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

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

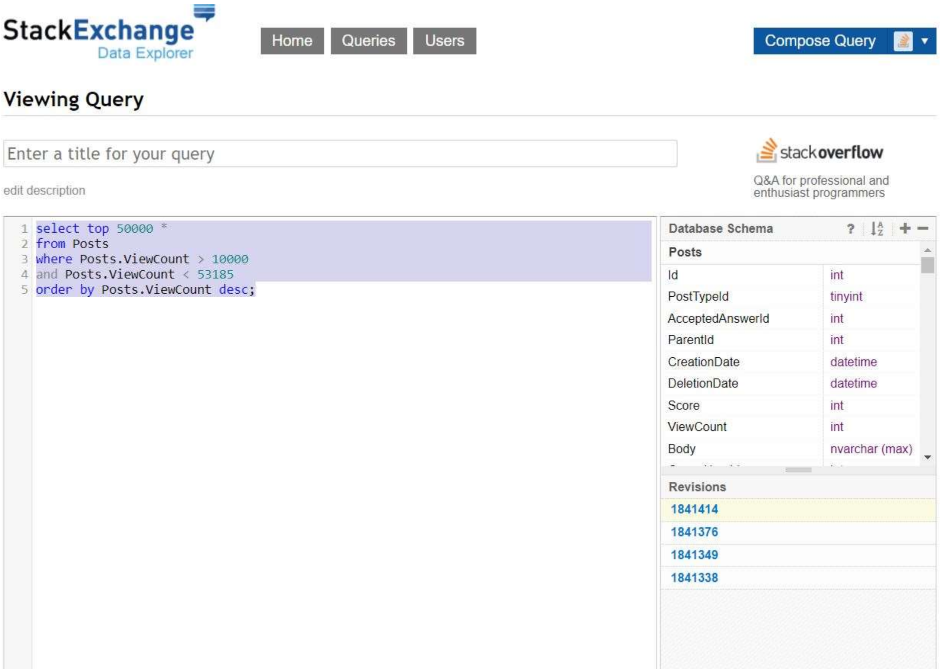
Top 10 posts by score? The top 10 users by post score? The number of distinct users, who used the word 'cloud' in one of their posts?

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;



```
select top 50000 *
from Posts
where Posts.ViewCount > 10000
and Posts.ViewCount < 53185
order by Posts.ViewCount desc;
```

Database Schema

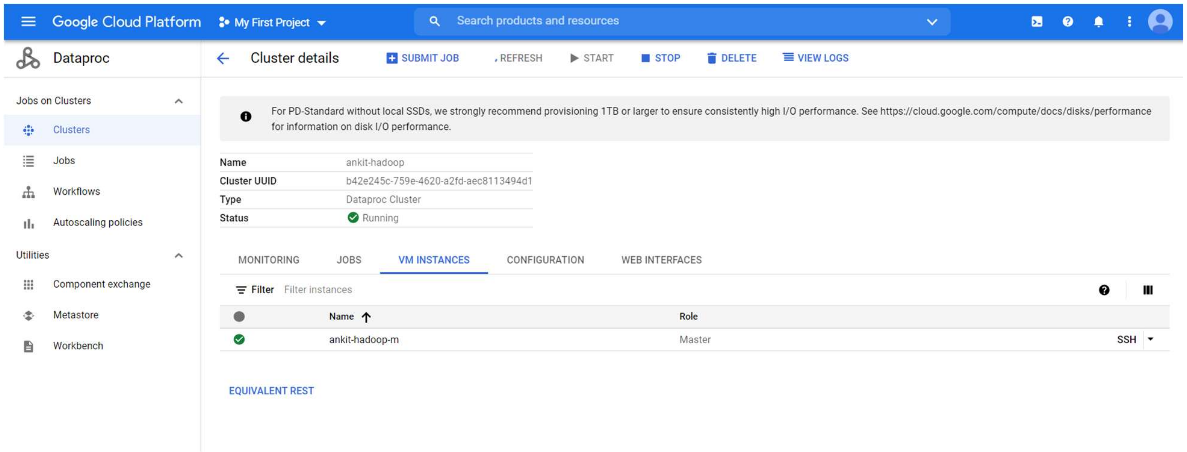
Posts	
Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar (max)

Revisions

- 1841414
- 1841376
- 1841349
- 1841338

Task 2-Load Data

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

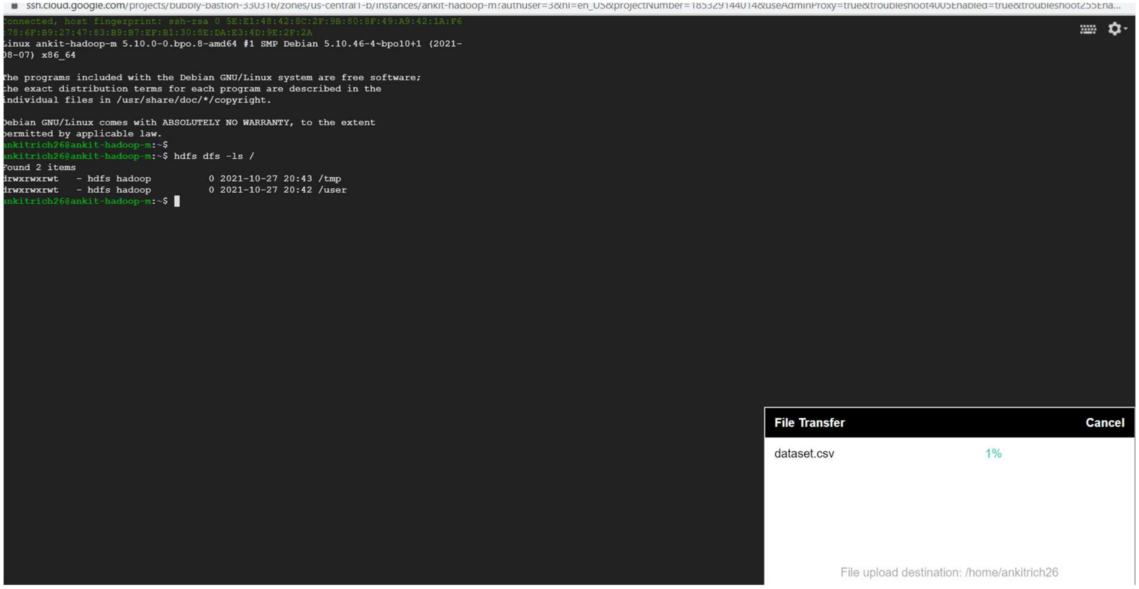


For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

Name	ankit-hadoop
Cluster UUID	b42e245c-759e-4620-a2fd-aec8113494d1
Type	Dataproc Cluster
Status	Running

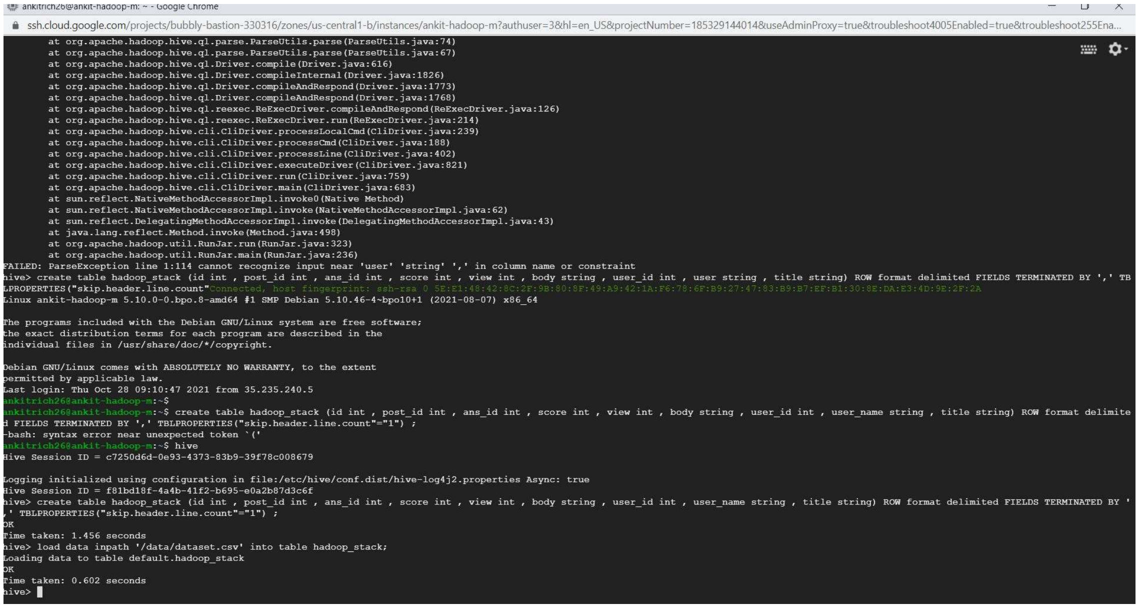
	Name	Role	
✓	ankit-hadoop-m	Master	SSH

File Uploaded


ssh.cloud.google.com/projects/bubbly-bastion-330316/zones/us-central1-b/instances/ankit-hadoop-m?authuser=3&hl=en_US&projectNumber=185329144014&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Ena...
File Transfer
Cancel
dataset.csv
1%
File upload destination: /home/ankitrich26

```
Connected, host fingerprint: ssh-rsa 0 5E:E1:48:42:8C:2F:9B:80:8F:49:A9:42:1A:F6
:78:6F:B9:27:47:83:B9:B7:EF:B1:30:8E:DA:E3:4D:9E:2F:2A
Linux ankit-hadoop-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4~bpo10+1 (2021-
08-07) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
ankitrich26@ankit-hadoop-m:~$
ankitrich26@ankit-hadoop-m:~$ hdfs dfs -ls /
Found 2 items
drwxrwxrwt   - hdfs hadoop          0 2021-10-27 20:43 /tmp
drwxrwxrwt   - hdfs hadoop          0 2021-10-27 20:42 /user
ankitrich26@ankit-hadoop-m:~$
```

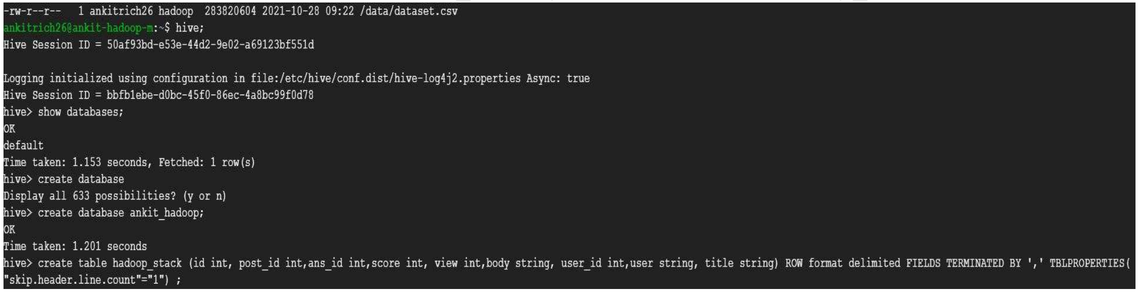
Data loaded through Hive


ankitrich26@ankit-hadoop-m: ~ - Google Chrome
ssh.cloud.google.com/projects/bubbly-bastion-330316/zones/us-central1-b/instances/ankit-hadoop-m?authuser=3&hl=en_US&projectNumber=185329144014&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Ena...

```
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:114 cannot recognize input near 'user' 'string' ',' in column name or constraint
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 ',' TB
LPROPERTIES("skip.header.line.count"Connected, host fingerprint: ssh-rsa 0 5E:E1:48:42:8C:2F:9B:80:8F:49:A9:42:1A:F6:78:6F:B9:27:47:83:B9:B7:EF:B1:30:8E:DA:E3:4D:9E:2F:2A
Linux ankit-hadoop-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4~bpo10+1 (2021-08-07) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Oct 28 09:10:47 2021 from 35.235.240.5
ankitrich26@ankit-hadoop-m:~$
ankitrich26@ankit-hadoop-m:~$ create table hadoop_stack (id int , post_id int , ans_id int , score int , view int , body string , user_id int , user_name string , title string) ROW format delimite
d FIELDS TERMINATED BY ',' TBLPROPERTIES("skip.header.line.count"="1") ;
-bash: syntax error near unexpected token `('
ankitrich26@ankit-hadoop-m:~$ hive
Hive Session ID = c7250d6d-0e93-4373-83b9-39f78c008679

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = f81bd18f-4a4b-41f2-b695-e0a2b87d3c6f
hive> create table hadoop_stack (id int , post_id int , ans_id int , score int , view int , body string , user_id int , user_name string , title string) ROW format delimited FIELDS TERMINATED BY '
,' TBLPROPERTIES("skip.header.line.count"="1") ;
OK
Time taken: 1.456 seconds
hive> load data inpath '/data/dataset.csv' into table hadoop_stack;
Loading data to table default.hadoop_stack
OK
Time taken: 0.602 seconds
hive>
```

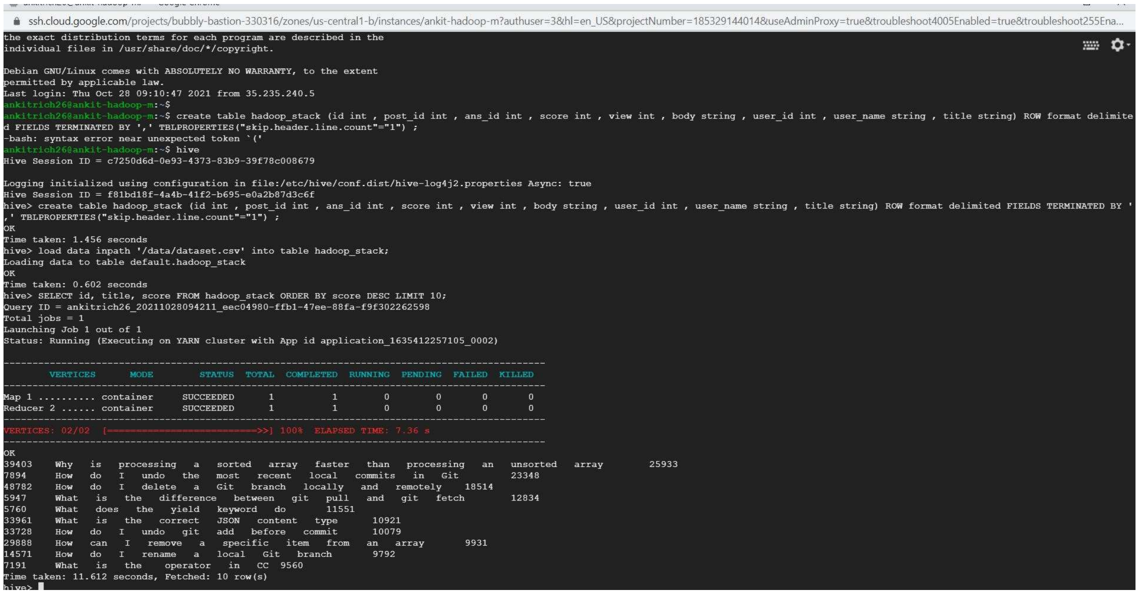
Task 3- Query with Hive

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



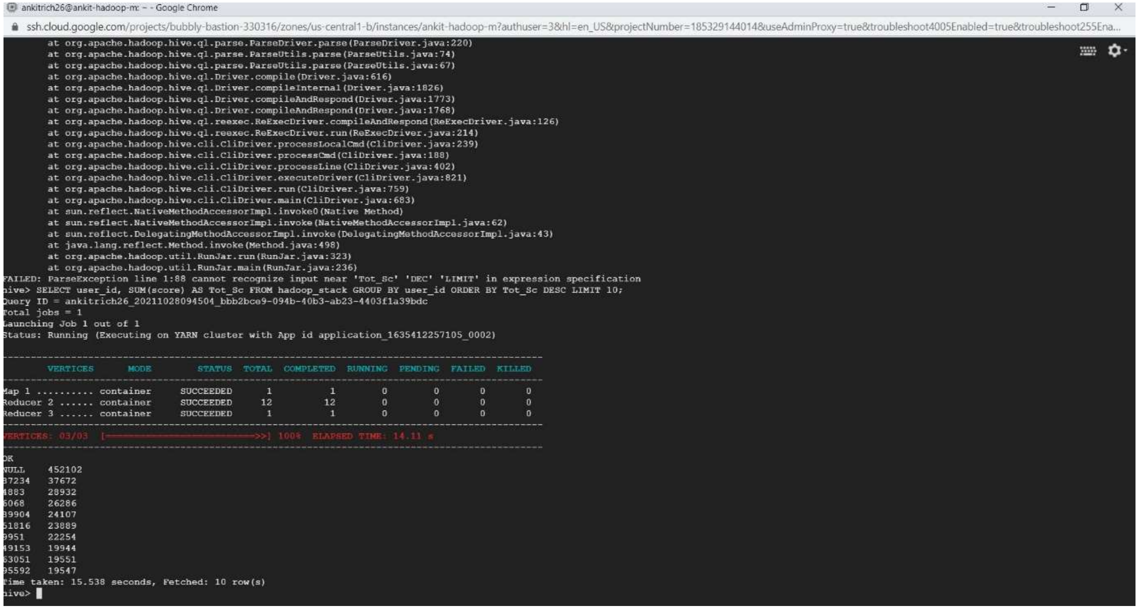
```
-rw-r--r--   1 ankitrich26 hadoop  283820604 2021-10-28 09:22 /data/dataset.csv
ankitrich26@ankit-hadoop-m:~$ hive;
Hive Session ID = 50af93bd-e53e-44d2-9e02-a69123bf551d

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;


ssh.cloud.google.com/projects/bubbly-bastion-330316/zones/us-central1-b/instances/ankit-hadoop-m?authuser=3&hl=en_US&projectNumber=185329144014&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Ena...

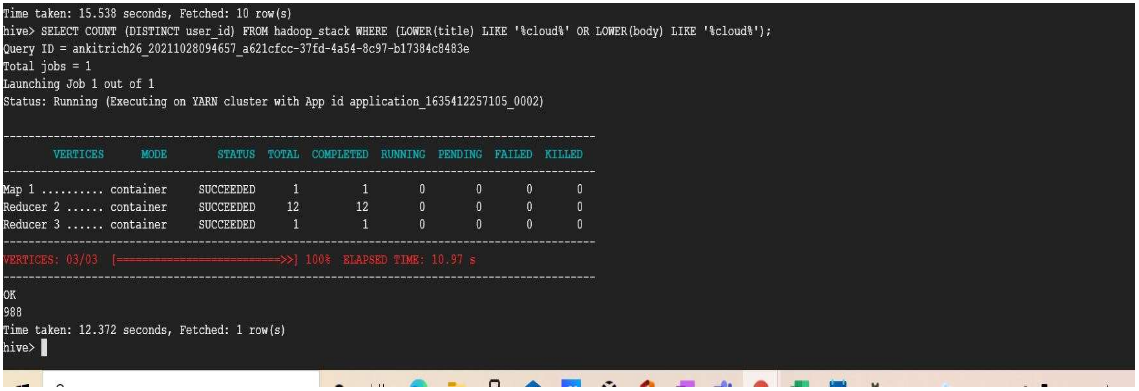
```
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Oct 28 09:10:47 2021 from 35.235.240.5
ankitrich26@ankit-hadoop-m:~$
ankitrich26@ankit-hadoop-m:~$ create table hadoop_stack (id int , post_id int , ans_id int , score int , view int , body string , user_id int , user_name string , title string) ROW format delimite
d FIELDS TERMINATED BY ',' TBLPROPERTIES("skip.header.line.count"="1") ;
-bash: syntax error near unexpected token `('
ankitrich26@ankit-hadoop-m:~$ hive
Hive Session ID = c7250d6d-0e93-4373-83b9-39f78c008679

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = f81bd18f-4a4b-41f2-b695-e0a2b87d3c6f
hive> create table hadoop_stack (id int , post_id int , ans_id int , score int , view int , body string , user_id int , user_name string , title string) ROW format delimited FIELDS TERMINATED BY '
,' TBLPROPERTIES("skip.header.line.count"="1") ;
OK
Time taken: 1.456 seconds
hive> load data inpath '/data/dataset.csv' into table hadoop_stack;
Loading data to table default.hadoop_stack
OK
Time taken: 0.602 seconds
hive> SELECT id, title, score FROM hadoop_stack ORDER BY score DESC LIMIT 10;
Query ID = ankitrich26_20211028094211_eec04980-ffb1-47ee-88fa-f9f302262598
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635412257105_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.36 s
----------------------------------------------------------------------------------------------
OK
39403   Why   is   processing   a   sorted   array   faster   than   processing   an   unsorted   array        25933
7894    How   do   I   undo   the   most   recent   local   commits   in   Git          23348
48782   How   do   I   delete   a   Git   branch   locally   and   remotely    18514
5947    What   is   the   difference   between   git   pull   and   git   fetch         12834
5760    What   does   the   yield   keyword   do        11551
33961   What   is   the   correct   JSON   content   type        10921
33728   How   do   I   undo   git   add   before   commit        10079
29888   How   can   I   remove   a   specific   item   from   an   array        9931
14571   How   do   I   rename   a   local   Git   branch        9792
7191    What   is   the   operator   in   CC   9560
Time taken: 11.612 seconds, Fetched: 10 row(s)
hive> 
```

- 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


ankitrich26@ankit-hadoop-m: ~ - Google Chrome
ssh.cloud.google.com/projects/bubbly-bastion-330316/zones/us-central1-b/instances/ankit-hadoop-m?authuser=3&hl=en_US&projectNumber=185329144014&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Ena...

```
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:220)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:88 cannot recognize input near 'Tot_Sc' 'DEC' 'LIMIT' in expression specification
hive> SELECT user_id, SUM(score) AS Tot_Sc FROM hadoop_stack GROUP BY user_id ORDER BY Tot_Sc DESC LIMIT 10;
Query ID = ankitrich26_20211028094504_bbb2bce9-094b-40b3-ab23-4403f1a39bdc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635412257105_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     12         12        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.11 s
----------------------------------------------------------------------------------------------
OK
NULL    452102
37234   37672
4883    28932
6068    26286
39904   24107
51816   23889
9951    22254
49153   19944
63051   19551
95592   19547
Time taken: 15.538 seconds, Fetched: 10 row(s)
hive> 
```

- 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%');

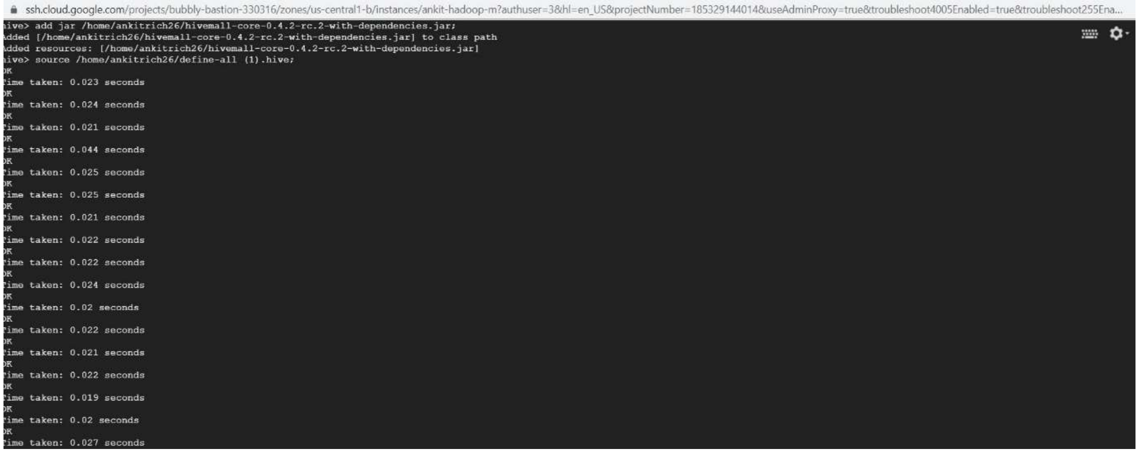


```
Time taken: 15.538 seconds, Fetched: 10 row(s)
hive> SELECT COUNT (DISTINCT user_id) FROM hadoop_stack WHERE (LOWER(title) LIKE '%cloud%' OR LOWER(body) LIKE '%cloud%');
Query ID = ankitrich26_20211028094657_a621cfcc-37fd-4a54-8c97-b17384c8483e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635412257105_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     12         12        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.97 s
----------------------------------------------------------------------------------------------
OK
988
Time taken: 12.372 seconds, Fetched: 1 row(s)
hive> 
```

Task 4-Calculate TF-IDF with Hive

Through Hive mall TF-IDF is calculated

Jar file is added



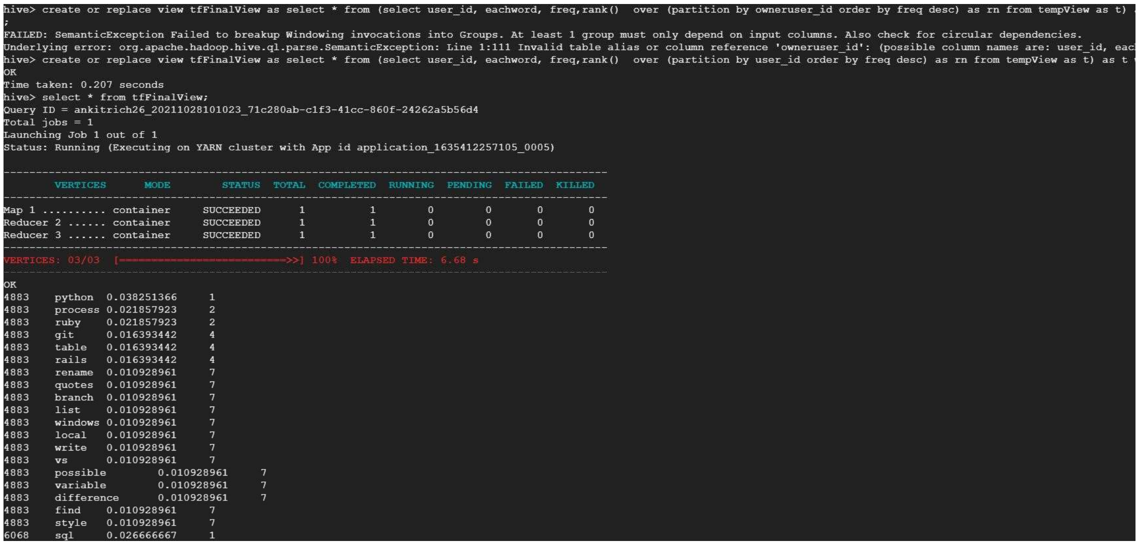
ssh.cloud.google.com/projects/bubbly-bastion-330316/zones/us-central1-b/instances/ankit-hadoop-m?authuser=3&hl=en_US&projectNumber=185329144014&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Ena...

```
ive> add jar /home/ankitrich26/hivemall-core-0.4.2-rc.2-with-dependencies.jar;
dded [/home/ankitrich26/hivemall-core-0.4.2-rc.2-with-dependencies.jar] to class path
dded resources: [/home/ankitrich26/hivemall-core-0.4.2-rc.2-with-dependencies.jar]
ive> source /home/ankitrich26/define-all (1).hive;
K
ime taken: 0.023 seconds
K
ime taken: 0.024 seconds
K
ime taken: 0.021 seconds
K
ime taken: 0.044 seconds
K
ime taken: 0.025 seconds
K
ime taken: 0.025 seconds
K
ime taken: 0.021 seconds
K
ime taken: 0.022 seconds
K
ime taken: 0.022 seconds
K
ime taken: 0.024 seconds
K
ime taken: 0.02 seconds
K
ime taken: 0.022 seconds
K
ime taken: 0.021 seconds
K
ime taken: 0.022 seconds
K
ime taken: 0.019 seconds
K
ime taken: 0.02 seconds
K
ime taken: 0.027 seconds
```

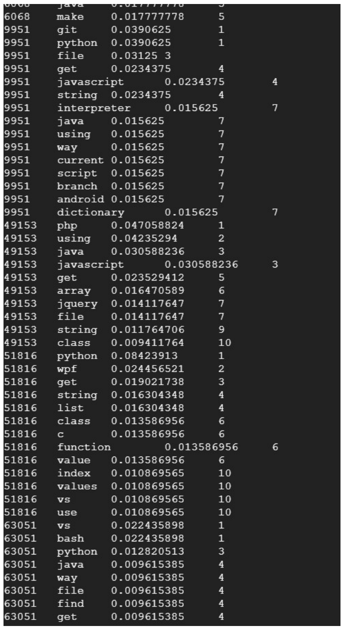
- Below are the queries to calculate TF-IDF

- 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;



```
hive> create or replace view tfFinalView as select * from (select user_id, eachword, freq,rank()  over (partition by owneruser_id order by freq desc) as rn from tempView as t)
;
FAILED: SemanticException Failed to breakup Windowing invocations into Groups. At least 1 group must only depend on input columns. Also check for circular dependencies.
Underlying error: org.apache.hadoop.hive.ql.parse.SemanticException: Line 1:111 Invalid table alias or column reference 'owneruser_id': (possible column names are: user_id, eac
hive> 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
OK
Time taken: 0.207 seconds
hive> select * from tfFinalView;
Query ID = ankitrich26_20211028101023_71c280ab-c1f3-41cc-860f-24262a5b56d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635412257105_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.68 s
----------------------------------------------------------------------------------------------
OK
4883    python  0.038251366     1
4883    process 0.021857923     2
4883    ruby    0.021857923     2
4883    git     0.016393442     4
4883    table   0.016393442     4
4883    rails   0.016393442     4
4883    rename  0.010928961     7
4883    quotes  0.010928961     7
4883    branch  0.010928961     7
4883    list    0.010928961     7
4883    windows 0.010928961     7
4883    local   0.010928961     7
4883    write   0.010928961     7
4883    vs      0.010928961     7
4883    possible        0.010928961     7
4883    variable        0.010928961     7
4883    difference      0.010928961     7
4883    find    0.010928961     7
4883    style   0.010928961     7
6068    sql     0.026666667     1
```



```
6068    java    0.017777778     5
6068    make    0.017777778     5
9951    git     0.0390625       1
9951    python  0.0390625       1
9951    file    0.03125 3
9951    get     0.0234375       4
9951    javascript      0.0234375       4
9951    string  0.0234375       4
9951    interpreter     0.015625        7
9951    java    0.015625        7
9951    using   0.015625        7
9951    way     0.015625        7
9951    current 0.015625        7
9951    script  0.015625        7
9951    branch  0.015625        7
9951    android 0.015625        7
9951    dictionary      0.015625        7
49153   php     0.047058824     1
49153   using   0.04235294      2
49153   java    0.030588236     3
49153   javascript      0.030588236     3
49153   get     0.023529412     5
49153   array   0.016470589     6
49153   jquery  0.014117647     7
49153   file    0.014117647     7
49153   string  0.011764706     9
49153   class   0.009411764     10
51816   python  0.08423913      1
51816   wpf     0.024456521     2
51816   get     0.019021738     3
51816   string  0.016304348     4
51816   list    0.016304348     4
51816   class   0.013586956     6
51816   c       0.013586956     6
51816   function        0.013586956     6
51816   value   0.013586956     6
51816   index   0.010869565     10
51816   values  0.010869565     10
51816   vs      0.010869565     10
51816   use     0.010869565     10
63051   vs      0.022435898     1
63051   bash    0.022435898     1
63051   python  0.012820513     3
63051   java    0.009615385     4
63051   way     0.009615385     4
63051   file    0.009615385     4
63051   find    0.009615385     4
63051   get     0.009615385     4
```

References:

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