## **CA675 Cloud Technologies**

|  |  |
| --- | --- |
| Name | Paritosh Gupta |
| Student Number | 18210686 |
| Programme | MCM – DA |
| Module Code | CA675 |
| Assignment Title | Cloud Technology Assignment 1 |
| Submission date | 09th March 2019 |
| Module coordinator | Alessandra Mileo |

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

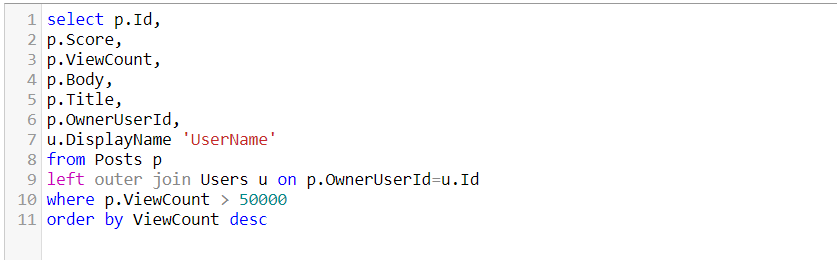
I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: Paritosh Gupta Date: 09th March 2019

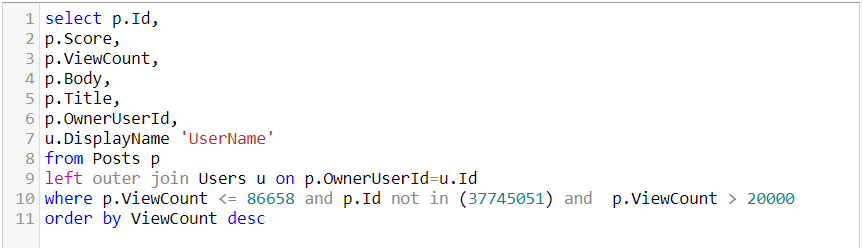
# Task 1 - Data Extraction:

The top 200,000 posts as per view count has been extracted from stackexchange portal. As only 50,000 records can be downloaded in single query run, the execution was performed in 4 batches. Below are the screenshots for the query execution on each batch.

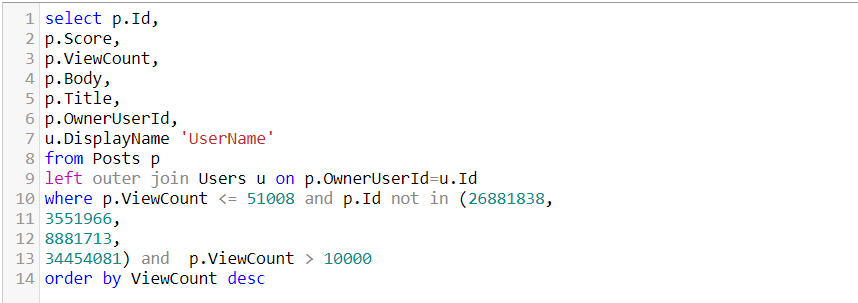
Batch1-



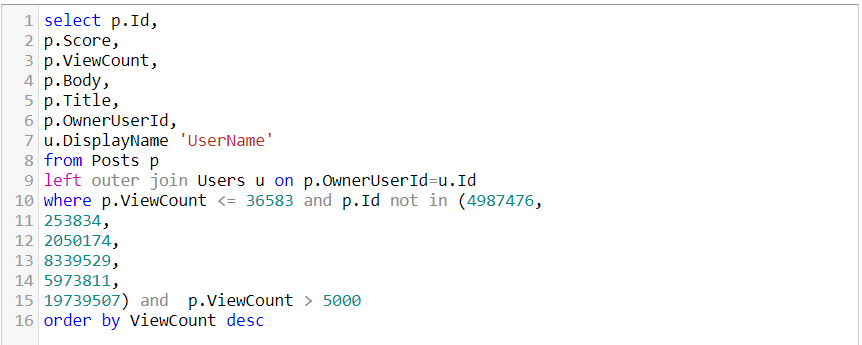
Batch2-



Batch3-

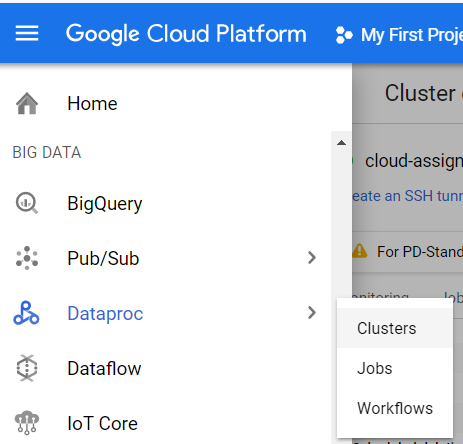


Batch4-

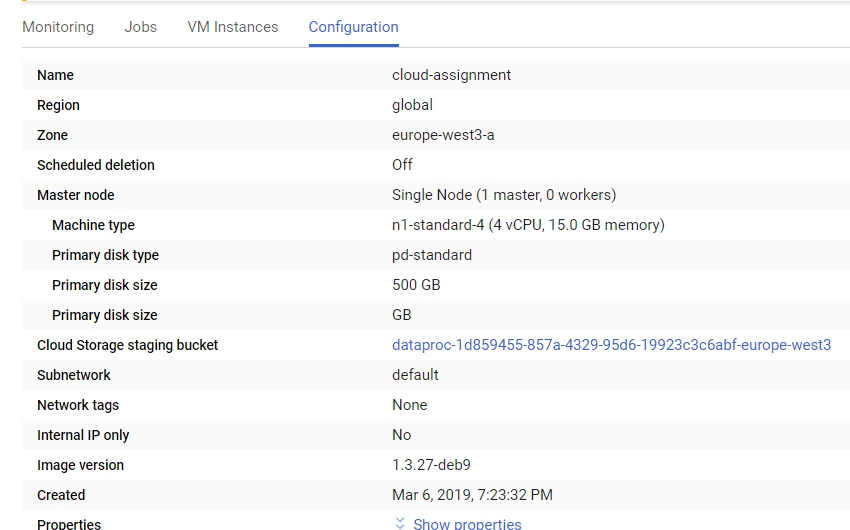


# Cluster creation in Google Dataproc:

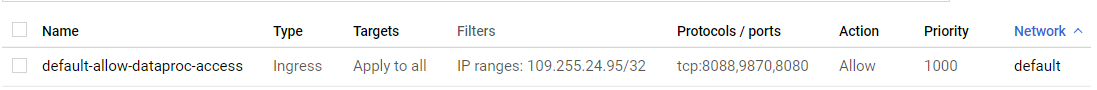
* Create account in Google Cloud Platform and once created, go to console section.
* On left Tab pane, go to BIG DATA 🡪 Dataproc 🡪 Clusters . Then Enable API and click on ‘Create Cluster’



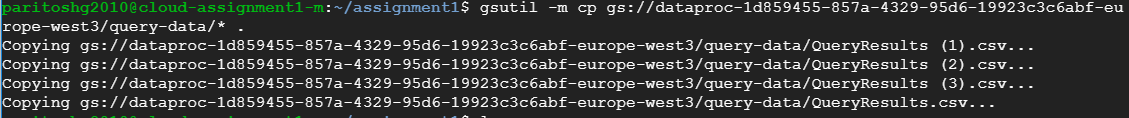
* Below configuration details need to be filled to create single node cluster.



* Create the Firewall rule from **VPC Network** > **Firewall rules.**



* Copy the query extracted files in Google cloud storage by drag and drop.
* Files fetched from google bucket to VM as below. Creating directory in Hadoop file system and further copying these files from VM to hdfs.

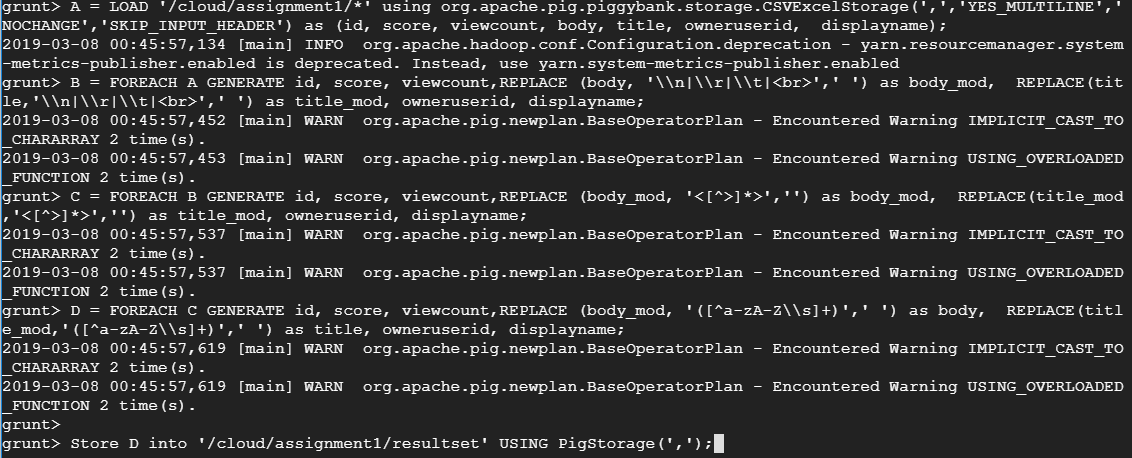




# Task 2 - PIG ETL:

Using pig or mapreduce, extract, transform & load the data as applicable

* To load the data in pig variable, UDF from piggybank has been used to fetch the csv data with multiline for a row, skipping the header.
* Sequence of pig commands has been executed to clean the data, such as removing html tags, new line and all type of special character from title & body (post) column.
* Once the data is cleaned, data has been stored with comma delimiter in output file.



# Task 3 - HIVE Queries:

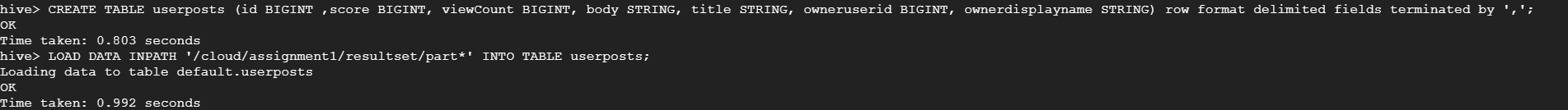
# 

* Create table in Hive.

***CREATE TABLE userposts (id BIGINT ,score BIGINT, viewCount BIGINT, body STRING, title STRING, owneruserid BIGINT, ownerdisplayname STRING) row format delimited fields terminated by ',';***

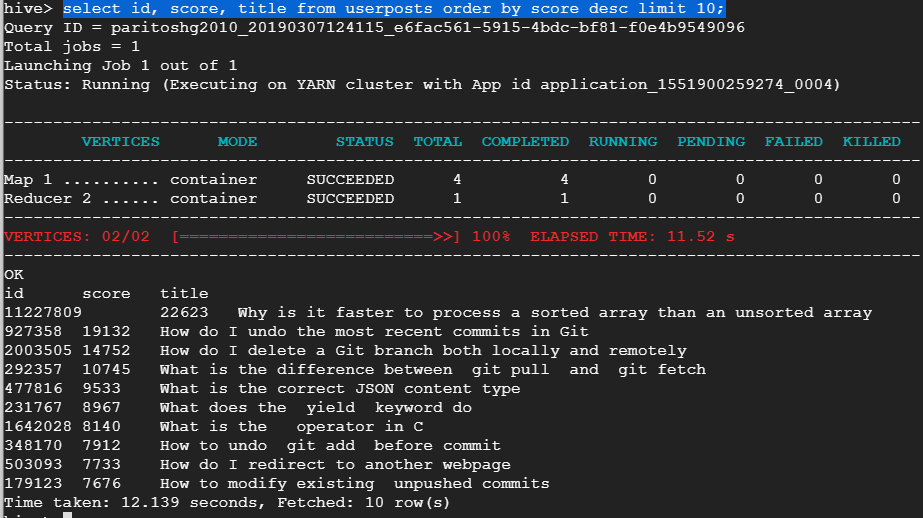
* Load the extracted data from pig queries into hive table.

***LOAD DATA INPATH '/cloud/assignment1/resultset/part\*' INTO TABLE userposts;***



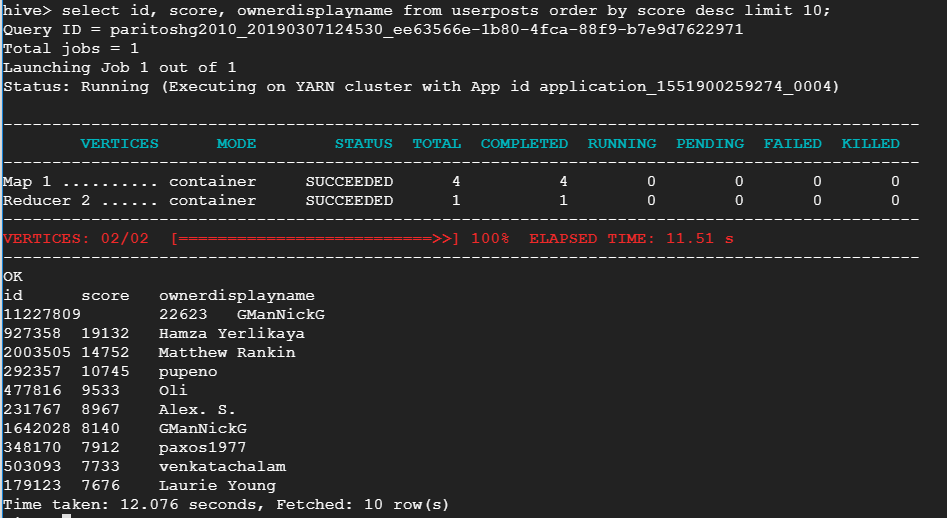
3.1- The top 10 post by score.

**select id, score, title from userposts order by score desc limit 10;**



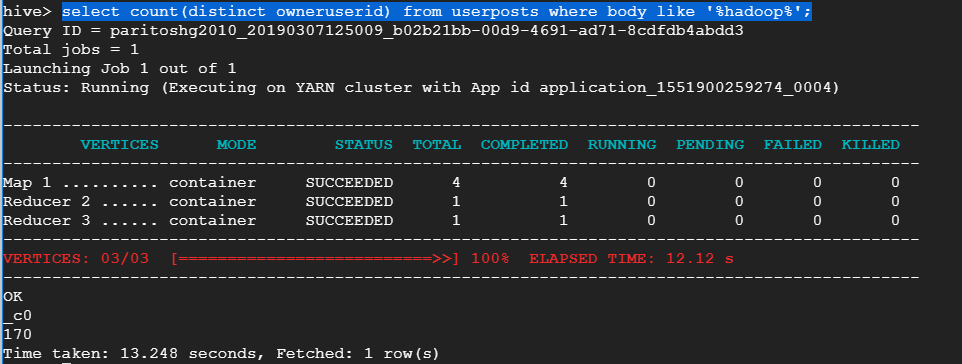
3.2- The top 10 users by post score- The top 10 users of the post with highest score has been selected.

**select owneruserid, ownerdisplayname, score from userposts order by score desc limit 10;**



3.3- The number of distinct users, who used the word ‘hadoop’ in one of their posts.

**select count(distinct owneruserid) from userposts where body like '%hadoop%'**



# Task 4 – TFIDF Calculation:

Map reduce paradigm was used to find TFIDF for the posts of top 10 users by score. Below are the sequence of commands were executed to create the temporary macro and views for calculation of term frequency and inverse document frequency for identified users. Below is the screenshot for calculated TFIDF for some of the users.

