Session 6

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 6 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 6 – Assignment 2** |
|  |  |
| **Start Date:** | 20/09/2017 |
|  |  |
|  |  |

# 

# Contents

[Contents 2](#_Toc493669438)

[Change History 3](#_Toc493669439)

[1. Problem Statement 4](#_Toc493669440)

[2. Solutions 5](#_Toc493669441)

[2.1. Fetch date and temperature from temperature\_data where zip code is greater than 300000 and less than 399999. 5](#_Toc493669442)

[2.2. Calculate maximum temperature corresponding to every year from temperature\_data table. 5](#_Toc493669443)

[2.3. Calculate maximum temperature from temperature\_data table corresponding to those years which have at least 2 entries in the table. 6](#_Toc493669444)

[2.4. Create a view on the top of last query, name it temperature\_data\_vw. 7](#_Toc493669445)

[2.5. Export contents from temperature\_data\_vw to a file in local file system, such that each file is '|' delimited. 8](#_Toc493669446)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 20/09/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* Fetch date and temperature from temperature\_data where zip code is greater than 300000 and less than 399999.
* Calculate maximum temperature corresponding to every year from temperature\_data table.
* Calculate maximum temperature from temperature\_data table corresponding to those years which have at least 2 entries in the table.
* Create a view on the top of last query, name it temperature\_data\_vw.
* Export contents from temperature\_data\_vw to a file in local file system, such that each file is '|' delimited.

**Data in temperature\_data**

*hive (custom)> select \* from 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: 0.042 seconds, Fetched: 20 row(s)*

# Solutions

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

*hive (custom)> SELECT \* FROM temperature\_data WHERE postcode >300000 and postcode <399999;*

*OK*

*10-03-1990 381920 15*

*10-01-1991 302918 22*

*12-02-1990 384902 9*

*10-03-1991 381920 16*

*10-01-1990 302918 23*

*12-02-1991 384902 10*

*10-03-1993 381920 16*

*10-01-1994 302918 23*

*12-02-1991 384902 10*

*10-03-1991 381920 16*

*10-01-1990 302918 23*

*12-02-1991 384902 10*

*Time taken: 0.084 seconds, Fetched: 12 row(s)*

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

*hive (custom)>* *select date, max(temperature)*

*from temperature\_data*

*group by date;*

**Results**

OK

10-01-1990      23

10-01-1991      22

10-01-1993      11

10-01-1994      23

10-03-1990      15

10-03-1991      16

10-03-1993      16

12-02-1990      9

12-02-1991      10

14-02-1990      12

14-02-1991      11

14-02-1994      12

Time taken: 20.951 seconds, Fetched: 12 row(s) hive (custom)>

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

*hive (custom)>* *SELECT date, max(temperature),count(1) AS cnt FROM temperature\_data GROUP BY date HAVING cnt > 1;*

**Results**

OK

10-01-1990 23 3

10-01-1991 22 3

10-03-1991 16 2

12-02-1991 10 3

14-02-1990 12 2

Time taken: 22.775 seconds, Fetched: 5 row(s)

**Just to check results**

*hive (custom)> select \* from temperature\_data order by date DESC;*

OK

14-02-1994 283901 12

14-02-1991 283901 11

14-02-1990 283901 12

14-02-1990 283901 12

12-02-1991 384902 10

12-02-1991 384902 10

12-02-1991 384902 10

12-02-1990 384902 9

10-03-1993 381920 16

10-03-1991 381920 16

10-03-1991 381920 16

10-03-1990 381920 15

10-01-1994 302918 23

10-01-1993 123112 11

10-01-1991 123112 11

10-01-1991 123112 11

10-01-1991 302918 22

10-01-1990 302918 23

10-01-1990 302918 23

10-01-1990 123112 10

Time taken: 23.037 seconds, Fetched: 20 row(s)

**Other way to present this without displaying the count**

hive (custom)> *SELECT date, max(temperature) from temperature\_data*

*GROUP BY date*

*having count(date) >1*

**Results**

OK

10-01-1990 23

10-01-1991 22

10-03-1991 16

12-02-1991 10

14-02-1990 12

Time taken: 22.342 seconds, Fetched: 5 row(s)

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

*hive (custom)> create view if not exists temperature\_data\_vw as SELECT date, max(temperature),count(1) AS cnt FROM temperature\_data GROUP BY date HAVING cnt > 1;*

*OK*

*Time taken: 0.077 seconds*

**Results**

hive (custom)> select \* from temperature\_data\_vw;

10-01-1990 23 3

10-01-1991 22 3

10-03-1991 16 2

12-02-1991 10 3

14-02-1990 12 2

Time taken: 19.511 seconds, Fetched: 5 row(s)

hive (custom)>

**View created.**

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

*INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hiveintro'*

*ROW FORMAT DELIMITED*

*FIELDS TERMINATED BY '|'*

**Results**

