ASSIGNMENT 6.1

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.

Solution:

Database creation query:

CREATE DATABASE custom;

```
hive> CREATE DATABASE custom;
OK
Time taken: 0.119 seconds
hive> show databases;
OK
b1
custom
default
Time taken: 0.126 seconds, Fetched: 3 row(s)
hive>
```

We can also access the physical location where the data of this database will be stored in HDFS.

```
[acadgild@localhost conf]$ hadoop fs -ls /user/hive/warehouse
17/12/15 00:14:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 3 items
drwxr-xr-x   - acadgild supergroup          0 2015-11-25 15:25 /user/hive/warehouse/b1.db
drwxr-xr-x   - acadgild supergroup          0 2017-12-15 00:00 /user/hive/warehouse/custom.db
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 13:14 /user/hive/warehouse/first
[acadgild@localhost conf]$
```

acadgild | acadgild@localhost:/u... | acadgild@localhost:/u...

Now, I have to create a table and load the given dataset. The date format used in the input dataset is MM-dd-yyyy which is different than the default date format, yyyy-MM-dd. We can't change the default date format in Hive while creating a table and adding a field of type 'DATE'. So I have followed two steps to load such data into a table:

Step 1: Create a table with date column of type **STRING** and load the data from input file.

Table creation query:

```
CREATE TABLE IF NOT EXISTS temperature_data_new
(
date STRING,
zip_code INT,
temperature INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

```
hive> CREATE TABLE IF NOT EXISTS temperature_data_new
    > (
    > date STRING,
    > zip_code INT,
    > temperature INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.289 seconds
```

Data load query:

```
LOAD DATA LOCAL INPATH '/home/acadgild/temperature_dataset.txt' INTO TABLE temperature_data_new;
```

```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/temperature_dataset.txt' INTO TABLE temperature_data_new;
Loading data to table custom.temperature_data_new
Table custom.temperature_data_new stats: [numFiles=1, totalSize=437]
OK
Time taken: 1.633 seconds
hive> SELECT * FROM temperature_data_new;
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: 0.18 seconds, Fetched: 20 row(s)
```

Here, I have loaded data from local path. We can also load data from HDFS.

Step 2: Create another table with date column of type **DATE** and insert data into it from the table we created in first step.

Table creation query:

CREATE TABLE IF NOT EXISTS temperature_data
(
date DATE,
zip_code INT,
temperature INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

```
hive> CREATE TABLE IF NOT EXISTS temperature_data
    > (
    > date DATE,
    > zip_code INT,
    > temperature INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.181 seconds
hive> SHOW TABLES;
OK
temperature_data
temperature_data_new
Time taken: 0.184 seconds, Fetched: 2 row(s)
```

Data load query:

INSERT INTO TABLE temperature_data SELECT CAST (TO_DATE (FROM_UNIXTIME (UNIX_TIMESTAMP (date, 'MM-dd-yyyy'))) AS date), zip_code, temperature FROM temperature_data_new;

```
hive> INSERT INTO TABLE temperature_data SELECT CAST(TO_DATE(FROM_UNIXTIME(UNIX_TIMESTAMP(date, 'MM-dd-yyyy'))) AS date), zip
_code, temperature FROM temperature_data_new;
Query ID = acadgild_20171215045656_2e77055c-30b5-44f0-9cd1-299b96e1a4fc
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1513259210539_0002, Tracking URL = http://localhost:8088/proxy/application_1513259210539_0002/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1513259210539_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2017-12-15 04:56:52,332 Stage-1 map = 0%,  reduce = 0%
2017-12-15 04:57:16,837 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.75 sec
MapReduce Total cumulative CPU time: 4 seconds 750 msec
Ended Job = job_1513259210539_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:9000/tmp/hive/acadgild/7a213af8-36df-4950-bb15-deb6aef20589/hive_2017-12-15_04-56-08_190_637
5479837078230521-1/-ext-10000
Loading data to table custom.temperature_data
Table custom.temperature_data stats: [numFiles=1, numRows=20, totalSize=419, rawDataSize=399]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.75 sec   HDFS Read: 687 HDFS Write: 499 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 750 msec
OK
Time taken: 73.899 seconds
hive>
```

In the above query, we are informing Hive that date string in the table, *temperature_data_new* is following different format (in this case, 'MM-dd-yyyy') and trying to cast those values from custom format to default one.

Output:

SELECT * FROM temperature_data;

```
hive> SELECT * FROM temperature_data;
OK
1990-10-01      123112  10
1992-02-02      283901  11
1990-10-03      381920  15
1991-10-01      302918  22
1990-12-02      384902  9
1991-10-01      123112  11
1991-02-02      283901  12
1991-10-03      381920  16
1990-10-01      302918  23
1991-12-02      384902  10
1993-10-01      123112  11
1995-02-02      283901  12
1993-10-03      381920  16
1994-10-01      302918  23
1991-12-02      384902  10
1991-10-01      123112  11
1991-02-02      283901  12
1991-10-03      381920  16
1990-10-01      302918  23
1991-12-02      384902  10
Time taken: 0.179 seconds, Fetched: 20 row(s)
hive> DESCRIBE temperature_data;
OK
date                    date
zip_code                int
temperature             int
Time taken: 0.336 seconds, Fetched: 3 row(s)
hive>
```