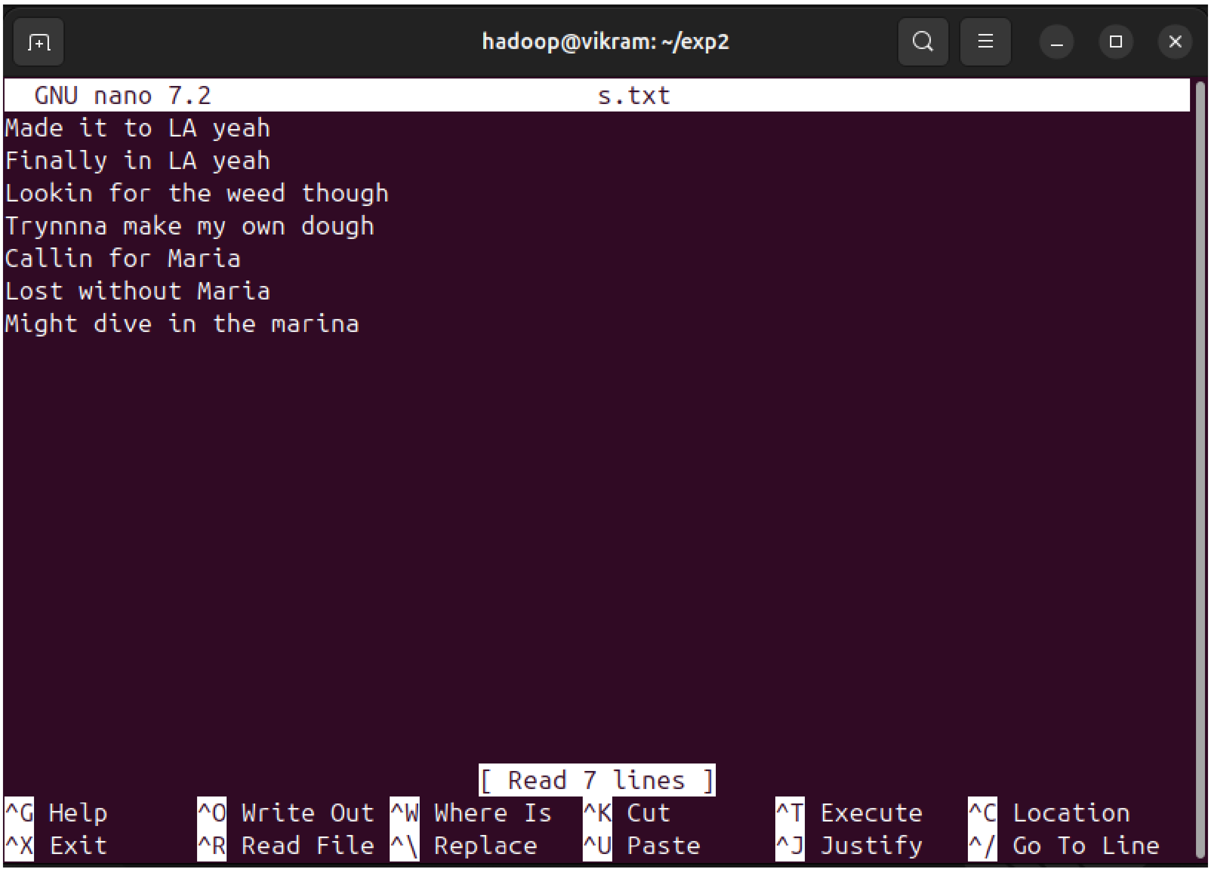
Exp. No : 2

Word Count Map Reduce program

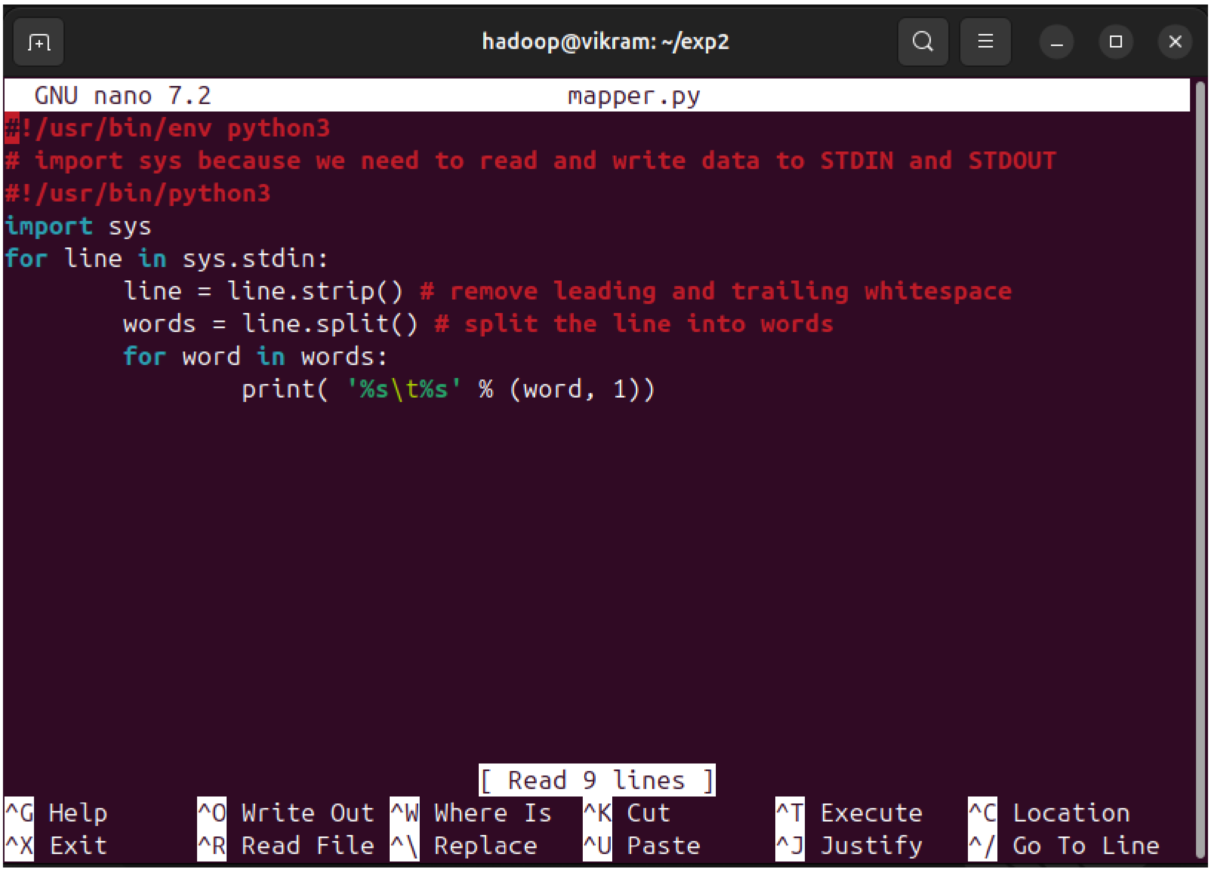
1. Create word_count.txt file



```
  GNU nano 7.2                       s.txt
Made it to LA yeah
Finally in LA yeah
Lookin for the weed though
Trynnna make my own dough
Callin for Maria
Lost without Maria
Might dive in the marina
```

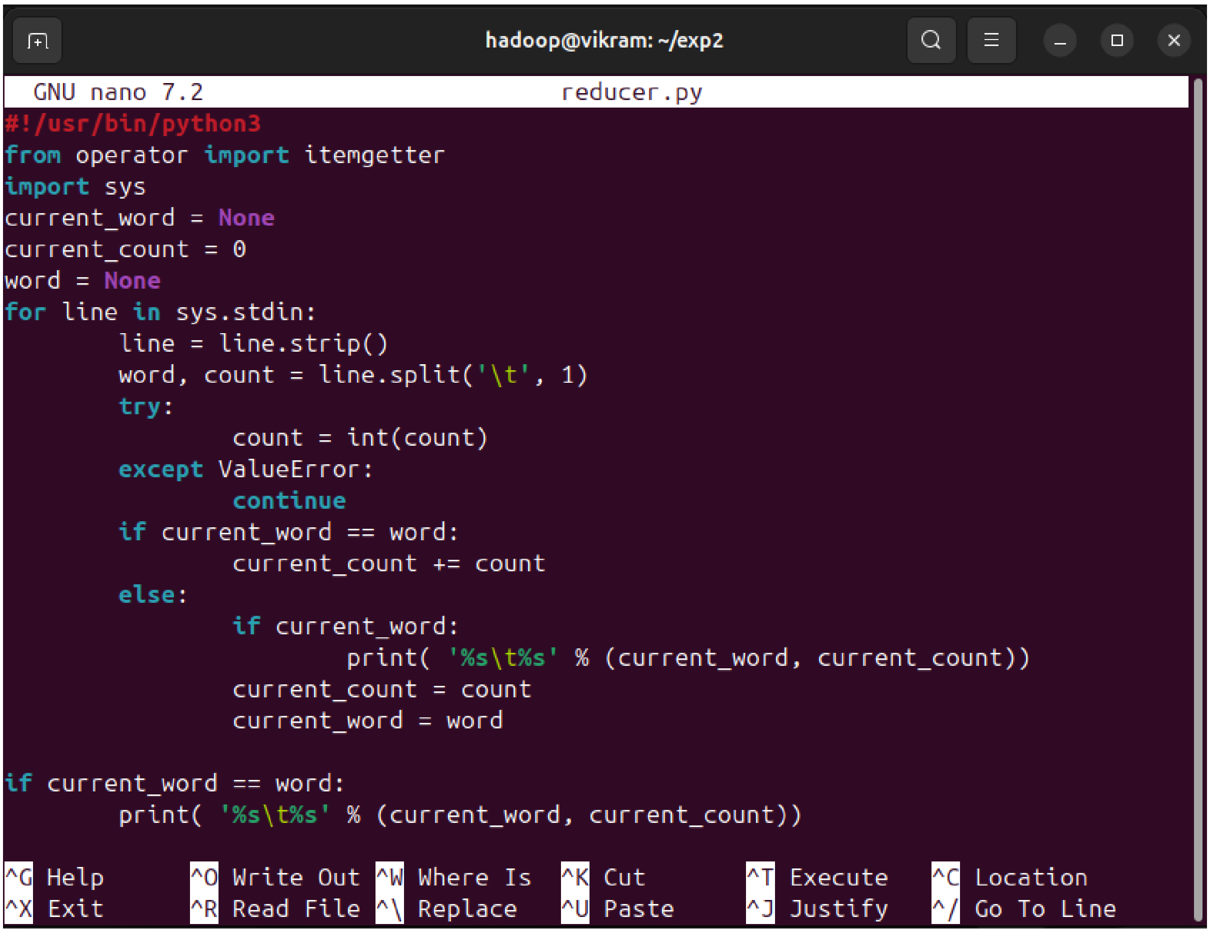
```
                                [ Read 7 lines ]
^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line
```

2. Create mapper.py program



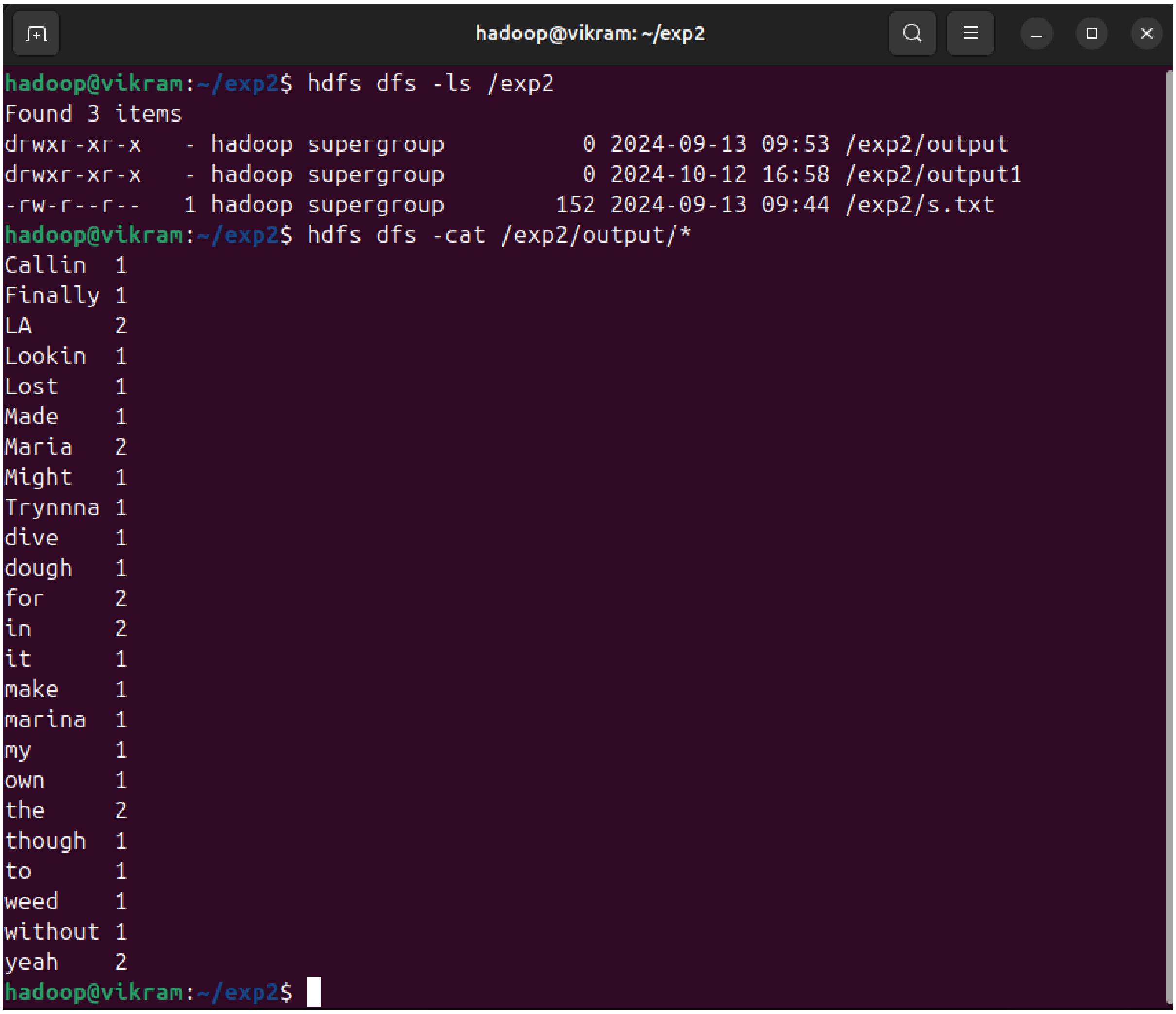
```
  GNU nano 7.2                              mapper.py
#!/usr/bin/env python3
# import sys because we need to read and write data to STDIN and STDOUT
#!/usr/bin/python3
import sys
for line in sys.stdin:
        line = line.strip() # remove leading and trailing whitespace
        words = line.split() # split the line into words
        for word in words:
                print( '%s\t%s' % (word, 1))

                                [ Read 9 lines ]
^G Help        ^O Write Out ^W Where Is  ^K Cut        ^T Execute   ^C Location
^X Exit        ^R Read File ^\ Replace   ^U Paste      ^J Justify   ^/ Go To Line
```

3. Create reducer.py program.



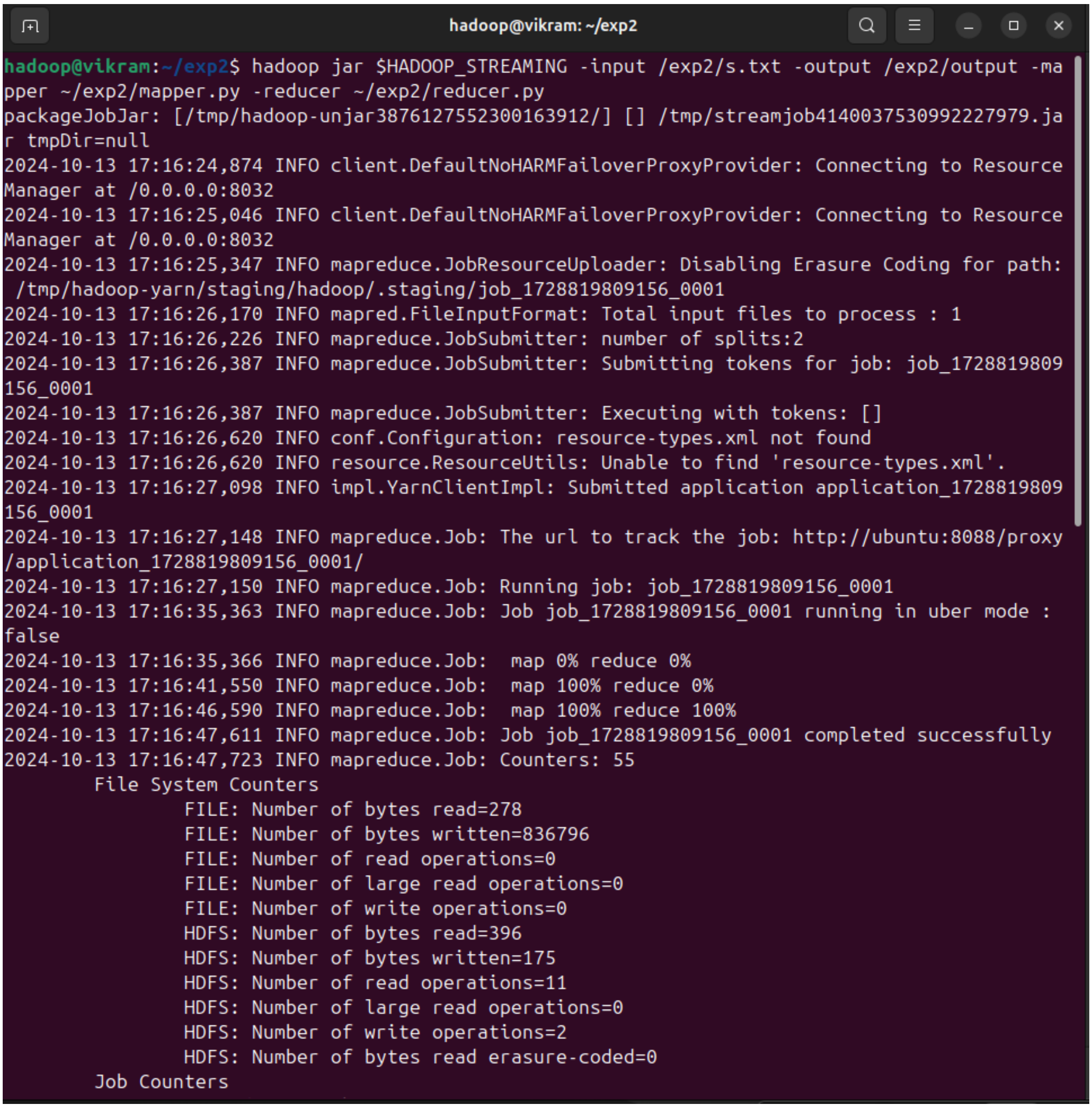
```
  GNU nano 7.2                         reducer.py
#!/usr/bin/python3
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
        line = line.strip()
        word, count = line.split('\t', 1)
        try:
                count = int(count)
        except ValueError:
                continue
        if current_word == word:
                current_count += count
        else:
                if current_word:
                        print( '%s\t%s' % (current_word, current_count))
                current_count = count
                current_word = word

if current_word == word:
        print( '%s\t%s' % (current_word, current_count))

^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line
```

4. Storing the word_count.txt in HDFS Storage.

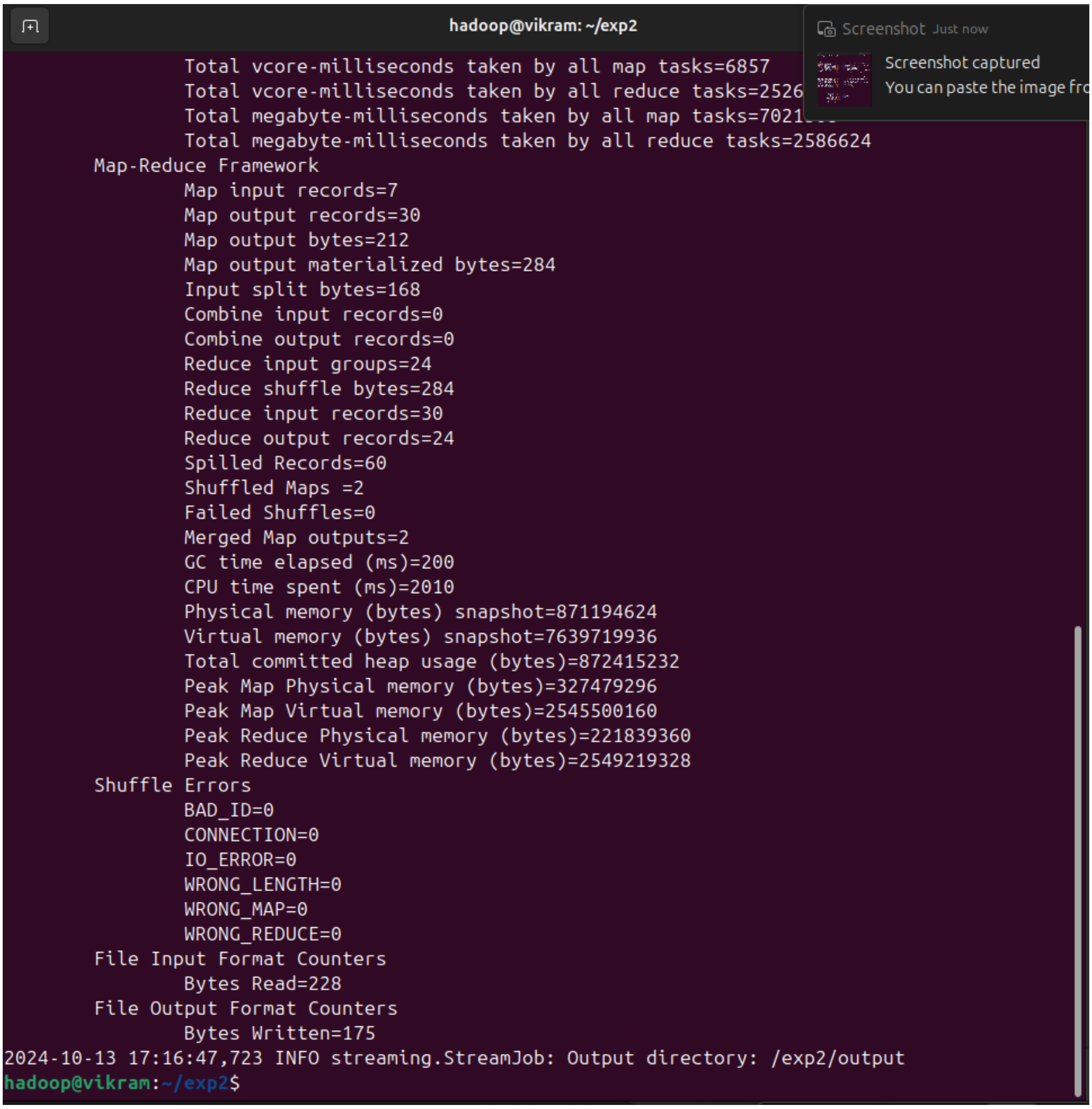


```
hadoop@vikram:~/exp2$ hdfs dfs -ls /exp2
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-13 09:53 /exp2/output
drwxr-xr-x   - hadoop supergroup          0 2024-10-12 16:58 /exp2/output1
-rw-r--r--   1 hadoop supergroup        152 2024-09-13 09:44 /exp2/s.txt
hadoop@vikram:~/exp2$ hdfs dfs -cat /exp2/output/*
Callin  1
Finally 1
LA      2
Lookin  1
Lost    1
Made    1
Maria   2
Might   1
Trynnna 1
dive    1
dough   1
for     2
in      2
it      1
make    1
marina  1
my      1
own     1
the     2
though  1
to      1
weed    1
without 1
yeah    2
hadoop@vikram:~/exp2$
```

5. Running the Word Count program using Hadoop Streaming.

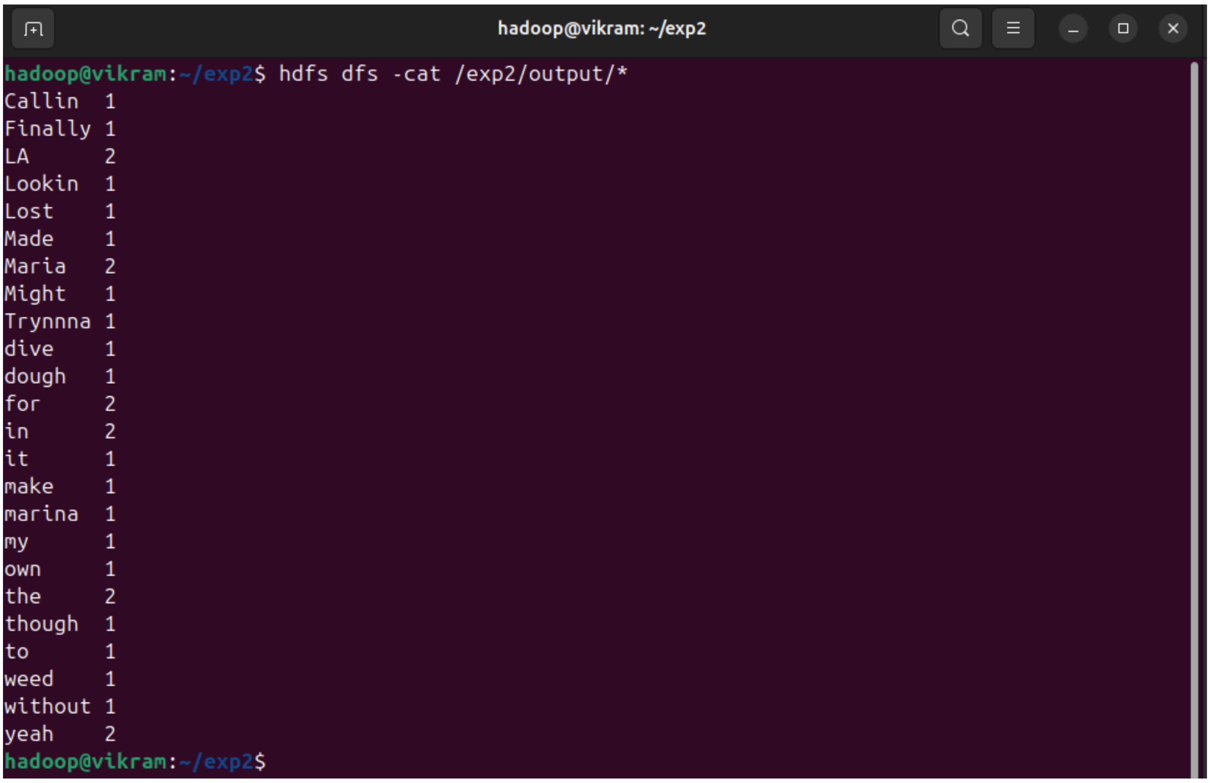


```
hadoop@vikram:~/exp2$ hadoop jar $HADOOP_STREAMING -input /exp2/s.txt -output /exp2/output -ma
pper ~/exp2/mapper.py -reducer ~/exp2/reducer.py
packageJobJar: [/tmp/hadoop-unjar3876127552300163912/] [] /tmp/streamjob4140037530992227979.ja
r tmpDir=null
2024-10-13 17:16:24,874 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to Resource
Manager at /0.0.0.0:8032
2024-10-13 17:16:25,046 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to Resource
Manager at /0.0.0.0:8032
2024-10-13 17:16:25,347 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
 /tmp/hadoop-yarn/staging/hadoop/.staging/job_1728819809156_0001
2024-10-13 17:16:26,170 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-13 17:16:26,226 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-13 17:16:26,387 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728819809
156_0001
2024-10-13 17:16:26,387 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-13 17:16:26,620 INFO conf.Configuration: resource-types.xml not found
2024-10-13 17:16:26,620 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-13 17:16:27,098 INFO impl.YarnClientImpl: Submitted application application_1728819809
156_0001
2024-10-13 17:16:27,148 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy
/application_1728819809156_0001/
2024-10-13 17:16:27,150 INFO mapreduce.Job: Running job: job_1728819809156_0001
2024-10-13 17:16:35,363 INFO mapreduce.Job: Job job_1728819809156_0001 running in uber mode :
false
2024-10-13 17:16:35,366 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-13 17:16:41,550 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-13 17:16:46,590 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-13 17:16:47,611 INFO mapreduce.Job: Job job_1728819809156_0001 completed successfully
2024-10-13 17:16:47,723 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=278
                FILE: Number of bytes written=836796
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=396
                HDFS: Number of bytes written=175
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
```



```
                Total vcore-milliseconds taken by all map tasks=6857
                Total vcore-milliseconds taken by all reduce tasks=2526
                Total megabyte-milliseconds taken by all map tasks=7021
                Total megabyte-milliseconds taken by all reduce tasks=2586624
        Map-Reduce Framework
                Map input records=7
                Map output records=30
                Map output bytes=212
                Map output materialized bytes=284
                Input split bytes=168
                Combine input records=0
                Combine output records=0
                Reduce input groups=24
                Reduce shuffle bytes=284
                Reduce input records=30
                Reduce output records=24
                Spilled Records=60
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=200
                CPU time spent (ms)=2010
                Physical memory (bytes) snapshot=871194624
                Virtual memory (bytes) snapshot=7639719936
                Total committed heap usage (bytes)=872415232
                Peak Map Physical memory (bytes)=327479296
                Peak Map Virtual memory (bytes)=2545500160
                Peak Reduce Physical memory (bytes)=221839360
                Peak Reduce Virtual memory (bytes)=2549219328
        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=228
        File Output Format Counters
                Bytes Written=175
2024-10-13 17:16:47,723 INFO streaming.StreamJob: Output directory: /exp2/output
hadoop@vikram:~/exp2$
```

Output :



```
hadoop@vikram:~/exp2$ hdfs dfs -cat /exp2/output/*
Callin  1
Finally 1
LA      2
Lookin  1
Lost    1
Made    1
Maria   2
Might   1
Trynnna 1
dive    1
dough   1
for     2
in      2
it      1
make    1
marina  1
my      1
own     1
the     2
though  1
to      1
weed    1
without 1
yeah    2
hadoop@vikram:~/exp2$
```