BIG DATA HADOOP & SPARK TRAINING

ACADGILD

ASSIGNMENT 6.1

BY :-

SAHIL KHURANA

PROBLEM STATEMENT

Create a database named 'custom'.
Create a table named temperature_data inside custom having below fields:

1. date (mm-dd-yyyy) format
2. zip code
3. temperature

The table will be loaded from comma-delimited file.
Load the dataset.txt (which is ',' delimited) in the table.

Associated Data Files

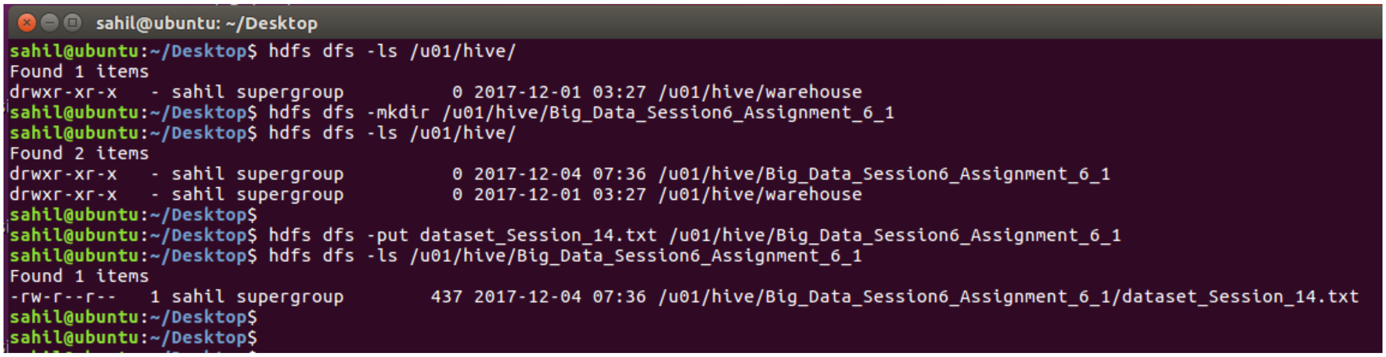
https://drive.google.com/file/d/0Bxr27gVaXO5sa0JBamZXdkpYUFk/view?usp=sharing

dataset_Session_14.txt

```
10-01-1990,123112,10
14-02-1991,283901,11
10-03-1990,381920,15
10-01-1991,302918,22
12-02-1990,384902,9
10-01-1991,123112,11
14-02-1990,283901,12
10-03-1991,381920,16
10-01-1990,302918,23
12-02-1991,384902,10
10-01-1993,123112,11
14-02-1994,283901,12
10-03-1993,381920,16
10-01-1994,302918,23
12-02-1991,384902,10
10-01-1991,123112,11
14-02-1990,283901,12
10-03-1991,381920,16
10-01-1990,302918,23
12-02-1991,384902,10
```

Note:- To solve the Assignment 6.1, I have created a VM with Ubuntu 16.04 OS and configured Hadoop 2.8.2 and hive-2.3.2 on the same.

Step 1:- Put the dataset in HDFS location

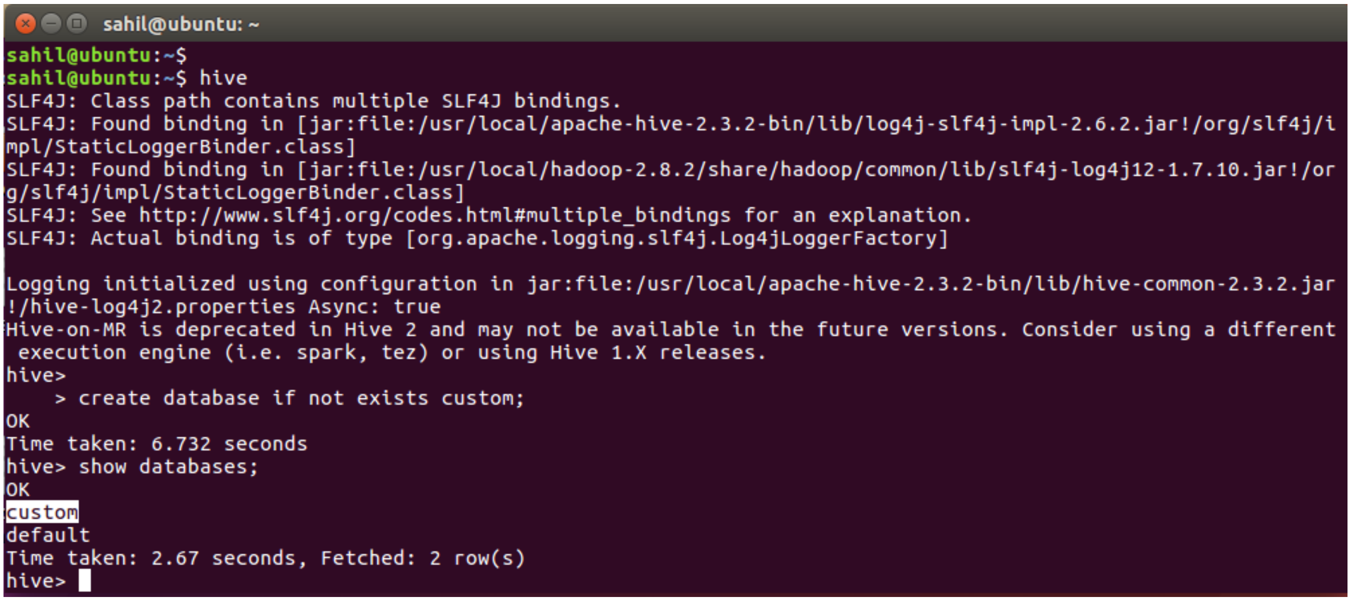


```
sahil@ubuntu: ~/Desktop
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/
Found 1 items
drwxr-xr-x   - sahil supergroup          0 2017-12-01 03:27 /u01/hive/warehouse
sahil@ubuntu:~/Desktop$ hdfs dfs -mkdir /u01/hive/Big_Data_Session6_Assignment_6_1
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/
Found 2 items
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:36 /u01/hive/Big_Data_Session6_Assignment_6_1
drwxr-xr-x   - sahil supergroup          0 2017-12-01 03:27 /u01/hive/warehouse
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$ hdfs dfs -put dataset_Session_14.txt /u01/hive/Big_Data_Session6_Assignment_6_1
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/Big_Data_Session6_Assignment_6_1
Found 1 items
-rw-r--r--   1 sahil supergroup        437 2017-12-04 07:36 /u01/hive/Big_Data_Session6_Assignment_6_1/dataset_Session_14.txt
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$
```

Step 2:- Open the Hive Shell and CREATE the DATABASE.

Commands used in Step 2

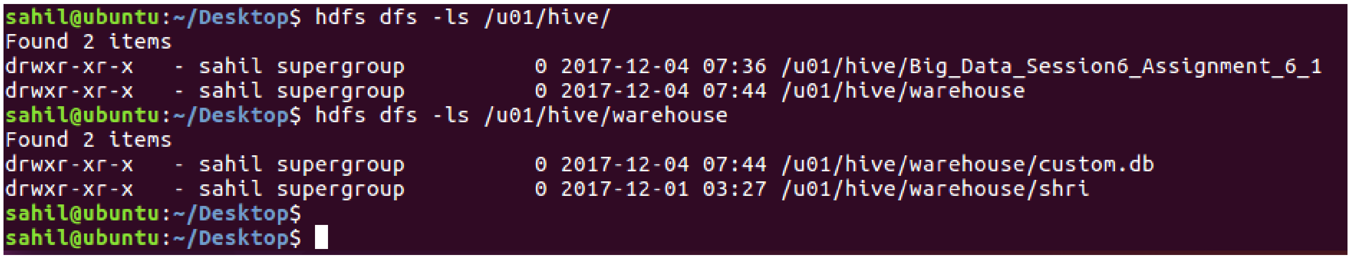
```
hive                                     -- open the hive shell
hive> create database if not exists custom;  -- create database
hive> show databases;                    -- check whether database is created or not
```



```
sahil@ubuntu:~$
sahil@ubuntu:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.8.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/or
g/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar
!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different
 execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
    > create database if not exists custom;
OK
Time taken: 6.732 seconds
hive> show databases;
OK
custom
default
Time taken: 2.67 seconds, Fetched: 2 row(s)
hive>
```

Custom database created in default directory in hive

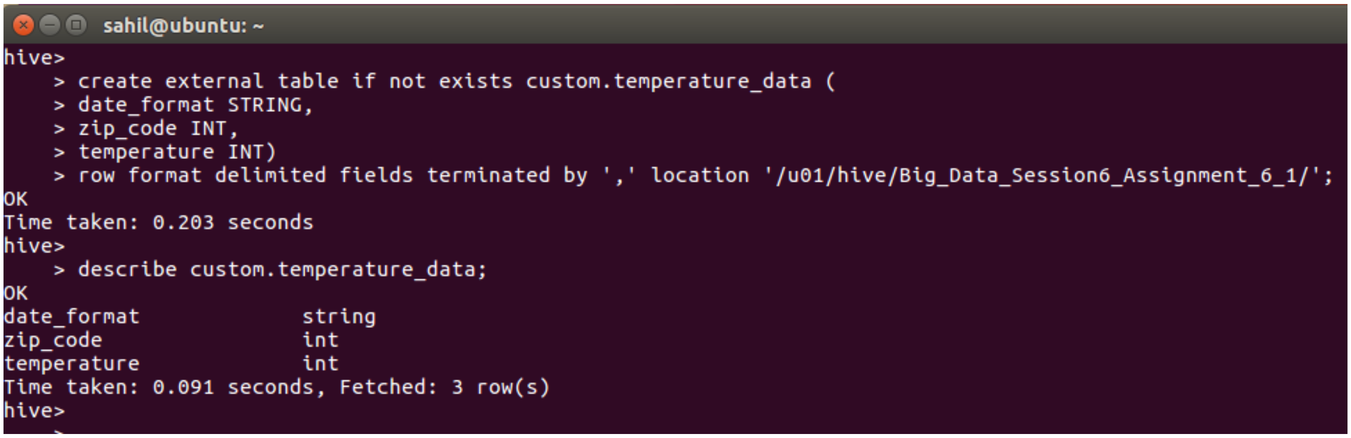


```
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/
Found 2 items
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:36 /u01/hive/Big_Data_Session6_Assignment_6_1
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:44 /u01/hive/warehouse
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/warehouse
Found 2 items
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:44 /u01/hive/warehouse/custom.db
drwxr-xr-x   - sahil supergroup          0 2017-12-01 03:27 /u01/hive/warehouse/shri
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$
```

Step 3:- CREATE EXTERNAL TABLE

Commands used in Step 3

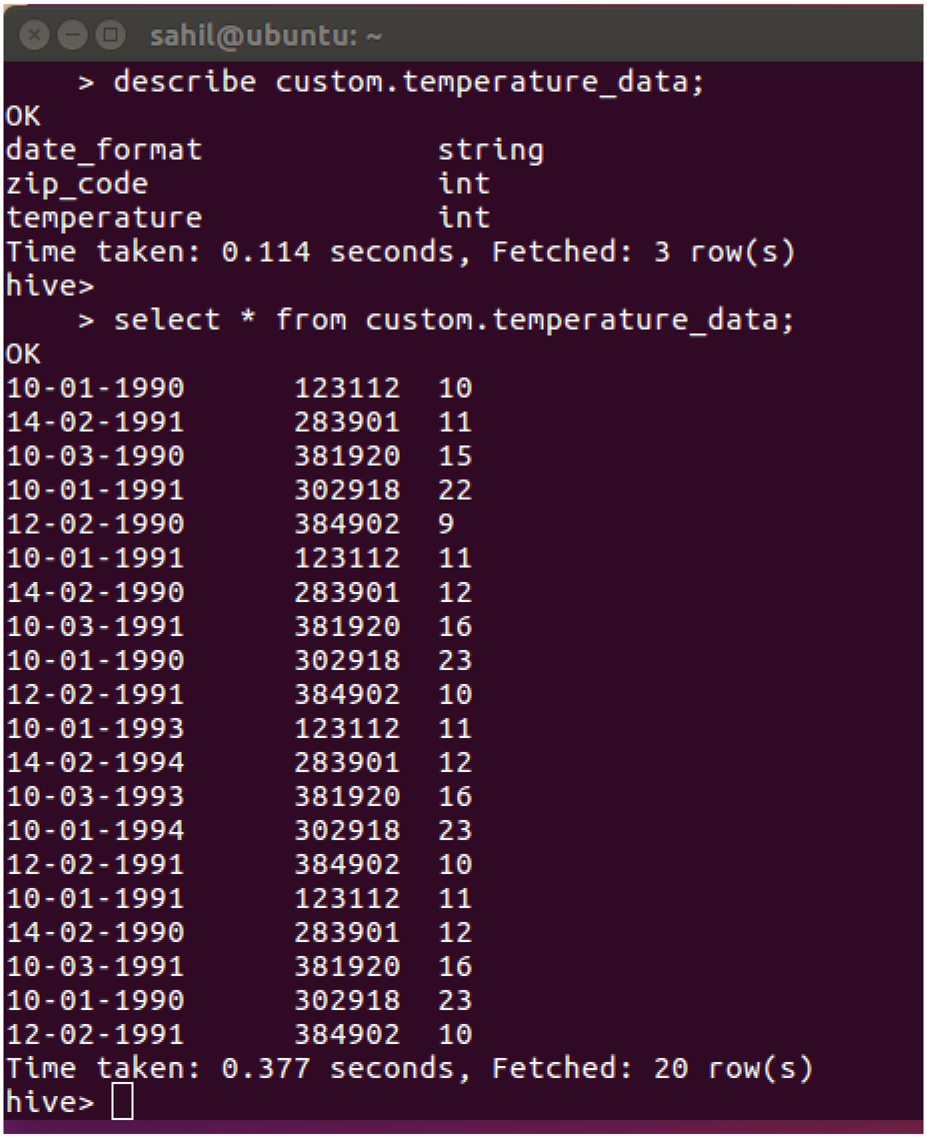
```
hive>
    > create external table if not exists custom.temperature_data (
    > date_format STRING,
    > zip_code INT,
    > temperature INT)
    > row format delimited fields terminated by ',' location
'/u01/hive/Big_Data_Session6_Assignment_6_1/';
hive> describe custom.temperature_data;
```



```
hive>
    > create external table if not exists custom.temperature_data (
    > date_format STRING,
    > zip_code INT,
    > temperature INT)
    > row format delimited fields terminated by ',' location '/u01/hive/Big_Data_Session6_Assignment_6_1/';
OK
Time taken: 0.203 seconds
hive>
    > describe custom.temperature_data;
OK
date_format             string
zip_code                int
temperature             int
Time taken: 0.091 seconds, Fetched: 3 row(s)
hive>
```

Step 4 :- Check whether the dataset is imported in the hive table or not.
Commands used in Step 4

```
hive>
    > select * from custom.temperature_data;
```



```
    > describe custom.temperature_data;
OK
date_format             string
zip_code                int
temperature             int
Time taken: 0.114 seconds, Fetched: 3 row(s)
hive>
    > select * from custom.temperature_data;
OK
10-01-1990      123112  10
14-02-1991      283901  11
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902  9
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
10-01-1993      123112  11
14-02-1994      283901  12
10-03-1993      381920  16
10-01-1994      302918  23
12-02-1991      384902  10
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 0.377 seconds, Fetched: 20 row(s)
hive> 
```