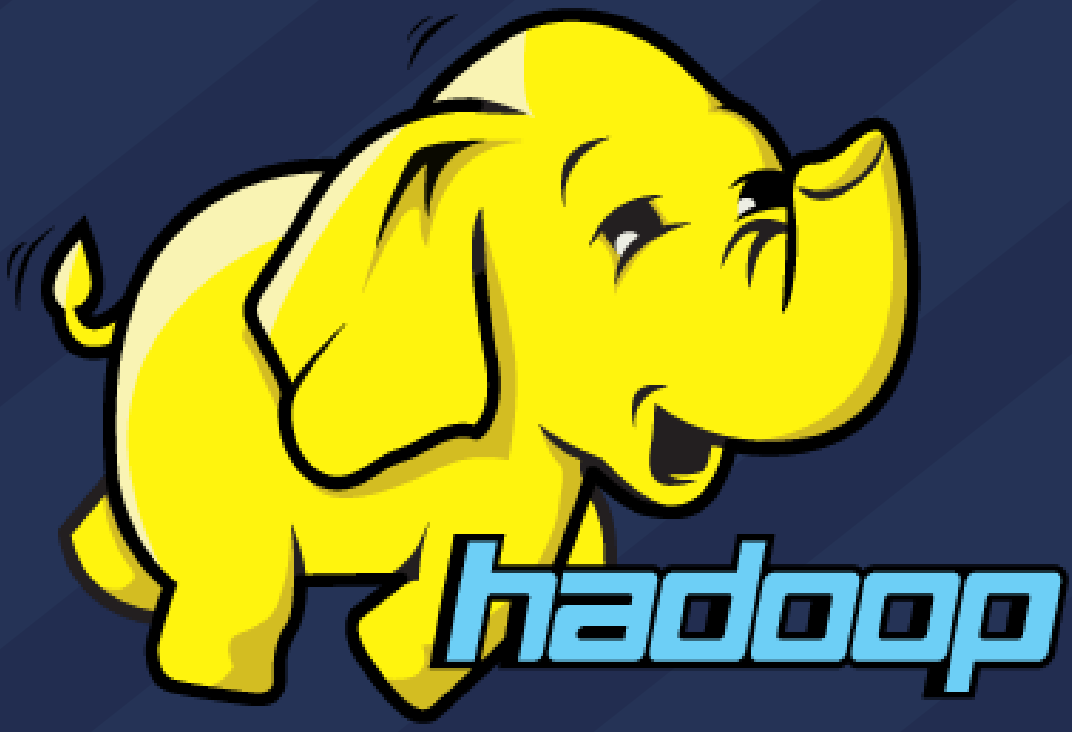
Thiago Panini

2020-Feb-15

PEER-GRADED EXERCISE

LOAD DATA FROM AMAZON S3



S3 bucket:

training-coursera2

folder:

tbm_sf_la

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

After understanding the business problem to be solved with this exercise, the first action is to look at the files located on the amazon s3 bucket. For that, both of the following commands can be used:

1. Searching for Files on the AWS S3 Bucket

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

```
Found 3 items
drwxrwxrwx   - training training          0 2020-02-15 06:57 s3a://training-coursera2/tbm_sf_la/centra
drwxrwxrwx   - training training          0 2020-02-15 06:57 s3a://training-coursera2/tbm_sf_la/north
drwxrwxrwx   - training training          0 2020-02-15 06:57 s3a://training-coursera2/tbm_sf_la/south
```

With AWS S3: aws s2 ls s3://training-coursera2/tbm_sf_la/

```
[training@localhost ~]$ aws s3 ls s3://training-coursera2/tbm_sf_la/
                           PRE central/
                           PRE north/
                           PRE south/
```

With the *ls* command it was possible to see that files were organized on three different folders, one for each TBM machine. The next step was to verify the content of those files and analyze the differences between them.

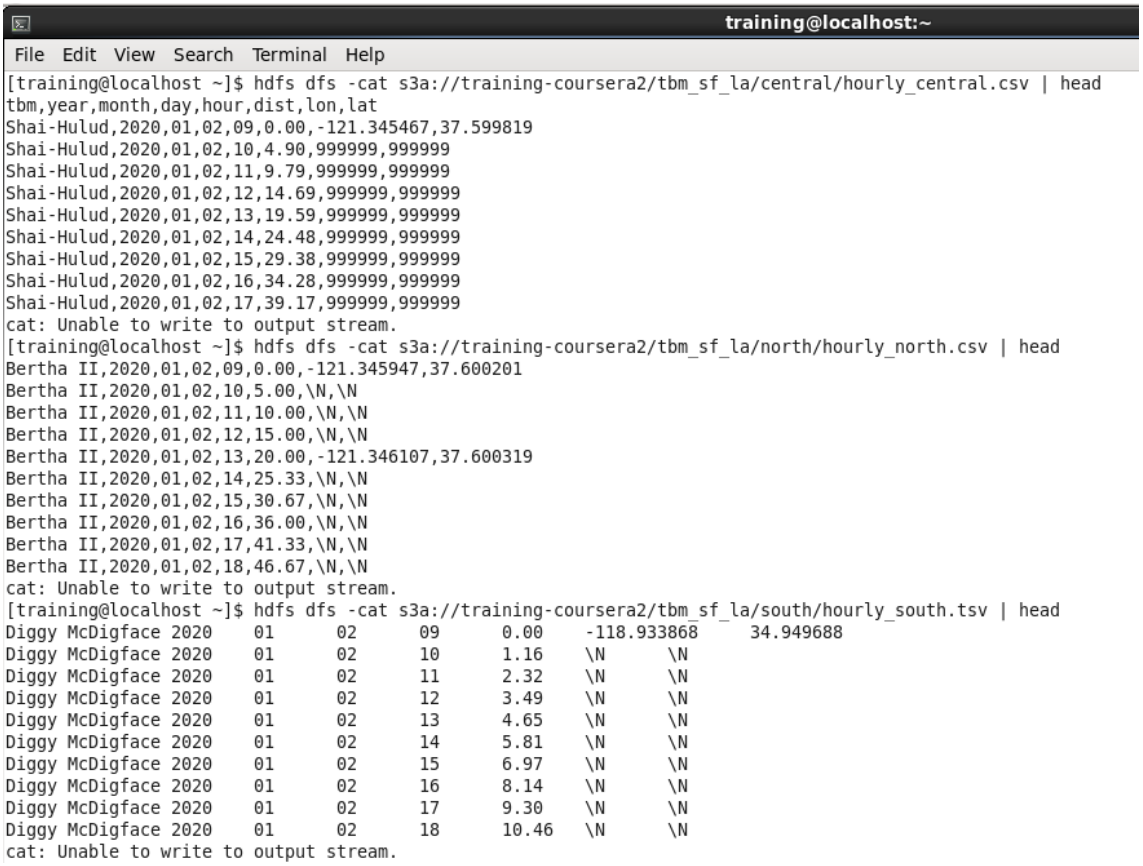
2. Analyzing Files

With the *ls* command it was possible to see that files were organized on three different folders, one for each TBM machine. The next step was to verify the content of those files and analyze the differences between them.

Searching inside the three paths found on the step 1, it was possible to look at the three files for this exercise:

- s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv
- s3a://training-coursera2/tbm_sf_la/north/north_central.csv
- s3a://training-coursera2/tbm_sf_la/south/south_central.tsv

The *hdfs dfs -cat* command can be used for looking at the content of each file. The | *head* statement were used to print just the first then rows:



It's important to see that all the three files have some unique features like the way they threat null values or the field separator. These are things the solution developer must handle during the exercise.

3. Loading Data Into Tables

The flow adopted to load data into tables are shown below:

a) Create a database names *dig*
b) Use a create table statement for each of the three tbm machine (central, north and south)
c) Copy files from s3 bucket to hdfs storage directory created for the tables in the dig database
d) Run a Load data statement to load the files into the hdfs directories created for the tables created recently

Below there is a example of the complete flow for the central tbm using Hue's interface with Impala query engine:

```
CREATE DATABASE IF NOT EXISTS dig;

CREATE EXTERNAL TABLE IF NOT EXISTS dig.tbm_sf_la_central (
    tbm STRING,
    year SMALLINT,
    month TINYINT,
    day TINYINT,
    hour TINYINT,
    dist DECIMAL(11, 2),
    lon DOUBLE,
    lat DOUBLE
) ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES (
    'serialization.null.format'='999999',
    'skip.header.line.count'='1'
);
```

$ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv /user/hive/warehouse/dig.db/tbm_sf_la_central

```
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la_central/hourly_central.csv'
    OVERWRITE INTO TABLE dig.tbm_sf_la_central;

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

Query History | Saved Queries | Results (10)

	tbm	year	month	day	hour	dist	lon	lat
1	Shai-Hulud	2020	1	2	9	0.00	-121.345467	37.599818999999997
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

After running the statements shown below, it was possible to create a “temporary” table for storing the central tbm data with its specific threatments, like the null format values and the presence of the header.

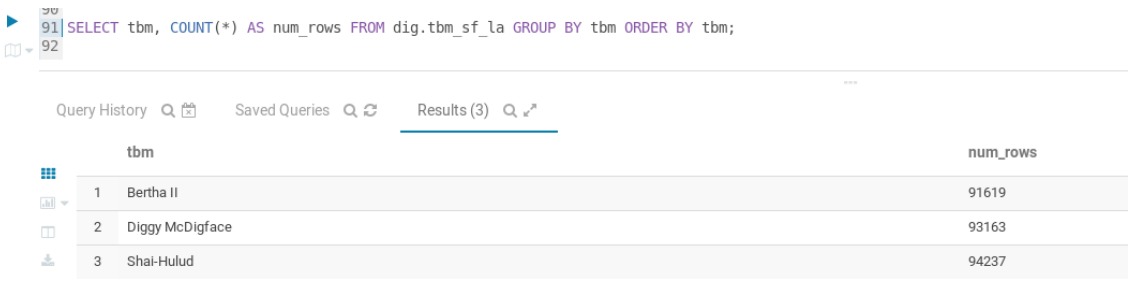
So after repeating the flow for each tbm file, the *dig* database can have three different tables for the north, central and south tbm data:

4. Creating and Loading Data for a Unique TBM Table

For creating a tbm_sf_la table it was necessary to run the following statement:

```
CREATE TABLE IF NOT EXISTS dig.tbm_sf_la
    STORED AS PARQUET AS
    (SELECT
        *
    FROM dig.tbm_sf_la_central
    UNION ALL
    SELECT
        *
    FROM dig.tbm_sf_la_north
    UNION ALL
    SELECT
        *
    FROM dig.tbm_sf_la_south);
```

5. Testing



6. Script

All the scripts can be found below:

```
-- 1. Creating a database names dig
CREATE DATABASE IF NOT EXISTS dig;

-- 2. Central TBM
-- 2.1 Creating table tbm_sf_la_central
CREATE EXTERNAL TABLE IF NOT EXISTS dig.tbm_sf_la_central (
   tbm STRING,
   year SMALLINT,
   month TINYINT,
   day TINYINT,
   hour TINYINT,
   dist DECIMAL(11, 2),
   lon DOUBLE,
   lat DOUBLE
) ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES (
   'serialization.null.format'='999999',
   'skip.header.line.count'='1'
);
```

```
-- 2.2 Run the statement on the command line: $ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/central/hourly_central.csv /user/hive/warehouse/dig.db/tbm_sf_la_central

-- 2.3 Loading data copied from the s3 bucket to the hdfs storage directory
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la_central/hourly_central.csv'
   OVERWRITE INTO TABLE dig.tbm_sf_la_central;

-- 3. North TBM
-- 3.1 Creating table tbm_sf_la_north
CREATE EXTERNAL TABLE IF NOT EXISTS dig.tbm_sf_la_north (
   tbm STRING,
   year SMALLINT,
   month TINYINT,
   day TINYINT,
   hour TINYINT,
   dist DECIMAL(11, 2),
   lon DOUBLE,
   lat DOUBLE
) ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

-- 3.2 Run the statement on the command line: $ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/north/hourly_north.csv /user/hive/warehouse/dig.db/tbm_sf_la_north

-- 3.3 Loading data copied from the s3 bucket to the hdfs storage directory
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la_north/hourly_north.csv'
   OVERWRITE INTO TABLE dig.tbm_sf_la_north;

-- 4 South TBM
-- 4.1 Creating table tbm_sf_la_south
CREATE EXTERNAL TABLE IF NOT EXISTS dig.tbm_sf_la_south (
   tbm STRING,
   year SMALLINT,
   month TINYINT,
   day TINYINT,
   hour TINYINT,
   dist DECIMAL(11, 2),
   lon DOUBLE,
   lat DOUBLE
) ROW FORMAT DELIMITED
   FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE;

-- 4.2 Run the statement on the command line: $ hdfs dfs -cp s3a://training-coursera2/tbm_sf_la/south/hourly_south.tsv /user/hive/warehouse/dig.db/tbm_sf_la_south

-- 4.3 Loading data copied from the s3 bucket to the hdfs storage directory
LOAD DATA INPATH '/user/hive/warehouse/dig.db/tbm_sf_la_south/hourly_south.tsv'
   OVERWRITE INTO TABLE dig.tbm_sf_la_south;

-- 5. Creating a tbm table for storing the data for all three tbm machines
CREATE TABLE IF NOT EXISTS dig.tbm_sf_la
   STORED AS PARQUET AS
   (SELECT
      *
```

```
  FROM dig.tbm_sf_la_central
  UNION ALL
  SELECT
    *
  FROM dig.tbm_sf_la_north
  UNION ALL
  SELECT
    *
  FROM dig.tbm_sf_la_south);

SELECT * FROM dig.tbm_sf_la
  WHERE tbm = 'Diggy McDigface' LIMIT 10;

SELECT DISTINCT tbm FROM dig.tbm_sf_la;
```