Assignment 1 – Data Analysis

Arif Rafeek Khan
Student Id – 20210435
MSc in Computing – Data Analytics
Cloud Technologies (CA675)

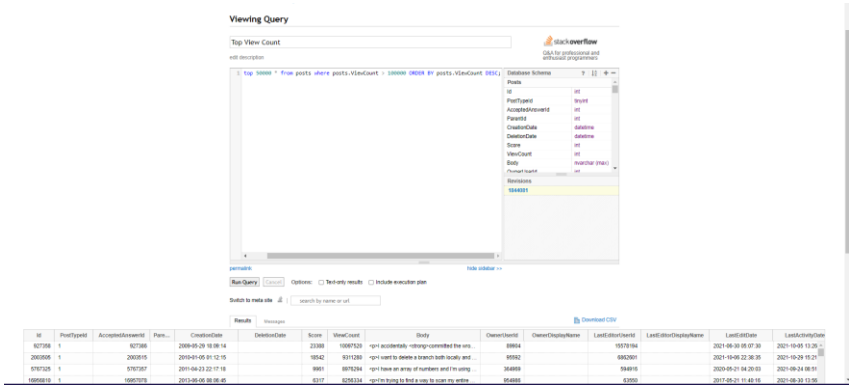
This is a report for the Data Analysis tasks on the Stack Exchange which includes:

- Task 1: Get data from Stack Exchange (Data Acquisition/Collection)
- Task 2: Load data into chosen cloud technology (MapReduce/Pig/Hive)
- Task 3: Query data using MapReduce/Pig/Hive
- Task 4: Calculate TF-IDF with MapReduce/Pig/Hive

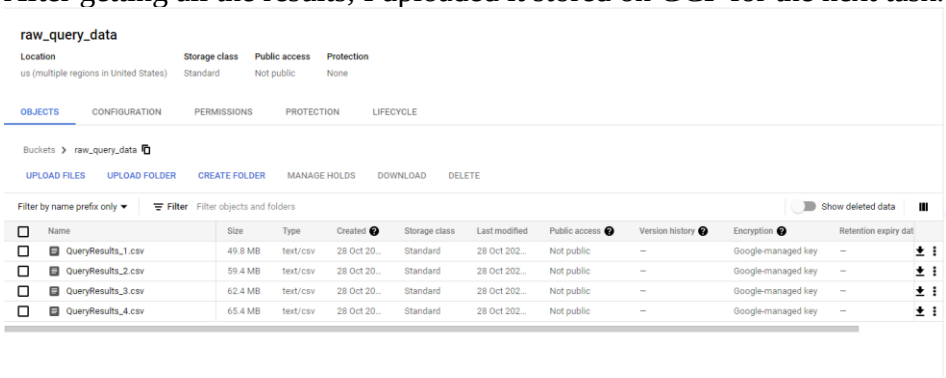
GitHub Repo: https://github.com/Ichigo-lab/Cloud_Assignment

Task 1: Get data from Stack Exchange

For data collection task https://data.stackexchange.com/stackoverflow/query/new is used. Task was to collect 200000 data by top view count but on this site, query returns only 50000 data at a time. So below 4 queries are used to get the desired result. First highest count set to greater than 100000 and then from the result. The queries is committed on git.



After getting all the results, I uploaded it stored on GCP for the next task.



Task 2: Load data into chosen cloud technologies (MapReduce/Pig/Hive)

For this task, I chose GCP because of the ease of Dataproc. First, I created dataproc cluster for Pig and Hive. For this task, I have used pig for the data cleaning and transformation. The script is committed on git.

Instances of cluster

INSTANCES | INSTANCE SCHEDULE

VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

Filter Enter property name or value

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
✓	cluster-5d95-m	us-central1-b			10.128.0.3 (nic0)	34.136.84.211	SSH
✓	cluster-5d95-w-0	us-central1-b			10.128.0.4 (nic0)	104.154.77.98	SSH
✓	cluster-5d95-w-1	us-central1-b			10.128.0.2 (nic0)	35.226.99.100	SSH

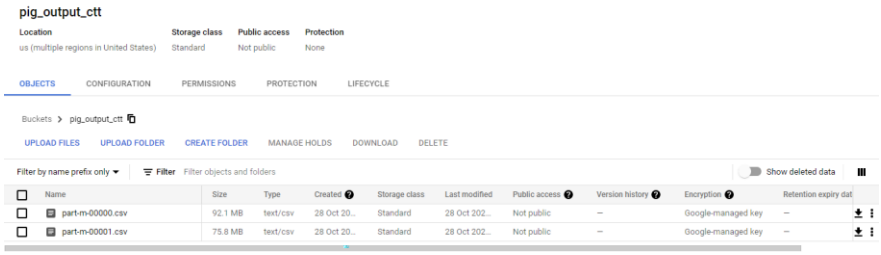
SSH into master node

```
arif_khan8@cluster-6341-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2021-10-28 21:58 /tmp
drwxrwxrwt   - hdfs hadoop          0 2021-10-28 21:58 /user
drwx-wx-wx   - hive hadoop          0 2021-10-28 21:58 /var
arif_khan8@cluster-6341-m:~$ hdfs dfs -cp gs://raw_query_data /raw_data
arif_khan8@cluster-6341-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - arif_khan8 hadoop          0 2021-10-28 22:09 /raw_data
drwxrwxrwt   - hdfs       hadoop          0 2021-10-28 21:58 /tmp
drwxrwxrwt   - hdfs       hadoop          0 2021-10-28 21:58 /user
drwx-wx-wx   - hive       hadoop          0 2021-10-28 21:58 /var
arif_khan8@cluster-6341-m:~$ hdfs dfs -cat /raw_data/QueryResults_1.csv /raw_data/QueryResults_2.csv /raw_data/Quer
yResults_3.csv /raw_data/QueryResults_4.csv > /CombinedResult.csv
```

```
2021-10-28 22:29:49,662 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is complete
d. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 22:29:49,712 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
 - 100% complete
2021-10-28 22:29:49,714 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.2   0.18.0-SNAPSHOT arif_khan8      2021-10-28 22:29:17     2021-10-28 22:29:49     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTi
me      AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1635458285888_0003  2       0       17      15      16      16      0       0       0       0       Allclean,Lo
adPig,NewPig,ReplaceHTML        MAP_ONLY        /outputss,

Input(s):
Successfully read 200000 records (243872844 bytes) from: "/PigData"

Output(s):
Successfully stored 200000 records (175983771 bytes) in: "/outputss"

Counters:
Total records written : 200000
Total bytes written : 175983771
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1635458285888_0003
```

```
arif_khan8@cluster-6341-m:~$ hdfs dfs -ls /
Found 6 items
-rw-r--r--   2 arif_khan8 hadoop  243868036 2021-10-28 22:13 /PigData
drwxr-xr-x   - arif_khan8 hadoop          0 2021-10-28 22:29 /outputss
drwxr-xr-x   - arif_khan8 hadoop          0 2021-10-28 22:09 /raw_data
drwxrwxrwt   - hdfs       hadoop          0 2021-10-28 22:29 /tmp
drwxrwxrwt   - hdfs       hadoop          0 2021-10-28 21:58 /user
drwx-wx-wx   - hive       hadoop          0 2021-10-28 21:58 /var
arif_khan8@cluster-6341-m:~$ hdfs dfs -ls /outputss
Found 3 items
-rw-r--r--   2 arif_khan8 hadoop          0 2021-10-28 22:29 /outputss/_SUCCESS
-rw-r--r--   2 arif_khan8 hadoop   96523477 2021-10-28 22:29 /outputss/part-m-00000
-rw-r--r--   2 arif_khan8 hadoop   79460294 2021-10-28 22:29 /outputss/part-m-00001
arif_khan8@cluster-6341-m:~$
```

After cleaning and transformation, I moved the data into the cloud storage.



Task 3: Query data using MapReduce/Pig/Hive

For this task, Hive is being used as it is best for querying data. The query is committed on git.

Top 10 posts by score

```
hive> Select Id, Score FROM hiveTable ORDER BY Score desc limit 10;
Query ID = arif_khan8_20211028154453_0758681c-257e-4f07-aaf7-c9260e4bb9d3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635422770976_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      8          8        0        0       0       0
Reduce 2 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.25 s
--------------------------------------------------------------------------------
OK
11227809        25903
11227809        25903
927358  23303
927358  23303
2003505 18475
2003505 18475
292357  12812
292357  12812
231767  11528
231767  11528
Time taken: 18.366 seconds, Fetched: 10 row(s)
```

The top 10 users by post score: For the first add all the score of users then take the result.

```
hive> Select SUM(Score) AS AddedScore, OwnerUserId
    > FROM hiveTable
    > WHERE OwnerUserId IS NOT NULL
    > GROUP BY OwnerUserId
    > ORDER BY AddedScore DESC
    > LIMIT 10;
Query ID = arif_khan8_20211028154947_e69781a8-dd2e-42fb-8323-266333bc3ba7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635422770976_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.57 s
--------------------------------------------------------------------------------
OK
75248   87234
57558   4883
53528   9951
51778   6068
47956   89904
47360   51816
40366   49153
38966   179736
38880   95592
38632   63051
Time taken: 17.293 seconds, Fetched: 10 row(s)
```

The number of distinct users, who used the word “cloud” in one of their posts: The word ‘cloud’ can be lower or upper or mixed case so lowercase Body, Title and tag and then get the results.

```
hive> SELECT COUNT(DISTINCT OwnerUserId) FROM hiveTable WHERE (Lower(Body) like '%cloud%' OR Lower(Title) like '%cloud%' OR Lower(Tags) like '%cloud%');
Query ID = arif_khan8_20211028212857_8f1b03ab-6591-454f-945a-2c7c98a15c8c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635422770976_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [========================>>] 100%  ELAPSED TIME: 14.98 s
--------------------------------------------------------------------------------
OK
909
Time taken: 15.79 seconds, Fetched: 1 row(s)
```

Task 4: Calculate TF-IDF with MapReduce/Pig/Hive

To calculate tf-idf of top 10 users of top 10 words. First, I filtered the data based on sum of score to get top user then used hivemall to create views for different scenario and for calculations. The code is committed on git.

```
hive> Create view topuser as select OwnerUserId, SUM(Score) AS AddedScore from hiveTable
    > WHERE OwnerUserId IS NOT NULL
    > GROUP BY OwnerUserId
    > ORDER BY AddedScore DESC
    > LIMIT 10;
OK
Time taken: 0.181 seconds
hive> create or replace view topquery as select OwnerUserId, Body from hiveTable WHERE OwnerUserId IN (select OwnerUserId

    > from topuser);
OK
Time taken: 0.267 seconds
hive> create or replace view query_exploded
    > as
    > select
    >   OwnerUserId,
    >   word
    > from
    >   topquery LATERAL VIEW explode(tokenize(Body,true)) t as word
    > where
    >   not is_stopword(word) and LENGTH(word) > 3;
OK
Time taken: 0.12 seconds
```

```
hive> select * from topuser;
Query ID = arifkhan89837_20211031205023_b4005846-4d89-482b-9671-91b423cd6c83
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635705362501_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      7          7        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.76 s
--------------------------------------------------------------------------------
OK
87234   37624
4883    28779
9951    26764
6068    25889
89904   23978
51816   23680
49153   20183
179736  19483
95592   19440
63051   19316
Time taken: 12.567 seconds, Fetched: 10 row(s)
```

```
9951    ["like:8.082165229954115E-4","string:9.042747856487248E-4","using:0.0013167120900964507","file:2.4868201854877614E
-4","python:7.900272721087671E-4","want:4.351935058259933E-4","would:4.351935058259933E-4","know:8.082165229954115E-4"]
89904   ["possible:0.0019459841244553666","like:1.837650154862481E-4","know:1.837650154862481E-4","want:5.512950464587442E
-4","would:1.837650154862481E-4","file:0.0011025900929174885"]
95592   ["file:1.5329142402211997E-4","python:0.0021102682222290236","know:1.5329142402211997E-4","like:2.2993713603317996
E-4","want:2.2993713603317996E-4","would:1.5329142402211997E-4","possible:3.246566565163485E-4","using:4.869849847745227E-
4"]
179736  ["using:3.981921281758795E-4","string:0.0010254950018240708","possible:1.4932205934776516E-4","return:0.0021645711
589069034","want:9.400614023665549E-4","would:3.290215121357862E-4","file:2.820184420174584E-4","python:5.972882373910607E
-4","like:3.290215121357862E-4","know:3.055199664228763E-4"]
51816   ["using:7.764094049773339E-4","like:8.66492939651627E-4","file:2.3328654592409746E-4","python:0.001341070872464925
","want:6.332063724200376E-4","would:3.332664972212095E-4","know:1.6663324861060475E-4","possible:9.175748169764731E-4","r
eturn:0.0030694973455712197","string:0.0021005345011022194"]
4883    ["using:5.266848631149252E-4","know:8.289400618292537E-5","like:2.4868201854877614E-4","file:1.6578801236585074E-4
","python:0.0014044929683064671","want:2.4868201854877614E-4","would:4.973640370975523E-4","possible:8.778080600643005E-4"
,"return:0.0016360326156038144"]
63051   ["possible:4.2073810412965876E-4","string:9.631638611931812E-4","want:0.0011257269609055295","would:1.324384634821
2206E-4","know:3.9731539044636613E-4","file:9.932885400383913E-4","python:8.414762082593175E-4","like:1.9865769522318306E-
4","using:2.8049206941977247E-4"]
6068    ["using:8.846557579303885E-4","like:5.316223078983251E-4","want:2.278381350003525E-4","would:1.1391906750017624E-4
","file:6.835143836935655E-4","python:1.608464963137932E-4","know:7.594604322449317E-5","possible:1.608464963137932E-4","r
eturn:4.996348528390188E-4","string:0.0011046411143925514"]
49153   ["know:4.080025945148418E-5","using:0.0010801383935441194","possible:3.456442760061251E-4","return:0.0010736691872
980523","string:0.0016814216767994416","file:2.6520167444918293E-4","python:4.320553450076564E-5","want:5.304033488983659E
-4","would:5.304033488983659E-4","like:2.6520167444918293E-4"]
Time taken: 29.648 seconds, Fetched: 9 row(s)
hive>
```

References:

1. https://data.stackexchange.com/stackoverflow/query/new
2. https://hivemall.incubator.apache.org/userguide/ft_engineering/tfidf.html
3. https://github.com/myui/hivemall/releases/tag/v0.4.2-rc.2
4. https://github.com/Khalees2/MapReduce-Pig-and-Hive-on-Vagrant-and-Google-Dataproc