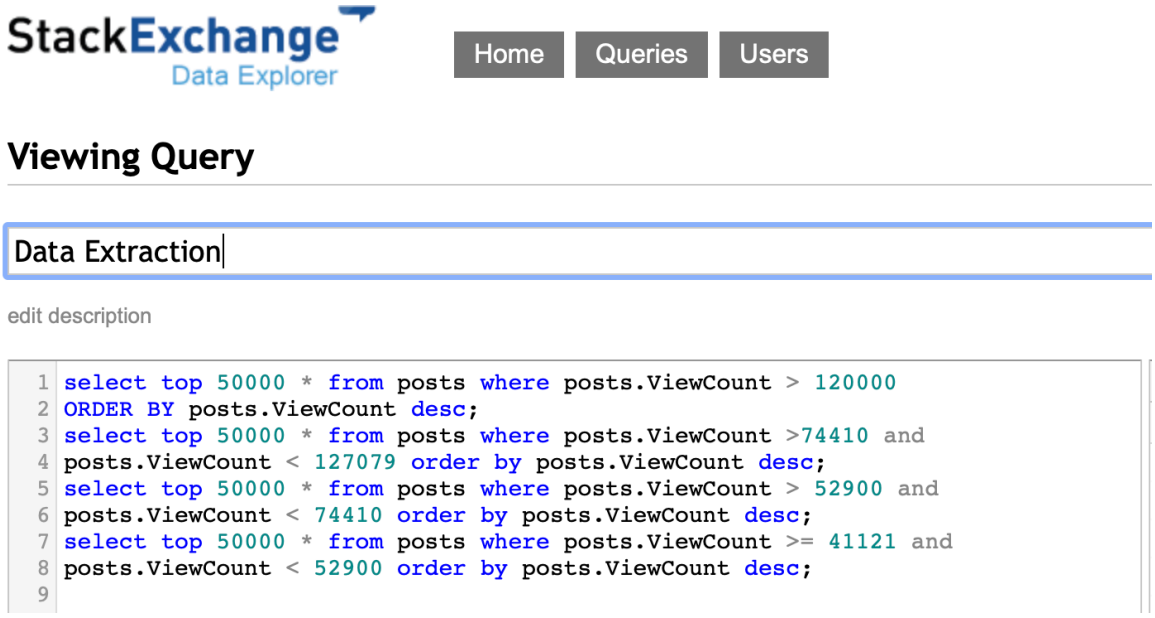
CA675 Cloud Technologies

Name	Atif Shahab
Student Number	21260968
Module Coordinator	Manoj Kesavulu
Gitlab	https://gitlab.computing.dcu.ie/shahaba2/ca675-assignment-1.git

DATA ACQUISITION

Data explorer feature from stack exchange was used to extract stackoverflow data. Since the requirement was for extracting top 200,000 posts, however only 50,000 records could be downloaded at a time. This meant data needed to be downloaded in four sets of 50,000 records. A count query was run to find the range of the first 50,000 records by ViewCount and then a query was run to fetch the data. Same process was followed to extract the next three sets of 50,000 records(See Appendix A).



Extract, Load and Transform (ELT)

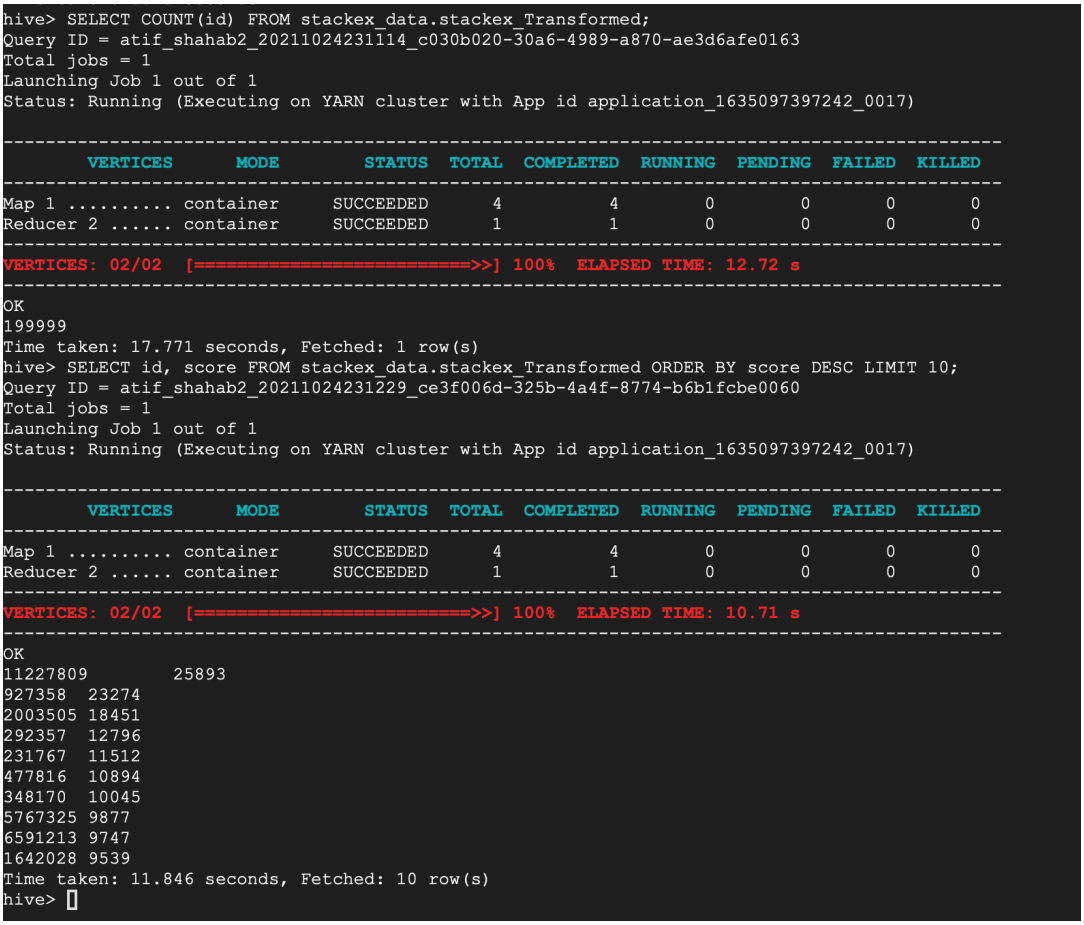
The extracted data set was then uploaded to the hdfs directory. The dataset had 23 columns of which body column had posts written in natural language, had special characters, html tags etc. Pig was used to perform the cleanup tasks, the data was loaded to pig using CSVExcelStorage and "YES_MULTILINE" argument. The four csv files were then combined to one using the UNION command. A count query was run to verify the records were loaded successfully(See Appendix B).
The new lines, html tags, special characters and duplicate entries were removed from the body column. The cleaned data was then uploaded to hive for further processing.

Hive Querying

Hive was the clear choice for task 3 since the three questions asked could be easily answered using HiveQL queries. Firstly a database and a table was created in hive and the cleaned data was then loaded to that table from pig using HCatStorer.

1. The top 10 posts by score

   SELECT id, score FROM stackex_data.stackex_Transformed ORDER BY score DESC LIMIT 10;



```
hive> SELECT COUNT(id) FROM stackex_data.stackex_Transformed;
Query ID = atif_shahab2_20211024231114_c030b020-30a6-4989-a870-ae3d6afe0163
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635097397242_0017)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.72 s
--------------------------------------------------------------------------------
OK
199999
Time taken: 17.771 seconds, Fetched: 1 row(s)
hive> SELECT id, score FROM stackex_data.stackex_Transformed ORDER BY score DESC LIMIT 10;
Query ID = atif_shahab2_20211024231229_ce3f006d-325b-4a4f-8774-b6b1fcbe0060
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635097397242_0017)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.71 s
--------------------------------------------------------------------------------
OK
11227809        25893
927358  23274
2003505 18451
292357  12796
231767  11512
477816  10894
348170  10045
5767325 9877
6591213 9747
1642028 9539
Time taken: 11.846 seconds, Fetched: 10 row(s)
hive>
```

2. The top 10 users by post score

   SELECT owneruserid, SUM(score) FROM stackex_data.stackex_Transformed WHERE owneruserid is not NULL GROUP BY owneruserid ORDER BY SUM(score) DESC LIMIT 10;

```
hive> SELECT owneruserid, SUM(score) FROM stackex_data.stackex_Transformed where owneruserid is not NULL GROUP BY owneruserid ORDER BY SUM(score) DESC LIMIT 10;
Query ID = atif_shahab2_20211027234626_097cca2b-4e13-4778-abb9-618da7e7153a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635378200367_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      8          8        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.97 s
--------------------------------------------------------------------------------
OK
87234   37606
4883    28739
9951    26728
6068    25860
89904   23949
51816   23632
49153   20156
179736  19454
95592   19413
63051   19295
Time taken: 17.284 seconds, Fetched: 10 row(s)
hive>
```

3. The number of distinct users, who used the word “cloud” in one of their posts.

   I approached this problem in two ways:

   i. I searched for the word ‘cloud’ regardless of it being a sub-string of words like cloudera.

      SELECT COUNT(DISTINCT(owneruserid)) FROM stackex_data.stackex_Transformed WHERE LCASE(body) LIKE '%cloud%' OR LCASE(title)LIKE '%cloud%' OR LCASE(tags) LIKE '%cloud%';

```
hive> SELECT COUNT(DISTINCT(owneruserid)) FROM stackex_data.stackex_Transformed_new WHERE LCASE(body) LIKE '%cloud%' OR LCASE(title) LIKE '%cloud%' OR LCASE(tags) LIKE '%cloud%';
Query ID = atif_shahab2_20211028110744_68a48922-130e-46f7-b236-ff04effab2b3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635418938236_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      9          9        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.75 s
--------------------------------------------------------------------------------
OK
950
```

   ii. The second scenario I searched only for keyword ‘cloud’ and not including words were it’s a sub-string like cloudera, cloudspace etc.

      select count(distinct owneruserid) from stackex_data.stackex_Transformed_new where owneruserid is not null and (body rlike '\\bcloud\\b' or title rlike '\\bcloud\\b' or tags rlike '\\bcloud\\b');

```
hive> select count(distinct owneruserid) from stackex_data.stackex_Transformed_new where owneruserid is not null and (body rlike '\\bcloud\\b' or title rlike '\\bcloud\\b' or tags rlik
e '\\bcloud\\b');
Query ID = atif_shahab2_20211028112040_cfe09160-62cf-4204-8c7f-012dd3f243ca
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635418938236_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      9          9        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.88 s
--------------------------------------------------------------------------------
OK
487
Time taken: 22.588 seconds, Fetched: 1 row(s)
hive>
```

TF-IDF

Data is loaded to Pig for thorough cleaning in body column. All the line breaks, HTML tags, special characters, words with an apostrophe, single or double character text were removed and multiple spaces between words were reduced to single spaces. The data is filtered with pig to extract the top 10 users based on their score, and the result is grouped by post owner id and includes body column. The data is then exported into a CSV file to hdfs using pig.

I used 4 mappers and 3 reducers TF-IDF code in python, I ran the CSV file through first mapper and reducer and got the output(See Appendix C). The output then served as the input for the next set of mappers and reducers. The output of the fourth mapper was then saved to the local server in .txt format using -getmerge command. A python script was run to sort the top 10 results based on TF-IDF score.

```
atif_shahab2@hadoop-cluster-m:~$ python sort_Results.py
result Owner User ID

                word  tfidf_score
leaf           14069     0.199142
box           338204     0.132116
release       338204     0.132116
disregarded    18300     0.103321
cable          18300     0.103321
types          13161     0.085281
spider         13161     0.085281
clojure         6068     0.080378
loader        338204     0.075495
man            18300     0.073801



atif_shahab2@hadoop-cluster-m:~$ 4
```

Google DataProc

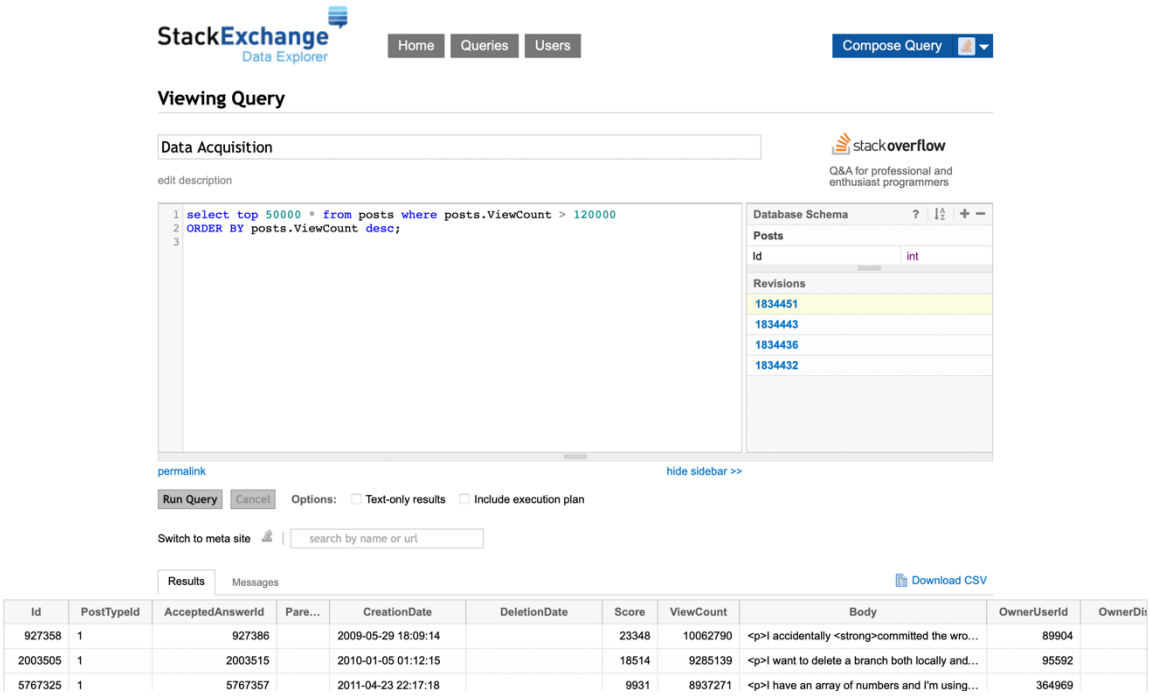
Apart from the local system, steps 2, 3 & 4 were performed on GCP using DataProc. Dataproc is a fully managed service by GCP built for Hadoop. A Dataproc cluster was set up with one name node and 3 data nodes.

References:

1. https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Term_frequency
2. https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python-/blob/master/README.md
3. https://datablog.roman-halliday.com/index.php/2019/10/26/rlike-in-hive-filtering-with-regular-expressions/

Appendix:

A. DATA ACQUISITION



Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDi
927358	1	927386		2009-05-29 18:09:14		23348	10062790	<p>I accidentally committed the wro…	89904	
2003505	1	2003515		2010-01-05 01:12:15		18514	9285139	<p>I want to delete a branch both locally and…	95592	
5767325	1	5767357		2011-04-23 22:17:18		9931	8937271	<p>I have an array of numbers and I'm using…	364969	

StackExchange Data Explorer

Home | Queries | Users

Compose Query

Viewing Query

Data Acquisition Query 2

edit description

stackoverflow
Q&A for professional and enthusiast programmers

```
1 select top 50000 * from posts where posts.ViewCount>74410 and
2 posts.ViewCount < 127079 order by posts.ViewCount desc;
3
```

Database Schema

Posts

Id	int

Revisions

1834477
1834451
1834443
1834436
1834432

permalink | hide sidebar >>

Run Query | Cancel | Options: Text-only results | Include execution plan

Switch to meta site | search by name or url

Results | Messages

Download CSV

Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerD
10858261	1	10858332		2012-06-01 22:48:52		175	127078	How could I abort a make/makefile exec...	561948	
481725	1	601186		2009-01-26 22:56:38		78	127077	I know VB.Net and am trying to brush up...	58375	Dan Appley
8345282	1	8345351		2011-12-01 17:01:07		60	127076	I have a table which has this schema</p...	30512	

StackExchange Data Explorer

Home | Queries | Users

Compose Query

Editing Query

Data Acquisition Query 3

edit description

stackoverflow
Q&A for professional and enthusiast programmers

```
1 select top 50000 * from posts where posts.ViewCount > 52900 and
2 posts.ViewCount < 74410 order by posts.ViewCount desc;
```

Database Schema

Posts

Id	int

Revisions

1834492
1834432 anonymous

permalink | hide sidebar >>

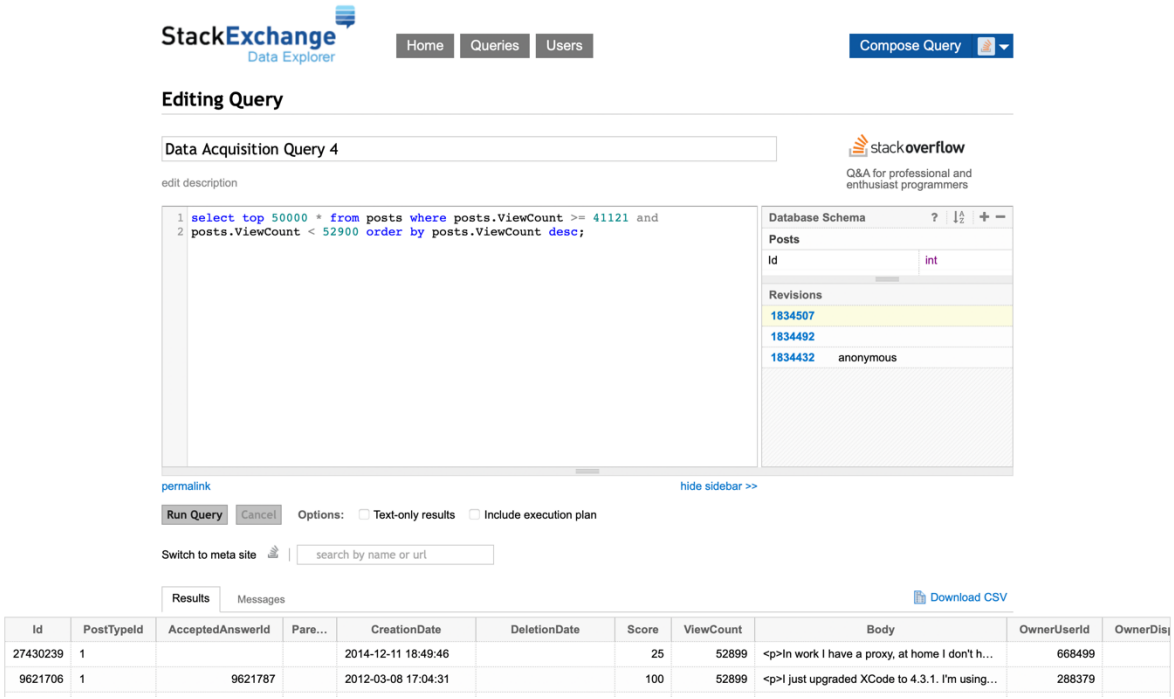
Run Query | Cancel | Options: Text-only results | Include execution plan

Switch to meta site | search by name or url

Results | Messages

Download CSV

Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDi
20629547	1	20629566		2013-12-17 08:45:20		24	74409	I have this kind of datatable: <pre><...	2923670	
14080758	1	14080830		2012-12-29 10:25:34		20	74409	The following code works fine when hea...	1397864	
42844778	1	42845100		2017-03-16 21:06:07		7	74409	I've looked around for this and it seems s...	6377101	



Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisp
27430239	1			2014-12-11 18:49:46		25	52899	<p>In work I have a proxy, at home I don't h...	668499	
9621706	1	9621787		2012-03-08 17:04:31		100	52899	<p>I just upgraded XCode to 4.3.1. I'm using...	288379	

B. Extract, Load and Transform (ELT)

Loading data in pig

```
atif_shahab2@hadoop-cluster-m:~$ pig
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2021-10-28 21:55:09,478 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-10-28 21:55:09,479 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2021-10-28 21:55:09,480 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-28 21:55:09,552 [main] INFO  org.apache.pig.Main - Apache Pig version 0.18.0-SNAPSHOT (r: unknown) compiled Dec 21 1969, 06:05:27
2021-10-28 21:55:09,552 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/atif_shahab2/pig_1635458109548.log
2021-10-28 21:55:09,585 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/atif_shahab2/.pigbootup not found
2021-10-28 21:55:10,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-28 21:55:10,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://hadoop-cluster-m
2021-10-28 21:55:11,362 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-922b49c2-037d-422f-a800-6864f429fd4d
2021-10-28 21:55:11,582 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: hadoop-cluster-m:8188
2021-10-28 21:55:11,977 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2021-10-28 21:55:12,008 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> data1= LOAD '/data_storage/QueryResults4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, acceptedanswerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserId:int, lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:chararray, communityowneddate:chararray, contentlicense:chararray);
2021-10-28 21:56:30,136 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> data2= LOAD '/data_storage/QueryResults4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, acceptedanswerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserId:int, lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:chararray, communityowneddate:chararray, contentlicense:chararray);
2021-10-28 21:56:53,023 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> data3= LOAD '/data_storage/QueryResults4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, acceptedanswerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserId:int, lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:chararray, communityowneddate:chararray, contentlicense:chararray);
2021-10-28 21:57:10,960 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> data4= LOAD '/data_storage/QueryResults4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, acceptedanswerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserId:int, lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:chararray, communityowneddate:chararray, contentlicense:chararray);
2021-10-28 21:57:27,074 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> stackex_data = UNION data1,data2,data3,data4;
grunt>
```

Verify count after loading the data in pig

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1635097397242_0012  4       1       36      15      26      26      3       3       3       3       Stackex_count,Stackex_data,Stackex_data_all,data1,data2,data3,data4      GROUP_BY
,COMBINER       hdfs://hadoop-cluster-m/tmp/temp-1257892967/tmp1979943736,

Input(s):
Successfully read 50000 records from: "/data_storage/QueryResults1.csv"
Successfully read 50000 records from: "/data_storage/QueryResults3.csv"
Successfully read 50000 records from: "/data_storage/QueryResults2.csv"
Successfully read 50000 records from: "/data_storage/QueryResults4.csv"

Output(s):
Successfully stored 1 records (16 bytes) in: "hdfs://hadoop-cluster-m/tmp/temp-1257892967/tmp1979943736"

Counters:
Total records written : 1
Total bytes written : 16
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1635097397242_0012


2021-10-24 22:15:38,455 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 22:15:38,456 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 22:15:38,459 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 22:15:38,493 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 22:15:38,494 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 22:15:38,498 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 22:15:38,523 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 22:15:38,524 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 22:15:38,528 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 22:15:38,570 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-10-24 22:15:38,579 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-me
trics-publisher.enabled
2021-10-24 22:15:38,587 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-10-24 22:15:38,622 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-10-24 22:15:38,625 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-10-24 22:15:38,622 [GetFileInfo #1] WARN  org.apache.hadoop.util.concurrent.ExecutorHelper - Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
(200000)
grunt>
```

Transform extracted data, remove new lines in body column and removes duplicate values.

```
grunt> stackex_data = UNION data1,data2,data3,data4;
grunt> stackex_data = FOREACH stackex_data GENERATE id, score, viewcount, body, owneruserid, title, tags;
grunt> stackex_data = FOREACH stackex_data GENERATE id, score, viewcount, owneruserid, title, tags, (REPLACE(body,'[\r\n]+','')) AS body;
grunt> stackex_data = DISTINCT stackex_data;
grunt>
```

Load data to hive.

```
r
2021-10-24 23:08:52,222 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 23:08:52,223 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 23:08:52,228 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 23:08:52,260 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 23:08:52,261 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 23:08:52,264 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 23:08:52,319 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-10-24 23:08:52,322 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.2   0.18.0-SNAPSHOT atif_shahab2    2021-10-24 23:07:44     2021-10-24 23:08:52     DISTINCT,UNION

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1635097397242_0016  4       1       34      18      26      26      11      11      11      11      Stackex_data,data1,data2,data3,data4    DISTINCT        stackex_data.stackex_Tra
nsformed,

Input(s):
Successfully read 50000 records from: "/data_storage/QueryResults4.csv"
Successfully read 50000 records from: "/data_storage/QueryResults2.csv"
Successfully read 50000 records from: "/data_storage/QueryResults3.csv"
Successfully read 50000 records from: "/data_storage/QueryResults1.csv"

Output(s):
Successfully stored 199999 records (183907166 bytes) in: "stackex_data.stackex_Transformed"

Counters:
Total records written : 199999
Total bytes written : 183907166
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1635097397242_0016


2021-10-24 23:08:52,329 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 23:08:52,334 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 23:08:52,338 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 23:08:52,380 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 23:08:52,381 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 23:08:52,393 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 23:08:52,423 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-24 23:08:52,423 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-24 23:08:52,427 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-24 23:08:52,452 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

C. Mapper & Reducer

Mapper and Reducer 1

```
atif_shahab2@hadoop-cluster-m:~$ hadoop jar hadoop-streaming-3.2.2.jar -files /home/atif_shahab2/Python_MapReduce_mapper1.py,/home/atif_shahab2/Python_MapReduce_reducer1.py -mapper 'py
thon Python_MapReduce_mapper1.py' -reducer 'python Python_MapReduce_reducer1.py' -input hdfs:///PostDataOutput2/part-r-00000 -output hdfs:///data/output1
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob3557391836079832287.jar tmpDir=null
2021-10-26 02:13:34,540 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:13:34,799 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:13:35,472 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:13:35,473 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:13:35,709 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/atif_shahab2/.staging/job_1635199698843_0003
2021-10-26 02:13:36,211 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
2021-10-26 02:13:36,213 INFO mapred.FileInputFormat: Total input files to process : 1
2021-10-26 02:13:36,324 INFO mapreduce.JobSubmitter: number of splits:15
2021-10-26 02:13:36,527 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635199698843_0003
2021-10-26 02:13:36,529 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-26 02:13:36,825 INFO conf.Configuration: resource-types.xml not found
2021-10-26 02:13:36,826 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-26 02:13:36,909 INFO impl.YarnClientImpl: Submitted application application_1635199698843_0003
2021-10-26 02:13:37,039 INFO mapreduce.Job: The url to track the job: http://hadoop-cluster-m:8088/proxy/application_1635199698843_0003/
2021-10-26 02:13:37,040 INFO mapreduce.Job: Running job: job_1635199698843_0003
2021-10-26 02:13:46,153 INFO mapreduce.Job: Job job_1635199698843_0003 running in uber mode : false
2021-10-26 02:13:46,155 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-26 02:13:52,242 INFO mapreduce.Job:  map 7% reduce 0%
2021-10-26 02:13:55,262 INFO mapreduce.Job:  map 20% reduce 0%
2021-10-26 02:13:56,268 INFO mapreduce.Job:  map 40% reduce 0%
2021-10-26 02:14:00,292 INFO mapreduce.Job:  map 47% reduce 0%
2021-10-26 02:14:03,312 INFO mapreduce.Job:  map 60% reduce 0%
2021-10-26 02:14:04,318 INFO mapreduce.Job:  map 80% reduce 0%
2021-10-26 02:14:07,332 INFO mapreduce.Job:  map 87% reduce 0%
2021-10-26 02:14:11,351 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-26 02:14:18,390 INFO mapreduce.Job:  map 100% reduce 20%
2021-10-26 02:14:20,400 INFO mapreduce.Job:  map 100% reduce 60%
2021-10-26 02:14:21,405 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-26 02:14:23,422 INFO mapreduce.Job: Job job_1635199698843_0003 completed successfully
2021-10-26 02:14:23,523 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=100977
                FILE: Number of bytes written=5173044
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=111838
                HDFS: Number of bytes written=40606
                HDFS: Number of read operations=70
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=15
```

```
                HDFS: Number of read operations=70
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=15
                Launched reduce tasks=5
                Data-local map tasks=9
                Rack-local map tasks=6
                Total time spent by all maps in occupied slots (ms)=268655616
                Total time spent by all reduces in occupied slots (ms)=90448896
                Total time spent by all map tasks (ms)=87453
                Total time spent by all reduce tasks (ms)=29443
                Total vcore-milliseconds taken by all map tasks=87453
                Total vcore-milliseconds taken by all reduce tasks=29443
                Total megabyte-milliseconds taken by all map tasks=268655616
                Total megabyte-milliseconds taken by all reduce tasks=90448896
        Map-Reduce Framework
                Map input records=92
                Map output records=6114
                Map output bytes=88719
                Map output materialized bytes=101397
                Input split bytes=1560
                Combine input records=0
                Combine output records=0
                Reduce input groups=2666
                Reduce shuffle bytes=101397
                Reduce input records=6114
                Reduce output records=2666
                Spilled Records=12228
                Shuffled Maps =75
                Failed Shuffles=0
                Merged Map outputs=75
                GC time elapsed (ms)=3195
                CPU time spent (ms)=15910
                Physical memory (bytes) snapshot=9046663168
                Virtual memory (bytes) snapshot=86943686656
                Total committed heap usage (bytes)=7664566272
                Peak Map Physical memory (bytes)=532127744
                Peak Map Virtual memory (bytes)=4349390848
                Peak Reduce Physical memory (bytes)=261910528
                Peak Reduce Virtual memory (bytes)=4350861312
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=110278
        File Output Format Counters
                Bytes Written=40606
2021-10-26 02:14:23,523 INFO streaming.StreamJob: Output directory: hdfs:///data/output1
```

Mapper and Reducer 2

```
atif_shahab2@hadoop-cluster-m:~$ hadoop jar hadoop-streaming-3.2.2.jar -files /home/atif_shahab2/Python_MapReduce_mapper2.py,/home/atif_shahab2/Python_MapReduce_reducer2.py -mapper 'py
thon Python_MapReduce_mapper2.py' -reducer 'python Python_MapReduce_reducer2.py ' -input hdfs:///data/output1/ -output hdfs:///data/output2
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob9167267055307438154.jar tmpDir=null
2021-10-26 02:19:14,160 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:19:14,488 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:19:15,076 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:19:15,077 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:19:15,285 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/atif_shahab2/.staging/job_1635199698843_0004
2021-10-26 02:19:15,699 INFO mapred.FileInputFormat: Total input files to process : 5
2021-10-26 02:19:15,792 INFO mapreduce.JobSubmitter: number of splits:16
2021-10-26 02:19:15,996 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635199698843_0004
2021-10-26 02:19:15,999 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-26 02:19:16,315 INFO conf.Configuration: resource-types.xml not found
2021-10-26 02:19:16,316 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-26 02:19:16,396 INFO impl.YarnClientImpl: Submitted application application_1635199698843_0004
2021-10-26 02:19:16,433 INFO mapreduce.Job: The url to track the job: http://hadoop-cluster-m:8088/proxy/application_1635199698843_0004/
2021-10-26 02:19:16,435 INFO mapreduce.Job: Running job: job_1635199698843_0004
2021-10-26 02:19:25,649 INFO mapreduce.Job: Job job_1635199698843_0004 running in uber mode : false
2021-10-26 02:19:25,650 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-26 02:19:31,719 INFO mapreduce.Job:  map 6% reduce 0%
2021-10-26 02:19:34,742 INFO mapreduce.Job:  map 31% reduce 0%
2021-10-26 02:19:36,754 INFO mapreduce.Job:  map 38% reduce 0%
2021-10-26 02:19:40,778 INFO mapreduce.Job:  map 44% reduce 0%
2021-10-26 02:19:41,786 INFO mapreduce.Job:  map 50% reduce 0%
2021-10-26 02:19:42,794 INFO mapreduce.Job:  map 69% reduce 0%
2021-10-26 02:19:44,806 INFO mapreduce.Job:  map 75% reduce 0%
2021-10-26 02:19:48,828 INFO mapreduce.Job:  map 88% reduce 0%
2021-10-26 02:19:50,839 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-26 02:19:56,875 INFO mapreduce.Job:  map 100% reduce 20%
2021-10-26 02:19:58,885 INFO mapreduce.Job:  map 100% reduce 60%
2021-10-26 02:19:59,902 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-26 02:20:01,920 INFO mapreduce.Job: Job job_1635199698843_0004 completed successfully
2021-10-26 02:20:02,020 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=45968
                FILE: Number of bytes written=5311271
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=76737
                HDFS: Number of bytes written=53017
                HDFS: Number of read operations=73
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=16
                Launched reduce tasks=5
                Data-local map tasks=16
                Total time spent by all maps in occupied slots (ms)=279837696
                Total time spent by all reduces in occupied slots (ms)=83177472
                Total time spent by all map tasks (ms)=91093
                Total time spent by all reduce tasks (ms)=27076
                Total vcore-milliseconds taken by all map tasks=91093
```

```
                HDFS: Number of bytes read=76737
                HDFS: Number of bytes written=53017
                HDFS: Number of read operations=73
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=16
                Launched reduce tasks=5
                Data-local map tasks=16
                Total time spent by all maps in occupied slots (ms)=279837696
                Total time spent by all reduces in occupied slots (ms)=83177472
                Total time spent by all map tasks (ms)=91093
                Total time spent by all reduce tasks (ms)=27076
                Total vcore-milliseconds taken by all map tasks=91093
                Total vcore-milliseconds taken by all reduce tasks=27076
                Total megabyte-milliseconds taken by all map tasks=279837696
                Total megabyte-milliseconds taken by all reduce tasks=83177472
        Map-Reduce Framework
                Map input records=2666
                Map output records=2666
                Map output bytes=40606
                Map output materialized bytes=46418
                Input split bytes=1584
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=46418
                Reduce input records=2666
                Reduce output records=2666
                Spilled Records=5332
                Shuffled Maps =80
                Failed Shuffles=0
                Merged Map outputs=80
                GC time elapsed (ms)=3068
                CPU time spent (ms)=14000
                Physical memory (bytes) snapshot=9427419136
                Virtual memory (bytes) snapshot=91300077568
                Total committed heap usage (bytes)=7973371904
                Peak Map Physical memory (bytes)=535650304
                Peak Map Virtual memory (bytes)=4347183104
                Peak Reduce Physical memory (bytes)=271048704
                Peak Reduce Virtual memory (bytes)=4355829760
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=75153
        File Output Format Counters
                Bytes Written=53017
2021-10-26 02:20:02,020 INFO streaming.StreamJob: Output directory: hdfs:///data/output2
```

Mapper and Reducer 3

```
2021-10-26 02:20:02,020 INFO streaming.StreamJob: Output directory: hdfs:///data/output2
atif_shahab2@hadoop-cluster-m:~$ hadoop jar hadoop-streaming-3.2.2.jar -files /home/atif_shahab2/Python_MapReduce_mapper3.py,/home/atif_shahab2/Python_MapReduce_reducer3.py -mapper 'py
thon Python_MapReduce_mapper3.py' -reducer 'python Python_MapReduce_reducer3.py ' -input hdfs:///data/output2 -output hdfs:///data/output3
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob5867168988210842249.jar tmpDir=null
2021-10-26 02:20:55,784 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:20:56,099 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:20:56,662 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:20:56,663 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:20:56,883 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/atif_shahab2/.staging/job_1635199698843_0005
2021-10-26 02:20:57,321 INFO mapred.FileInputFormat: Total input files to process : 5
2021-10-26 02:20:57,326 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
2021-10-26 02:20:57,399 INFO mapreduce.JobSubmitter: number of splits:18
2021-10-26 02:20:57,584 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635199698843_0005
2021-10-26 02:20:57,587 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-26 02:20:57,909 INFO conf.Configuration: resource-types.xml not found
2021-10-26 02:20:57,910 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-26 02:20:58,006 INFO impl.YarnClientImpl: Submitted application application_1635199698843_0005
2021-10-26 02:20:58,060 INFO mapreduce.Job: The url to track the job: http://hadoop-cluster-m:8088/proxy/application_1635199698843_0005/
2021-10-26 02:20:58,061 INFO mapreduce.Job: Running job: job_1635199698843_0005
2021-10-26 02:21:07,179 INFO mapreduce.Job: Job job_1635199698843_0005 running in uber mode : false
2021-10-26 02:21:07,180 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-26 02:21:13,265 INFO mapreduce.Job:  map 6% reduce 0%
2021-10-26 02:21:15,285 INFO mapreduce.Job:  map 17% reduce 0%
2021-10-26 02:21:16,296 INFO mapreduce.Job:  map 28% reduce 0%
2021-10-26 02:21:17,302 INFO mapreduce.Job:  map 33% reduce 0%
2021-10-26 02:21:21,336 INFO mapreduce.Job:  map 39% reduce 0%
2021-10-26 02:21:23,348 INFO mapreduce.Job:  map 61% reduce 0%
2021-10-26 02:21:26,373 INFO mapreduce.Job:  map 67% reduce 0%
2021-10-26 02:21:30,397 INFO mapreduce.Job:  map 83% reduce 0%
2021-10-26 02:21:31,403 INFO mapreduce.Job:  map 94% reduce 0%
2021-10-26 02:21:34,419 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-26 02:21:40,618 INFO mapreduce.Job:  map 100% reduce 20%
2021-10-26 02:21:42,629 INFO mapreduce.Job:  map 100% reduce 40%
2021-10-26 02:21:43,636 INFO mapreduce.Job:  map 100% reduce 80%
2021-10-26 02:21:44,641 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-26 02:21:45,654 INFO mapreduce.Job: Job job_1635199698843_0005 completed successfully
2021-10-26 02:21:45,753 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=63711
                FILE: Number of bytes written=5843927
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=98738
                HDFS: Number of bytes written=39563
                HDFS: Number of read operations=79
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
```

```
                HDFS: Number of read operations=79
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=15
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=18
                Launched reduce tasks=5
                Other local map tasks=1
                Data-local map tasks=17
                Total time spent by all maps in occupied slots (ms)=306392064
                Total time spent by all reduces in occupied slots (ms)=84968448
                Total time spent by all map tasks (ms)=99737
                Total time spent by all reduce tasks (ms)=27659
                Total vcore-milliseconds taken by all map tasks=99737
                Total vcore-milliseconds taken by all reduce tasks=27659
                Total megabyte-milliseconds taken by all map tasks=306392064
                Total megabyte-milliseconds taken by all reduce tasks=84968448
        Map-Reduce Framework
                Map input records=2666
                Map output records=2666
                Map output bytes=58349
                Map output materialized bytes=64221
                Input split bytes=1782
                Combine input records=0
                Combine output records=0
                Reduce input groups=1764
                Reduce shuffle bytes=64221
                Reduce input records=2666
                Reduce output records=1764
                Spilled Records=5332
                Shuffled Maps =90
                Failed Shuffles=0
                Merged Map outputs=90
                GC time elapsed (ms)=3234
                CPU time spent (ms)=15890
                Physical memory (bytes) snapshot=10359029760
                Virtual memory (bytes) snapshot=100001132544
                Total committed heap usage (bytes)=8779202560
                Peak Map Physical memory (bytes)=534241280
                Peak Map Virtual memory (bytes)=4347543552
                Peak Reduce Physical memory (bytes)=272543744
                Peak Reduce Virtual memory (bytes)=4358041600
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=96956
        File Output Format Counters
                Bytes Written=39563
2021-10-26 02:21:45,759 INFO streaming.StreamJob: Output directory: hdfs:///data/output3
```

Mapper 4

```
atif_shahab2@hadoop-cluster-m:~$ hadoop jar hadoop-streaming-3.2.2.jar -files /home/atif_shahab2/Python_MapReduce_mapper4.py -numReduceTasks 0 -input hdfs:///data/output3/ -output hdfs
:///data/output4 -mapper 'python Python_MapReduce_mapper4.py'
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob2200861851020617955.jar tmpDir=null
2021-10-26 02:26:24,090 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:26:24,342 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:26:24,987 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-m/10.128.0.3:8032
2021-10-26 02:26:24,987 INFO client.AHSProxy: Connecting to Application History server at hadoop-cluster-m/10.128.0.3:10200
2021-10-26 02:26:25,235 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/atif_shahab2/.staging/job_1635199698843_0006
2021-10-26 02:26:25,595 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
2021-10-26 02:26:25,601 INFO mapred.FileInputFormat: Total input files to process : 5
2021-10-26 02:26:25,685 INFO mapreduce.JobSubmitter: number of splits:16
2021-10-26 02:26:25,872 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635199698843_0006
2021-10-26 02:26:25,875 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-26 02:26:26,135 INFO conf.Configuration: resource-types.xml not found
2021-10-26 02:26:26,135 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-26 02:26:26,246 INFO impl.YarnClientImpl: Submitted application application_1635199698843_0006
2021-10-26 02:26:26,303 INFO mapreduce.Job: The url to track the job: http://hadoop-cluster-m:8088/proxy/application_1635199698843_0006/
2021-10-26 02:26:26,304 INFO mapreduce.Job: Running job: job_1635199698843_0006
2021-10-26 02:26:34,526 INFO mapreduce.Job: Job job_1635199698843_0006 running in uber mode : false
2021-10-26 02:26:34,527 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-26 02:26:40,611 INFO mapreduce.Job:  map 6% reduce 0%
2021-10-26 02:26:43,635 INFO mapreduce.Job:  map 25% reduce 0%
2021-10-26 02:26:44,645 INFO mapreduce.Job:  map 31% reduce 0%
2021-10-26 02:26:45,651 INFO mapreduce.Job:  map 38% reduce 0%
2021-10-26 02:26:50,680 INFO mapreduce.Job:  map 44% reduce 0%
2021-10-26 02:26:51,687 INFO mapreduce.Job:  map 69% reduce 0%
2021-10-26 02:26:54,707 INFO mapreduce.Job:  map 75% reduce 0%
2021-10-26 02:26:57,725 INFO mapreduce.Job:  map 81% reduce 0%
2021-10-26 02:26:58,731 INFO mapreduce.Job:  map 88% reduce 0%
2021-10-26 02:26:59,737 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-26 02:27:01,757 INFO mapreduce.Job: Job job_1635199698843_0006 completed successfully
2021-10-26 02:27:01,850 INFO mapreduce.Job: Counters: 34
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=3962534
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=75364
                HDFS: Number of bytes written=39819
                HDFS: Number of read operations=112
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=48
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=16
                Data-local map tasks=15
                Rack-local map tasks=1
```

```
2021-10-26 02:26:26,304 INFO mapreduce.Job: Running job: job_1635199698843_0006
2021-10-26 02:26:34,526 INFO mapreduce.Job: Job job_1635199698843_0006 running in uber mode : false
2021-10-26 02:26:34,527 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-26 02:26:40,611 INFO mapreduce.Job:  map 6% reduce 0%
2021-10-26 02:26:43,635 INFO mapreduce.Job:  map 25% reduce 0%
2021-10-26 02:26:44,645 INFO mapreduce.Job:  map 31% reduce 0%
2021-10-26 02:26:45,651 INFO mapreduce.Job:  map 38% reduce 0%
2021-10-26 02:26:50,680 INFO mapreduce.Job:  map 44% reduce 0%
2021-10-26 02:26:51,687 INFO mapreduce.Job:  map 69% reduce 0%
2021-10-26 02:26:54,707 INFO mapreduce.Job:  map 75% reduce 0%
2021-10-26 02:26:57,725 INFO mapreduce.Job:  map 81% reduce 0%
2021-10-26 02:26:58,731 INFO mapreduce.Job:  map 88% reduce 0%
2021-10-26 02:26:59,737 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-26 02:27:01,757 INFO mapreduce.Job: Job job_1635199698843_0006 completed successfully
2021-10-26 02:27:01,850 INFO mapreduce.Job: Counters: 34
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=3962534
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=75364
                HDFS: Number of bytes written=39819
                HDFS: Number of read operations=112
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=48
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=16
                Data-local map tasks=15
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=288457728
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=93899
                Total vcore-milliseconds taken by all map tasks=93899
                Total megabyte-milliseconds taken by all map tasks=288457728
        Map-Reduce Framework
                Map input records=1764
                Map output records=1764
                Input split bytes=1584
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2413
                CPU time spent (ms)=11630
                Physical memory (bytes) snapshot=4103008256
                Virtual memory (bytes) snapshot=69618614272
                Total committed heap usage (bytes)=2930769920
                Peak Map Physical memory (bytes)=268533760
                Peak Map Virtual memory (bytes)=4355706880
        File Input Format Counters
                Bytes Read=73780
        File Output Format Counters
                Bytes Written=39819
2021-10-26 02:27:01,850 INFO streaming.StreamJob: Output directory: hdfs:///data/output4
```

Merge Mapper Reducer output to text file.

```
atif_shahab2@hadoop-cluster-m:~$ hadoop fs -getmerge /data/output4 /home/atif_shahab2/tfidResults/result.txt
atif_shahab2@hadoop-cluster-m:~$ more /home/atif_shahab2/tfidResults/result.txt
proper  18300   0.020634
import pandas as pd;
property        12870   0.000985
provide 12870   0.003769
provides        6068    0.001182
public  6068    0.002822
publickeytoken  338204  0.018874
pushed  12870   0.001884
pushing 14069   0.012446
python  12870   0.001702
query   12870   0.001317
queryset        12870   0.003769
question        12870   0.001254
quick   95592   0.001777
ran     87234   0.00468
rand    6068    0.001182
readline        6068    0.001182
readtoend       6068    0.001182
really  12870   0.001317
rear    6068    0.001182
recommended     6068    0.000826
redirecttoaction        6068    0.001182
referenced      6068    0.007092
reflection      87234   0.00468
region  95592   0.003553
remotely        6068    0.002364
remove  6068    0.000826
removed 12870   0.001884
removing        12870   0.002634
rest    89904   0.005443
revert  12870   0.001884
```

Python sorting script

```
atif_shahab2@hadoop-cluster-m:~$ python sort_Results.py
result Owner User ID

                word  tfidf_score
leaf           14069     0.199142
box           338204     0.132116
release       338204     0.132116
disregarded    18300     0.103321
cable          18300     0.103321
types          13161     0.085281
spider         13161     0.085281
clojure         6068     0.080378
loader        338204     0.075495
man            18300     0.073801



atif_shahab2@hadoop-cluster-m:~$ 4
```