CA-675 || CLOUD TECHNOLOGY || ASSIGNMENT 1 – DATA ANALYSI

NAME	ANKIT SHARMA
STUDENT ID	20211119
EMAIL	ankit.sharma28@mail.dcu.ie
PROGRAMME OF STUDY	MSC. IN COMPUTING (DATA ANALYTICS)
GITHUB LINK	https://github.com/ankitapril/CA675-Assignment-1/tree/main

Task 1-Data Extraction
We must acquire top 2,00,000 posts from stack exchange

Task 2-Load Data
Data is load with hive

Task 3-Query with Hive

Top 10 posts by score? The top 10 users by post score? The number of distinct users, who used the word 'cloud' in one of their posts?

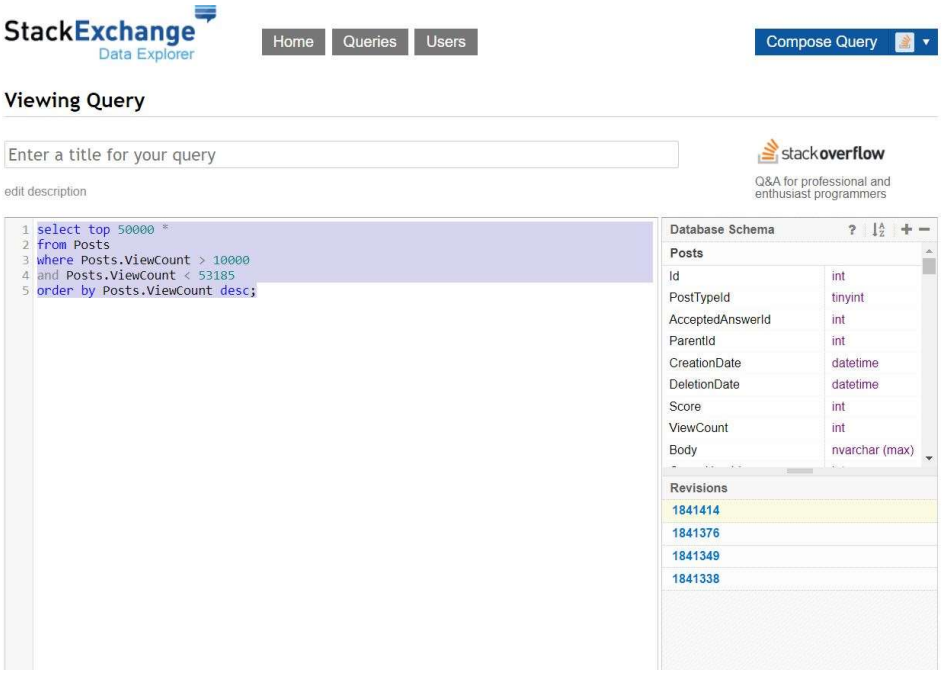
Task 4- Calculate TF-IDF with Hive

Find Top 10 terms used for each of the top 10 users by post score

Task 1-Extarcting Data

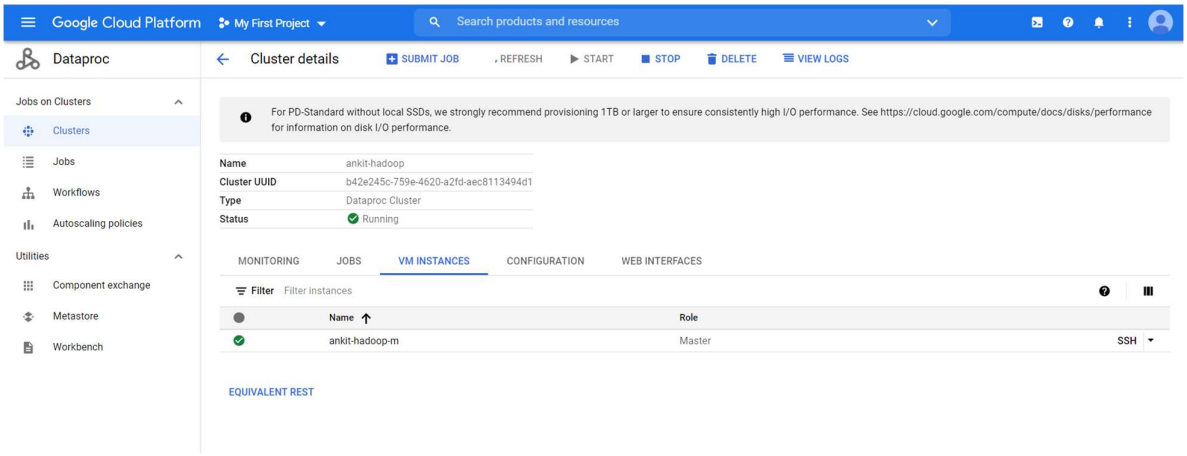
- Extracting top 2,00,000 posts from Stack Exchange
- Data Acquisition from Stack Exchange Data Explorer (SEDE).
- We can download only 50,000 posts at one time. So, we will run 4 query to extract data
- Below are the queries

- select top 50000 * from posts where Posts.ViewCount < 10030841 order by posts.ViewCount DESC;
- select top 50000 * from Posts where Posts.ViewCount >62000 and Posts.ViewCount < 127042 order by Posts.ViewCount desc;
- select top 50000 * from Posts where Posts.ViewCount >20000 and Posts.ViewCount < 74480 order by Posts.ViewCount desc;
- select top 50000 * from Posts where Posts.ViewCount > 10000 and Posts.ViewCount < 53185 order by Posts.ViewCount desc;

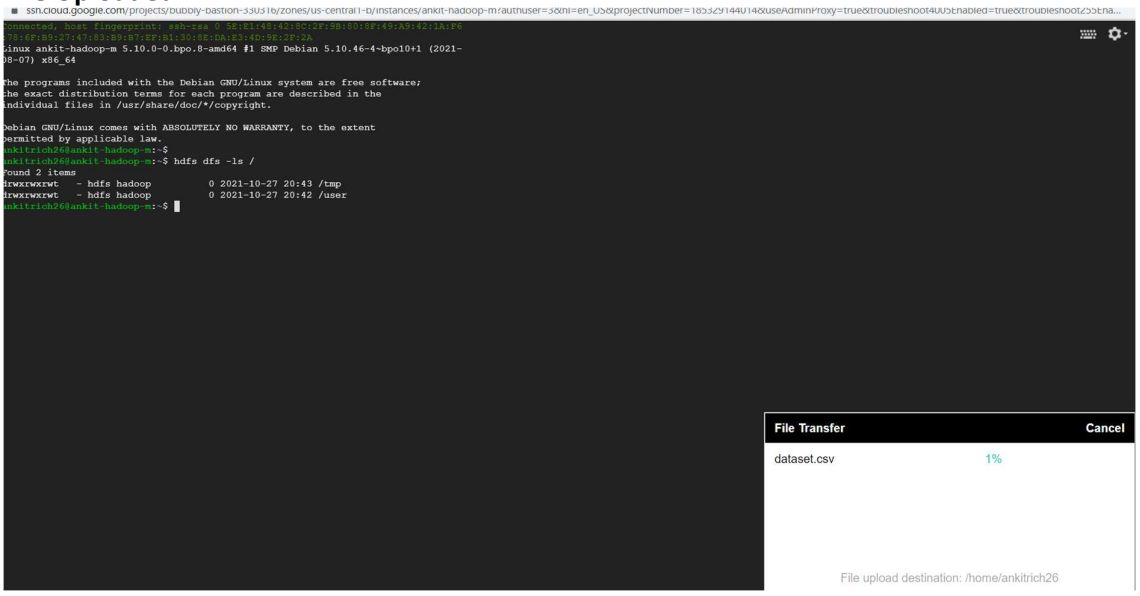


Task 2-Load Data

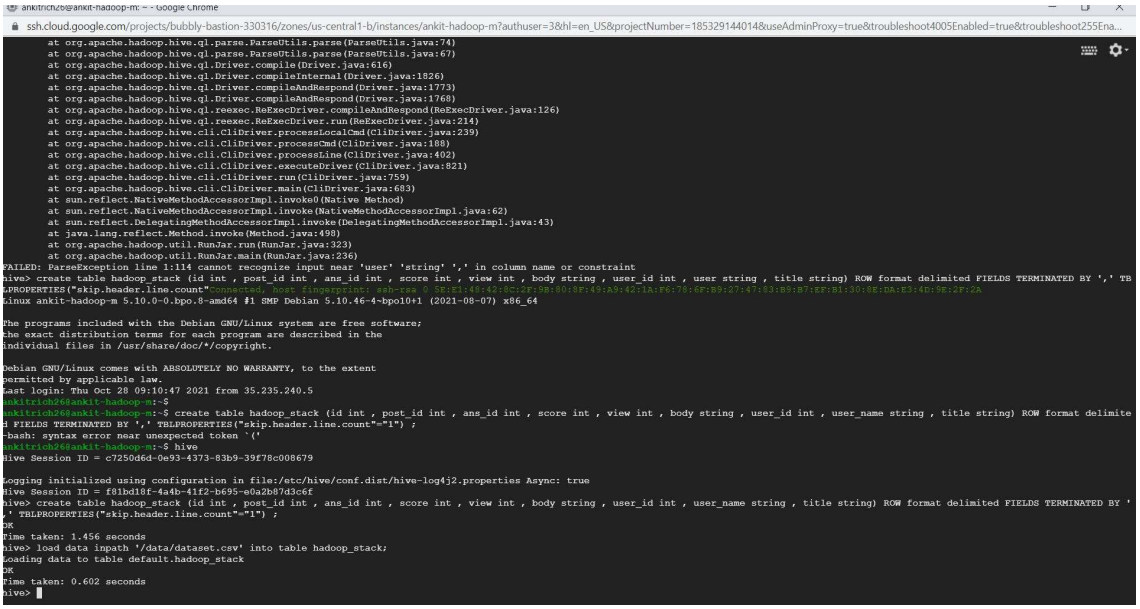
For loading data. Cluster is created in GCP and cleaned csv file is uploaded and then data is loaded



File Uploaded



Data loaded through Hive



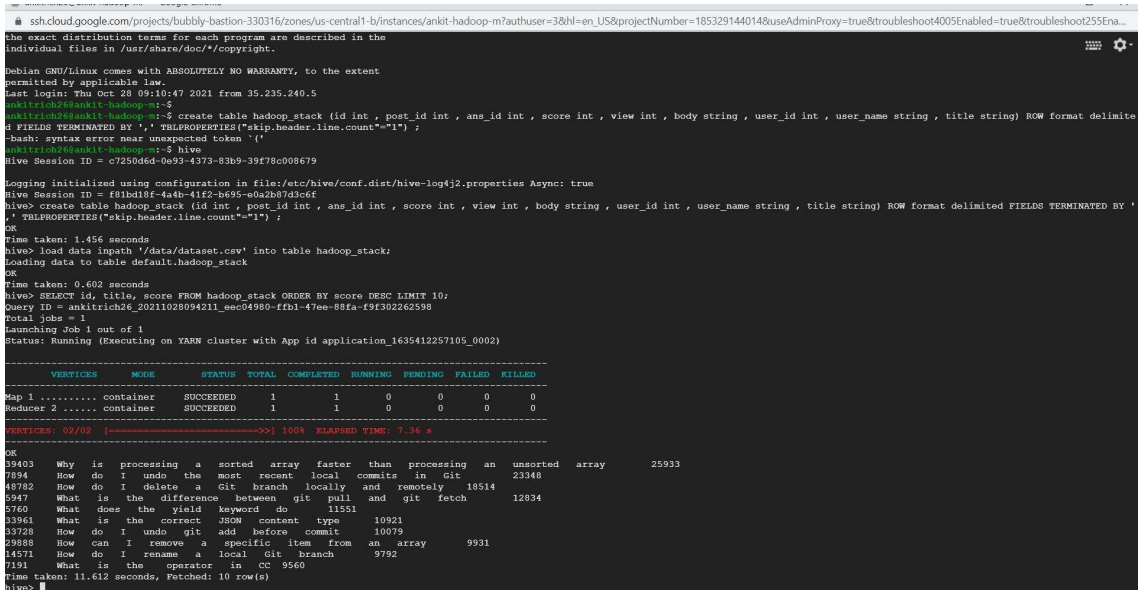
Task 3- Query with Hive

- Hive->create Database->ankit_hadoop->create table->hadoop_stack….(so on)

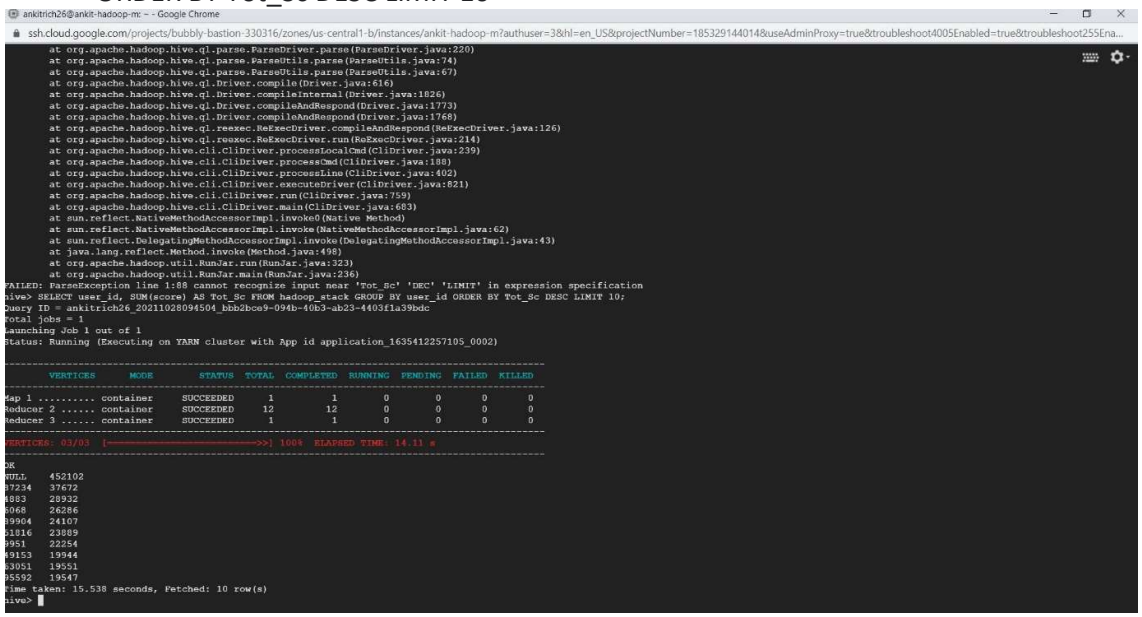
```
-rw-r--r--   1 ankitrich26 hadoop  283820604 2021-10-28 09:22 /data/dataset.csv
ankitrich26@ankit-hadoop-m:~$ hive;
Hive Session ID = 50af93bd-e53e-44d2-9e02-a69123bf551d

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = bbfb1ebe-d0bc-45f0-86ec-4a8bc99f0d78
hive> show databases;
OK
default
Time taken: 1.153 seconds, Fetched: 1 row(s)
hive> create database
Display all 633 possibilities? (y or n)
hive> create database ankit_hadoop;
OK
Time taken: 1.201 seconds
hive> create table hadoop_stack (id int, post_id int,ans_id int,score int, view int,body string, user_id int,user string, title string) ROW format delimited FIELDS TERMINATED BY ',' TBLPROPERTIES(
"skip.header.line.count"="1") ;
```

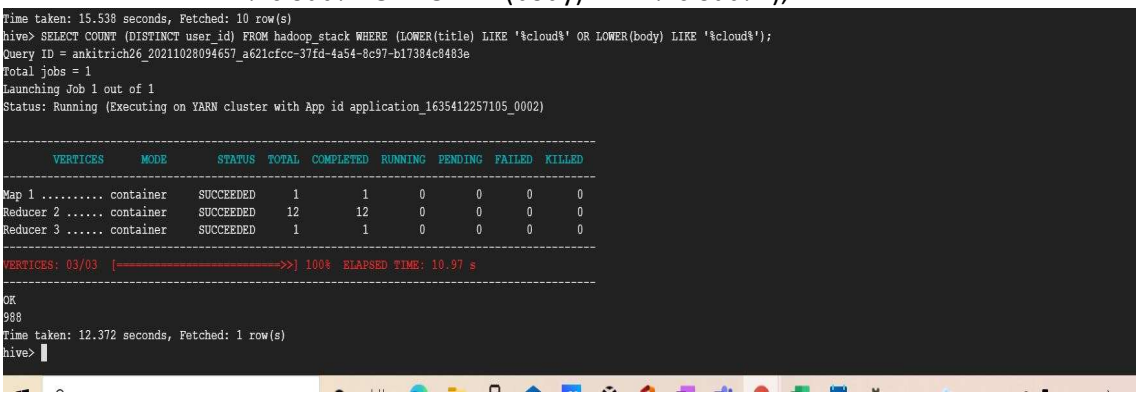
- 3.1-Top 10 Post by Score
    - SELECT id, title, score FROM hadoop_stack ORDER BY score DESC LIMIT 10;



- 3.2-Top 10 Users by Post Score
- SELECT user_id, SUM(score) AS Tot_Sc FROM hadoop_stack GROUP BY user_id ORDER BY Tot_Sc DESC LIMIT 10



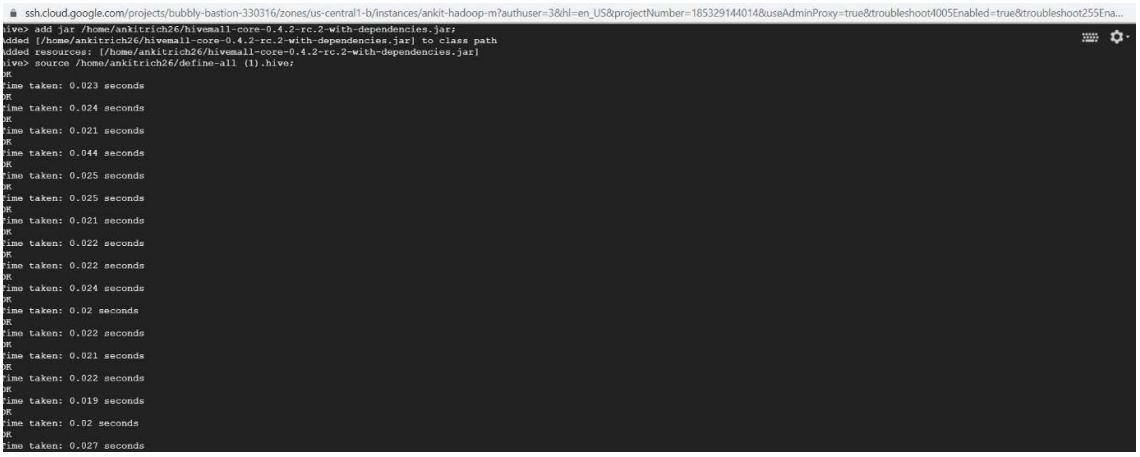
- 3.3- The number of distinct users, who used the word ‘cloud’ in one of their posts?
    - SELECT COUNT (DISTINCT user_id) FROM hadoop_stack WHERE (LOWER(title) LIKE '%cloud%' OR LOWER(body) LIKE '%cloud%');



Task 4-Calculate TF-IDF with Hive

Through Hive mall TF-IDF is calculated

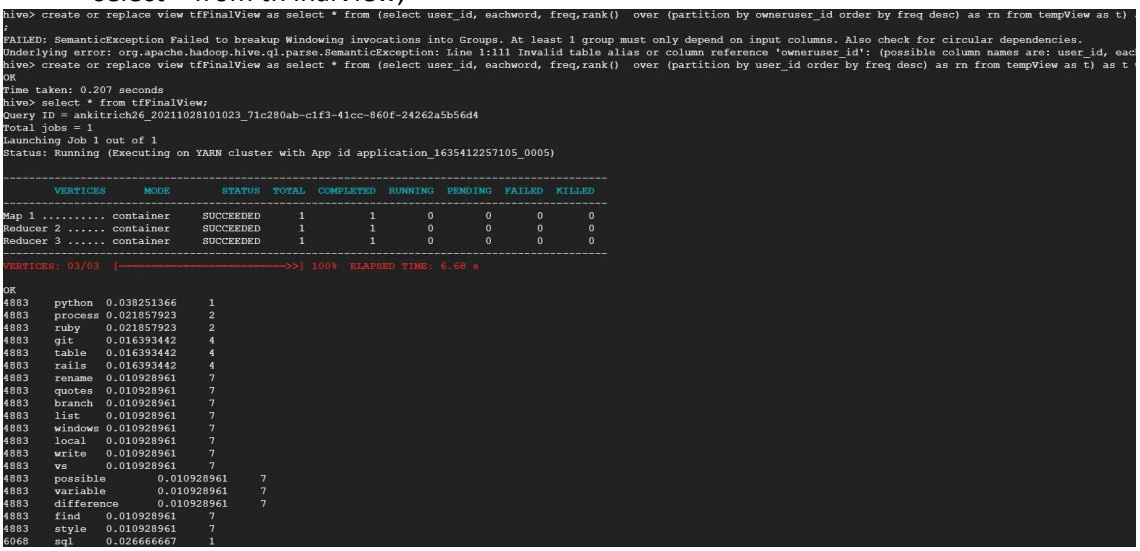
Jar file is added



- Below are the queries to calculate TF-IDF

- create temporary macro max2(x INT, y INT) if(x>y,x,y);
- create temporary macro tfidf(tf FLOAT, df_t INT, n_docs INT) tf * (log(10, CAST(n_docs as FLOAT)/max2(1,df_t)) + 1.0);
- create table distOwnerIDs as SELECT user_id, SUM(score) AS TotalScore FROM hadoop_stack GROUP BY user_id ORDER BY TotalScore DESC LIMIT 10;
- create table mainUSRData as Select HT.user_id,title from hadoop_stack HT JOIN distOwnerIDs DO on HT.user_id = DO.user_id
- create or replace view mainUSRView as select user_id, eachword from mainUSRData LATERAL VIEW explode(tokenize(title, True)) t as eachword where not is_stopword(eachword);

- create or replace view tempView as select user_id, eachword, freq from (select user_id, tf(eachword) as word2freq from mainUSRView group by user_id) t LATERAL VIEW explode(word2freq) t2 as eachword, freq;
- create or replace view tfFinalView as select * from (select user_id, eachword, freq,rank() over (partition by user_id order by freq desc) as rn from tempView as t) as t where t.rn<=10 ;
- select * from tfFinalView;



References:

1.) https://www.tutorialspoint.com/hive/index.htm