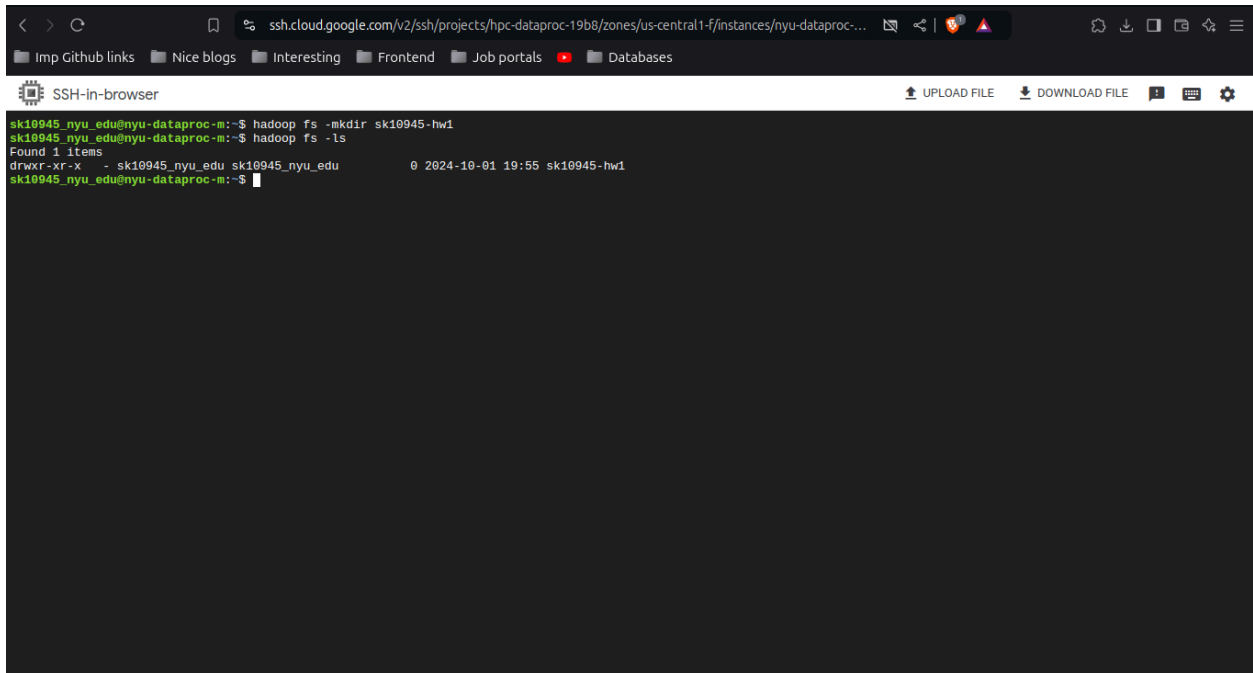


Big Data HW1

1. HDFS

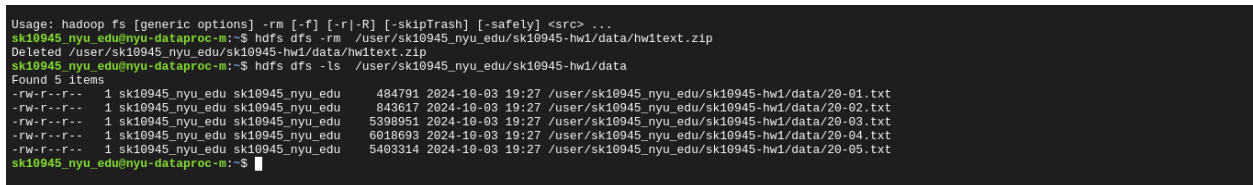
20 points

- a) create a directory in HDFS with this format: netid-hw1



```
ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-...
Imp Github links Nice blogs Interesting Frontend Job portals Databases
SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
sk10945_nyu_edu@nyu-dataproc-m:~$ hadoop fs -mkdir sk10945-hw1
sk10945_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls
Found 1 items
drwxr-xr-x - sk10945_nyu_edu sk10945_nyu_edu 0 2024-10-01 19:55 sk10945-hw1
sk10945_nyu_edu@nyu-dataproc-m:~$
```

- b) Create a subdirectory in HDFS, `netid-hw1/data` and extract all input files into it.



```
Usage: hadoop fs [generic options] -rm [-f] [-r]-R] [-skipTrash] [-safely] <src> ...
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -rm /user/sk10945_nyu_edu/sk10945-hw1/data/hw1text.zip
Deleted /user/sk10945_nyu_edu/sk10945-hw1/data/hw1text.zip
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/sk10945_nyu_edu/sk10945-hw1/data
Found 5 items
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 484791 2024-10-03 19:27 /user/sk10945_nyu_edu/sk10945-hw1/data/20-01.txt
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 843617 2024-10-03 19:27 /user/sk10945_nyu_edu/sk10945-hw1/data/20-02.txt
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 5398951 2024-10-03 19:27 /user/sk10945_nyu_edu/sk10945-hw1/data/20-03.txt
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 6018693 2024-10-03 19:27 /user/sk10945_nyu_edu/sk10945-hw1/data/20-04.txt
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 5403314 2024-10-03 19:27 /user/sk10945_nyu_edu/sk10945-hw1/data/20-05.txt
sk10945_nyu_edu@nyu-dataproc-m:~$
```

2. Beginner's Language Models with MapReduce

2.1 10 Most likely words, 100 points:

Running Job 1 :

```
SSH-in-browser UPLOAD FILE DOWNLOAD FILE 🗨 📄 ⚙

sk10945 nyu_edu@nyu-dataproc-m1:~$ hdfs dfs -rm -r /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output
Deleted /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output
sk10945 nyu_edu@nyu-dataproc-m1:~$ mapred streaming -input /user/sk10945_nyu_edu/sk10945-hw1/data/* -output /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output/ -map
per "python3 job1_mapper.py" -reducer "python3 job1_reducer.py" -file job1_mapper.py -file job1_reducer.py -numReduceTasks 5
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-10-06 15:10:01,380 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [job1_mapper.py, job1_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.6.6.jar] /tmp/streamjob16230015125937805572.jar tmpDir=null
2024-10-06 15:10:02,478 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:10:02,648 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:10:03,455 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:10:03,455 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:10:03,699 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sk10945_nyu_edu/.staging/job_1724767128407_4254
2024-10-06 15:10:04,114 INFO mapred.FileInputFormat: Total input files to process : 5
2024-10-06 15:10:04,186 INFO mapreduce.JobSubmitter: number of splits:8
2024-10-06 15:10:04,364 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724767128407_4254
2024-10-06 15:10:04,365 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-06 15:10:04,550 INFO conf.Configuration: resource-types.xml not found
2024-10-06 15:10:04,550 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-06 15:10:04,705 INFO impl.YarnClientImpl: Submitted application application_1724767128407_4254
2024-10-06 15:10:04,740 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1724767128407_4254/
2024-10-06 15:10:04,742 INFO mapreduce.Job: Running job: job_1724767128407_4254
2024-10-06 15:10:04,884 INFO mapreduce.Job: Job job_1724767128407_4254 running in uber mode : false
2024-10-06 15:10:14,884 INFO mapreduce.Job: map 0% reduce 0%
2024-10-06 15:10:32,084 INFO mapreduce.Job: map 25% reduce 0%
2024-10-06 15:10:33,090 INFO mapreduce.Job: map 88% reduce 0%
2024-10-06 15:10:34,096 INFO mapreduce.Job: map 100% reduce 0%
2024-10-06 15:10:42,140 INFO mapreduce.Job: map 100% reduce 20%
2024-10-06 15:10:43,146 INFO mapreduce.Job: map 100% reduce 48%
2024-10-06 15:10:44,151 INFO mapreduce.Job: map 100% reduce 69%
2024-10-06 15:10:45,164 INFO mapreduce.Job: map 100% reduce 100%
2024-10-06 15:10:46,180 INFO mapreduce.Job: Job job_1724767128407_4254 completed successfully
2024-10-06 15:10:46,282 INFO mapreduce.Job: Counters: 56
  File System Counters
    FILE: Number of bytes read=33803336
    FILE: Number of bytes written=71413580
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=18162622
```

Running Job 2:

```
SSH-in-browser UPLOAD FILE DOWNLOAD FILE 🗨 📄 ⚙

Total megabyte-milliseconds taken by all map tasks=508305408
Total megabyte-milliseconds taken by all reduce tasks=147517440
Map-Reduce Framework
  Map input records=4459
  Map output records=3814636
  Map output bytes=26174934
  Map output materialized bytes=33803546
  Input split bytes=968
  Combine input records=0
  Combine output records=0
  Reduce input groups=81724
  Reduce shuffle bytes=33803546
  Reduce input records=3814636
  Reduce output records=500
  Spilled Records=7629272
  Shuffled Maps =40
  Failed Shuffles=0
  Merged Map outputs=40
  GC time elapsed (ms)=987
  CPU time spent (ms)=58859
  Physical memory (bytes) snapshot=7765975040
  Virtual memory (bytes) snapshot=65063616512
  Total committed heap usage (bytes)=12840861696
  Peak Map Physical memory (bytes)=716083200
  Peak Map Virtual memory (bytes)=5016780800
  Peak Reduce Physical memory (bytes)=479371264
  Peak Reduce Virtual memory (bytes)=5032259584
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=18161654
File Output Format Counters
  Bytes Written=5264
2024-10-06 15:10:46,282 INFO streaming.StreamJob: Output directory: /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output/
```

```

sk10945_nyu_edu@nyu-dataproc-m:~$ mapred streaming -input /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output/" -output /user/sk10945_nyu_edu/sk10945-hw1/final_output
/ -mapper "python3 job2_mapper.py" -reducer "python3 job2_reducer.py" -file job2_mapper.py -file job2_reducer.py -numReduceTasks 1
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-10-06 15:12:31,025 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [job2_mapper.py, job2_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob10906794937796972844.jar tmpDir=null
2024-10-06 15:12:32,149 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:12:32,324 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:12:33,126 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:12:33,127 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:12:33,380 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sk10945_nyu_edu/.staging/job_1724767128407_4255
2024-10-06 15:12:33,796 INFO mapred.FileInputFormat: Total input files to process : 5
2024-10-06 15:12:33,873 INFO mapreduce.JobSubmitter: number of splits:10
2024-10-06 15:12:34,065 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724767128407_4255
2024-10-06 15:12:34,065 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-06 15:12:34,236 INFO conf.Configuration: resource-types.xml not found
2024-10-06 15:12:34,237 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-06 15:12:34,297 INFO impl.YarnClientImpl: Submitted application application_1724767128407_4255
2024-10-06 15:12:34,334 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1724767128407_4255/
2024-10-06 15:12:34,335 INFO mapreduce.Job: Running job: job_1724767128407_4255
2024-10-06 15:12:43,434 INFO mapreduce.Job: Job job_1724767128407_4255 running in uber mode : false
2024-10-06 15:12:43,435 INFO mapreduce.Job: map 0% reduce 0%
2024-10-06 15:12:55,595 INFO mapreduce.Job: map 20% reduce 0%
2024-10-06 15:12:56,601 INFO mapreduce.Job: map 100% reduce 0%
2024-10-06 15:13:04,643 INFO mapreduce.Job: map 100% reduce 100%
2024-10-06 15:13:05,657 INFO mapreduce.Job: Job job_1724767128407_4255 completed successfully
2024-10-06 15:13:05,748 INFO mapreduce.Job: Counters: 55
  File System Counters
    FILE: Number of bytes read=6270
    FILE: Number of bytes written=3233273
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0

```

Final Output :

```

Peak Map Physical memory (bytes)=650129408
Peak Map Virtual memory (bytes)=4975009792
Peak Reduce Physical memory (bytes)=468299776
Peak Reduce Virtual memory (bytes)=5037862912
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=6143
File Output Format Counters
  Bytes Written=90
2024-10-06 15:13:05,748 INFO streaming.StreamJob: Output directory: /user/sk10945_nyu_edu/sk10945-hw1/final_output/
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/sk10945_nyu_edu/sk10945-hw1/final_output
Found 2 items
-rw-r--r--  1 sk10945_nyu_edu sk10945_nyu_edu      0 2024-10-06 15:13 /user/sk10945_nyu_edu/sk10945-hw1/final_output/_SUCCESS
-rw-r--r--  1 sk10945_nyu_edu sk10945_nyu_edu    90 2024-10-06 15:13 /user/sk10945_nyu_edu/sk10945-hw1/final_output/part-00000
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -cat /user/sk10945_nyu_edu/sk10945-hw1/final_output/part-00000
@      151205
      145248
the    142957
,      141936
to     87866
>      84913
<      84913
p      78464
of     75062
and    70811
sk10945_nyu_edu@nyu-dataproc-m:~$

```

Q2.2

Running Job 1 :

```
SSH-in-browser
[Icons] [UPLOAD FILE] [DOWNLOAD FILE] [Flags] [Keyboard] [Settings]

sk10945_nyu_edu@nyu-dataproc-m:~$ mapred streaming -input /user/sk10945_nyu_edu/sk10945-hw1/data/* -output /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output_q2/ -m
apper "python3 q2_job1_mapper.py" -reducer "python3 q2_job1_reducer.py" -file q2_job1_mapper.py -file q2_job1_reducer.py -numReduceTasks 5
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-10-06 15:29:48,937 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [q2_job1_mapper.py, q2_job1_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob14904786112575765718.jar tmpDir=null
2024-10-06 15:29:50,033 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:29:50,210 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:29:51,035 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:29:51,038 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:29:51,294 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sk10945_nyu_edu/.staging/job_1724767128407_4266
2024-10-06 15:29:51,716 INFO mapred.FileInputFormat: Total input files to process : 5
2024-10-06 15:29:51,788 INFO mapreduce.JobSubmitter: number of splits:8
2024-10-06 15:29:51,980 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724767128407_4266
2024-10-06 15:29:51,990 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-06 15:29:52,152 INFO conf.Configuration: resource-types.xml not found
2024-10-06 15:29:52,152 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-06 15:29:52,208 INFO impl.YarnClientImpl: Submitted application application_1724767128407_4266
2024-10-06 15:29:52,242 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1724767128407_4266/
2024-10-06 15:29:52,244 INFO mapreduce.Job: Running job: job_1724767128407_4266
2024-10-06 15:30:01,364 INFO mapreduce.Job: Job job_1724767128407_4266 running in uber mode : false
2024-10-06 15:30:01,365 INFO mapreduce.Job: map 0% reduce 0%
2024-10-06 15:30:19,585 INFO mapreduce.Job: map 13% reduce 0%
2024-10-06 15:30:20,596 INFO mapreduce.Job: map 25% reduce 0%
2024-10-06 15:30:21,604 INFO mapreduce.Job: map 100% reduce 0%
2024-10-06 15:30:28,648 INFO mapreduce.Job: map 100% reduce 20%
2024-10-06 15:30:29,653 INFO mapreduce.Job: map 100% reduce 40%
2024-10-06 15:30:30,661 INFO mapreduce.Job: map 100% reduce 100%
2024-10-06 15:30:32,680 INFO mapreduce.Job: Job job_1724767128407_4266 completed successfully
2024-10-06 15:30:32,779 INFO mapreduce.Job: Counters: 56
File System Counters
  FILE: Number of bytes read=33803336
  FILE: Number of bytes written=71413931
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=18162622
  HDFS: Number of bytes written=850125
  HDFS: Number of read operations=49
  HDFS: Number of large read operations=0
```

Running Job 2:

```
SSH-in-browser
[Icons] [UPLOAD FILE] [DOWNLOAD FILE] [Flags] [Keyboard] [Settings]

Bytes Read=18161654
File Output Format Counters
  Bytes Written=850125
2024-10-06 15:30:32,779 INFO streaming.StreamJob: Output directory: /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output_q2/
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/sk10945_nyu_edu/sk10945-hw1/final_output_q2
ls: /user/sk10945_nyu_edu/sk10945-hw1/final_output_q2: No such file or directory
sk10945_nyu_edu@nyu-dataproc-m:~$ mapred streaming -input /user/sk10945_nyu_edu/sk10945-hw1/intermediate_output_q2/* -output /user/sk10945_nyu_edu/sk10945-hw1/final_out
put_q2/ -mapper "python3 q2_job2_mapper.py" -reducer "python3 q2_job2_reducer.py" -file q2_job2_mapper.py -file q2_job2_reducer.py -numReduceTasks 1
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-10-06 15:31:00,186 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [q2_job2_mapper.py, q2_job2_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob7009764030594486692.jar tmpDir=null
2024-10-06 15:31:01,312 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:31:01,490 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:31:02,352 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:8032
2024-10-06 15:31:02,352 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m.c.hpc-dataproc-19b8.internal./192.168.1.25:10200
2024-10-06 15:31:02,597 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sk10945_nyu_edu/.staging/job_1724767128407_4269
2024-10-06 15:31:02,994 INFO mapred.FileInputFormat: Total input files to process : 5
2024-10-06 15:31:03,068 INFO mapreduce.JobSubmitter: number of splits:10
2024-10-06 15:31:03,260 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724767128407_4269
2024-10-06 15:31:03,260 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-06 15:31:03,437 INFO conf.Configuration: resource-types.xml not found
2024-10-06 15:31:03,437 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-06 15:31:03,492 INFO impl.YarnClientImpl: Submitted application application_1724767128407_4269
2024-10-06 15:31:03,525 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal.:8088/proxy/application_1724767128407_4269/
2024-10-06 15:31:03,526 INFO mapreduce.Job: Running job: job_1724767128407_4269
2024-10-06 15:31:13,628 INFO mapreduce.Job: Job job_1724767128407_4269 running in uber mode : false
2024-10-06 15:31:13,629 INFO mapreduce.Job: map 0% reduce 0%
2024-10-06 15:31:30,750 INFO mapreduce.Job: map 60% reduce 0%
2024-10-06 15:31:31,765 INFO mapreduce.Job: map 100% reduce 0%
2024-10-06 15:31:37,797 INFO mapreduce.Job: map 100% reduce 100%
2024-10-06 15:31:38,811 INFO mapreduce.Job: Job job_1724767128407_4269 completed successfully
2024-10-06 15:31:38,900 INFO mapreduce.Job: Counters: 56
File System Counters
  FILE: Number of bytes read=1730109
  FILE: Number of bytes written=6681270
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=872005
```

Final output : Most Frequent words with count ID

```
SSH-in-browser  UPLOAD FILE  DOWNLOAD FILE  [Icons]

Found 2 items
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 0 2024-10-06 15:31 /user/sk10945_nyu_edu/sk10945-hw1/final_output_q2/_SUCCESS
-rw-r--r-- 1 sk10945_nyu_edu sk10945_nyu_edu 1146929 2024-10-06 15:31 /user/sk10945_nyu_edu/sk10945-hw1/final_output_q2/part-00000
sk10945_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -head /user/sk10945_nyu_edu/sk10945-hw1/final_output_q2/part-00000
1 @
2 .
3 the
4 ,
5 to
6 >
7 <
8 p
9 of
10 and
11 -
12 in
13 a
14 "
15 '
16 for
17 that
18 is
19 on
20 s
21 The
22 are
23 with
24 be
25 ;
26 have
27 as
28 at
29 said
30 from
31 it
32 will
33 has
34 by
35 :
```