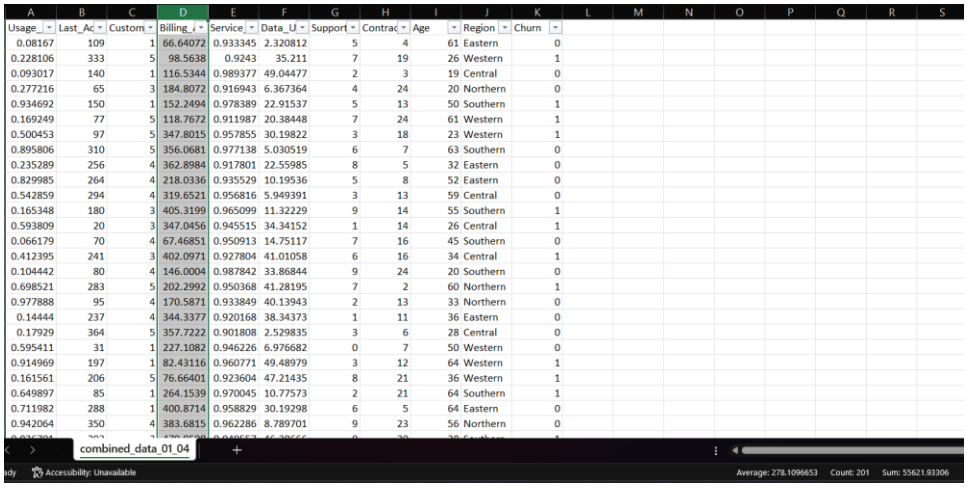
WQD 7007 BDM LAB TEST

Name	Matric No	Group
Safwan Shamsir	S2195293	2
Kar Hong Sam	S2191926	1

Part 1:

1. Download 2 dataset from Google Classroom, combine them and import to HDFS
    a. *Set01* and *Set04* dataset were downloaded and combined using Microsoft Excel. There are 200 rows from the combined dataset.



Usage_	Last_Ac	Custom	Billing_	Service	Data_U	Support	Contrac	Age	Region	Churn
0.08167	109	1	66.64072	0.933345	2.320812	5	4	61	Eastern	0
0.228106	333	5	98.5638	0.9243	35.211	7	19	26	Western	1
0.093017	140	1	116.5344	0.989377	49.04477	2	3	19	Central	0
0.277216	65	3	184.8072	0.916943	6.367364	4	24	20	Northern	0
0.934692	150	1	152.2494	0.978389	22.91537	5	13	50	Southern	1
0.169249	77	5	118.7672	0.911987	20.38448	7	24	61	Western	1
0.500453	97	5	347.8015	0.957855	30.19822	3	18	23	Western	1
0.895806	310	5	356.0681	0.977138	5.030519	6	7	63	Southern	0
0.235289	256	4	362.8984	0.917801	22.55985	8	5	32	Eastern	0
0.829985	264	4	218.0336	0.935529	10.19536	5	8	52	Eastern	0
0.542859	294	4	319.6521	0.956816	5.949391	3	13	59	Central	0
0.165348	180	3	405.3199	0.965099	11.32229	9	14	55	Southern	1
0.593809	20	3	347.0456	0.945515	34.34152	1	14	26	Central	1
0.066179	70	4	67.46851	0.950913	14.75117	7	16	45	Southern	0
0.412395	241	3	402.0971	0.927804	41.01058	6	16	34	Central	1
0.104442	80	4	146.0004	0.987842	33.86844	9	24	20	Southern	0
0.698521	283	5	202.2992	0.950368	41.28195	7	2	60	Northern	1
0.977888	95	4	170.5871	0.933849	40.13943	2	13	33	Northern	0
0.14444	237	4	344.3377	0.920168	38.34373	1	11	36	Eastern	0
0.17929	364	5	357.7222	0.901808	2.529835	3	6	28	Central	0
0.595411	31	1	227.1082	0.946226	6.976682	0	7	50	Western	0
0.914969	197	1	82.43116	0.960771	49.48979	3	12	64	Western	1
0.161561	206	5	76.66401	0.923604	47.21435	8	21	36	Western	1
0.649897	85	1	264.1539	0.970045	10.77573	2	21	64	Southern	1
0.711982	288	1	400.8714	0.958829	30.19298	6	5	64	Eastern	0
0.942064	350	4	383.6815	0.962286	8.789701	9	23	56	Northern	0

    b. Combined dataset imported to HDFS:

```
student@student-VirtualBox:~$ hdfs dfs -put ~/Downloads/churn.csv /user/hdfs/
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs/
Found 7 items
drwxr-xr-x   - student supergroup          0 2024-01-03 16:20 /user/hdfs/Batting
-rw-r--r--   1 student supergroup    6398990 2024-01-03 15:49 /user/hdfs/Batting.csv
-rw-r--r--   1 student supergroup    6398990 2019-04-30 00:14 /user/hdfs/batting.csv
-rw-r--r--   1 student supergroup      14591 2024-01-06 09:23 /user/hdfs/churn.csv
drwxr-xr-x   - student supergroup          0 2024-01-05 10:52 /user/hdfs/geolocation
-rw-r--r--   1 student supergroup     526677 2024-01-05 09:45 /user/hdfs/geolocation.csv
-rw-r--r--   1 student supergroup        235 2019-05-07 15:19 /user/hdfs/student_details.txt
student@student-VirtualBox:~$ hdfs dfs -mkdir /user/hdfs/churn
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs/
Found 8 items
drwxr-xr-x   - student supergroup          0 2024-01-03 16:20 /user/hdfs/Batting
-rw-r--r--   1 student supergroup    6398990 2024-01-03 15:49 /user/hdfs/Batting.csv
-rw-r--r--   1 student supergroup    6398990 2019-04-30 00:14 /user/hdfs/batting.csv
drwxr-xr-x   - student supergroup          0 2024-01-06 09:24 /user/hdfs/churn
-rw-r--r--   1 student supergroup      14591 2024-01-06 09:23 /user/hdfs/churn.csv
drwxr-xr-x   - student supergroup          0 2024-01-05 10:52 /user/hdfs/geolocation
-rw-r--r--   1 student supergroup     526677 2024-01-05 09:45 /user/hdfs/geolocation.csv
-rw-r--r--   1 student supergroup        235 2019-05-07 15:19 /user/hdfs/student_details.txt
student@student-VirtualBox:~$ hdfs dfs -put ~/Downloads/churn.csv /user/hdfs/churn/
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs/churn/
Found 1 items
-rw-r--r--   1 student supergroup      14591 2024-01-06 09:25 /user/hdfs/churn/churn.csv
student@student-VirtualBox:~$
```

c. Sample data from terminal:

```
student@student-VirtualBox:~$ hadoop fs -cat /user/hdfs/churn/churn.csv
Usage_Frequency,Last_Activity,Customer_Satisfaction,Billing_Amount,Service_Uptime,Data_Usage,Support_Contacts,Contract_Length,Age,Region,Churn
0.081669601,109,1,66.64071627,0.93334481,2.320811678,5,4,61,Eastern,0
0.228105935,333,5,98.56380129,0.924299801,35.21100313,7,19,26,Western,1
0.093017072,140,1,116.5343845,0.989376776,49.04476889,2,3,19,Central,0
0.277215909,65,3,184.8072108,0.916943306,6.367364067,4,24,20,Northern,0
0.934691574,150,1,152.2493843,0.9783889,22.91537166,5,13,50,Southern,1
0.169249382,77,5,118.767199,0.911986874,20.38448261,7,24,61,Western,1
0.500452515,97,5,347.8015008,0.957855283,30.19822045,3,18,23,Western,1
0.895806178,310,5,356.068064,0.977137999,5.030518696,6,7,63,Southern,0
0.235289051,256,4,362.8984103,0.917800879,22.5598548,8,5,32,Eastern,0
0.829985114,264,4,218.03364,0.935529053,10.19536136,5,8,52,Eastern,0
0.54285897,294,4,319.6521133,0.956816175,5.949390564,3,13,59,Central,0
0.16534793,180,3,405.3199251,0.965098934,11.32228933,9,14,55,Southern,1
0.593808847,20,3,347.0456475,0.945514507,34.3415165,1,14,26,Central,1
0.066178777,70,4,67.46850975,0.950913351,14.75116959,7,16,45,Southern,0
0.412394683,241,3,402.0970504,0.927804063,41.01057706,6,16,34,Central,1
0.104441718,80,4,146.0004202,0.987841793,33.86844386,9,24,20,Southern,0
0.698521129,283,5,202.2992431,0.950367519,41.28195007,7,2,60,Northern,1
0.977887586,95,4,170.5870528,0.93384872,40.13942807,2,13,33,Northern,0
0.144439708,237,4,344.3376733,0.920168083,38.34372995,1,11,36,Eastern,0
0.179290216,364,5,357.7221711,0.90180754,5.529834662,3,6,28,Central,0
0.595411196,31,1,227.1082111,0.946226469,6.976682267,0,7,50,Western,0
0.914968825,197,1,82.43115544,0.960770966,49.4897866,3,12,64,Western,1
0.161561166,206,5,76.66401005,0.92360399,47.21435101,8,21,36,Western,1
0.649896572,85,1,264.1538627,0.970044707,10.77572645,2,21,64,Southern,1
0.711982322,288,1,400.8714437,0.958829485,30.19298427,6,5,64,Eastern,0
0.942063505,350,4,383.6814659,0.962285915,8.789700797,9,23,56,Northern,0
```

d. Table churn_01 created in Hive:

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS churn_01(
    > Usage_Frequency DOUBLE, Last_Activity INT, Customer_Satisfaction INT, Billing_Amount DOUBLE,
    > Service_Uptime DOUBLE, Data_Usage DOUBLE, Support_Contacts INT, Contract_Length INT,
    > Age INT, Region STRING, Churn INT)
    > COMMENT 'Customer churn'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hdfs/churn';
OK
Time taken: 0.466 seconds
```

2. By using Hive or Pig, identify:
    a. 10 users that have the highest billing amount.
        i. SELECT * FROM churn_01 ORDER BY Biling_Amount DESC LIMIT 10;

```
hive> SELECT * FROM churn_01 ORDER BY Billing_Amount DESC LIMIT 10;
Query ID = student_20240106101046_6567822f-4301-4e17-b183-4ff2d5a7c224
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1704355577006_0004, Tracking URL = http://student-VirtualBox:8088/proxy/application_1704355577006_0004/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1704355577006_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-01-06 10:10:58,142 Stage-1 map = 0%,  reduce = 0%
2024-01-06 10:11:15,707 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.55 sec
2024-01-06 10:11:32,254 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.42 sec
MapReduce Total cumulative CPU time: 4 seconds 420 msec
Ended Job = job_1704355577006_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.42 sec   HDFS Read: 24383 HDFS Write: 712 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 420 msec
OK
0.105353938     302     4       498.6066252     0.984372547     13.71951703     10      23      57      Northern        0
0.195513816     38      1       491.8963747     0.969490447     18.71646587     7       9       41      Central 1
0.499909376     265     2       486.7871229     0.972367894     6.262822075     7       24      58      Western 0
0.02521134      298     3       484.1446921     0.956908783     7.178390207     4       13      59      Northern        1
0.361992666     318     5       479.9746906     0.906474763     14.53174742     4       23      57      Eastern 0
0.936701113     292     3       479.9597748     0.949556911     46.28665593     0       20      28      Southern        1
0.670961316     87      5       479.1690703     0.972586644     36.11769113     9       10      50      Central 0
0.42420237      72      5       478.6455165     0.921999752     35.72008905     7       11      27      Western 0
0.70489576      99      2       478.4431882     0.9905455       39.07493761     2       15      60      Eastern 1
0.432363972     74      1       477.447117      0.911710295     41.39624344     4       12      60      Eastern 1
Time taken: 48.27 seconds, Fetched: 10 row(s)
```

b. 3 central and 3 eastern users that have the lowest data usage. (i for central and ii for eastern)
    i. SELECT * FROM churn_01 WHERE Region IN (‘Central’) ORDER BY Data_Usage DESC LIMIT 3;

```
hive> SELECT * FROM churn_01 WHERE Region IN ('Central') ORDER BY Data_Usage LIMIT 3;
Query ID = student_20240106101533_5d28544b-b2ff-403f-88ae-b01808824b2c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1704355577006_0005, Tracking URL = http://student-VirtualBox:8088/proxy/application_1704355577006_0005/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1704355577006_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-01-06 10:15:46,673 Stage-1 map = 0%,  reduce = 0%
2024-01-06 10:16:27,726 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.18 sec
2024-01-06 10:16:41,881 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.99 sec
MapReduce Total cumulative CPU time: 6 seconds 990 msec
Ended Job = job_1704355577006_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.99 sec   HDFS Read: 25202 HDFS Write: 212 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 990 msec
OK
0.711570516     145     5       433.4193783     0.978722991     0.609654917     4       6       35      Central 0
0.179290216     364     5       357.7221711     0.901807545     2.529834662     3       6       28      Central 0
0.146174915     7       2       417.5168202     0.932615807     2.792959592     9       18      61      Central 1
Time taken: 74.58 seconds, Fetched: 3 row(s)
```

    ii. SELECT * FROM churn_01 WHERE Region IN (Eastern) ORDER BY Data_Usage DESC LIMIT 3;

```
hive> SELECT * FROM churn_01 WHERE Region IN ('Eastern') ORDER BY Data_Usage LIMIT 3;
Query ID = student_20240106103434_367db1c5-4a65-4c2d-983d-f978d786a89f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1704508175616_0002, Tracking URL = http://student-VirtualBox:8088/proxy/application_1704508175616_0002/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1704508175616_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-01-06 10:34:44,537 Stage-1 map = 0%,  reduce = 0%
2024-01-06 10:34:52,333 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.58 sec
2024-01-06 10:34:58,874 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.87 sec
MapReduce Total cumulative CPU time: 2 seconds 870 msec
Ended Job = job_1704508175616_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.87 sec   HDFS Read: 25202 HDFS Write: 211 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 870 msec
OK
0.15334581      261     4       405.9460404     0.925429628     0.017120621     7       1       20      Eastern 1
0.153654139     346     4       149.0885906     0.970123889     1.748349401     2       4       55      Eastern 1
0.081669601     109     1       66.64071627     0.93334481      2.320811678     5       4       61      Eastern 0
Time taken: 26.192 seconds, Fetched: 3 row(s)
```

c. 5 users that shows the biggest contrast from their customer satisfaction and churn decision. Justify from your data why the contrast happens. Suspecting the billing amount is too expensive leads to customers not willing to continue the services.
    i. SELECT * FROM churn_01 WHERE Churn = 0 AND Customer_Satisfaction = 5 ORDER BY Customer_Satisfaction DESC LIMIT 5;

```
hive> SELECT * FROM churn_01 WHERE Churn = 1 AND Customer_Satisfaction = 5 LIMIT 5;
OK
0.228105935     333     5       98.56380129     0.924299801     35.21100313     7       19      26      Western 1
0.169249382     77      5       118.767199      0.911986874     20.38448261     7       24      61      Western 1
0.500452515     97      5       347.8015008     0.957855283     30.19822045     3       18      23      Western 1
0.698521129     283     5       202.2992431     0.950367519     41.28195007     7       2       60      Northern
0.161561166     206     5       76.66401005     0.92360399      47.21435101     8       21      36      Western 1
Time taken: 0.384 seconds, Fetched: 5 row(s)
```

Part 2:

1. Import two sets of text from the specified web link in Google classroom to HDFS.

```
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs
Found 2 items
-rw-r--r--   1 student supergroup    6398990 2019-04-30 00:14 /user/hdfs/batting
.csv
-rw-r--r--   1 student supergroup        235 2019-05-07 15:19 /user/hdfs/student
_details.txt
student@student-VirtualBox:~$ hdfs dfs -put set1.txt /user/hdfs/set1.txt
student@student-VirtualBox:~$ hdfs dfs -put set2.txt /user/hdfs/set2.txt
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs/
Found 4 items
-rw-r--r--   1 student supergroup    6398990 2019-04-30 00:14 /user/hdfs/batting
.csv
-rw-r--r--   1 student supergroup    1276266 2024-01-06 10:37 /user/hdfs/set1.tx
t
-rw-r--r--   1 student supergroup     448965 2024-01-06 10:37 /user/hdfs/set2.tx
t
-rw-r--r--   1 student supergroup        235 2019-05-07 15:19 /user/hdfs/student
_details.txt
```

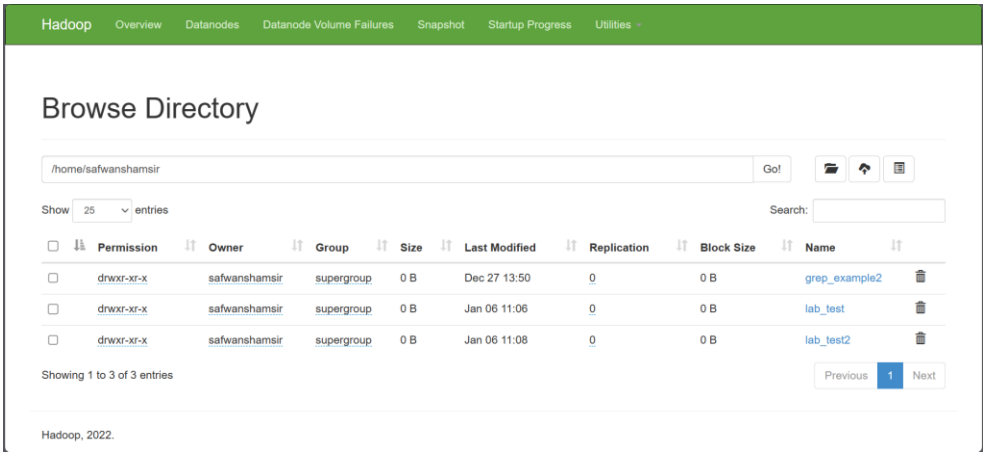
2. Run a word count program using Hadoop MapReduce concept to count the word occurrence of the imported texts as in step 1. Save the results in HDFS.
    a. Set1.txt

```
safwanshamsir@LAPTOP-F1RNLB18:~$ /home/safwanshamsir/hadoop/bin/hadoop jar /home/safwanshamsir/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.2.jar wordcoun
t /user/hdfs/set1.txt ~/lab_test
24/01/06 11:06:18 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
24/01/06 11:06:19 INFO input.FileInputFormat: Total input files to process : 1
24/01/06 11:06:20 INFO mapreduce.JobSubmitter: number of splits:1
24/01/06 11:06:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704509301219_0002
24/01/06 11:06:20 INFO conf.Configuration: resource-types.xml not found
24/01/06 11:06:20 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
24/01/06 11:06:20 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/01/06 11:06:20 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
24/01/06 11:06:22 INFO impl.YarnClientImpl: Submitted application application_1704509301219_0002
24/01/06 11:06:22 INFO mapreduce.Job: The url to track the job: http://LAPTOP-F1RNLB18.:8088/proxy/application_1704509301219_0002/
24/01/06 11:06:22 INFO mapreduce.Job: Running job: job_1704509301219_0002
24/01/06 11:06:32 INFO mapreduce.Job: Job job_1704509301219_0002 running in uber mode : false
24/01/06 11:06:32 INFO mapreduce.Job:  map 0% reduce 0%
24/01/06 11:06:39 INFO mapreduce.Job:  map 100% reduce 0%
24/01/06 11:06:46 INFO mapreduce.Job:  map 100% reduce 100%
24/01/06 11:06:49 INFO mapreduce.Job: Job job_1704509301219_0002 completed successfully
24/01/06 11:06:49 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=504574
                FILE: Number of bytes written=1429895
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1276366
                HDFS: Number of bytes written=372668
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11564
```

b. Set4.txt

```
safwanshamsir@LAPTOP-F1RNLB18:~$ /home/safwanshamsir/hadoop/bin/hadoop jar /home/safwanshamsir/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.2.jar wordcoun
t /user/hdfs/set4.txt ~/lab_test2
24/01/06 11:08:13 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
24/01/06 11:08:13 INFO input.FileInputFormat: Total input files to process : 1
24/01/06 11:08:13 INFO mapreduce.JobSubmitter: number of splits:1
24/01/06 11:08:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704509301219_0003
24/01/06 11:08:14 INFO conf.Configuration: resource-types.xml not found
24/01/06 11:08:14 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
24/01/06 11:08:14 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/01/06 11:08:14 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
24/01/06 11:08:14 INFO impl.YarnClientImpl: Submitted application application_1704509301219_0003
24/01/06 11:08:14 INFO mapreduce.Job: The url to track the job: http://LAPTOP-F1RNLB18.:8088/proxy/application_1704509301219_0003/
24/01/06 11:08:14 INFO mapreduce.Job: Running job: job_1704509301219_0003
24/01/06 11:08:21 INFO mapreduce.Job: Job job_1704509301219_0003 running in uber mode : false
24/01/06 11:08:21 INFO mapreduce.Job:  map 0% reduce 0%
24/01/06 11:08:37 INFO mapreduce.Job:  map 100% reduce 0%
24/01/06 11:08:43 INFO mapreduce.Job:  map 100% reduce 100%
24/01/06 11:08:44 INFO mapreduce.Job: Job job_1704509301219_0003 completed successfully
24/01/06 11:08:44 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=177344
                FILE: Number of bytes written=775437
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=449065
                HDFS: Number of bytes written=129564
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
```

Result:



3. Import the result from step 2 to Hive or Pig. Display:
    a. 10 words with 5 counts in ascending alphabetical order.
    b. 10 words with lowest counts in descending alphabetical order.