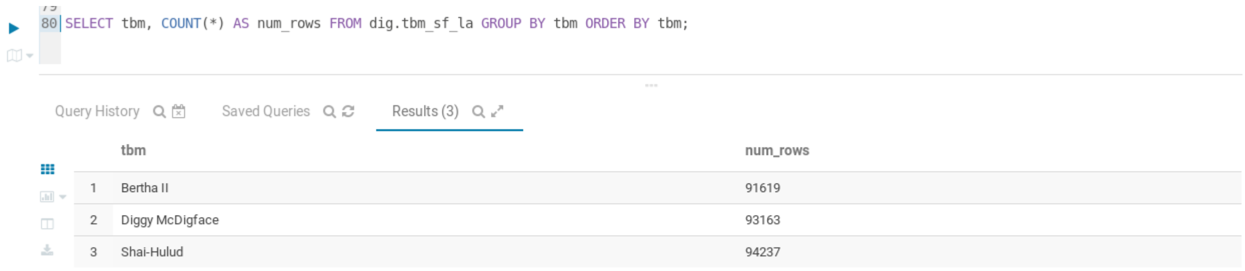
7. Following query will be a basic query to draw a sample view from the combined table.



```
SELECT * FROM dig.tbm_sf_la LIMIT 10 ;
```

	tbm	year	month	day	hour	dist	lon	lat
1	Bertha II	2025	2	8	21	57327.63	NULL	NULL
2	Diggy McDigface	2021	6	7	4	17639.53	NULL	NULL
3	Shai-Hulud	2024	7	16	11	154749.50	NULL	NULL
4	Diggy McDigface	2025	6	10	2	67981.33	NULL	NULL
5	Bertha II	2028	12	14	13	96252.50	NULL	NULL
6	Diggy McDigface	2028	9	20	23	108963.00	NULL	NULL
7	Diggy McDigface	2029	4	28	7	116026.36	NULL	NULL
8	Diggy McDigface	2027	6	12	9	92973.45	NULL	NULL
9	Bertha II	2030	3	24	20	108652.14	NULL	NULL
10	Bertha II	2027	4	20	11	79350.00	NULL	NULL

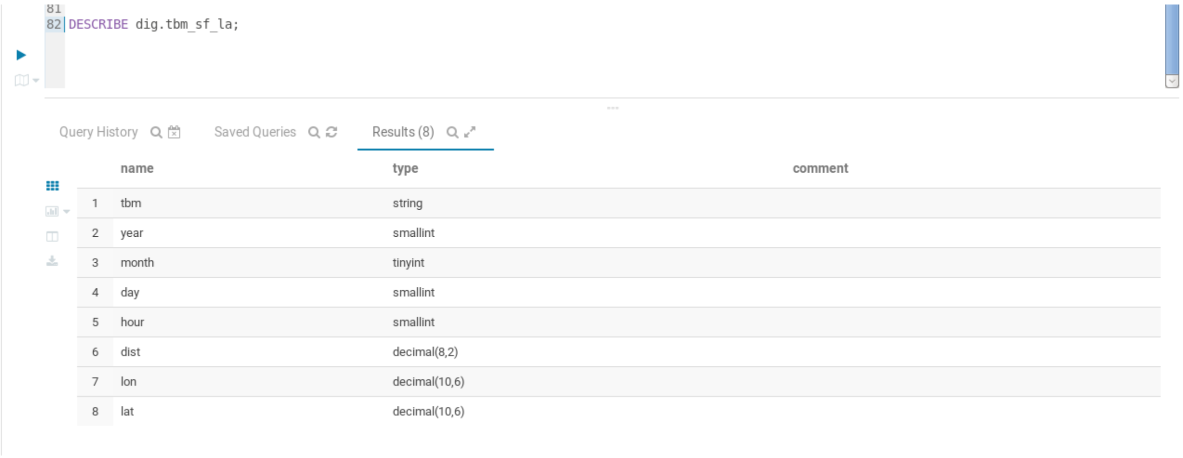
8. SELECT tbm, COUNT(*) AS num_rows FROM dig.tbm_sf_la GROUP BY tbm ORDER BY tbm;



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

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

9. DESCRIBE dig.tbm_sf_la;



```
DESCRIBE dig.tbm_sf_la;
```

	name	type	comment
1	tbm	string	
2	year	smallint	
3	month	tinyint	
4	day	smallint	
5	hour	smallint	
6	dist	decimal(8,2)	
7	lon	decimal(10,6)	
8	lat	decimal(10,6)	

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	smallint
Hour	Smallint
Dist	Decimal(8,2)
Lon	Decimal(10,6)
Lat	Decimal(10,6)