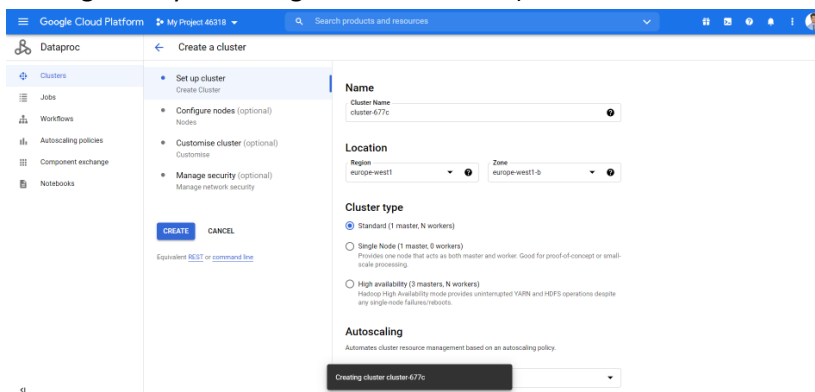
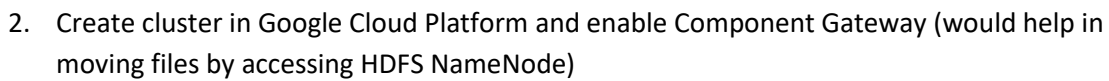


Srijith Unni
DCU Student ID: 20211114
MSc in Computing – Data Analytics
Cloud Technologies CA675

1. Extract input data from Stack Exchange using query provided in Source Code.



3. Create required folders and move the input files (in tab delimited) into those folders.

```
ssh.umn2@cluster-677c-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-677c-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true

System load: 0.37      Processes: 126
Usage of /: 46.3% of 14.37GB    Users logged in: 0
Memory usage: 63%      IP address for ens4: 10.132.0.11
Swap usage: 0%

* Introducing self-healing high availability clustering for MicroK8s!
Super simple, hardened and opinionated Kubernetes for production.

https://microk8s.io/high-availability

16 packages can be updated.
16 updates are security updates.

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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x - mapred hadoop 0 2020-11-20 07:49 /hadoop
drwxr-xr-x - hdfs hadoop 0 2020-11-20 07:50 /tmp
drwxr-xr-x - hdfs hadoop 0 2020-11-20 07:49 /user
srijith_umn2@cluster-677c-m:~$ hadoop fs -mkdir /stackex
srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - mapred hadoop 0 2020-11-20 07:49 /hadoop
drwxr-xr-x - srijith_umn2 hadoop 0 2020-11-20 07:54 /stackex
drwxr-xr-x - hdfs hadoop 0 2020-11-20 07:50 /tmp
drwxr-xr-x - hdfs hadoop 0 2020-11-20 07:49 /user
srijith_umn2@cluster-677c-m:~$ hadoop fs -mkdir /stackex/results
srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /stackex
Found 1 items
drwxr-xr-x - srijith_umn2 hadoop 0 2020-11-20 07:55 /stackex/results
srijith_umn2@cluster-677c-m:~$
```

```
ssh.umn2@cluster-677c-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-677c-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true

System information as of Fri Nov 20 13:10:43 UTC 2020

System load: 0.0      Processes: 130
Usage of /: 46.5% of 14.37GB    Users logged in: 1
Memory usage: 66%      IP address for ens4: 10.132.0.14
Swap usage: 0%

* Introducing self-healing high availability clustering for MicroK8s!
Super simple, hardened and opinionated Kubernetes for production.

https://microk8s.io/high-availability

16 packages can be updated.
16 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Fri Nov 20 12:49:07 2020 from 35.235.240.98
srijith_umn2@cluster-677c-m:~$ ls
ViewCount1.csv ViewCount2.csv ViewCount3.csv csvupload ViewCount1.1.csv csvupload
srijith_umn2@cluster-677c-m:~$ rm ViewCount3
rm: cannot remove 'ViewCount3': No such file or directory
srijith_umn2@cluster-677c-m:~$ rm ViewCount3.csv csvupload
rm: cannot remove 'ViewCount3.csv': No such file or directory
rm: cannot remove 'csvupload': No such file or directory
srijith_umn2@cluster-677c-m:~$ ls
ViewCount1.csv ViewCount2.csv ViewCount3.1.csv csvupload
srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /stackex
Found 3 items
-rw-r--r-- 2 srijith_umn2 hadoop 9652198 2020-11-20 12:59 /stackex/ViewCount1.csv
-rw-r--r-- 2 srijith_umn2 hadoop 8747424 2020-11-20 13:03 /stackex/ViewCount2.csv
-rw-r--r-- 2 srijith_umn2 hadoop 0 2020-11-20 12:50 /stackex/results
srijith_umn2@cluster-677c-m:~$ hadoop fs -put /home/srijith_umn2/ViewCount1.1.csv /stackex
put: /home/srijith_umn2/ViewCount1.1.csv: No such file or directory
srijith_umn2@cluster-677c-m:~$ hadoop fs -put /home/srijith_umn2/ViewCount3.1.csv /stackex
put: /home/srijith_umn2/ViewCount3.1.csv: No such file or directory
srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /stackex
Found 4 items
-rw-r--r-- 2 srijith_umn2 hadoop 9652198 2020-11-20 12:59 /stackex/ViewCount1.csv
-rw-r--r-- 2 srijith_umn2 hadoop 8747424 2020-11-20 13:03 /stackex/ViewCount2.csv
-rw-r--r-- 2 srijith_umn2 hadoop 9184095 2020-11-20 13:16 /stackex/ViewCount3.1.csv
-rw-r--r-- 2 srijith_umn2 hadoop 0 2020-11-20 12:50 /stackex/results
srijith_umn2@cluster-677c-m:~$ hadoop fs -put /home/srijith_umn2/ViewCount1.csv /stackex
put: /home/srijith_umn2/ViewCount1.csv: No such file or directory
srijith_umn2@cluster-677c-m:~$ hadoop fs -ls /stackex
Found 5 items
-rw-r--r-- 2 srijith_umn2 hadoop 9652198 2020-11-20 12:59 /stackex/ViewCount1.csv
-rw-r--r-- 2 srijith_umn2 hadoop 8747424 2020-11-20 13:03 /stackex/ViewCount2.csv
-rw-r--r-- 2 srijith_umn2 hadoop 9184095 2020-11-20 13:16 /stackex/ViewCount3.1.csv
-rw-r--r-- 2 srijith_umn2 hadoop 9652198 2020-11-20 13:32 /stackex/ViewCount4.csv
-rw-r--r-- 2 srijith_umn2 hadoop 0 2020-11-20 12:50 /stackex/results
srijith_umn2@cluster-677c-m:~$
```

File Transfer

Cancel

ViewCount5.csv 14%

File upload destination: /home/srijith_umn2

4. Perform ETL actions on the input data in Pig and merge the output files into final clean data file. This output file can then be downloaded from the HDFS NameNode.

```
ssh.umn2@cluster-677c-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-677c-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true

srijith_umn2@cluster-677c-m:~$ pig
20/11/20 17:53:10 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/11/20 17:53:10 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
20/11/20 17:53:10 INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r: unknown) compiled Oct 18 2020, 18:13:54
20/11/20 17:53:10,593 [main] INFO org.apache.pig.Main - Logging error messages to: /home/srijith_umn2/pig1605894790495.log
20/11/20 17:53:10,542 [main] INFO org.apache.pig.impl.util.UDUtils - Default bootstrap file /home/srijith_umn2/pigbootstrap not found
20/11/20 17:53:11,165 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker.address
20/11/20 17:53:11,165 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://cluster-677c-m
20/11/20 17:53:12,404 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: FFD-default-d8263236-8d37-4d48-9d31-d3d46b12e528
20/11/20 17:53:12,459 [main] INFO org.apache.hadoop.yarn.client.api.impl.TIMELinesClientImpl - Timeline service address: null
20/11/20 17:53:13,030 [main] INFO org.apache.pig.backend.hadoop.PigATClient - Created ATS Hook
20/11/20 17:53:13,083 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grant A = LOAD '/stackex/ViewCount1.txt' USING PigStorage('A') AS id(chararray, PostTypeId(chararray, AcceptedAnswerId(chararray, ParentId(chararray, CreationDate(chararray, DeletionDate(chararray, Score(chararray, ViewCount(chararray, OwnerUserId(chararray, OwnerDisplayname(chararray, LastEditorUserId(chararray, LastEditorDisplayName(chararray, LastEditorTime(chararray, Title(chararray, Tag(chararray, AnswerCount(chararray, CommentCount(chararray, FavouriteCount(chararray, ClosedDate(chararray, CommunityOwnedDate(chararray, ContentLicense(chararray);
20/11/20 17:53:14,437 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grant B = LOAD '/stackex/ViewCount2.txt' USING PigStorage('A') AS id(chararray, PostTypeId(chararray, AcceptedAnswerId(chararray, ParentId(chararray, CreationDate(chararray, DeletionDate(chararray, Score(chararray, ViewCount(chararray, OwnerUserId(chararray, OwnerDisplayname(chararray, LastEditorUserId(chararray, LastEditorDisplayName(chararray, LastEditorTime(chararray, Title(chararray, Tag(chararray, AnswerCount(chararray, CommentCount(chararray, FavouriteCount(chararray, ClosedDate(chararray, CommunityOwnedDate(chararray, ContentLicense(chararray);
20/11/20 17:53:15,143 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grant C = LOAD '/stackex/ViewCount3.txt' USING PigStorage('A') AS id(chararray, PostTypeId(chararray, AcceptedAnswerId(chararray, ParentId(chararray, CreationDate(chararray, DeletionDate(chararray, Score(chararray, ViewCount(chararray, OwnerUserId(chararray, OwnerDisplayname(chararray, LastEditorUserId(chararray, LastEditorDisplayName(chararray, LastEditorTime(chararray, Title(chararray, Tag(chararray, AnswerCount(chararray, CommentCount(chararray, FavouriteCount(chararray, ClosedDate(chararray, CommunityOwnedDate(chararray, ContentLicense(chararray);
20/11/20 17:53:16,091 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grant D = LOAD '/stackex/ViewCount4.txt' USING PigStorage('A') AS id(chararray, PostTypeId(chararray, AcceptedAnswerId(chararray, ParentId(chararray, CreationDate(chararray, DeletionDate(chararray, Score(chararray, ViewCount(chararray, OwnerUserId(chararray, OwnerDisplayname(chararray, LastEditorUserId(chararray, LastEditorDisplayName(chararray, LastEditorTime(chararray, Title(chararray, Tag(chararray, AnswerCount(chararray, CommentCount(chararray, FavouriteCount(chararray, ClosedDate(chararray, CommunityOwnedDate(chararray, ContentLicense(chararray);
20/11/20 17:54:03,063 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grant E =
```

```

@ sshcloud@cluster-67fc-m - Google Chrome
# sshcloud.com/projects/located-mate-2927f5/zones/europe-west-1/buckets/cluter-677c-m/authUser=18l-nn_GBq8projNumber=61159050418useAdminProxy=true

grunt> # = FORKACA A GENERARE $0,$2,$4,$6,$7,$8,$10,$12,$13,$14,$15,$16,$17,$18$
grunt> # = FORKACA B GENERARE $0,$2,$4,$6,$7,$8,$10,$12,$13,$14,$15,$16,$17,$18$
grunt> # = FORKACA C GENERARE $0,$2,$4,$6,$7,$8,$10,$12,$13,$14,$15,$16,$17,$18$
grunt> # = FORKACA D GENERARE $0,$2,$4,$6,$7,$8,$10,$12,$13,$14,$15,$16,$17,$18$
grunt> # = UNION I,J,K,L,M,N$
grunt> # STORO I FILE /stackes/output_final.txt" USING Pigscripte ("*)")
2020-11-20 17:55:34,453 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
er.enabled
2020-11-20 17:55:34,462 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapped.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2020-11-20 17:55:34,496 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
2020-11-20 17:55:34,707 [main] INFO org.apache.data.schema.pigbackend.kmp [key:schematuple] was not set... will not generate code.
2020-11-20 17:55:34,825 [main] INFO org.apache.hadoop.hive.executionengine.plan.HiveExecutionEnginePlan - HiveExecutionEnginePlan: HiveExecutionEnginePlan: ConstantCalculator, GroupByConstantFilter, Alter,
MultiJoinFilter, MergeJoinFilter, MapReduceJoinFilter, PartitionFilter, PredicatePushdownOptimizer, PushDownForSparkPlanner, PushDownFilter, SplitFilter, Stack
TraceCastInserted)
2020-11-20 17:55:34,864 [main] INFO org.apache.pig.mapplan.logical.rules.ColumnFromWriter - Columns pruned for A: $1, $3, $5, $9, $11, $19, $20, $21
2020-11-20 17:55:34,868 [main] INFO org.apache.pig.mapplan.logical.rules.ColumnFromWriter - Columns pruned for B: $1, $3, $5, $9, $11, $19, $20, $21
2020-11-20 17:55:34,868 [main] INFO org.apache.pig.mapplan.logical.rules.ColumnFromWriter - Columns pruned for C: $1, $3, $5, $9, $11, $19, $20, $21
2020-11-20 17:55:34,868 [main] INFO org.apache.pig.mapplan.logical.rules.ColumnFromWriter - Columns pruned for D: $1, $3, $5, $9, $11, $19, $20, $21
2020-11-20 17:55:34,869 [main] INFO org.apache.pig.mapplan.logical.rules.ColumnFromWriter - Columns pruned for E: $1, $3, $5, $9, $11, $19, $20, $21
2020-11-20 17:55:34,955 [main] INFO org.apache.pig.impl.util.SqlLibrariesManager - Selected hsqldb (Removed cmd) of size 49907512 to monitor. collectionheapthreshold = 489350752, usageThresho
ld = 489350752
2020-11-20 17:55:35,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCCompiler - File concatenation threshold: 100 optimistic false
2020-11-20 17:55:35,132 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to ResourceManager at cluster-677c-m/w10.132.0.16:10200
2020-11-20 17:55:35,173 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryCompiler - MR plan size before optimization: 1
2020-11-20 17:55:35,212 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-11-20 17:55:35,301 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to ResourceManager at cluster-677c-m/w10.132.0.16:10200
2020-11-20 17:55:35,379 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to Application History server at cluster-677c-m/w10.132.0.16:10200
2020-11-20 17:55:35,653 [main] INFO org.apache.hadoop.tools.pslogstash.MapReduceMDCScriptTable - Pig script settings are added to the job
2020-11-20 17:55:35,665 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapped.job.reduce.markrest.buffer.percent is deprecated. Instead, use mapreduce.reduce.markrest.buffer.per
cent
2020-11-20 17:55:35,669 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - mapped.job.reduce.markrest.buffer.percent is not set, set to default 0.3
2020-11-20 17:55:35,679 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - This job cannot be converted run-in-process
2020-11-20 17:55:35,709 [main] INFO org.apache.pig.backend.hadoop.conf.Configuration.deprecation - mapped.submit.replication is deprecated. Instead, use mapreduce.client.repliation
2020-11-20 17:55:36,043 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/usr/lib/pig/lib/pig-0.17.0-core-hd.jar to DistributedCache through
amp /tmp/tmp-3811732x/tmp-493856074/pig-0.17.0-core-hd.jar
2020-11-20 17:55:36,126 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/usr/lib/pig/lib/automaton-1.11-8.jar to DistributedCache thro
ugh /tmp/tmp-3811732x/tmp-392828647/automaton-1.11-8.jar
2020-11-20 17:55:36,123 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/usr/lib/pig/lib/antlr-runtime-3.4.jar to DistributedCache thr
ough /tmp/tmp-3811732x/antlr-runtime-3.4.jar
2020-11-20 17:55:36,491 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/usr/lib/hive/hive-exec-2.3.7.jar to DistributedCache thro
ugh /tmp/tmp-3811732x/hive-exec-2.3.7.jar
2020-11-20 17:55:36,520 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Setting up single store Job
2020-11-20 17:55:36,543 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig:schematuple] is false, will not generate code.
2020-11-20 17:55:36,549 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2020-11-20 17:55:36,549 [main] INFO org.apache.pig.data.schema.pigbackend - Setting key [pig:schemauple.classname] with classes to deerialize []

@ sshcloud@cluster-67fc-m - Google Chrome
# sshcloud.com/projects/located-mate-2927f5/zones/europe-west-1/buckets/cluter-677c-m/authUser=18l-nn_GBq8projNumber=61159050418useAdminProxy=true

2020-11-20 17:56:19,053 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to ResourceManager at cluster-677c-m/w10.132.0.16:10302
2020-11-20 17:56:19,054 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to Application History server at cluster-677c-m/w10.132.0.16:10200
2020-11-20 17:56:19,055 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to Application History server at cluster-677c-m/w10.132.0.16:10302
2020-11-20 17:56:19,074 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapped.reduce.tasks is deprecated. Instead, use mapreduce.job.reducers
2020-11-20 17:56:19,075 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to ResourceManager at cluster-677c-m/w10.132.0.16:10302
2020-11-20 17:56:19,089 [main] INFO org.apache.hadoop.yarn.client.AMRProxy - Connecting to Application History server at cluster-677c-m/w10.132.0.16:10200
2020-11-20 17:56:19,084 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-11-20 17:56:19,162 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-11-20 17:56:19,162 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-11-20 17:56:19,162 [main] INFO org.apache.hadoop.tools.pslogstash.MapReduceMDCScriptTable - Script Statistics:
+-----+-----+
|Job Stats (time in seconds):|
+-----+-----+
|JobId|MapTime|ReduceMaxMapTime|MinMapTime|AvgMapTime|MedianMapTime|MaxReduceTime|MinReduceTime|AvgReduceTime|MedianReduceTime|Alias|Feature Outputs|
|job_1605876410595_0004|5|0|23|14|20|23|0|0|0|0|H_A,C,D,E,I,J,K,L,R_M,_MAP_ONLY|/stackes/output_final.txt,

Input(s):
+-----+-----+
|Successfully read 16295 records from|"/stackes/VinCount.txt"|
+-----+-----+
|Successfully read 48155 records from|"/stackes/VinCount.txt"|
+-----+-----+
|Successfully read 44477 records from|"/stackes/VinCount.txt"|
+-----+-----+
|Successfully read 46024 records from|"/stackes/VinCount.txt"|
+-----+-----+
|Successfully read 44777 records from|"/stackes/VinCount.txt"|
+-----+-----+

Output(s):
+-----+-----+
|Successfully stored 200004 records (34673280 bytes) int|"/stackes/output_final.txt"|
+-----+-----+

Counters:
Total records written | 200004
Total bytes written | 34673280
Spillable Memory Manager spill count : 0
Total bytes proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
graph TD
    job_1605876410595_0004 --> r1[ ]
    style r1 fill:none,stroke:none
    r1 --> j1[job_1605876410595_0004]
    style j1 fill:none,stroke:none
    j1 --> m1[mapper]
    style m1 fill:none,stroke:none
    m1 --> r2[reducer]
    style r2 fill:none,stroke:none
    r2 --> j2[job_1605876410595_0004]
    style j2 fill:none,stroke:none
    j2 --> m2[mapper]
    style m2 fill:none,stroke:none
    m2 --> r3[reducer]
    style r3 fill:none,stroke:none
    r3 --> j3[job_1605876410595_0004]
    style j3 fill:none,stroke:none
    j3 --> m3[mapper]
    style m3 fill:none,stroke:none
    m3 --> r4[reducer]
    style r4 fill:none,stroke:none
    r4 --> j4[job_1605876410595_0004]
    style j4 fill:none,stroke:none
    j4 --> m4[mapper]
    style m4 fill:none,stroke:none
    m4 --> r5[reducer]
    style r5 fill:none,stroke:none
    r5 --> j5[job_1605876410595_0004]
    style j5 fill:none,stroke:none
    j5 --> m5[mapper]
    style m5 fill:none,stroke:none
    m5 --> r6[reducer]
    style r6 fill:none,stroke:none
    r6 --> j6[job_1605876410595_0004]
    style j6 fill:none,stroke:none
    j6 --> m6[mapper]
    style m6 fill:none,stroke:none
    m6 --> r7[reducer]
    style r7 fill:none,stroke:none
    r7 --> j7[job_1605876410595_0004]
    style j7 fill:none,stroke:none
    j7 --> m7[mapper]
    style m7 fill:none,stroke:none
    m7 --> r8[reducer]
    style r8 fill:none,stroke:none
    r8 --> j8[job_1605876
```

- Download this file to check the result. On completion of ETL steps, we login to Hive and load the data in a table create in a database.
Then query as required (10 records each for top 10 posts and users by score, and 295 distinct users with Hadoop in their posts)

```
@ syth_unn2@cluster-677c-m - Google Chrome  
■ sshcloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-677c-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true  
[1] [1] [1]  
Loading initialized using configuration in file:/etc/hive/conf/dist/hive-logtj2.properties Async: true  
hive> CREATE DATABASE stackdb;  
OK  
Time taken: 0.649 seconds  
hive> CREATE EXTERNAL TABLE stackdb.posts ( id int comment 'Id', acceptedanwserid int comment 'AcceptedAnswerId', creationdate string comment 'CreationDate', score int comment 'Score', viewcount int comment 'ViewCount', owneruserid int comment 'ownerUserid', lasteditoruserid int comment 'LastEditorUserId', lasttidate string comment 'lastEditDate', lastactivitytime string comment 'LastActivityTime', title string comment 'title', tags string comment "tags", answercount int comment 'AnswersCount', commentcount int comment 'CommentCount', favouritecount int comment 'FavouriteCount') ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LOCATION '/stackdb';  
OK  
Time taken: 0.429 seconds  
hive> LOAD DATA LOCAL INPATH 'result_final.txt' INTO TABLE stackdb.posts;  
Loading data to table stackdb.posts  
OK  
Time taken: 1.818 seconds  
hive>
```

```
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/author=18hl=en_GB/projectNumber=6111590504198useAdminProxy=true
hive> CREATE TABLE stackadb.top10_post ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, score, viewcount, answercount, commentcount, favouritecount, title FROM stackadb.posts
ORDER BY score DESC LIMIT 10
Query ID = erijth_unml2_20201121084719_0a5c1451-0a29-4490-a21b-cfd343c4200f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)

VERTICES      MODE        STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0 0
-----
VERTICES: 03/03 (=====) 100% ELAPSED TIME: 11.39 s
-----
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackadb.db/top10_post
OK
id      score viewcount answercount commentcount favouritecount title
Time taken: 13.27 seconds
hive> SELECT * FROM stackadb.top10_post;
OK
top10_post.id      top10_post.score      top10_post.viewcount      top10_post.answercount      top10_post.commentcount      top10_post.favouritecount      top10_post.title
11227809      24963      1541163      26      3      11149      "Why is processing a sorted array faster than processing an unsorted array?"
927358      21777      900893      86      14      6970      "How do I undo the most recent local commits in Git?"
2083505      17385      8435222      40      7      5477      "How do I delete a Git branch locally and remotely?"
238357      12200      2868645      36      9      2382      "What is the difference between 'git pull' and 'git fetch'?"
231767      10627      2333118      42      0      5902      "What does the 'yield' keyword do?"
477816      10467      2946855      36      0      1463      "What is the correct JSON content type?"
148170      9309      3316301      37      11      1590      "How do I undo 'git add' before commit?"
1442028      9174      820505      24      26      2108      "What is the '--->' operator in C++?"
6581213      8919      3122723      34      0      1293      "How do I rename a local Git branch?"
5767325      8762      7275338      97      1      1345      "How can I remove a specific item from an array?"
Time taken: 0.289 seconds, Fetched: 10 row(s)
hive>

ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/author=18hl=en_GB/projectNumber=6111590504198useAdminProxy=true
hive> CREATE TABLE stackadb.top10_user ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, score, viewcount, owneruserid, title FROM stackadb.posts ORDER BY score DESC LIMIT 10
Query ID = erijth_unml2_20201121084834_7e765a0e-afe9-4294-b093-ace04b632f05
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)

VERTICES      MODE        STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0 0
-----
VERTICES: 03/03 (=====) 100% ELAPSED TIME: 11.61 s
-----
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackadb.db/top10_user
OK
id      score viewcount owneruserid title
Time taken: 13.469 seconds
hive> SELECT * FROM stackadb.top10_user;
OK
top10_user.id      top10_user.score      top10_user.viewcount      top10_user.owneruserid      top10_user.title
11227809      24963      1541163      87234      "Why is processing a sorted array faster than processing an unsorted array?"
927358      21777      900893      89904      "How do I undo the most recent local commits in Git?"
2083505      17385      8435222      95592      "How do I delete a Git branch locally and remotely?"
238357      12200      2868645      6068      "What is the difference between 'git pull' and 'git fetch'?"
231767      10627      2333118      19300      "What does the 'yield' keyword do?"
477816      10467      2946855      12870      "What is the correct JSON content type?"
148170      9309      3316301      14049      "How do I undo 'git add' before commit?"
1442028      9174      820505      87234      "What is the '--->' operator in C++?"
6581213      8919      3122723      338204      "How do I rename a local Git branch?"
5767325      8762      7275338      364969      "How can I remove a specific item from an array?"
Time taken: 0.301 seconds, Fetched: 10 row(s)
hive>

ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/author=18hl=en_GB/projectNumber=6111590504198useAdminProxy=true
hive> CREATE TABLE stackadb.users_hadoop ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, owneruserid, title, tags FROM stackadb.posts where tags like '%<hadoop>%'
Query ID = erijth_unml2_20201121084156_001eb096-2b62-4a80-a879-2f64173f5e48
Total jobs = 1
Launching Job 1 out of 1
Pre session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)

VERTICES      MODE        STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0 0
-----
VERTICES: 02/02 (=====) 100% ELAPSED TIME: 15.58 s
-----
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackadb.db/users_hadoop
OK
id      owneruserid title tags
Time taken: 30.578 seconds
hive> SELECT * FROM stackadb.users_hadoop;
OK
users_hadoop.id      users_hadoop.owneruserid      users_hadoop.title      users_hadoop.tags
24179      2588      How does Hive compare to HBase? <hadoop><hbase><hive>
139344      10861      Is there a HMR equivalent to Apache Hadoop? <cf><.net><hadoop><mapreduce>
1152712      114196      How does the MapReduce sort algorithm work? <algorithm><sorting><parallel-processing><hadoop><mapreduce>
1482282      123067      Java vs Python on Hadoop <java><python><hadoop>
1583330      41747      Writing data to Hadoop <hadoop><chdfs>
2354525      77088      What should be Hadoop tmp.dir? <hadoop><chdfs><config>
2358402      246477      Where HDFS stores files locally by default? <hadoop><chdfs>
4489585      141196      Chaining multiple MapReduce jobs in Hadoop <hadoop><mapreduce>
2469800      305105      Change block size of dfs file <hadoop>
2674421      152253      Free large datasets to experiment with Hadoop <resources><hadoop><copendata>
8231507      154586      How does Hadoop perform input splits? <hadoop><mapreduce><chdfs>
1207238      218900      Where does hadoop mapreduce framework send its System.out.print() statements? (stdout) <hadoop><mapreduce>
13546259      68920      "Difference between Pig and Hive? Why have both?" <hadoop><hive><apache-pig>
3515441      69920      Pig Latin: Load multiple files from a date range (part of the directory structure) <hadoop><apache-pig>
3548239      428475      Merging multiple files into one within Hadoop <hadoop><apache-pig>
4065999      215328      "Does Hive have a string split function?" <hadoop><hive>
4740961      215971      Hadoop copy a directory? <hadoop><chdfs>
5058400      2819      "Where does Hive store files in HDFS?" <hadoop><hive><chdfs>
5283446      21824      "Hdfs error: could only be replicated to 0 nodes, instead of 1" <amazon><cd2><hadoop>
5377118      647952      How to convert .txt file to Hadoop's sequence file format <java><file><hadoop><type-conversion><hive>
5385163      315013      Create temporary table in Hive? <hadoop><hive>
5571146      463286      "Hadoop, how to compress mapper output but not the reducer output" <compression><hadoop><chdfs>
5700068      655360      merge output files after reduce phase <hadoop><mapreduce>
5845489      572138      "How to fix "Task attempt 20110451139_0295_r_000006_0 failed to report status for 600 seconds."" <hadoop><mapreduce>
6183560      145360      Base client Connectionless for Hbase error <java><hbase><hadoop><hbase><hive>
6287533      324968      Search/Find a file and file content in Hadoop <file><filesystem><hadoop><distributed-computing>
6445319      478961      "COLLECT_SET() in Hive, keep duplicates?" <java><hadoop><user-defined-functions><hive>

ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/author=18hl=en_GB/projectNumber=6111590504198useAdminProxy=true
erijth_unml2@cluster-e7d2-m:~$ hadoop fs -ls /user/hive/warehouse/stackadb.db;
Found 3 items
drwxrwxrwt - erijth_unml2 hadoop 0 2020-11-21 08:47 /user/hive/warehouse/stackadb.db/top10_post
drwxrwxrwt - erijth_unml2 hadoop 0 2020-11-21 08:48 /user/hive/warehouse/stackadb.db/top10_user
drwxrwxrwt - erijth_unml2 hadoop 0 2020-11-21 08:42 /user/hive/warehouse/stackadb.db/users_hadoop
erijth_unml2@cluster-e7d2-m:~$
```

6. Using the Top10_User table, we shall perform the calculation of TF-IDF using MapReduce programs (Source: <https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python->)

We have changed the Hadoop commands from source according to our file path and jar version. Mapper and Reducer programs remain unchanged due to logic being consistent.

```
ssh,unni2@cluster-e7d2-mc -- Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/authorser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
ssh,unni2@cluster-e7d2-mc:~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-jar -file /home/srijith_unni2/mapper1.py -mapper /home/srijith_unni2/mapper1.py -reducer /home/srijith_unni2/reducer2.py -input /tfidf/Top10_User.txt -output /tfidf/output
20/11/22 10:52:19 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/srijith_unni2/mapper1.py, /home/srijith_unni2/reducer2.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2-jar] /tmp/streamjob545348360814087090.jar tmpDir=null
20/11/22 10:52:20 INFO client.AMRProxy: Connecting to ResourceManager at cluster-e7d2-m/10.132.0.31:8032
20/11/22 10:52:21 INFO client.AMRProxy: Connecting to Application History server at cluster-e7d2-m/10.132.0.31:10200
20/11/22 10:52:22 INFO mapred.FileInputFormat: Total input files to process : 1
20/11/22 10:52:22 INFO mapreduce.JobSubmitter: number of splits:16
20/11/22 10:52:22 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/11/22 10:52:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1606041009512_0001
20/11/22 10:52:24 INFO Impl.YarnClientImpl: Submitted application application_1606041009512_0001
20/11/22 10:52:24 INFO mapreduce.Job: The url to track the job: http://cluster-e7d2-m:8088/pxxy/application_1606041009512_0001/
20/11/22 10:52:24 INFO mapreduce.Job: Running job: job_1606041009512_0001
20/11/22 10:52:27 INFO mapreduce.Job: Job job_1606041009512_0001 running in uber mode : false
20/11/22 10:52:37 INFO mapreduce.Job: map 0% reduce 0%
20/11/22 10:52:51 INFO mapreduce.Job: map 13% reduce 0%
20/11/22 10:52:59 INFO mapreduce.Job: map 31% reduce 0%
20/11/22 10:53:01 INFO mapreduce.Job: map 38% reduce 0%
20/11/22 10:53:02 INFO mapreduce.Job: map 44% reduce 0%
20/11/22 10:53:12 INFO mapreduce.Job: map 50% reduce 0%
20/11/22 10:53:13 INFO mapreduce.Job: map 56% reduce 0%
20/11/22 10:53:15 INFO mapreduce.Job: map 75% reduce 0%
20/11/22 10:53:23 INFO mapreduce.Job: map 81% reduce 0%
20/11/22 10:53:24 INFO mapreduce.Job: map 88% reduce 0%
20/11/22 10:53:28 INFO mapreduce.Job: map 94% reduce 0%
20/11/22 10:53:29 INFO mapreduce.Job: map 100% reduce 0%
20/11/22 10:53:37 INFO mapreduce.Job: map 100% reduce 100%
20/11/22 10:53:38 INFO mapreduce.Job: Job job_1606041009512_0001 completed successfully
20/11/22 10:53:38 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=5206
  FILE: Number of bytes written=3835944
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=7812
  HDFS: Number of bytes written=4282
  HDFS: Number of read operations=56
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=6
Job Counters
  Launched map tasks=16
  Launched reduce tasks=2
  Data-local map tasks=16
  Total time spent by all maps in occupied slots (ms)=825820
  Total time spent by all reduces in occupied slots (ms)=92000
  Total time spent by all map tasks (ms)=206455
  Total time spent by all reduce tasks (ms)=11500
  Total vcore-milliseconds taken by all map tasks=206455
  Total vcore-milliseconds taken by all reduce tasks=92000

ssh,unni2@cluster-e7d2-mc -- Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m/authorser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
ssh,unni2@cluster-e7d2-mc:~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-jar -file /home/srijith_unni2/mapper2.py -mapper /home/srijith_unni2/mapper2.py -reducer /home/srijith_unni2/reducer2.py -input /tfidf/input2/part-00000 -input /tfidf/input2/part-00001 -output /tfidf/output2
20/11/22 11:00:11 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/srijith_unni2/mapper2.py, /home/srijith_unni2/reducer2.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2-jar] /tmp/streamjob724702119924086392.jar tmpDir=null
20/11/22 11:00:12 INFO client.AMRProxy: Connecting to ResourceManager at cluster-e7d2-m/10.132.0.31:8032
20/11/22 11:00:13 INFO client.AMRProxy: Connecting to Application History server at cluster-e7d2-m/10.132.0.31:10200
20/11/22 11:00:14 INFO mapred.FileInputFormat: Total input files to process : 2
20/11/22 11:00:14 INFO mapreduce.JobSubmitter: number of splits:16
20/11/22 11:00:14 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/11/22 11:00:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1606041009512_0003
20/11/22 11:00:15 INFO Impl.YarnClientImpl: Submitted application application_1606041009512_0003
20/11/22 11:00:15 INFO mapreduce.Job: The url to track the job: http://cluster-e7d2-m:8088/pxxy/application_1606041009512_0003/
20/11/22 11:00:15 INFO mapreduce.Job: Running job: job_1606041009512_0003
20/11/22 11:00:25 INFO mapreduce.Job: Job job_1606041009512_0003 running in uber mode : false
20/11/22 11:00:25 INFO mapreduce.Job: map 0% reduce 0%
20/11/22 11:00:37 INFO mapreduce.Job: map 13% reduce 0%
20/11/22 11:00:44 INFO mapreduce.Job: map 31% reduce 0%
20/11/22 11:00:49 INFO mapreduce.Job: map 44% reduce 0%
20/11/22 11:00:59 INFO mapreduce.Job: map 50% reduce 0%
20/11/22 11:01:00 INFO mapreduce.Job: map 56% reduce 0%
20/11/22 11:01:01 INFO mapreduce.Job: map 75% reduce 0%
20/11/22 11:01:11 INFO mapreduce.Job: map 88% reduce 0%
20/11/22 11:01:12 INFO mapreduce.Job: map 100% reduce 0%
20/11/22 11:01:19 INFO mapreduce.Job: map 100% reduce 50%
20/11/22 11:01:20 INFO mapreduce.Job: map 100% reduce 100%
20/11/22 11:01:20 INFO mapreduce.Job: Job job_1606041009512_0003 completed successfully
20/11/22 11:01:21 INFO mapreduce.Job: Counters: 50
File System Counters
  FILE: Number of bytes read=4458
  FILE: Number of bytes written=3834916
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=19373
  HDFS: Number of bytes written=4528
  HDFS: Number of read operations=48
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=6
Job Counters
  Killed map tasks=1
  Launched map tasks=16
  Launched reduce tasks=2
  Data-local map tasks=16
  Total time spent by all maps in occupied slots (ms)=803656
  Total time spent by all reduces in occupied slots (ms)=78848
  Total time spent by all map tasks (ms)=200914
  Total time spent by all reduce tasks (ms)=9856
  Total vcore-milliseconds taken by all map tasks=200914
  Total vcore-milliseconds taken by all reduce tasks=9856
```