CLOUD ASSIGNMENT-1

# Name: Pallavi Prakash||Email: [pallavi.prakash3@mail.dcu.ie](mailto:pallavi.prakash3@mail.dcu.ie)

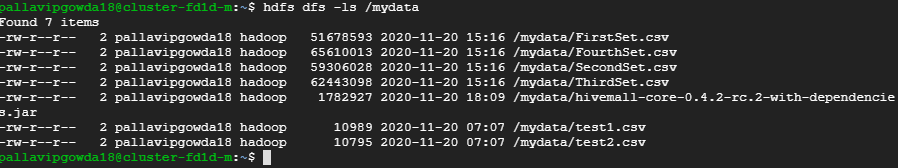
# This document contains only the tasks, steps followed and screenshots of the output. Detailed explanation of all the tasks, and the source code are available on GitHub: <https://github.com/pallavi-18-p/StackExchangeDataUsingPigHive>

# Task 1: Acquire the top 200,000 posts by ViewCount.

Five set of queries were used to fetch 2,00,000 posts, the fifth set consisted of 205 records that was used for testing purposes initially.

**Task 2: Using Pig or MapReduce, extract, transform and load the data as applicable**

I used pig to extract, transform and load the data as pig requires only few lines of code to implement, whereas with MapReduce, the lines of code is approximately 20 \*pigcode.

**HDFS:** Uploaded all the four datasets to GCP, created a directory on hdfs. Copied all the four files to the ‘myData’ directory. 

**PIG**

**Step 1: All the four datasets were loaded to Pig using LOAD operator and CSVExcelStorage function.**

**Step 2: Merged the datasets into one single file using UNION operator.**

**Step3: Removed all duplicate records using DISTINCT operator.**

**Step4: Removed all the records with blank OwnerUserId and Score using FILTER clause.**

**Optional Step: Counted the number of records after cleaning to verify the record count.**

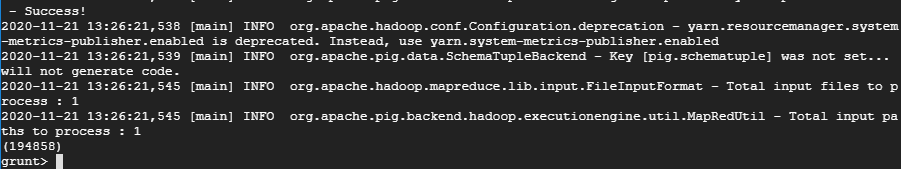
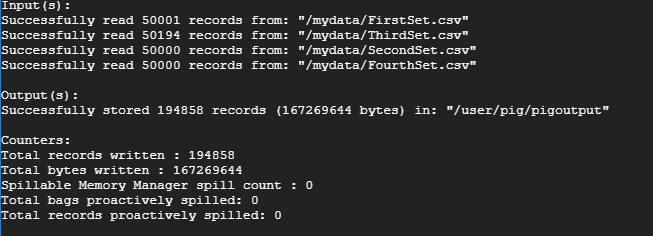
****

Fig: Record Count

**Step5: Removed all html tags, line breaks, commas, and special characters using REPLACE function.**

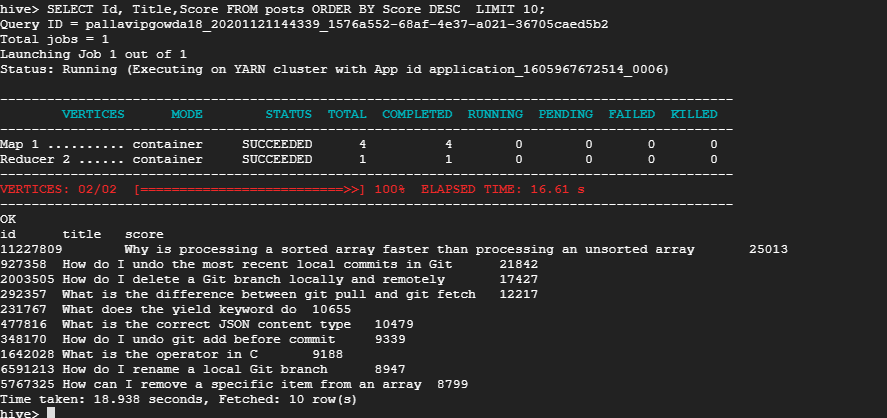
**Step6: Stored the output to HDFS using STORE operator and CSVExcelStorage function.**



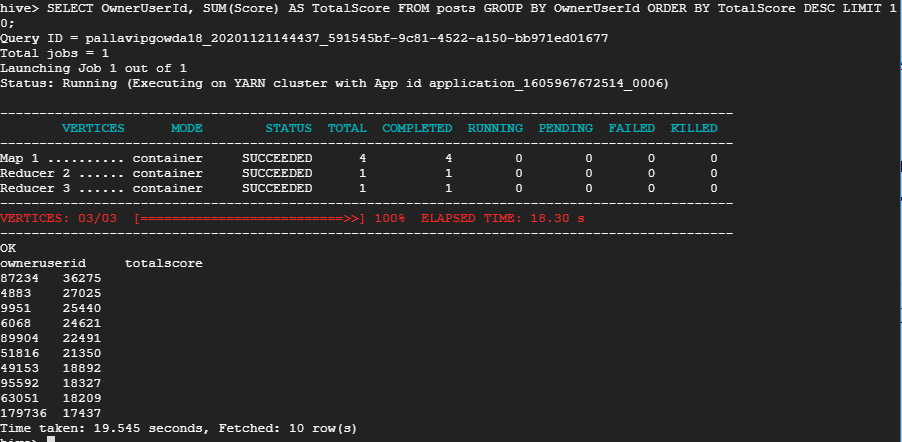
# Task3: Using Hive and/or MapReduce:

**Solution:** I have used HIVE to accomplish this task as I have experience in SQL. This was also considering the advantages of HIVE, which requires very few lines of code, when compared to MapReduce.

**1: The top 10 posts by score**



**2: The top 10 users by post score**



**3: The number of distinct users who used the word “Hadoop” in one of their posts**

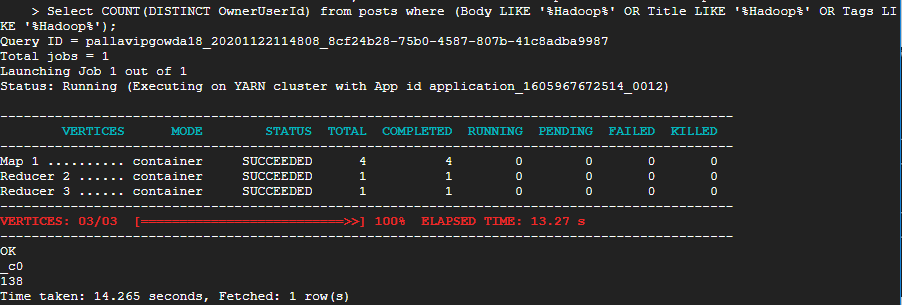


Fig: Hadoop count = 138

The count of Hadoop used by users in one of their posts when ignoring case gives 346.

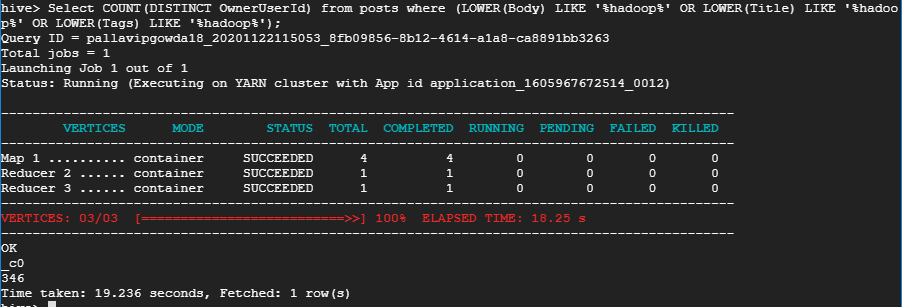


Fig: hadoop count=346

# Task4: Using Mapreduce/Pig/Hive calculate the per-user TF-IDF (just submit the top 10 terms for each of the top 10 users from Query 3.II)

**Solution:** I have used Hive to calculate tfidf of top 10 terms of top 10 users.

**Step1: Created a table to store the result of second query of the previous task.**

**Step2: Extracted Body column from the previously created table into another one for all the top ten**

**users.**

**Step 3: TFIDF calculation using Hivemall**

Downloaded Hivemall jar and define-all.hive source files from GitHub(<https://github.com/myui/hivemall/wiki/Installation>) Removed *drop temporary function sha1;* and *create temporary function sha1 as 'hivemall.ftvec.hashing.Sha1UDF';* lines from the define-all.hive file as these two lines were causing error during execution. Uploaded to GCP and copied these files to HDFS.

**Step4: Query to fetch top ten tfidfs for all the top ten users. The query fetched total of 100 records (top ten terms for all the ten users)**

