## **CA675- Cloud Technologies**

|  |  |
| --- | --- |
| Name | Harsha Kumble |
| Student ID | 21262485 |
| Email | harsha.kumble3@mail.dcu.ie |
| Git repository Link | <https://github.com/kumbleh3/CA675-assignment1> |

**Task 1:**

**Acquire the Dataset**

Acquired 2,000,00 records from the stack exchange using <http://data.stackexchange.com/stackoverflow/query/new>

*select top 50000 \* from posts where posts.ViewCount > 90000 ORDER BY ViewCount DESC;*

*select top 50000 \* from posts where posts.ViewCount <= 127754 ORDER BY ViewCount DESC;*

*select top 50000 \* from posts where posts.ViewCount <= 74786 ORDER BY ViewCount DESC;*

*select top 50000 \* from posts where posts.ViewCount <= 53348 ORDER BY ViewCount DESC;*

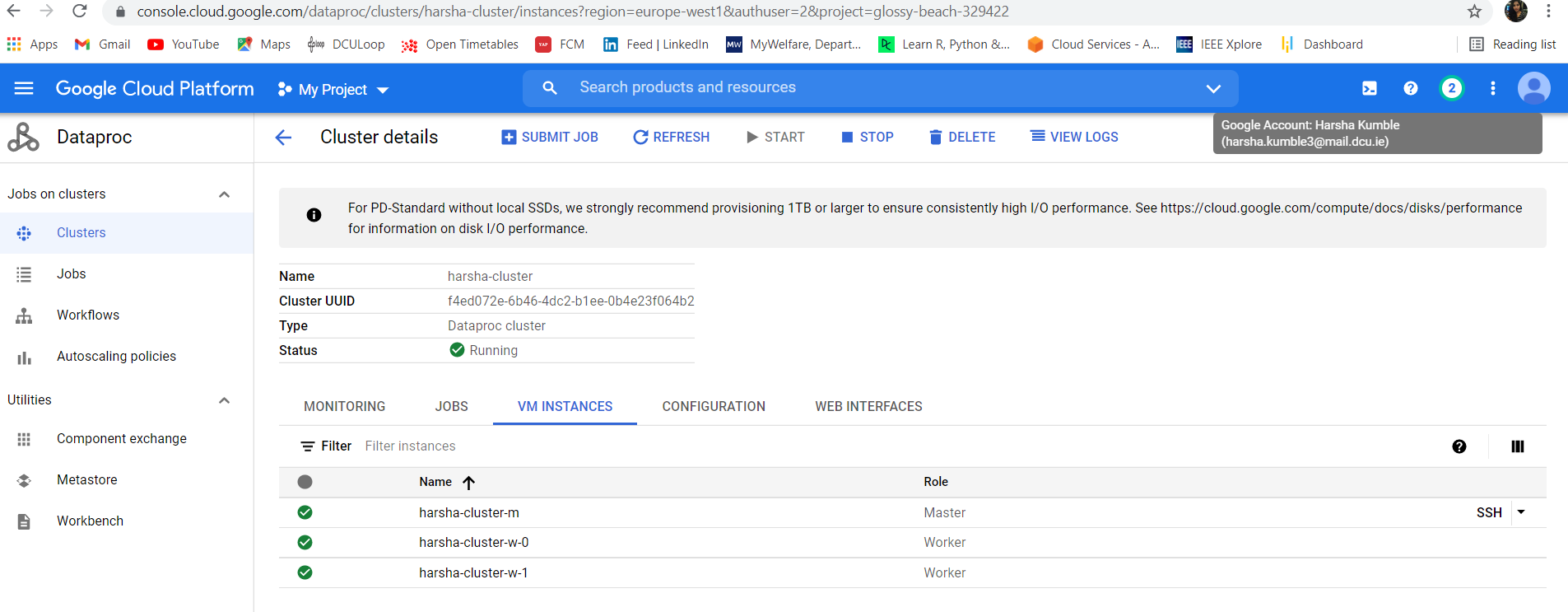
**Technologies Used:**

* Google Cloud Platform
* Hadoop
* Pig
* Hive

**Task 2.**

**Configuration of GCP:**

Used Dataproc and createdVM Instances with one master node and two worker nodes architecture. GCP is used here because it does not require any pre-installations for Hadoop, hive and Pig which makes it easier to work with.



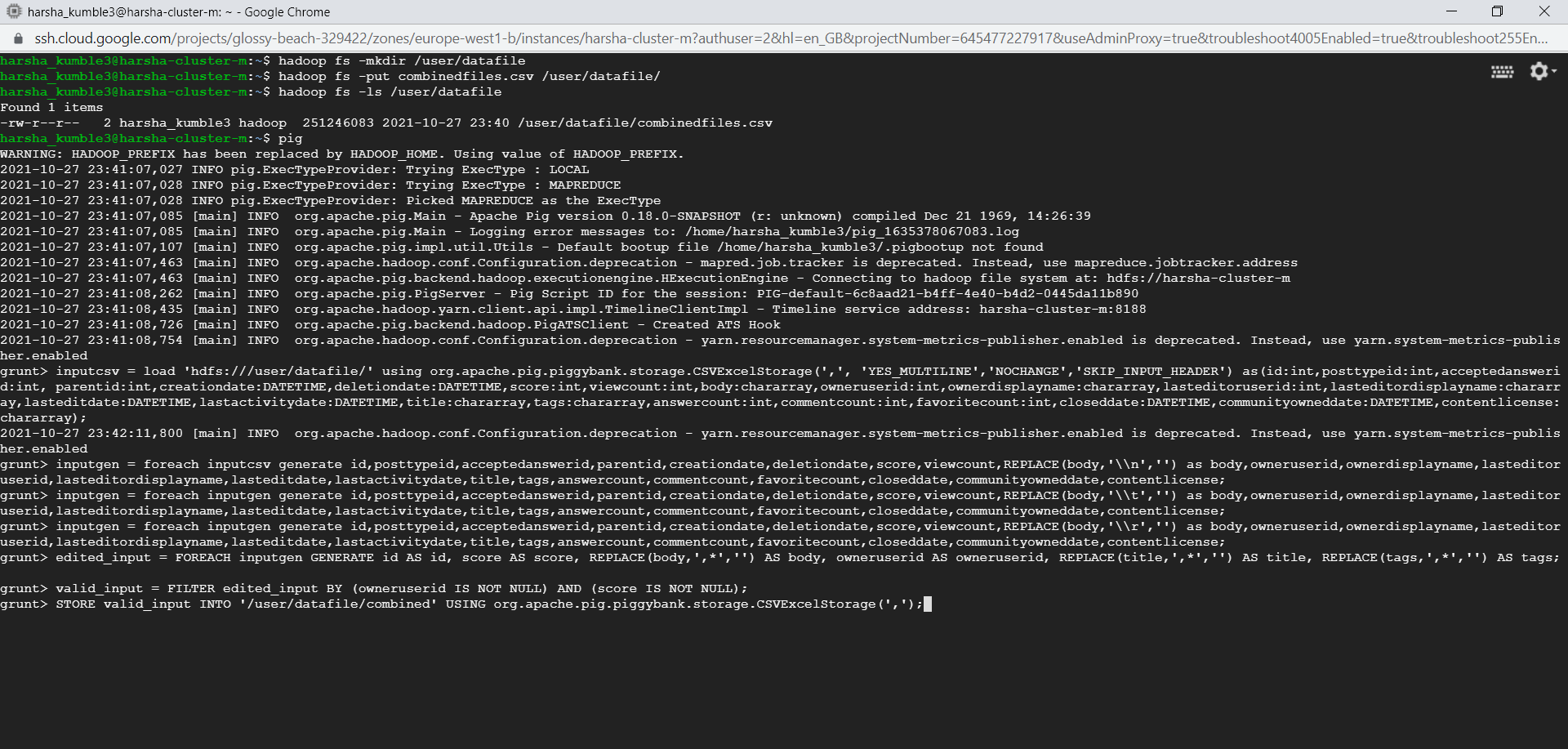
**Data Processing:**

Loaded combinedcsv file in VM, created a new directory /user/datafile in HDFS and loaded dataset to HDFS using PUT statement.

*hadoop fs -put combinedfiles.csv /user/datafile/*

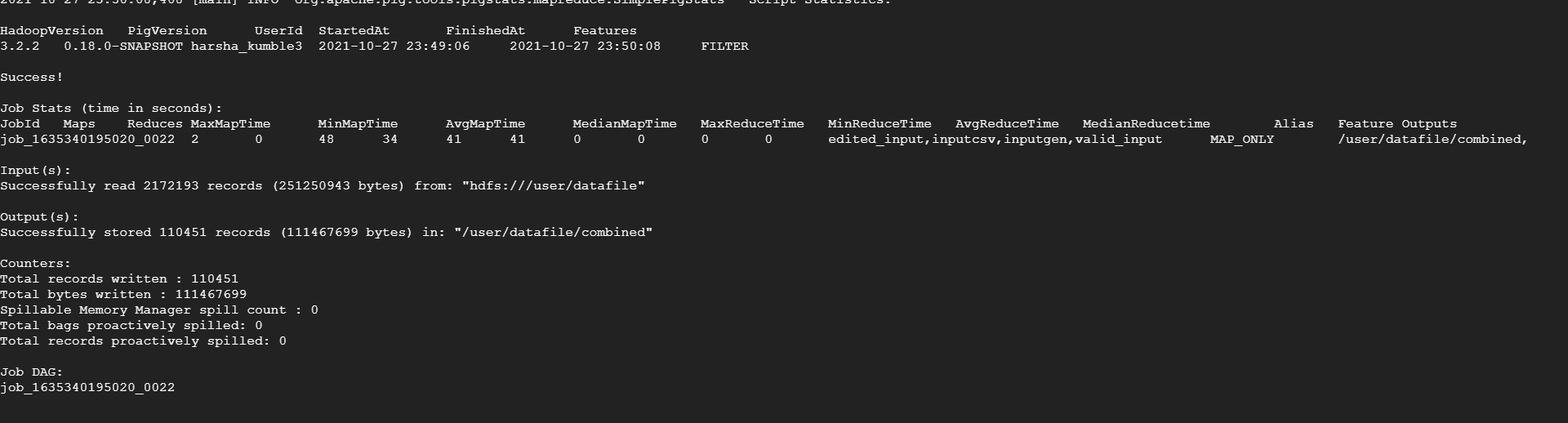
Pig is used to load, clean the data and remove null values. Pig is used for ETL processing because it is very similar to SQL and reduces effort of writing java based mapreduce codes.

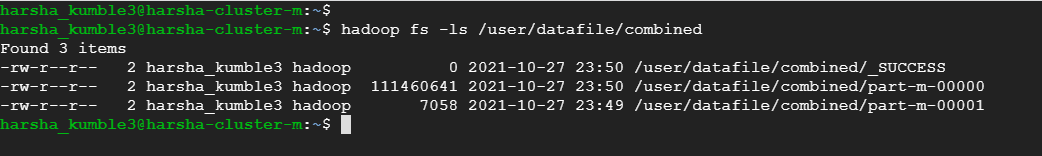
*loadposts = load 'hdfs:///user/harsha/' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE','NOCHANGE','SKIP\_INPUT\_HEADER') as(id:int,posttypeid:int,acceptedanswerid:int, parentid:int,creationdate:DATETIME,deletiondate:DATETIME,score:int,viewcount:int,body:chararray,owneruserid:int,ownerdisplayname:chararray,lasteditoruserid:int,lasteditordisplayname:chararray,lasteditdate:DATETIME,lastactivitydate:DATETIME,title:chararray,tags:chararray,answercount:int,commentcount:int,favoritecount:int,closeddate:DATETIME,communityowneddate:DATETIME,contentlicense:chararray);*



The cleaned csv file is stored in HDFS using below command

*store valid\_posts into '/user/harsha/combined' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',');*



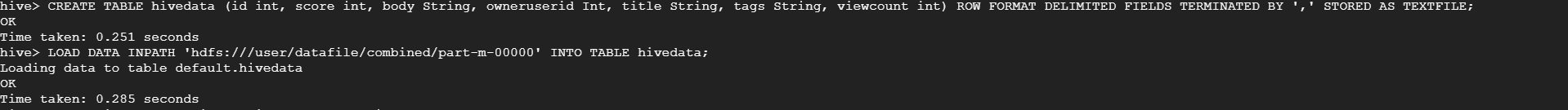


**CREATING TABLE IN HIVE**

Loaded output from Pig into table created using hive.

*CREATE TABLE hivedata (id int, score int, body String, owneruserid Int, title String, tags String, viewcount int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;*

*LOAD DATA INPATH 'hdfs:///user/harsha/combined/part-m-00000' INTO TABLE hivedata;*

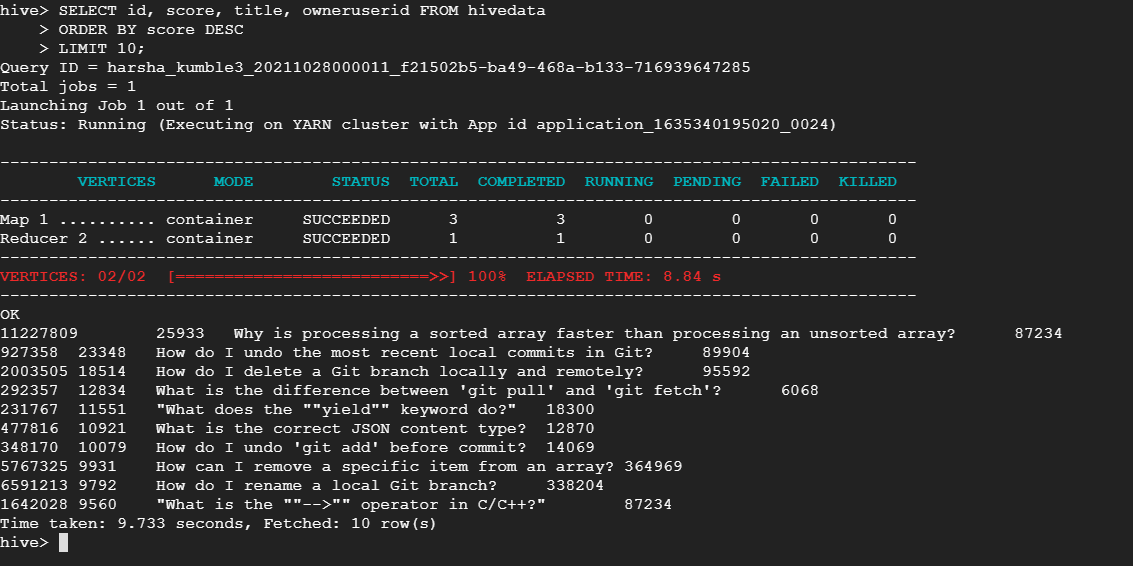


**Task 3: Query data using Hive**

**Results:**

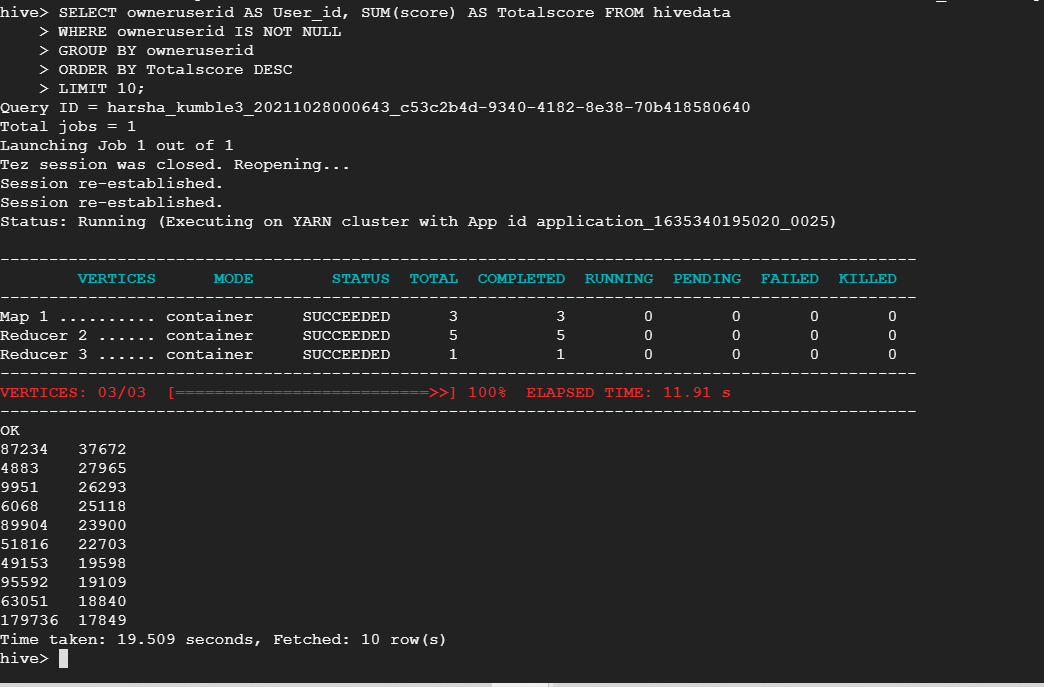
1. Top 10 Posts by Score

*SELECT id, score, title, owneruserid FROM hivedata ORDER BY score DESC LIMIT 10;*



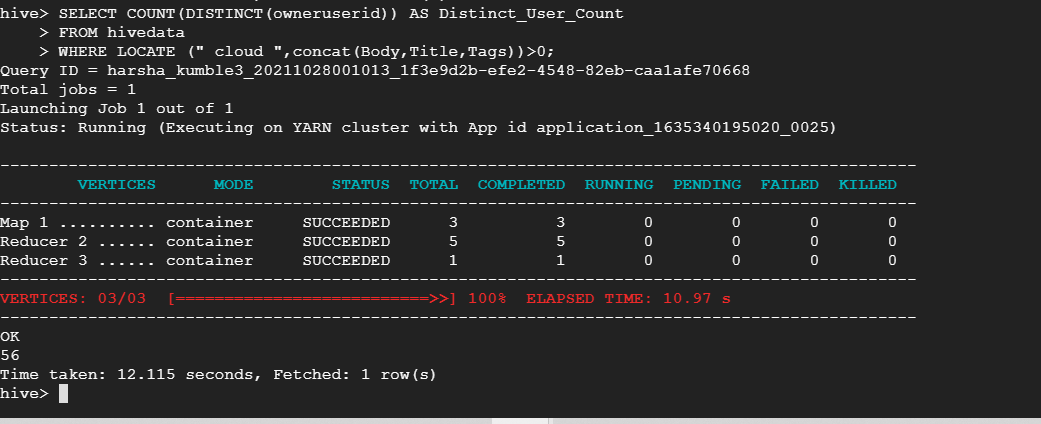
1. The top 10 users by post score

*select owneruserid,sum(score) as Total\_Score from hivedata group by owneruserid order by Total\_Score desc limit 10;*



1. The number of distinct users, who used the word “cloud” in one of their posts.

*select count(distinct(owneruserid)) AS Distinct\_User\_Count from hivedata WHERE locate (" cloud ",concat(Body,Title,Tags))>0;*



**References:**

<https://github.com/arunabellgutteramesh/PigHiveOnStackExchangeData/blob/master/code/ETL/fetchValidRecords.pig> <https://github.com/arunabellgutteramesh/PigHiveOnStackExchangeData/blob/master/code/Querying/hiveQueries.sql>