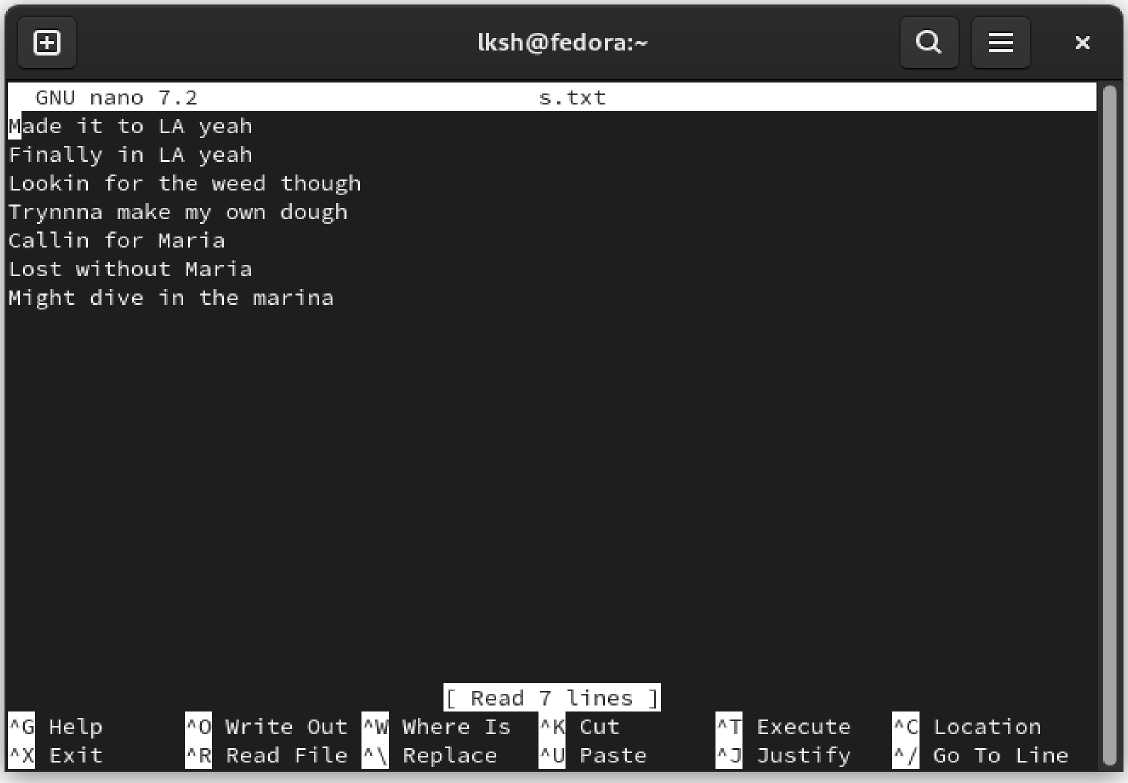
Exp. No : 2

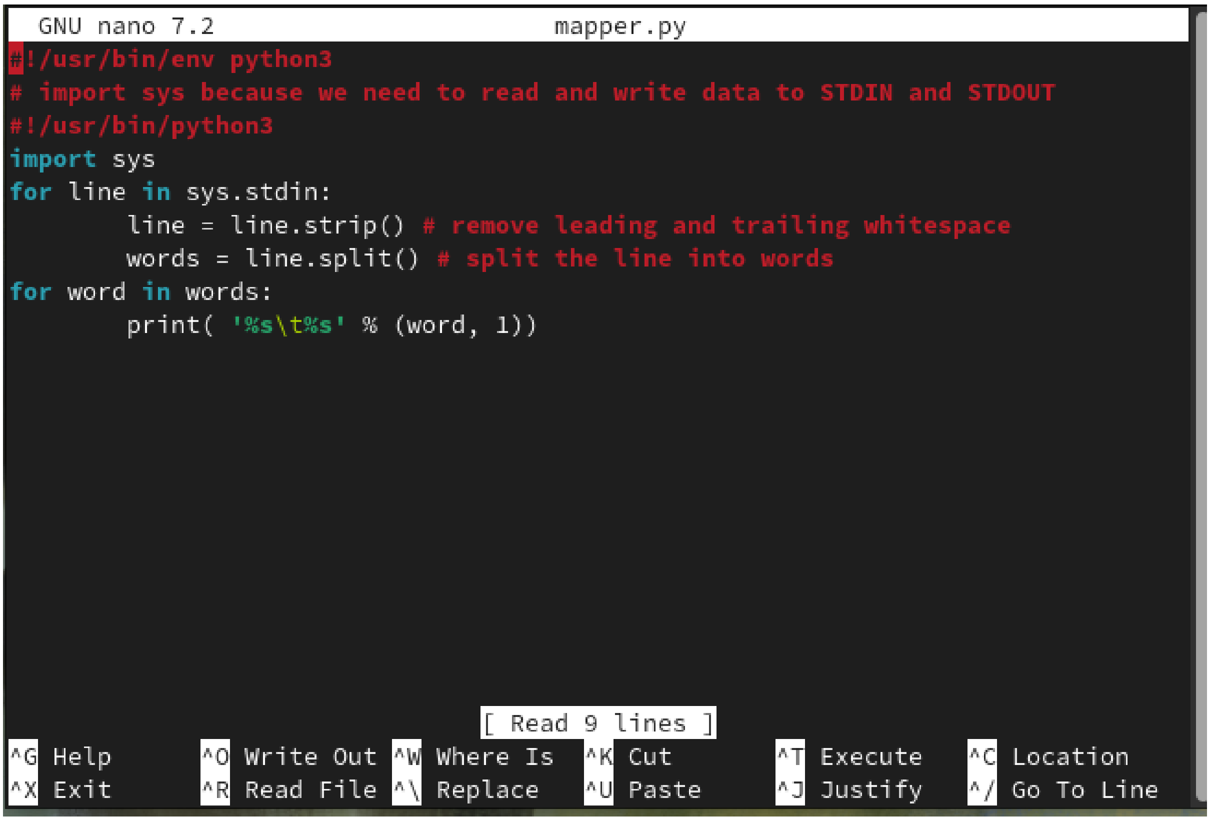
Word Count Map Reduce program

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
```

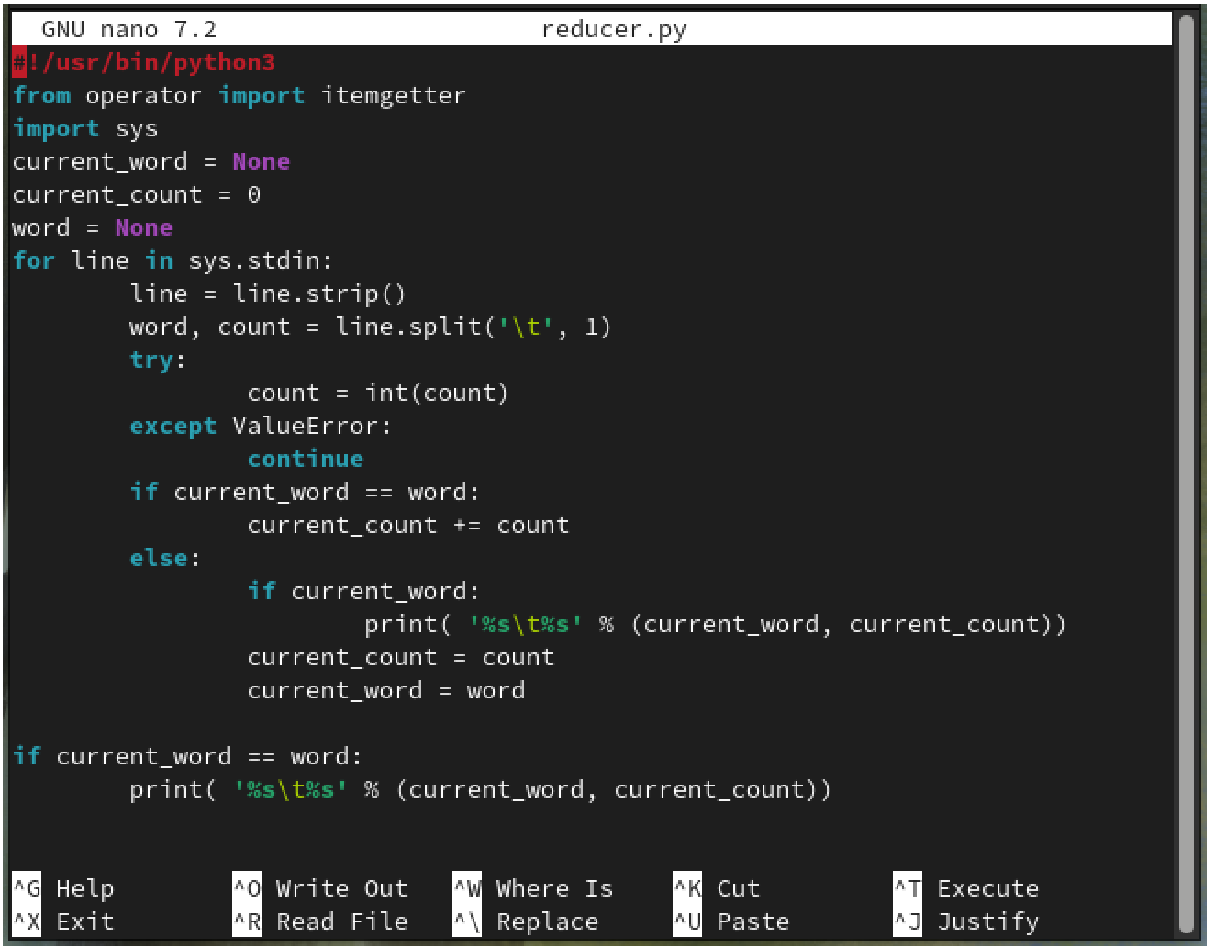
2. Create mapper.py program



```
#!/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))
```

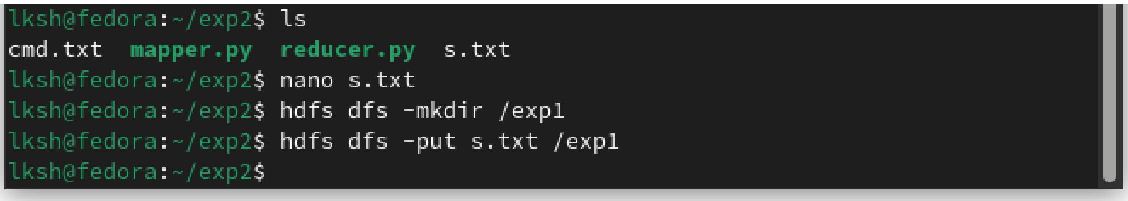
[Read 9 lines]

3. Create reducer.py program.



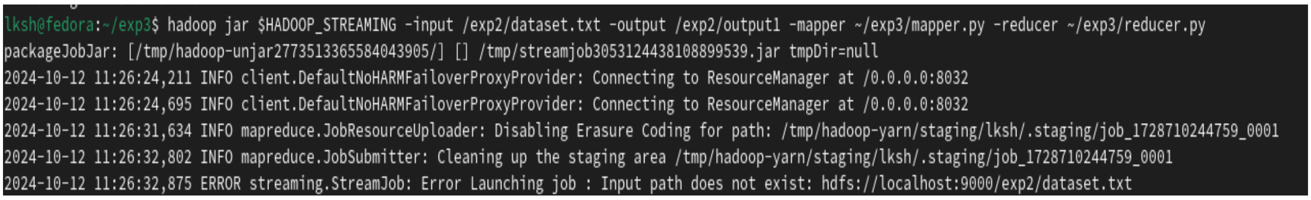
```
#!/usr/bin/python3
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
        line = line.strip()
        word, count = line.split('\t', 1)
        try:
                count = int(count)
        except ValueError:
                continue
        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))
```

4. Storing the word_count.txt in HDFS Storage.

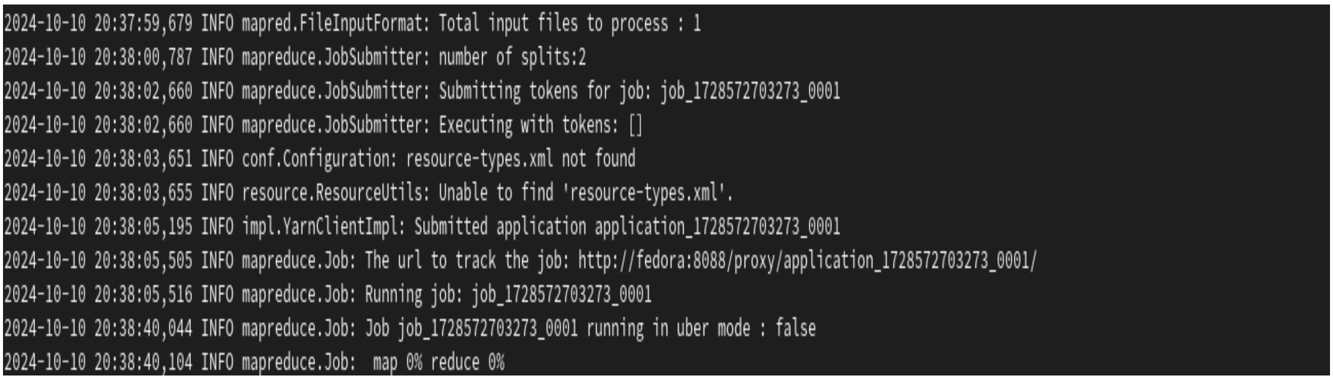


```
lksh@fedora:~/exp2$ ls
cmd.txt   mapper.py   reducer.py   s.txt
lksh@fedora:~/exp2$ nano s.txt
lksh@fedora:~/exp2$ hdfs dfs -mkdir /exp1
lksh@fedora:~/exp2$ hdfs dfs -put s.txt /exp1
lksh@fedora:~/exp2$
```

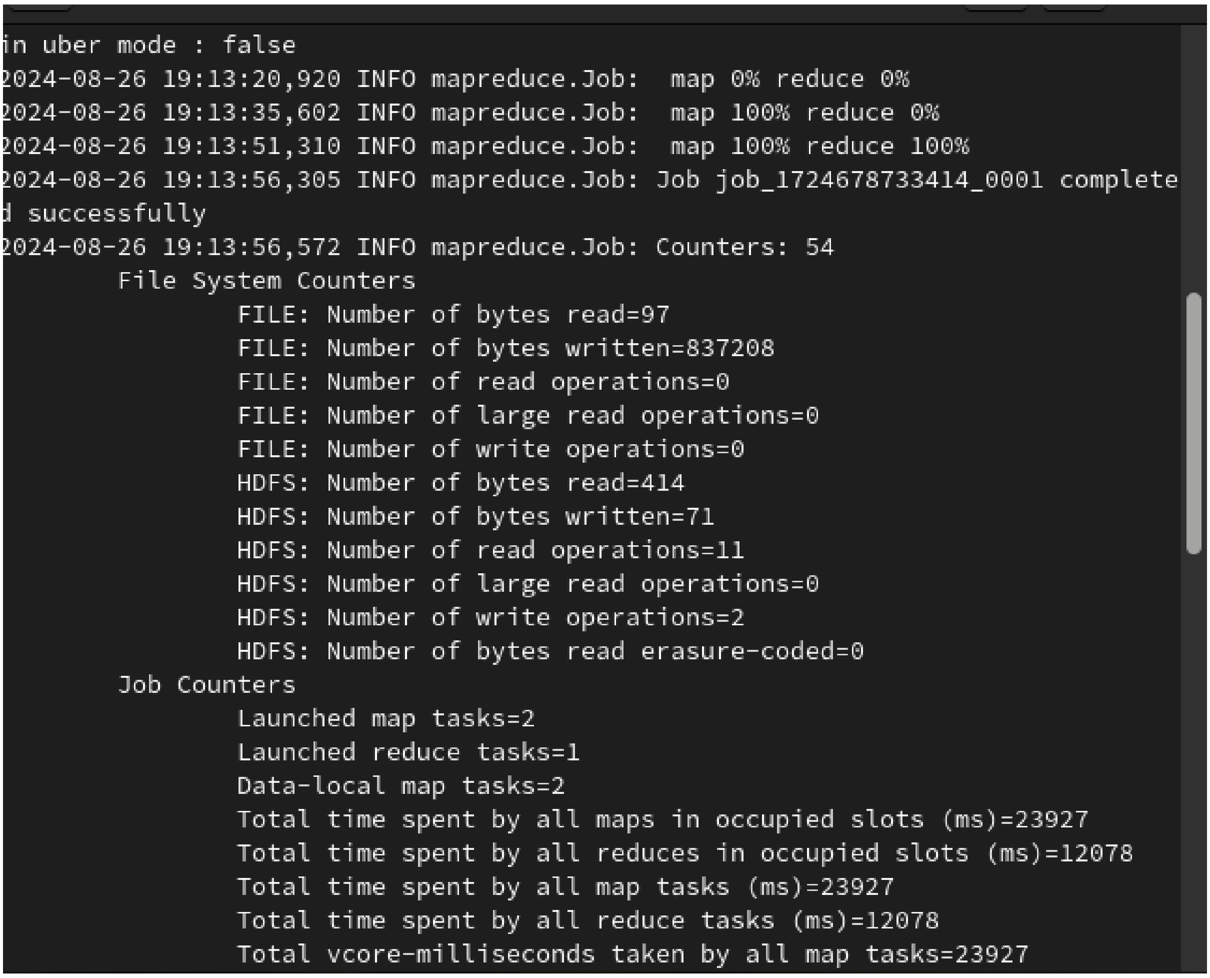
5. Running the Word Count program using Hadoop Streaming



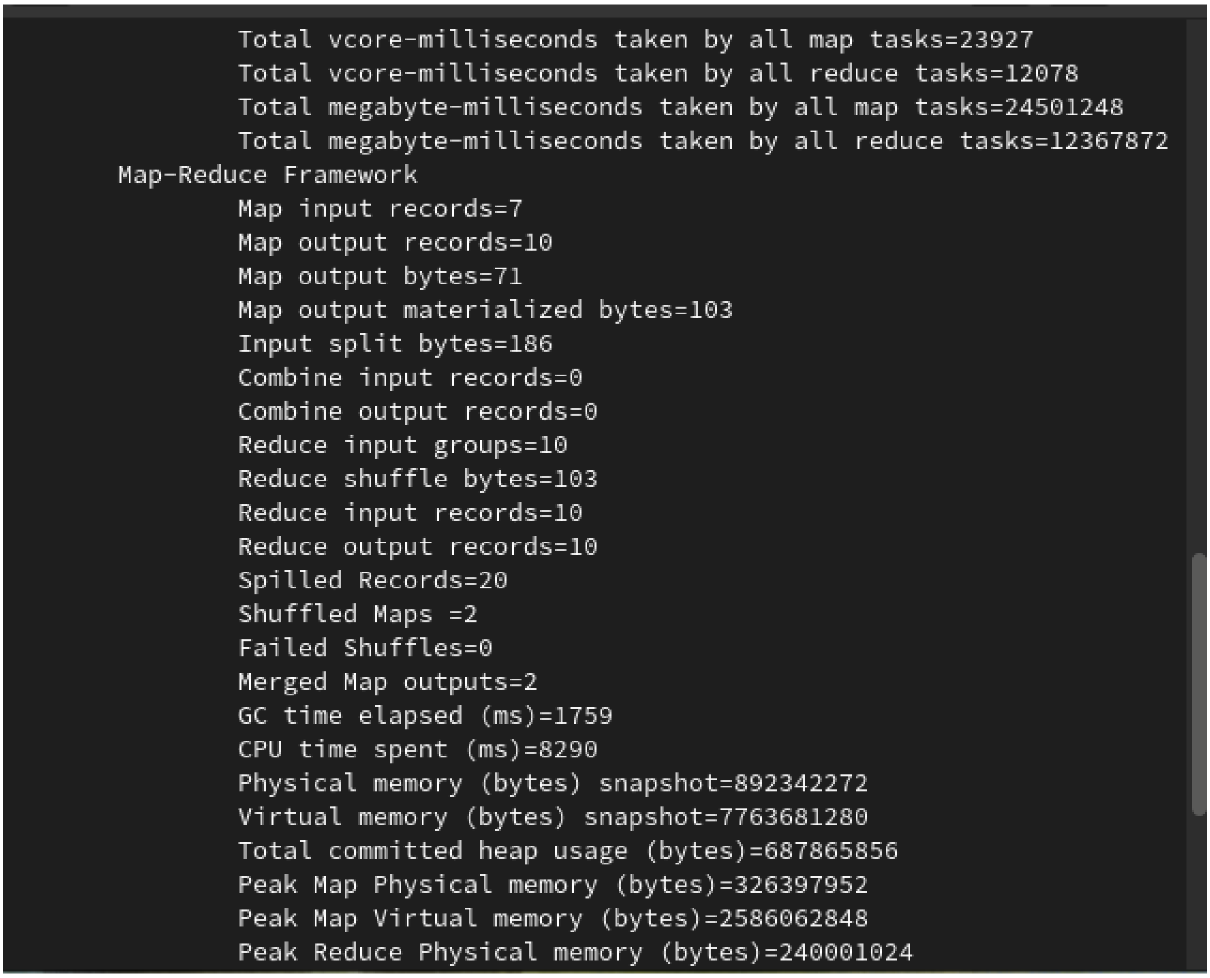
```
lksh@fedora:~/exp3$ hadoop jar $HADOOP_STREAMING -input /exp2/dataset.txt -output /exp2/output1 -mapper ~/exp3/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar2773513365584043905/] [] /tmp/streamjob3053124438108899539.jar tmpDir=null
2024-10-12 11:26:24,211 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 11:26:24,695 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 11:26:31,634 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/lksh/.staging/job_1728710244759_0001
2024-10-12 11:26:32,802 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/lksh/.staging/job_1728710244759_0001
2024-10-12 11:26:32,875 ERROR streaming.StreamJob: Error Launching job : Input path does not exist: hdfs://localhost:9000/exp2/dataset.txt
```



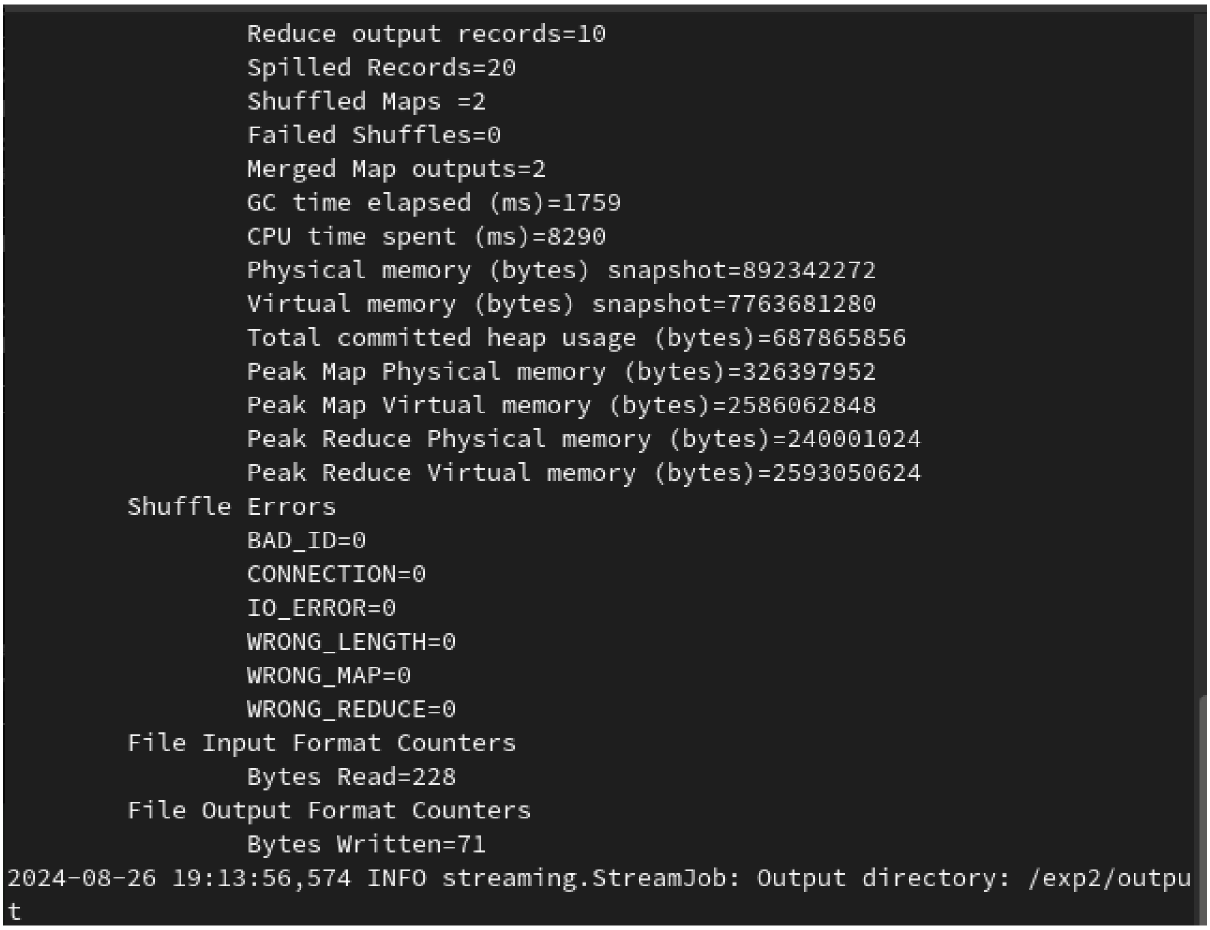
```
2024-10-10 20:37:59,679 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-10 20:38:00,787 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-10 20:38:02,660 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728572703273_0001
2024-10-10 20:38:02,660 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-10 20:38:03,651 INFO conf.Configuration: resource-types.xml not found
2024-10-10 20:38:03,655 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-10 20:38:05,195 INFO impl.YarnClientImpl: Submitted application application_1728572703273_0001
2024-10-10 20:38:05,505 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1728572703273_0001/
2024-10-10 20:38:05,516 INFO mapreduce.Job: Running job: job_1728572703273_0001
2024-10-10 20:38:40,044 INFO mapreduce.Job: Job job_1728572703273_0001 running in uber mode : false
2024-10-10 20:38:40,104 INFO mapreduce.Job:  map 0% reduce 0%
```



```
in uber mode : false
2024-08-26 19:13:20,920 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-26 19:13:35,602 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-26 19:13:51,310 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-26 19:13:56,305 INFO mapreduce.Job: Job job_1724678733414_0001 complete
d successfully
2024-08-26 19:13:56,572 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=97
                FILE: Number of bytes written=837208
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=414
                HDFS: Number of bytes written=71
                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
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=23927
                Total time spent by all reduces in occupied slots (ms)=12078
                Total time spent by all map tasks (ms)=23927
                Total time spent by all reduce tasks (ms)=12078
                Total vcore-milliseconds taken by all map tasks=23927
```

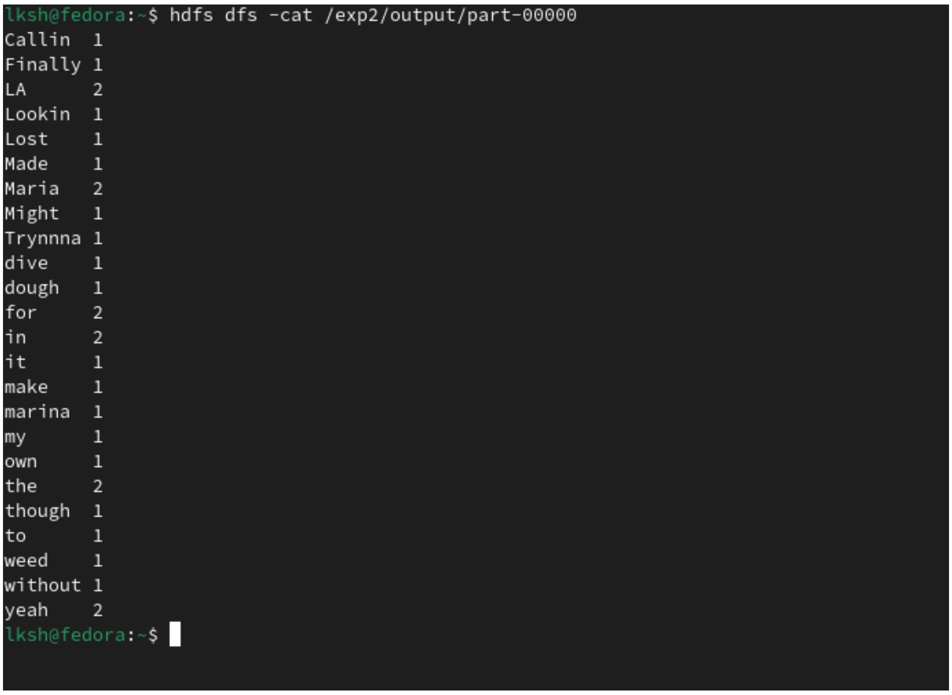


```
                Total vcore-milliseconds taken by all map tasks=23927
                Total vcore-milliseconds taken by all reduce tasks=12078
                Total megabyte-milliseconds taken by all map tasks=24501248
                Total megabyte-milliseconds taken by all reduce tasks=12367872
        Map-Reduce Framework
                Map input records=7
                Map output records=10
                Map output bytes=71
                Map output materialized bytes=103
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=103
                Reduce input records=10
                Reduce output records=10
                Spilled Records=20
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1759
                CPU time spent (ms)=8290
                Physical memory (bytes) snapshot=892342272
                Virtual memory (bytes) snapshot=7763681280
                Total committed heap usage (bytes)=687865856
                Peak Map Physical memory (bytes)=326397952
                Peak Map Virtual memory (bytes)=2586062848
                Peak Reduce Physical memory (bytes)=240001024
```



```
                Reduce output records=10
                Spilled Records=20
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1759
                CPU time spent (ms)=8290
                Physical memory (bytes) snapshot=892342272
                Virtual memory (bytes) snapshot=7763681280
                Total committed heap usage (bytes)=687865856
                Peak Map Physical memory (bytes)=326397952
                Peak Map Virtual memory (bytes)=2586062848
                Peak Reduce Physical memory (bytes)=240001024
                Peak Reduce Virtual memory (bytes)=2593050624
        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=71
2024-08-26 19:13:56,574 INFO streaming.StreamJob: Output directory: /exp2/outpu
t
```

Output :



```
lksh@fedora:~$ hdfs dfs -cat /exp2/output/part-00000
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
lksh@fedora:~$
```