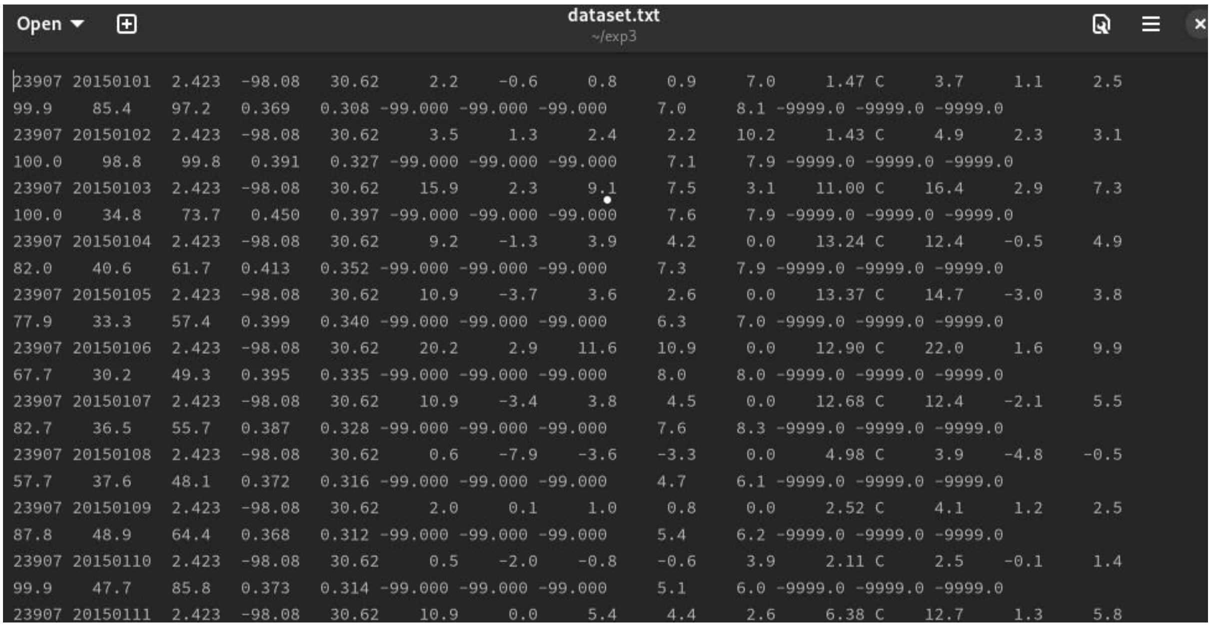
Exp. No : 3

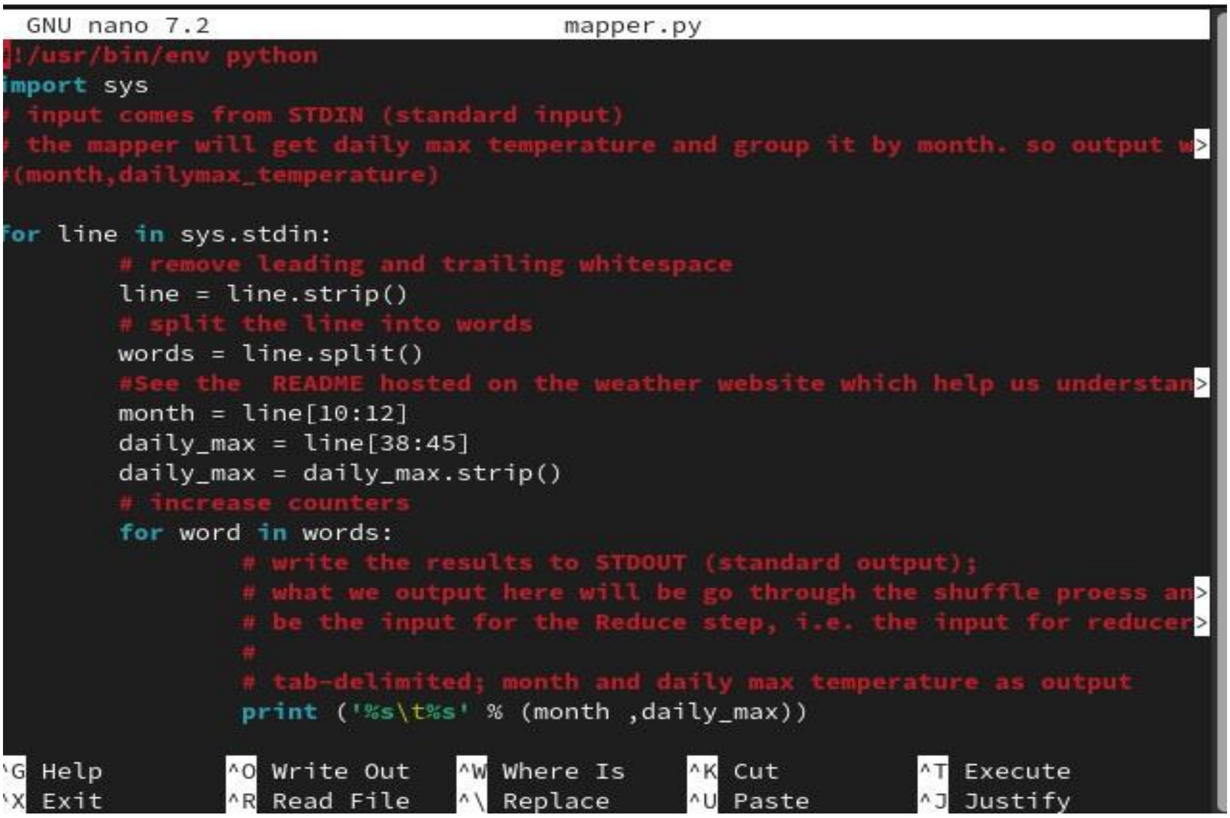
Map Reduce program to process Weather dataset

1. Download Weather dataset.



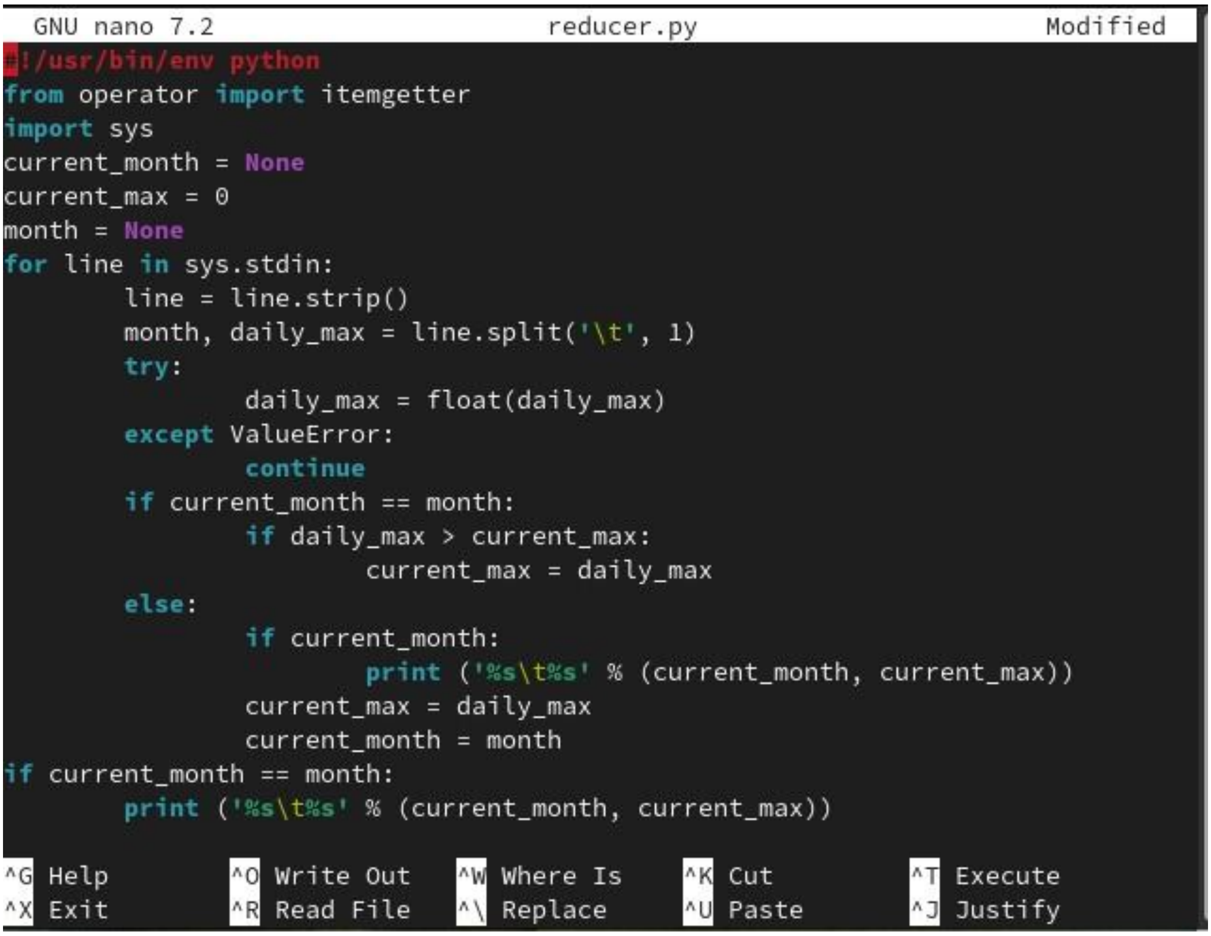
```
23907 20150101  2.423  -98.08   30.62     2.2    -0.6     0.8     0.9     7.0     1.47 C     3.7     1.1     2.5
99.9    85.4    97.2   0.369   0.308 -99.000 -99.000 -99.000     7.0     8.1 -9999.0 -9999.0 -9999.0
23907 20150102  2.423  -98.08   30.62     3.5     1.3     2.4     2.2    10.2     1.43 C     4.9     2.3     3.1
100.0    98.8    99.8   0.391   0.327 -99.000 -99.000 -99.000     7.1     7.9 -9999.0 -9999.0 -9999.0
23907 20150103  2.423  -98.08   30.62    15.9     2.3     9.1     7.5     3.1    11.00 C    16.4     2.9     7.3
100.0    34.8    73.7   0.450   0.397 -99.000 -99.000 -99.000     7.6     7.9 -9999.0 -9999.0 -9999.0
23907 20150104  2.423  -98.08   30.62     9.2    -1.3     3.9     4.2     0.0    13.24 C    12.4    -0.5     4.9
82.0    40.6    61.7   0.413   0.352 -99.000 -99.000 -99.000     7.3     7.9 -9999.0 -9999.0 -9999.0
23907 20150105  2.423  -98.08   30.62    10.9    -3.7     3.6     2.6     0.0    13.37 C    14.7    -3.0     3.8
77.9    33.3    57.4   0.399   0.340 -99.000 -99.000 -99.000     6.3     7.0 -9999.0 -9999.0 -9999.0
23907 20150106  2.423  -98.08   30.62    20.2     2.9    11.6    10.9     0.0    12.90 C    22.0     1.6     9.9
67.7    30.2    49.3   0.395   0.335 -99.000 -99.000 -99.000     8.0     8.0 -9999.0 -9999.0 -9999.0
23907 20150107  2.423  -98.08   30.62    10.9    -3.4     3.8     4.5     0.0    12.68 C    12.4    -2.1     5.5
82.7    36.5    55.7   0.387   0.328 -99.000 -99.000 -99.000     7.6     8.3 -9999.0 -9999.0 -9999.0
23907 20150108  2.423  -98.08   30.62     0.6    -7.9    -3.6    -3.3     0.0     4.98 C     3.9    -4.8    -0.5
57.7    37.6    48.1   0.372   0.316 -99.000 -99.000 -99.000     4.7     6.1 -9999.0 -9999.0 -9999.0
23907 20150109  2.423  -98.08   30.62     2.0     0.1     1.0     0.8     0.0     2.52 C     4.1     1.2     2.5
87.8    48.9    64.4   0.368   0.312 -99.000 -99.000 -99.000     5.4     6.2 -9999.0 -9999.0 -9999.0
23907 20150110  2.423  -98.08   30.62     0.5    -2.0    -0.8    -0.6     3.9     2.11 C     2.5    -0.1     1.4
99.9    47.7    85.8   0.373   0.314 -99.000 -99.000 -99.000     5.1     6.0 -9999.0 -9999.0 -9999.0
23907 20150111  2.423  -98.08   30.62    10.9     0.0     5.4     4.4     2.6     6.38 C    12.7     1.3     5.8
```

2. Create mapper.py program



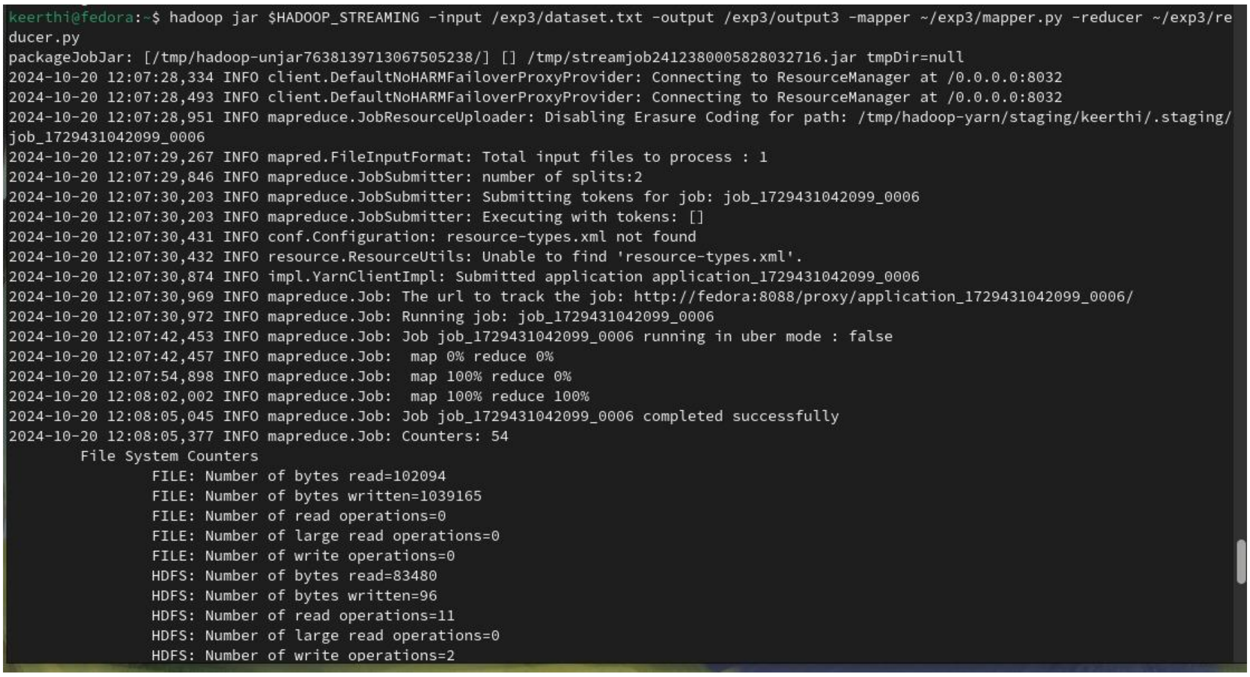
```
#!/usr/bin/env python
import sys
# input comes from STDIN (standard input)
# the mapper will get daily max temperature and group it by month. so output w>
#(month,dailymax_temperature)

for line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()
        # split the line into words
        words = line.split()
        #See the  README hosted on the weather website which help us understan>
        month = line[10:12]
        daily_max = line[38:45]
        daily_max = daily_max.strip()
        # increase counters
        for word in words:
                # write the results to STDOUT (standard output);
                # what we output here will be go through the shuffle proess an>
                # be the input for the Reduce step, i.e. the input for reducer>
                #
                # tab-delimited; month and daily max temperature as output
                print ('%s\t%s' % (month ,daily_max))
```

3. Create reducer.py

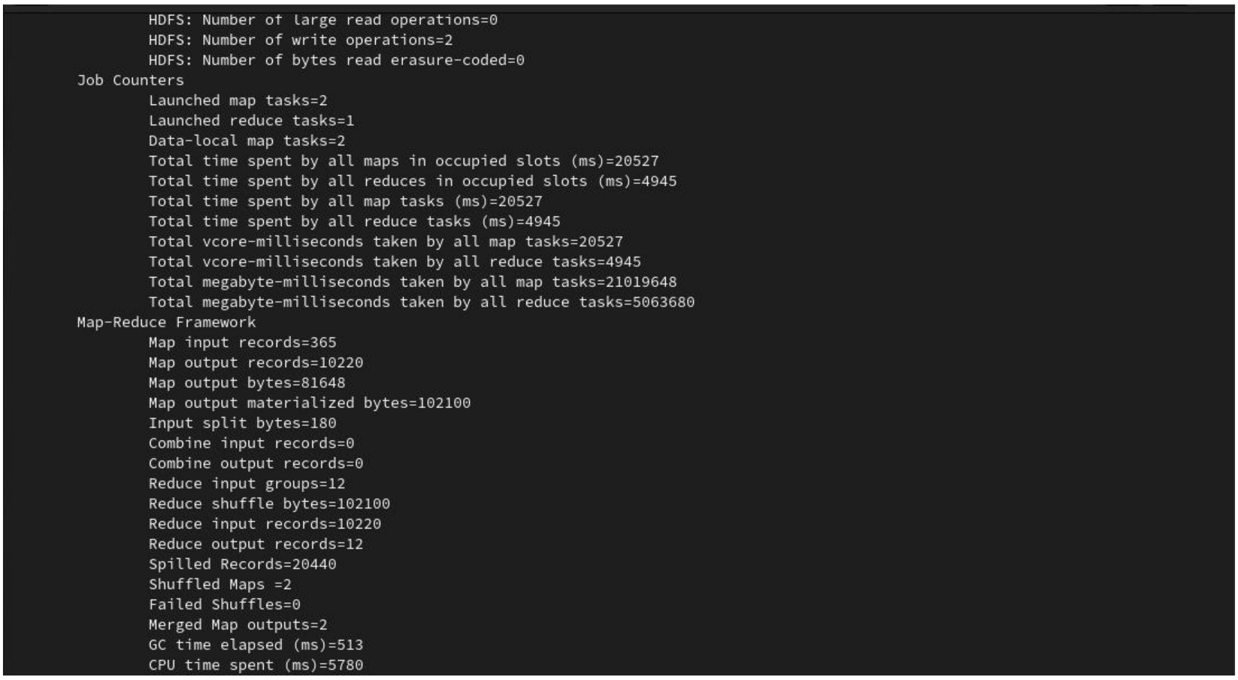


```
#!/usr/bin/env python
from operator import itemgetter
import sys
current_month = None
current_max = 0
month = None
for line in sys.stdin:
        line = line.strip()
        month, daily_max = line.split('\t', 1)
        try:
                daily_max = float(daily_max)
        except ValueError:
                continue
        if current_month == month:
                if daily_max > current_max:
                        current_max = daily_max
        else:
                if current_month:
                        print ('%s\t%s' % (current_month, current_max))
                current_max = daily_max
                current_month = month
if current_month == month:
        print ('%s\t%s' % (current_month, current_max))
```

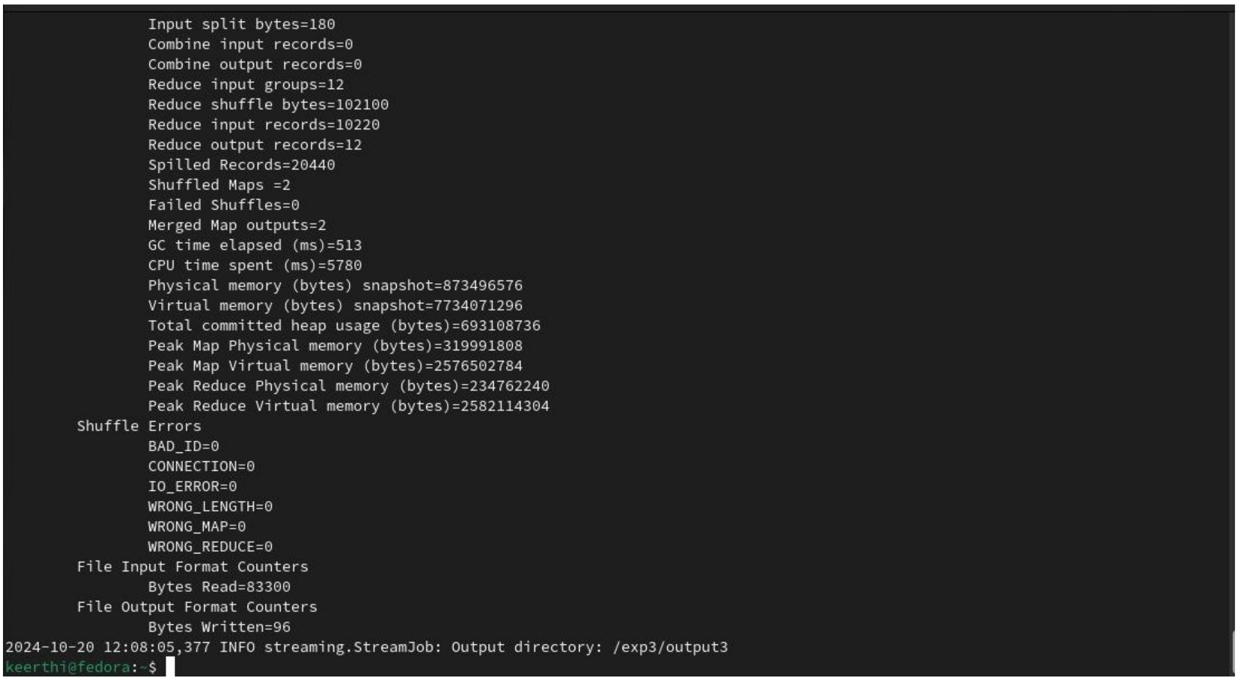
4. Run the Map reduce program using Hadoop Streaming.



```
keerthi@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp3/dataset.txt -output /exp3/output3 -mapper ~/exp3/mapper.py -reducer ~/exp3/re
ducer.py
packageJobJar: [/tmp/hadoop-unjar7638139713067505238/] [] /tmp/streamjob2412380005828032716.jar tmpDir=null
2024-10-20 12:07:28,334 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 12:07:28,493 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 12:07:28,951 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/keerthi/.staging/
job_1729431042099_0006
2024-10-20 12:07:29,267 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-20 12:07:29,846 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-20 12:07:30,203 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729431042099_0006
2024-10-20 12:07:30,203 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-20 12:07:30,431 INFO conf.Configuration: resource-types.xml not found
2024-10-20 12:07:30,432 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-20 12:07:30,874 INFO impl.YarnClientImpl: Submitted application application_1729431042099_0006
2024-10-20 12:07:30,969 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1729431042099_0006/
2024-10-20 12:07:30,972 INFO mapreduce.Job: Running job: job_1729431042099_0006
2024-10-20 12:07:42,453 INFO mapreduce.Job: Job job_1729431042099_0006 running in uber mode : false
2024-10-20 12:07:42,457 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-20 12:07:54,898 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-20 12:08:02,002 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-20 12:08:05,045 INFO mapreduce.Job: Job job_1729431042099_0006 completed successfully
2024-10-20 12:08:05,377 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=102094
                FILE: Number of bytes written=1039165
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83480
                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
```

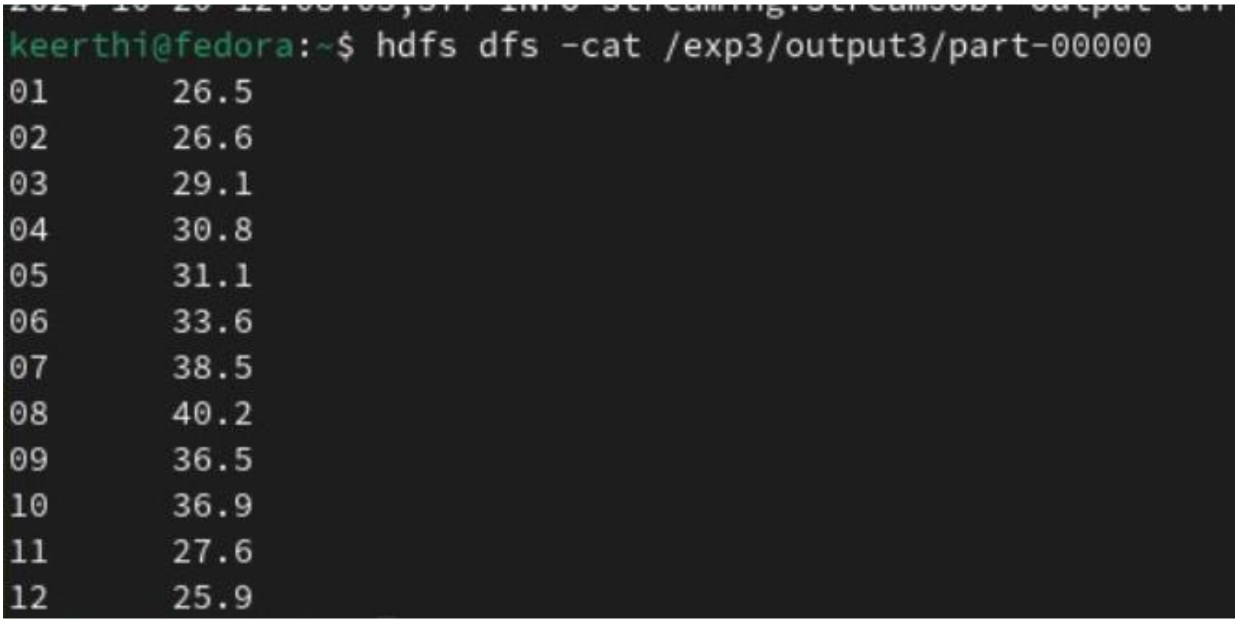


```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=20527
                Total time spent by all reduces in occupied slots (ms)=4945
                Total time spent by all map tasks (ms)=20527
                Total time spent by all reduce tasks (ms)=4945
                Total vcore-milliseconds taken by all map tasks=20527
                Total vcore-milliseconds taken by all reduce tasks=4945
                Total megabyte-milliseconds taken by all map tasks=21019648
                Total megabyte-milliseconds taken by all reduce tasks=5063680
        Map-Reduce Framework
                Map input records=365
                Map output records=10220
                Map output bytes=81648
                Map output materialized bytes=102100
                Input split bytes=180
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=102100
                Reduce input records=10220
                Reduce output records=12
                Spilled Records=20440
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=513
                CPU time spent (ms)=5780
```



```
                Input split bytes=180
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=102100
                Reduce input records=10220
                Reduce output records=12
                Spilled Records=20440
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=513
                CPU time spent (ms)=5780
                Physical memory (bytes) snapshot=873496576
                Virtual memory (bytes) snapshot=7734071296
                Total committed heap usage (bytes)=693108736
                Peak Map Physical memory (bytes)=319991808
                Peak Map Virtual memory (bytes)=2576502784
                Peak Reduce Physical memory (bytes)=234762240
                Peak Reduce Virtual memory (bytes)=2582114304
        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=83300
        File Output Format Counters
                Bytes Written=96
2024-10-20 12:08:05,377 INFO streaming.StreamJob: Output directory: /exp3/output3
keerthi@fedora:~$
```

Output :



```
2024-10-26 12:08:03,577 INFO streaming.StreamJob: Output di
keerthi@fedora:~$ hdfs dfs -cat /exp3/output3/part-00000
01      26.5
02      26.6
03      29.1
04      30.8
05      31.1
06      33.6
07      38.5
08      40.2
09      36.5
10      36.9
11      27.6
12      25.9
```