Session 8 - HIVE BASICS

Assignment 1

TASK 1:

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.

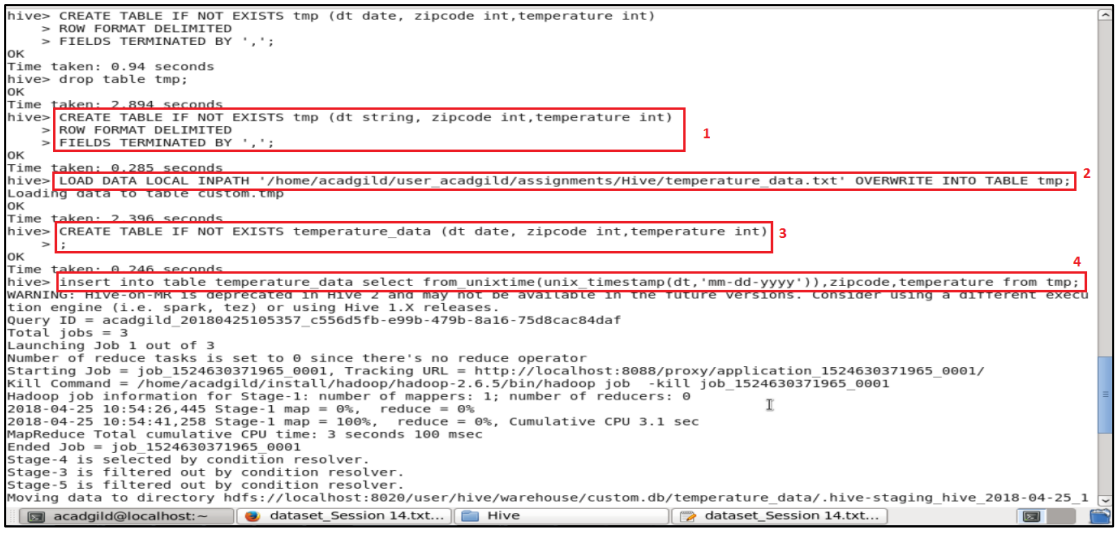
Load the dataset.txt (which is ',' delimited) in the table.

Database creation :

Command :

create database custom;

Table Creation :



With reference to the screenshot above,

1 : To convert the date to ‘mm-dd-yyyy’ format, a temporary table is used. Initially, the date is loaded from the file as a string.

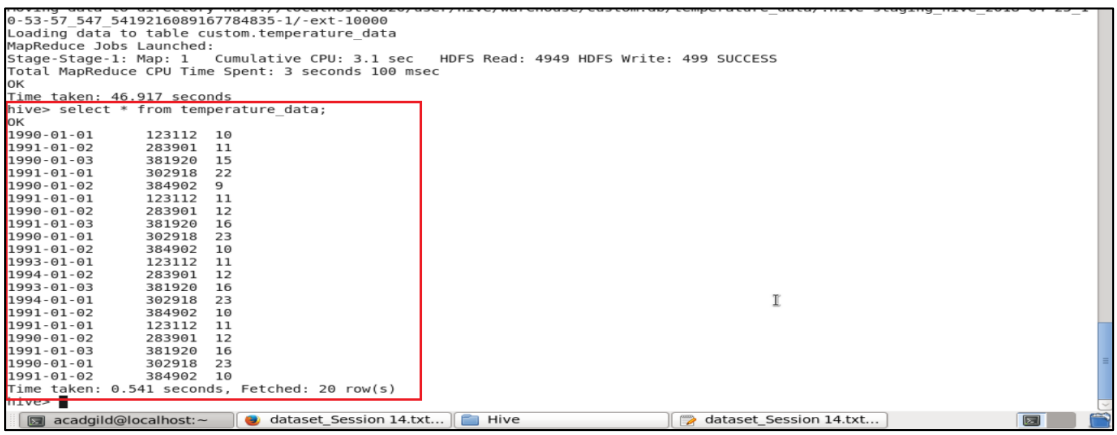
2 : The data from the text file is loaded into the temporary table.

LOAD DATA LOCAL INPATH ‘location/of/text/file’ OVERWRITE INTO TABLE table_name;

3 : The actual Table ‘temperature_data’ is created with the column data type as ‘date’.

4 : Now the date format transformation is applied on the date in string format in temporary table and inserted into the ‘temperature_data’ table.

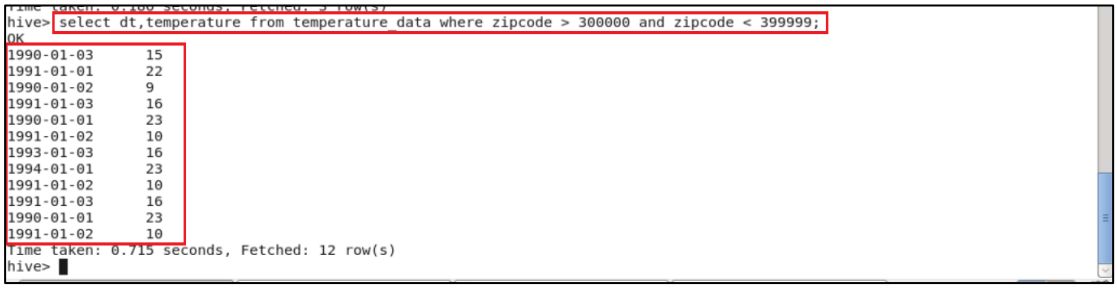
Insert into table temperature_data select ***from_unixtime(unix_time(dt,’mm-dd-yyyy’))****, zipcode, temperature from tmp;*



Data is loaded and on performing a select on the table ‘temperature_data’, the data in the table can be seen.

TASK 2:

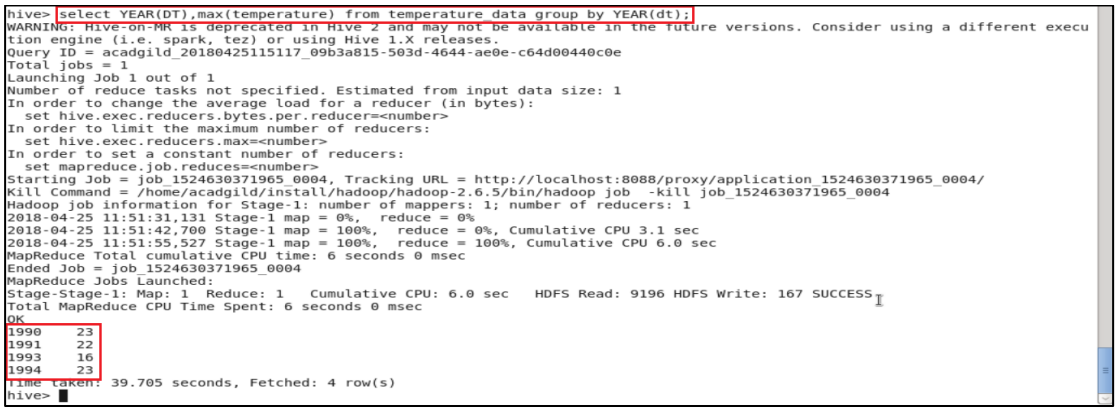
1. Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.



HQL statement :

select dt, temperature from temperature_data where zipcode > 300000 and zipcode < 399999;

2. Calculate maximum temperature corresponding to every year from temperature_data table.



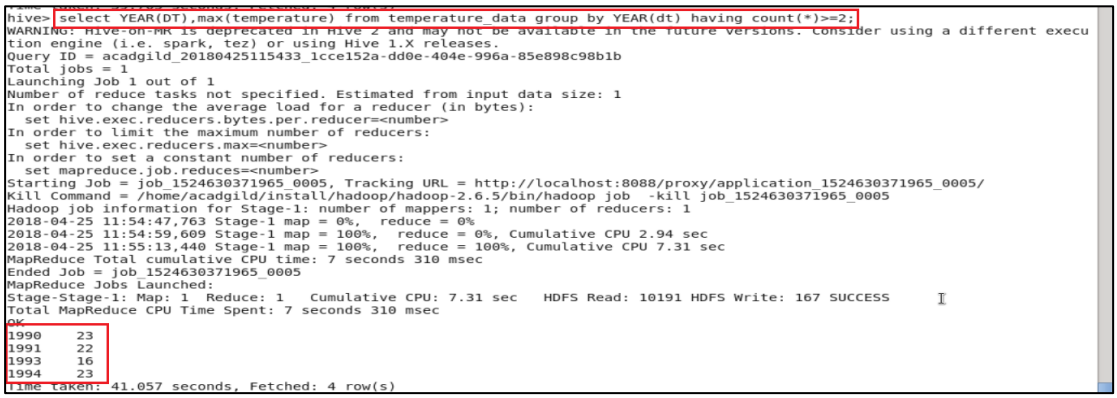
HQL statement :

select YEAR(dt), max(temperature) from temperature_data group by YEAR(dt);

YEAR(dt) : to extract the ‘yyyy’ part from the timestamp/date.

Group by : used to group together and perform max on each group.

3. Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.



HQL statement :

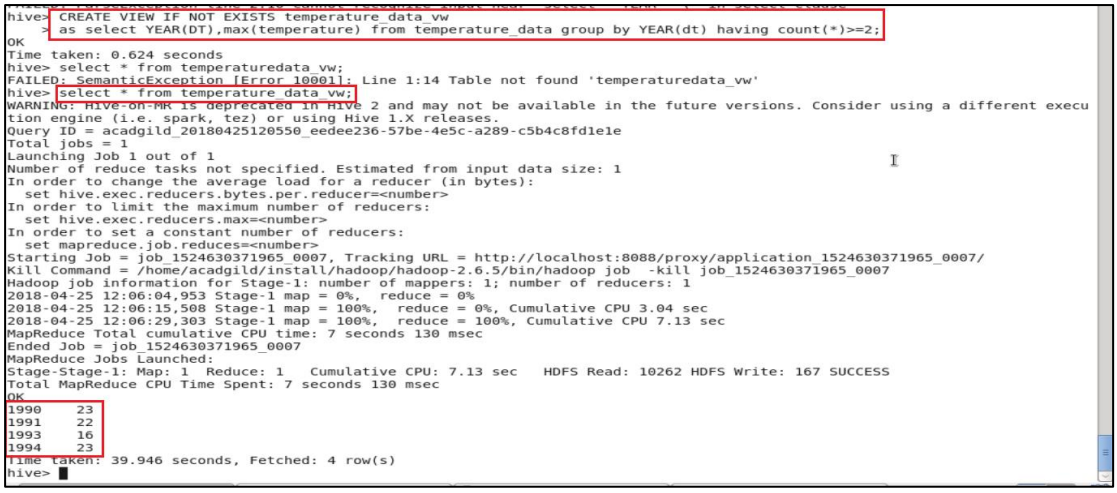
select YEAR(dt), max(temperature) from temperature_data group by YEAR(dt) having count(*) >= 2;

YEAR(dt) : to extract the ‘yyyy’ part from the timestamp/date.

group by : used to group together and perform max on each group.

having count(*) >=2 : to filter in the groups that have 2 or more values in it.

4. Create a view on the top of last query, name it temperature_data_vw.



HQL Statement :

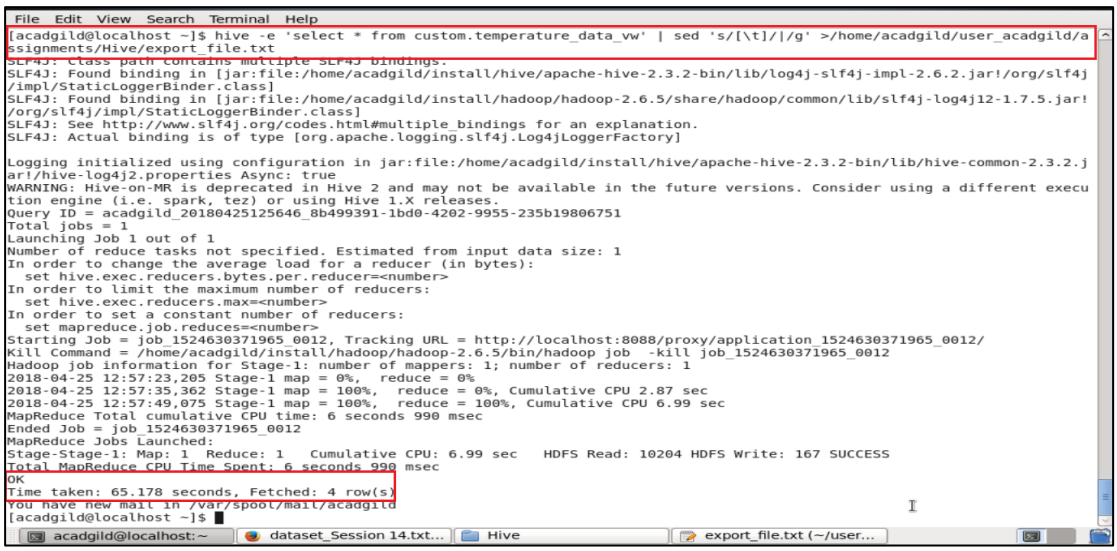
CREATE VIEW IF NOT EXISTS temperature_data_vw as

select YEAR(dt), max(temperature) from temperature_data group by YEAR(dt) having count(*) >= 2;

View named ‘temperature_data_vw’ is created.

When queried, the data can be seen in the view.

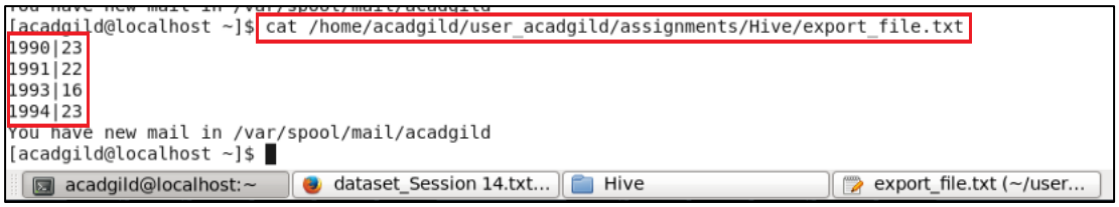
5. Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.



HQL Statement :

*hive -e 'select * from custom.temperature_data_vw' | sed 's/[\t]/|/g' >/home/acadgild/assignments/Hive/export_file.txt*

The query select * from custom.temperature_data_vw is run in hive and the resultant fields are seperated by '|' and the final result is put into the file 'export_file.txt'



When a cat is run on the exported file, |-separated values can be seen.