

## Assignment 6.2

### Create table and Load data:

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

Load data local inpath '/home/acadgild/dataset\_session.txt' overwrite into table temperature\_data ;

### Problem statement:

(a)

Fetch date and temperature from temperature\_data where zip code is greater than 300000 and less than 399999.

### Solution:

```
select distinct date,temperature from temperature_data where zip_code
BETWEEN 299999 AND 399999;
```

**// with distinct keyword**

```
Total mapreduce CPU Time Spent: 2 seconds 200 ms.  
OK  
10-01-1990      23  
10-01-1991      22  
10-01-1994      23  
10-03-1990      15  
10-03-1991      16  
10-03-1993      16  
12-02-1990       9  
12-02-1991      10  
Time taken: 21.571 seconds, Fetched: 8 row(s)
```

**// without distinct keyword**

```
...  
10-03-1990      15  
10-01-1991      22  
12-02-1990       9  
10-03-1991      16  
10-01-1990      23  
12-02-1991      10  
10-03-1993      16  
10-01-1994      23  
12-02-1991      10  
10-03-1991      16  
10-01-1990      23  
12-02-1991      10  
Time taken: 0.036 seconds, Fetched: 12 row(s)  
hive> █
```

---

**(B)**

Calculate maximum temperature corresponding to every year from temperature\_data table.

**Solution:**

```
select year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy')))  
,max(temperature) from temperature_data group by  
year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy')));
```

```
OK
1990      23
1991      22
1992      11
1993      16
1994      23
1995      12
Time taken: 19.272 seconds, Fetched: 6 row(s)
hive> █
```

---

(c)

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

**Solution:**

```
select year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-
yyyy'))),max(temperature) from temperature_data group by
year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy')) ) having
count(year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy')) )
>=2;
```

```
1990      23
1991      22
1993      16
Time taken: 20.544 seconds, Fetched: 3 row(s)
hive> █
```

---

(d)

Create a view on the top of last query, name it temperature\_data\_vw.

**Solution:**

**//creating view:**

```
create view temperature_data_vw as select  
year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-  
yyyy'))),max(temperature) from temperature_data group by  
year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy'))) having  
count(year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy'))) )  
>=2;
```

```
hive> create view temperature_data_vw as select year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy'))),max(temperature) from temperature_data group by year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy'))) having count(year(from_unixtime(UNIX_TIMESTAMP(date,'MM-dd-yyyy'))) ) >=2;  
OK  
Time taken: 0.075 seconds  
hive> select * from temperature_data_vw;
```

```
1990      23  
1991      22  
1993      16  
Time taken: 20.427 seconds, Fetched: 3 row(s)  
hive> █
```

**(e)**

Export contents from temperature\_data\_vw to a file in local file system, such that each file is '|' delimited.

**Solution:**

```
insert overwrite local directory '/home/acadgild/output' ROW FORMAT  
DELIMITED FIELDS TERMINATED BY '|' select * from  
temperature_data_vw;
```

```
[acadgild@localhost ~]$ ls /home/acadgild/output
000000_0
[acadgild@localhost ~]$ cat /home/acadgild/output/000000_0
1990|23
1991|22
1993|16
[acadgild@localhost ~]$
```

---