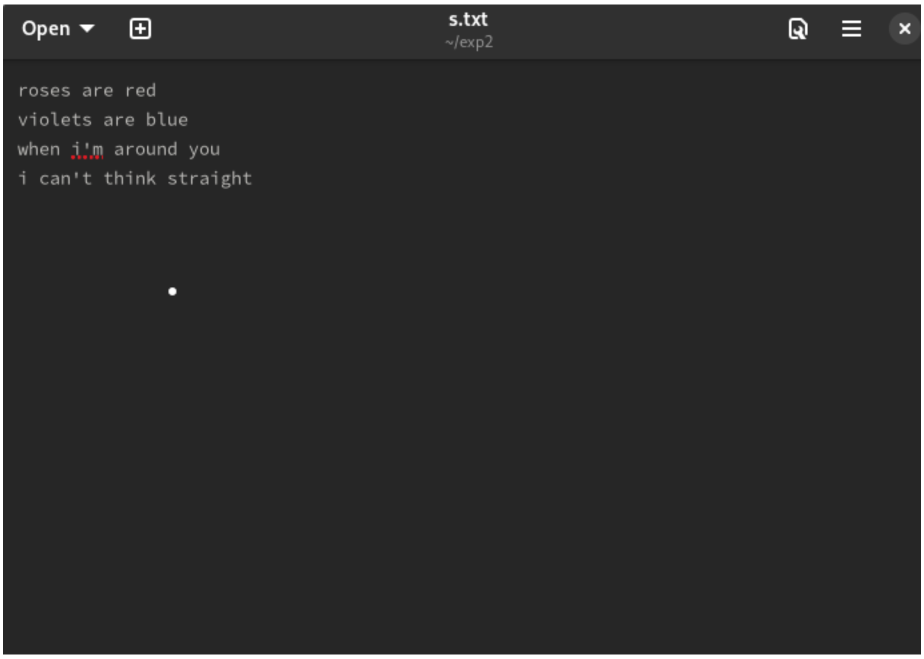
Exp. No : 2

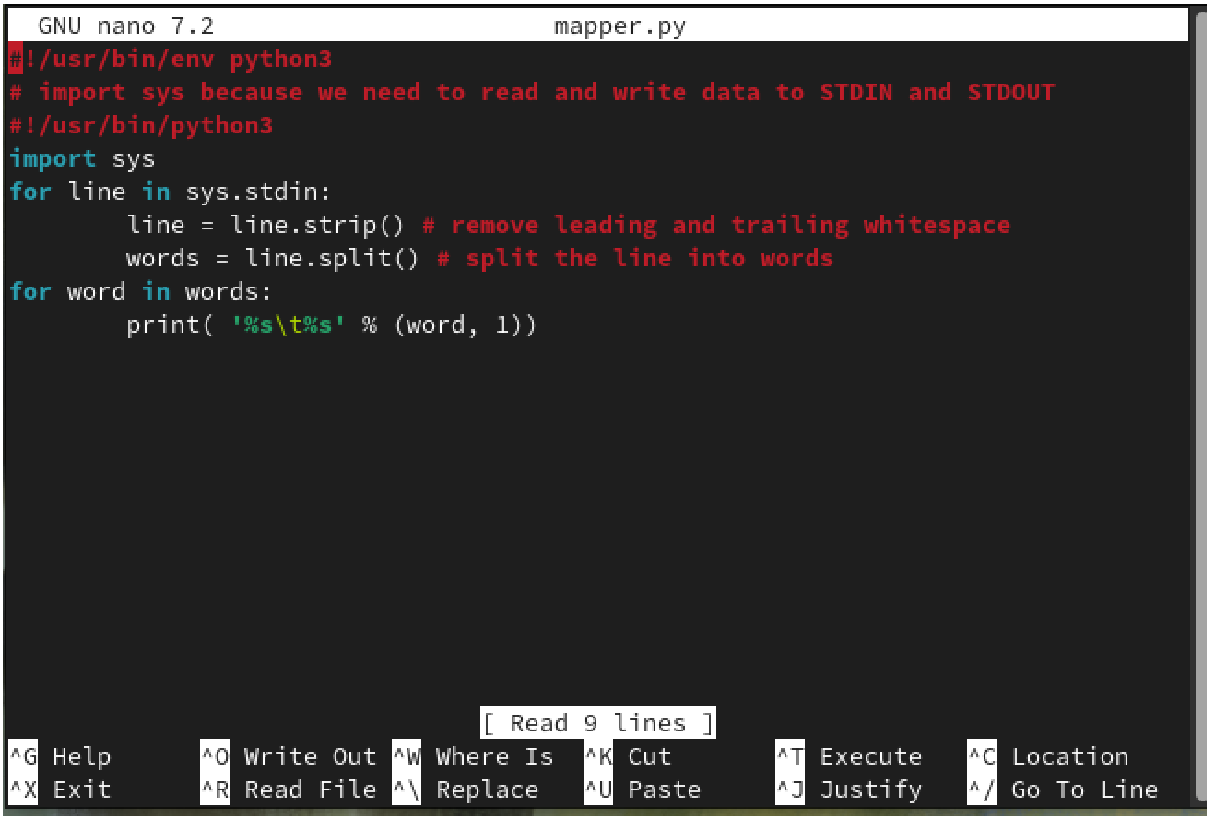
Word Count Map Reduce program

1. Create s.txt file



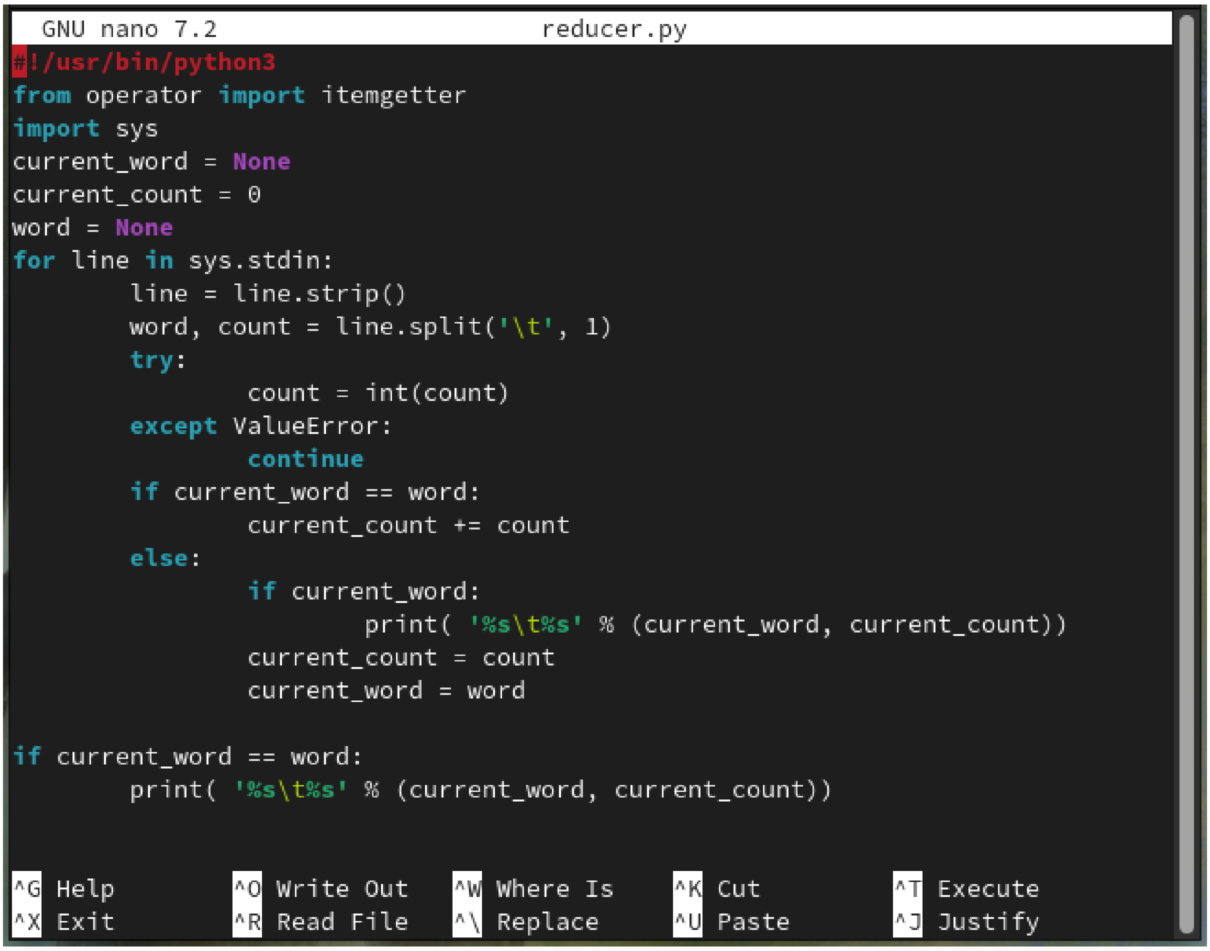
```
roses are red
violets are blue
when i'm around you
i can't think straight
```

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))
```

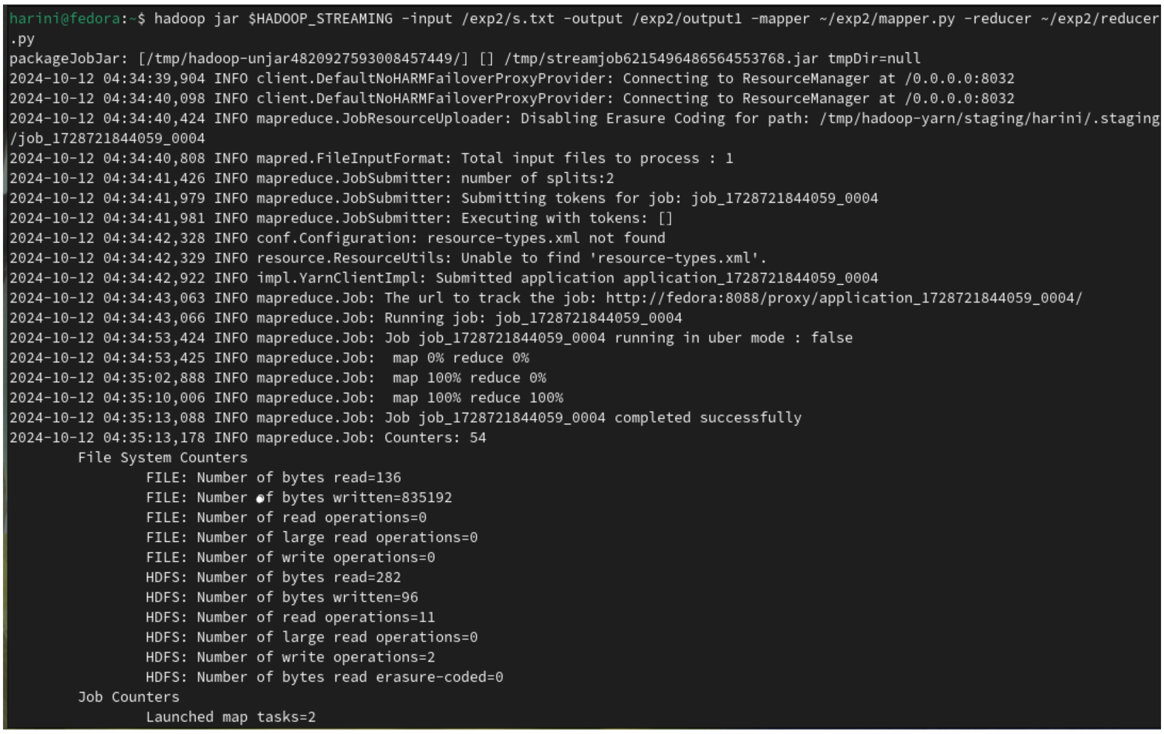
3. Create reducer.py program.



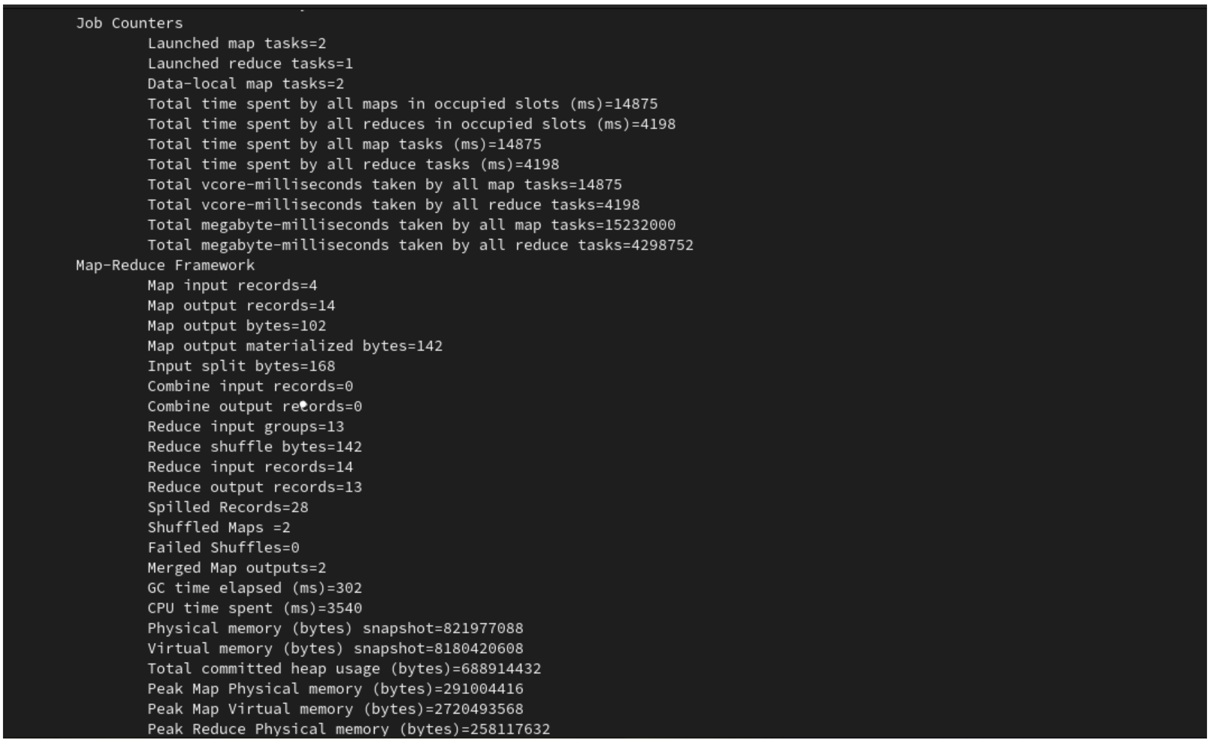
```
#!/usr/bin/python3
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
        line = line.strip()
        word, count = line.split('\t', 1)
        try:
                count = int(count)
        except ValueError:
                continue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

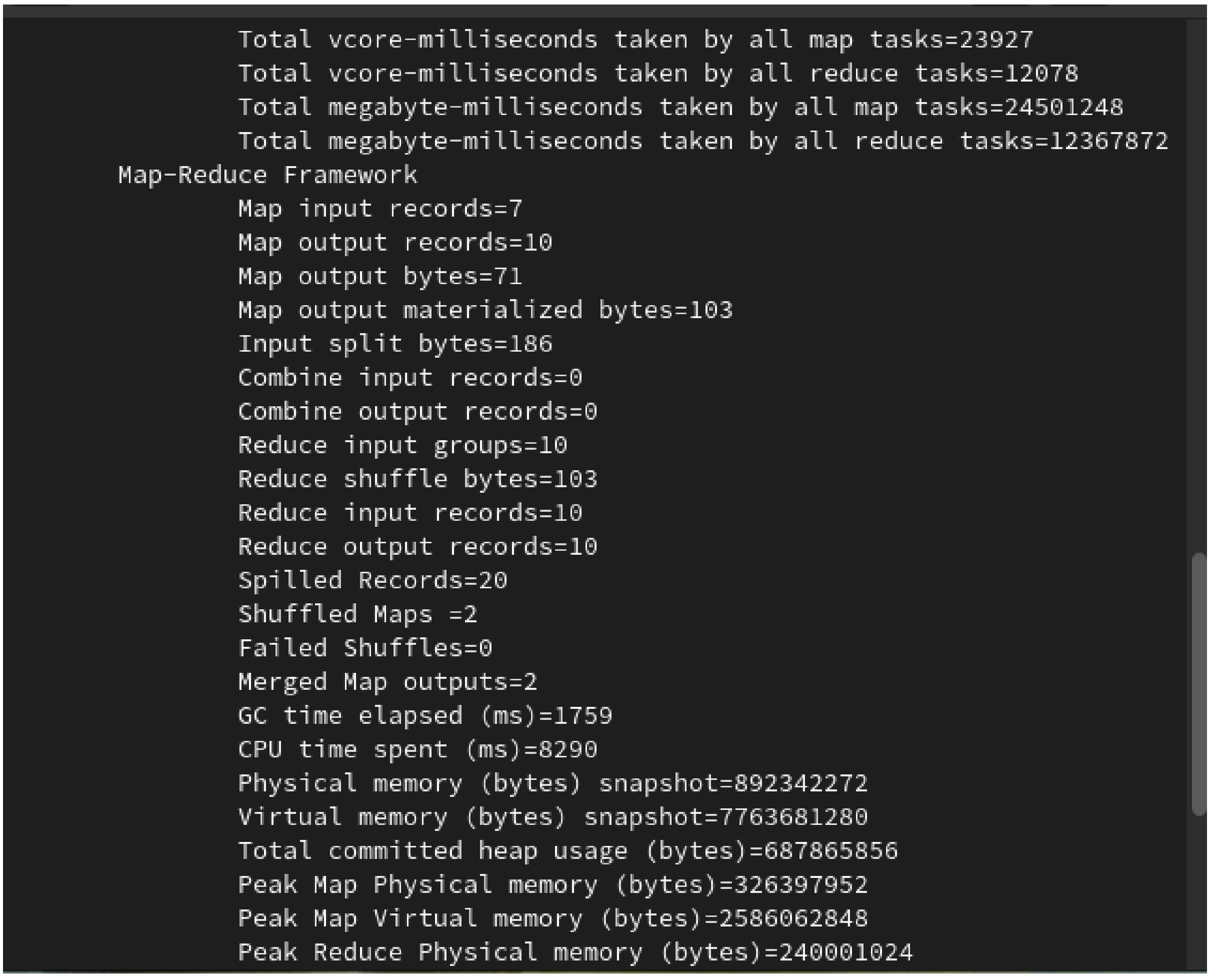
4.Running the Word Count program using Hadoop Streaming



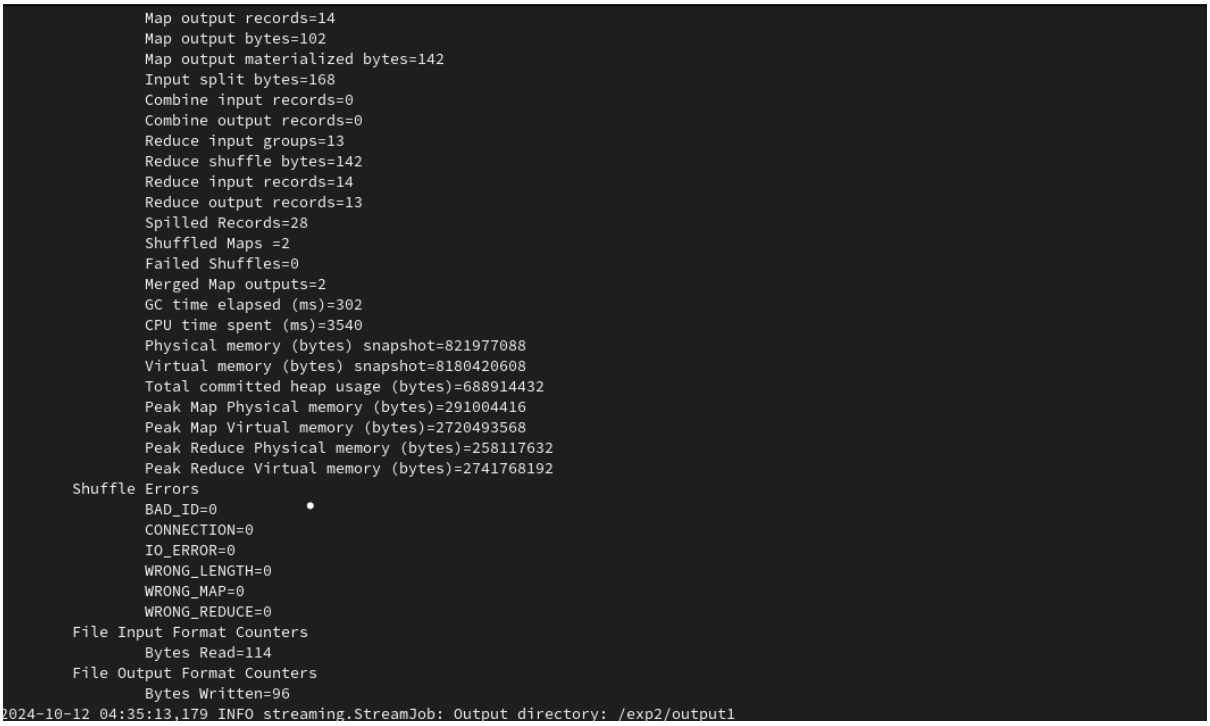
```
harini@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp2/s.txt -output /exp2/output1 -mapper ~/exp2/mapper.py -reducer ~/exp2/reducer
.py
packageJobJar: [/tmp/hadoop-unjar4820927593008457449/] [] /tmp/streamjob6215496486564553768.jar tmpDir=null
2024-10-12 04:34:39,904 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 04:34:40,098 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 04:34:40,424 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/harini/.staging
/job_1728721844059_0004
2024-10-12 04:34:40,808 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-12 04:34:41,426 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-12 04:34:41,979 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728721844059_0004
2024-10-12 04:34:41,981 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-12 04:34:42,328 INFO conf.Configuration: resource-types.xml not found
2024-10-12 04:34:42,329 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-12 04:34:42,922 INFO impl.YarnClientImpl: Submitted application application_1728721844059_0004
2024-10-12 04:34:43,063 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1728721844059_0004/
2024-10-12 04:34:43,066 INFO mapreduce.Job: Running job: job_1728721844059_0004
2024-10-12 04:34:53,424 INFO mapreduce.Job: Job job_1728721844059_0004 running in uber mode : false
2024-10-12 04:34:53,425 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-12 04:35:02,888 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-12 04:35:10,006 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-12 04:35:13,088 INFO mapreduce.Job: Job job_1728721844059_0004 completed successfully
2024-10-12 04:35:13,178 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=136
                FILE: Number of bytes written=835192
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=282
                HDFS: Number of bytes written=96
                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
```



```
        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)=14875
                Total time spent by all reduces in occupied slots (ms)=4198
                Total time spent by all map tasks (ms)=14875
                Total time spent by all reduce tasks (ms)=4198
                Total vcore-milliseconds taken by all map tasks=14875
                Total vcore-milliseconds taken by all reduce tasks=4198
                Total megabyte-milliseconds taken by all map tasks=15232000
                Total megabyte-milliseconds taken by all reduce tasks=4298752
        Map-Reduce Framework
                Map input records=4
                Map output records=14
                Map output bytes=102
                Map output materialized bytes=142
                Input split bytes=168
                Combine input records=0
                Combine output records=0
                Reduce input groups=13
                Reduce shuffle bytes=142
                Reduce input records=14
                Reduce output records=13
                Spilled Records=28
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=302
                CPU time spent (ms)=3540
                Physical memory (bytes) snapshot=821977088
                Virtual memory (bytes) snapshot=8180420608
                Total committed heap usage (bytes)=688914432
                Peak Map Physical memory (bytes)=291004416
                Peak Map Virtual memory (bytes)=2720493568
                Peak Reduce Physical memory (bytes)=258117632
```

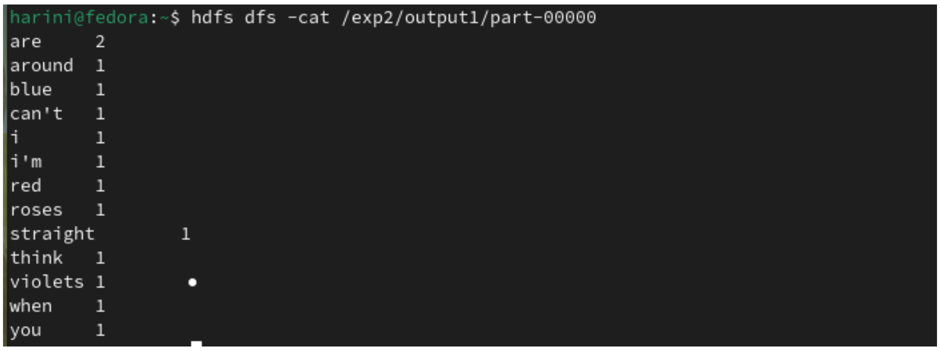


```
                Total vcore-milliseconds taken by all map tasks=23927
                Total vcore-milliseconds taken by all reduce tasks=12078
                Total megabyte-milliseconds taken by all map tasks=24501248
                Total megabyte-milliseconds taken by all reduce tasks=12367872
        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
```



```
                Map output records=14
                Map output bytes=102
                Map output materialized bytes=142
                Input split bytes=168
                Combine input records=0
                Combine output records=0
                Reduce input groups=13
                Reduce shuffle bytes=142
                Reduce input records=14
                Reduce output records=13
                Spilled Records=28
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=302
                CPU time spent (ms)=3540
                Physical memory (bytes) snapshot=821977088
                Virtual memory (bytes) snapshot=8180420608
                Total committed heap usage (bytes)=688914432
                Peak Map Physical memory (bytes)=291004416
                Peak Map Virtual memory (bytes)=2720493568
                Peak Reduce Physical memory (bytes)=258117632
                Peak Reduce Virtual memory (bytes)=2741768192
        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=114
        File Output Format Counters
                Bytes Written=96
2024-10-12 04:35:13,179 INFO streaming.StreamJob: Output directory: /exp2/output1
```

Output :



```
harini@fedora:~$ hdfs dfs -cat /exp2/output1/part-00000
are     2
around  1
blue    1
can't   1
i       1
i'm     1
red     1
roses   1
straight        1
think   1
violets 1
when    1
you     1
```