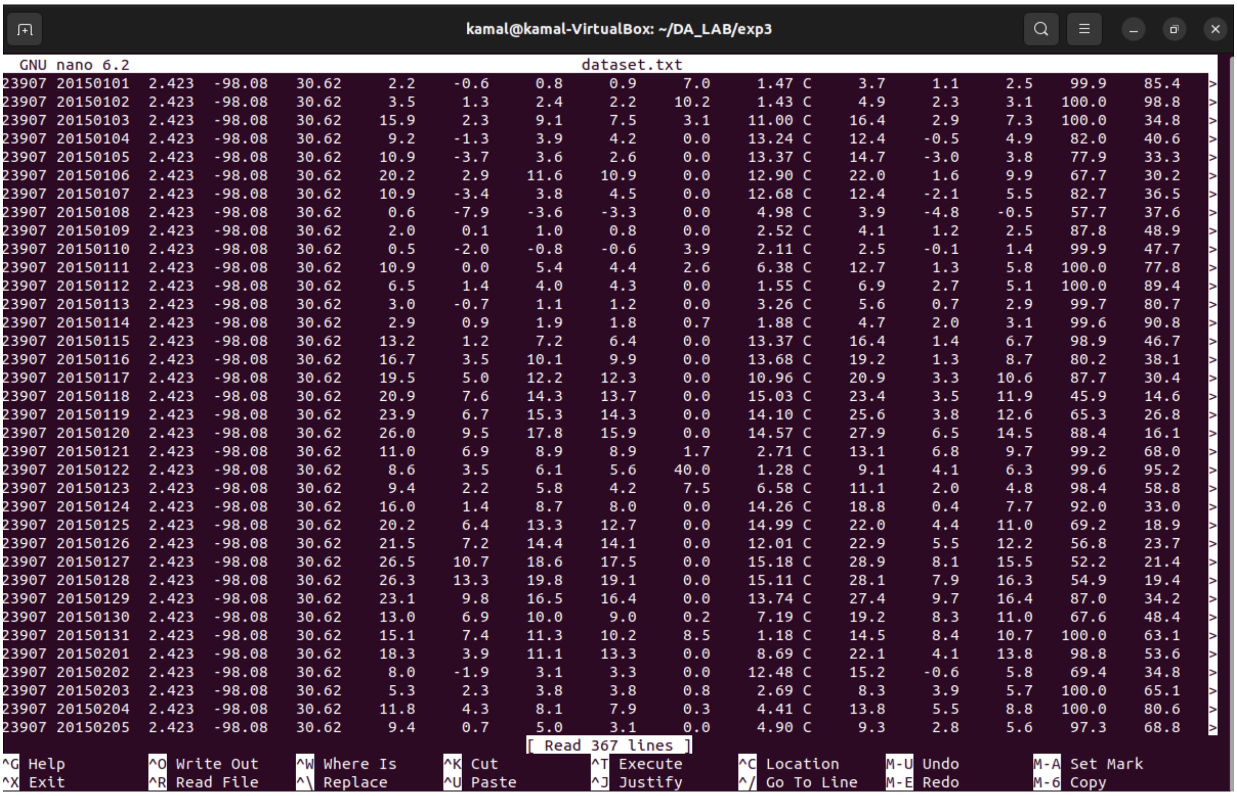
EXP NO: 3 MAP REDUCE PROGRAM TO PROCESS A WEATHER DATASET

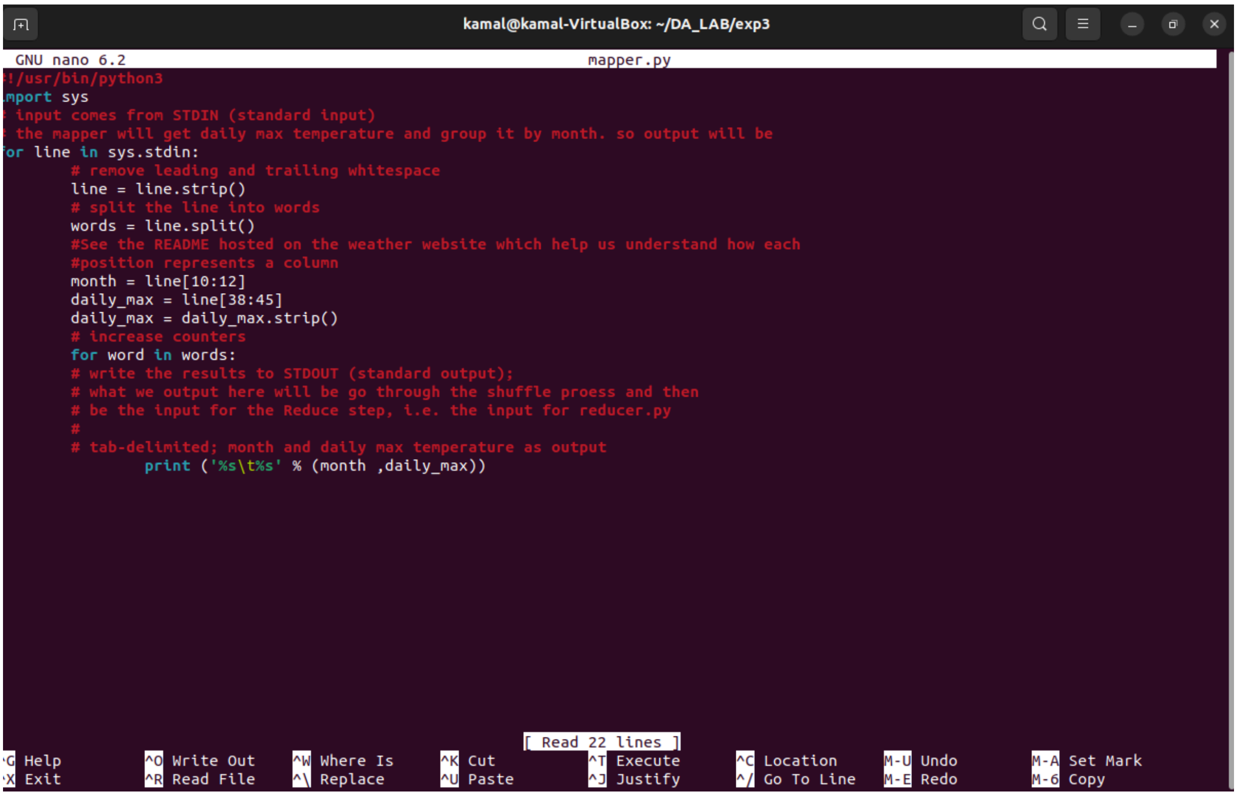
$cd DA-Lab
$mkdir exp3
$cd exp3

$nano dataset.txt



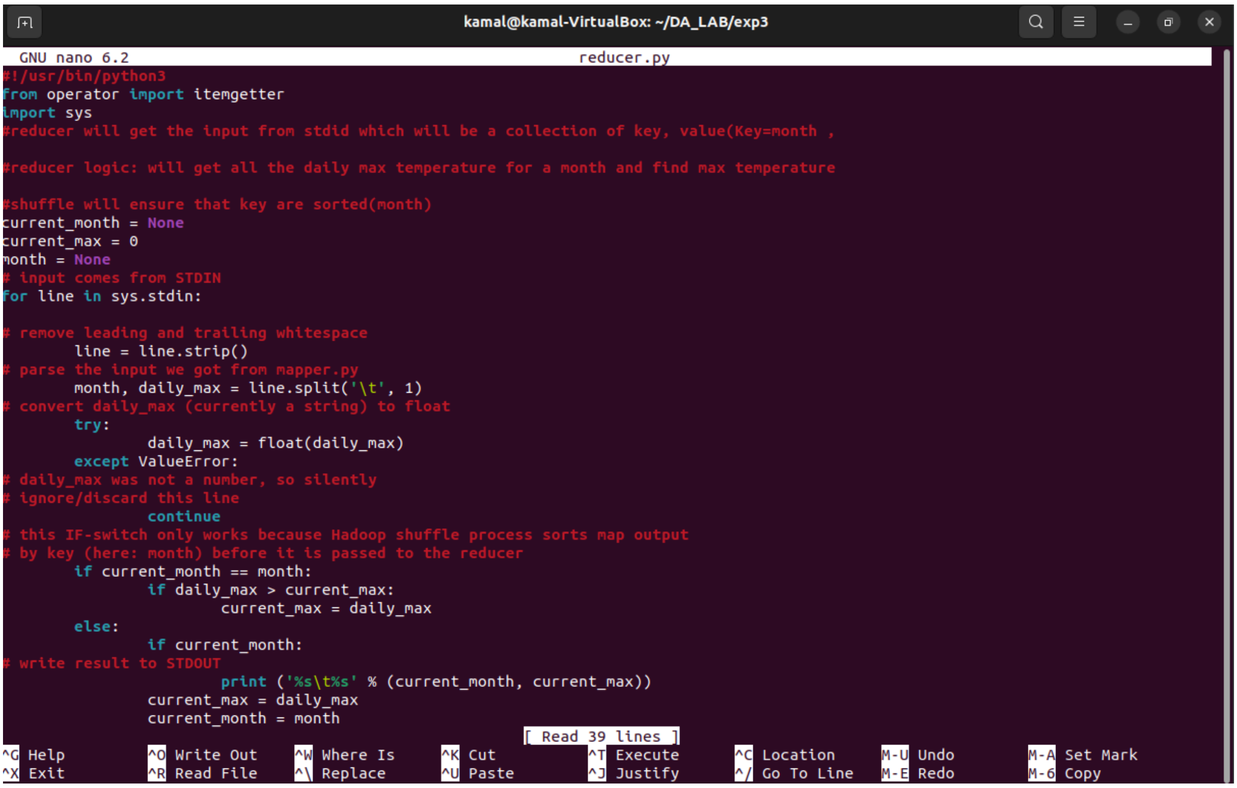
```
kamal@kamal-VirtualBox: ~/DA_LAB/exp3
 GNU nano 6.2                                   dataset.txt
23907 20150101  2.423  -98.08   30.62     2.2    -0.6     0.8     0.9     7.0     1.47 C     3.7     1.1     2.5    99.9    85.4 >
23907 20150102  2.423  -98.08   30.62     3.5     1.3     2.4     2.2    10.2     1.43 C     4.9     2.3     3.1   100.0    98.8 >
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 >
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 >
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 >
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 >
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 >
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 >
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 >
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 >
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   100.0    77.8 >
23907 20150112  2.423  -98.08   30.62     6.5     1.4     4.0     4.3     0.0     1.55 C     6.9     2.7     5.1   100.0    89.4 >
23907 20150113  2.423  -98.08   30.62     3.0    -0.7     1.1     1.2     0.0     3.26 C     5.6     0.7     2.9    99.7    80.7 >
23907 20150114  2.423  -98.08   30.62     2.9     0.9     1.9     1.8     0.7     1.88 C     4.7     2.0     3.1    99.6    90.8 >
23907 20150115  2.423  -98.08   30.62    13.2     1.2     7.2     6.4     0.0    13.37 C    16.4     1.4     6.7    98.9    46.7 >
23907 20150116  2.423  -98.08   30.62    16.7     3.5    10.1     9.9     0.0    13.68 C    19.2     1.3     8.7    80.2    38.1 >
23907 20150117  2.423  -98.08   30.62    19.5     5.0    12.2    12.3     0.0    10.96 C    20.9     3.3    10.6    87.7    30.4 >
23907 20150118  2.423  -98.08   30.62    20.9     7.6    14.3    13.7     0.0    15.03 C    23.4     3.5    11.9    45.9    14.6 >
23907 20150119  2.423  -98.08   30.62    23.9     6.7    15.3    14.3     0.0    14.10 C    25.6     3.8    12.6    65.3    26.8 >
23907 20150120  2.423  -98.08   30.62    26.0     9.5    17.8    15.9     0.0    14.57 C    27.9     6.5    14.5    88.4    16.1 >
23907 20150121  2.423  -98.08   30.62    11.0     6.9     8.9     8.9     1.7     2.71 C    13.1     6.8     9.7    99.2    68.0 >
23907 20150122  2.423  -98.08   30.62     8.6     3.5     6.1     5.6    40.0     1.28 C     9.1     4.1     6.3    99.6    95.2 >
23907 20150123  2.423  -98.08   30.62     9.4     2.2     5.8     4.2     7.5     6.58 C    11.1     2.0     4.8    98.4    58.8 >
23907 20150124  2.423  -98.08   30.62    16.0     1.4     8.7     8.0     0.0    14.26 C    18.8     0.4     7.7    92.0    33.0 >
23907 20150125  2.423  -98.08   30.62    20.2     6.4    13.3    12.7     0.0    14.99 C    22.0     4.4    11.0    69.2    18.9 >
23907 20150126  2.423  -98.08   30.62    21.5     7.2    14.4    14.1     0.0    12.01 C    22.9     5.5    12.2    56.8    23.7 >
23907 20150127  2.423  -98.08   30.62    26.5    10.7    18.6    17.5     0.0    15.18 C    28.9     8.1    15.5    52.2    21.4 >
23907 20150128  2.423  -98.08   30.62    26.3    13.3    19.8    19.1     0.0    15.11 C    28.1     7.9    16.3    54.9    19.4 >
23907 20150129  2.423  -98.08   30.62    23.1     9.8    16.5    16.4     0.0    13.74 C    27.4     9.7    16.4    87.0    34.2 >
23907 20150130  2.423  -98.08   30.62    13.0     6.9    10.0     9.0     0.2     7.19 C    19.2     8.3    11.0    67.6    48.4 >
23907 20150131  2.423  -98.08   30.62    15.1     7.4    11.3    10.2     8.5     1.18 C    14.5     8.4    10.7   100.0    63.1 >
23907 20150201  2.423  -98.08   30.62    18.3     3.9    11.1    13.3     0.0     8.69 C    22.1     4.1    13.8    98.8    53.6 >
23907 20150202  2.423  -98.08   30.62     8.0    -1.9     3.1     3.3     0.0    12.48 C    15.2    -0.6     5.8    69.4    34.8 >
23907 20150203  2.423  -98.08   30.62     5.3     2.3     3.8     3.8     0.8     2.69 C     8.3     3.9     5.7   100.0    65.1 >
23907 20150204  2.423  -98.08   30.62    11.8     4.3     8.1     7.9     0.3     4.41 C    13.8     5.5     8.8   100.0    80.6 >
23907 20150205  2.423  -98.08   30.62     9.4     0.7     5.0     3.1     0.0     4.90 C     9.3     2.8     5.6    97.3    68.8 >
                                                   [ Read 367 lines ]
^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location    M-U Undo       M-A Set Mark
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line  M-E Redo       M-6 Copy
```

$nano mapper.py



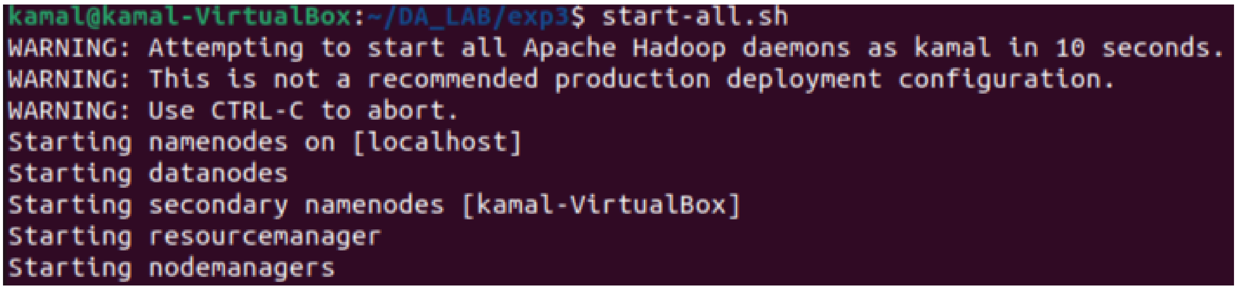
```
kamal@kamal-VirtualBox: ~/DA_LAB/exp3
 GNU nano 6.2                                   mapper.py
!/usr/bin/python3
mport sys
 input comes from STDIN (standard input)
 the mapper will get daily max temperature and group it by month. so output will be
or line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()
        # split the line into words
        words = line.split()
        #See the README hosted on the weather website which help us understand how each
        #position represents a column
        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 and then
        # be the input for the Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; month and daily max temperature as output
                print ('%s\t%s' % (month ,daily_max))

                                                   [ Read 22 lines ]
^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location    M-U Undo       M-A Set Mark
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line  M-E Redo       M-6 Copy
```

$nano reducer.py



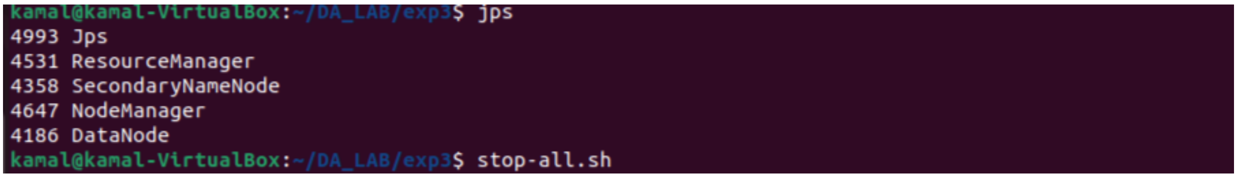
```
kamal@kamal-VirtualBox: ~/DA_LAB/exp3
  GNU nano 6.2                                                 reducer.py
#!/usr/bin/python3
from operator import itemgetter
import sys
#reducer will get the input from stdid which will be a collection of key, value(Key=month ,

#reducer logic: will get all the daily max temperature for a month and find max temperature

#shuffle will ensure that key are sorted(month)
current_month = None
current_max = 0
month = None
# input comes from STDIN
for line in sys.stdin:

# remove leading and trailing whitespace
        line = line.strip()
# parse the input we got from mapper.py
        month, daily_max = line.split('\t', 1)
# convert daily_max (currently a string) to float
        try:
                daily_max = float(daily_max)
        except ValueError:
# daily_max was not a number, so silently
# ignore/discard this line
                continue
# this IF-switch only works because Hadoop shuffle process sorts map output
# by key (here: month) before it is passed to the reducer
        if current_month == month:
                if daily_max > current_max:
                        current_max = daily_max
        else:
                if current_month:
# write result to STDOUT
                        print ('%s\t%s' % (current_month, current_max))
                current_max = daily_max
                current_month = month
                                                              [ Read 39 lines ]
^G Help          ^O Write Out     ^W Where Is      ^K Cut           ^T Execute       ^C Location      M-U Undo         M-A Set Mark
^X Exit          ^R Read File     ^\ Replace       ^U Paste         ^J Justify       ^/ Go To Line    M-E Redo         M-6 Copy
```

$start-all.sh



```
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kamal in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [kamal-VirtualBox]
Starting resourcemanager
Starting nodemanagers
```

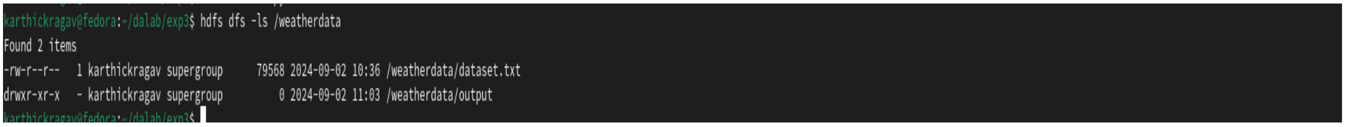
$ jps



```
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ jps
4993 Jps
4531 ResourceManager
4358 SecondaryNameNode
4647 NodeManager
4186 DataNode
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ stop-all.sh
```

$hdfs dfs -mkdir /exp3

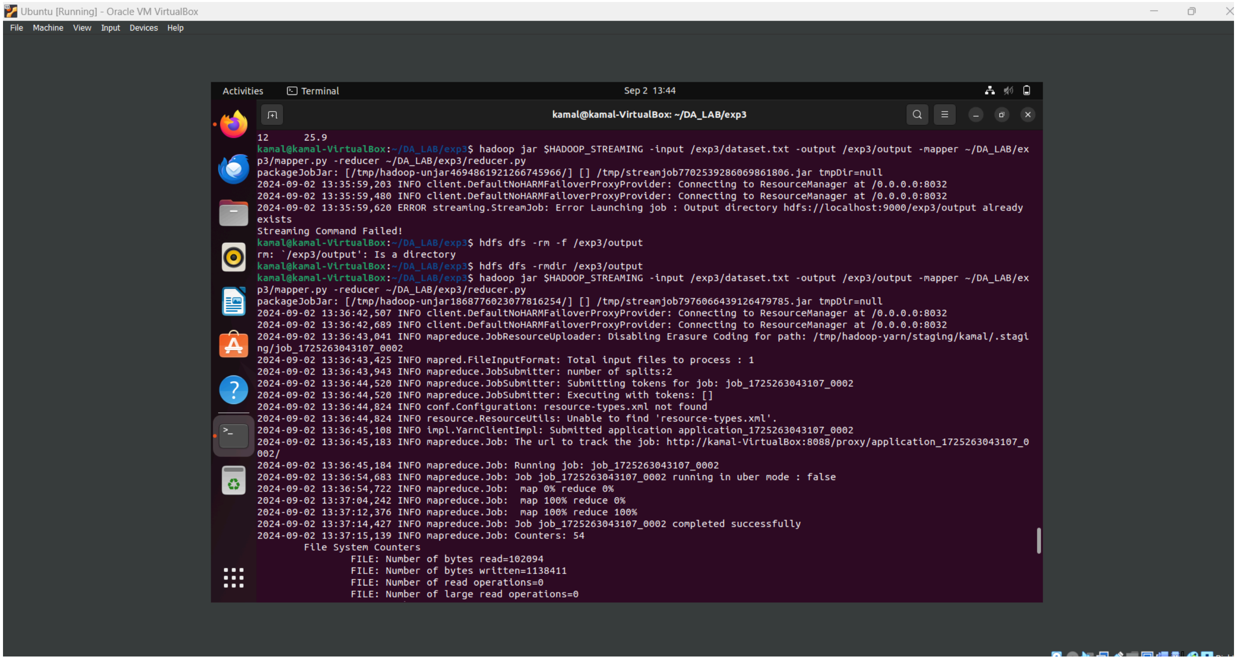
$hdfs dfs -copyFromLocal ~/DA-Lab/exp3/dataset.txt /exp2



```
karthickragav@fedora:~/dalab/exp3$ hdfs dfs -ls /weatherdata
Found 2 items
-rw-r--r--   1 karthickragav supergroup      79568 2024-09-02 10:36 /weatherdata/dataset.txt
drwxr-xr-x   - karthickragav supergroup          0 2024-09-02 11:03 /weatherdata/output
karthickragav@fedora:~/dalab/exp3$
```

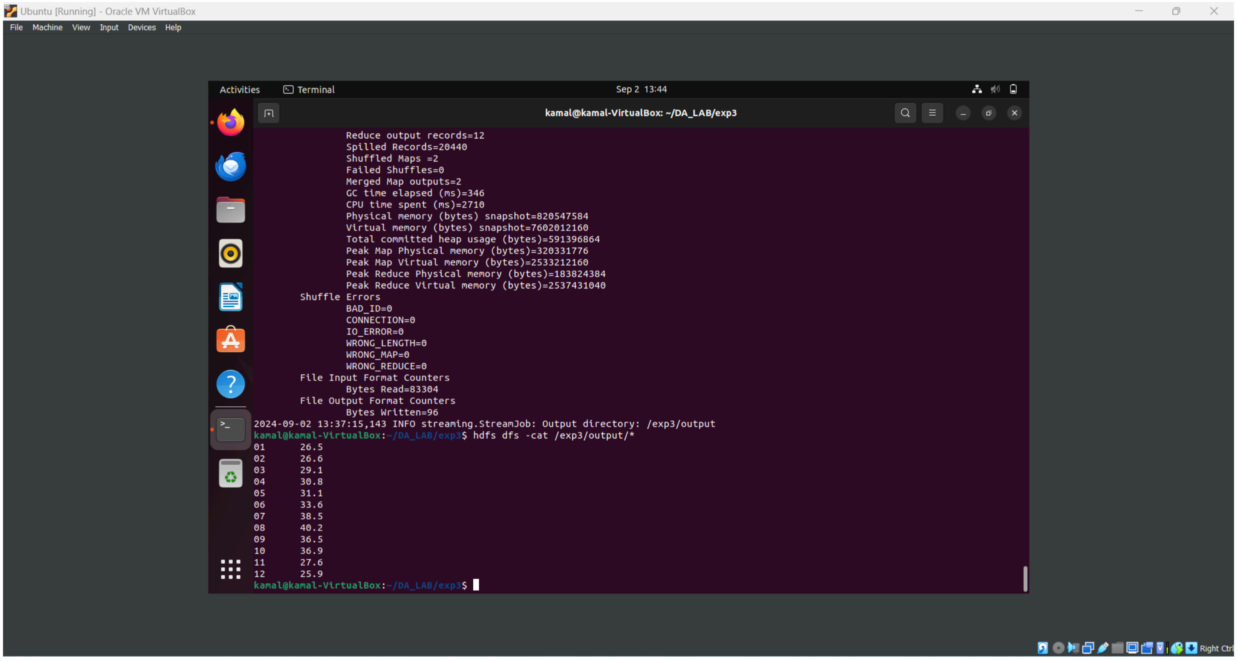
$chmod 777 mapper.py reducer.py

$hadoop jar $HADOOP_STREAMING -input /exp3/dataset.txt -output /exp3/output -mapper ~/DA-Lab/exp3/mapper.py -reducer ~/DA-Lab/exp3/reducer.py


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Sep 2 13:44
kamal@kamal-VirtualBox: ~/DA_LAB/exp3

```
12      25.9
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ hadoop jar $HADOOP_STREAMING -input /exp3/dataset.txt -output /exp3/output -mapper ~/DA_LAB/ex
p3/mapper.py -reducer ~/DA_LAB/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar4694861921266745966/] [] /tmp/streamjob7702539286069861806.jar tmpDir=null
2024-09-02 13:35:59,203 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-02 13:35:59,480 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-02 13:35:59,620 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://localhost:9000/exp3/output already
exists
Streaming Command Failed!
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ hdfs dfs -rm -f /exp3/output
rm: `/exp3/output': Is a directory
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ hdfs dfs -rmdir /exp3/output
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ hadoop jar $HADOOP_STREAMING -input /exp3/dataset.txt -output /exp3/output -mapper ~/DA_LAB/ex
p3/mapper.py -reducer ~/DA_LAB/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar1868776023077816254/] [] /tmp/streamjob7976066439126479785.jar tmpDir=null
2024-09-02 13:36:42,507 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-02 13:36:42,689 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-02 13:36:43,041 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/kamal/.stagi
ng/job_1725263043107_0002
2024-09-02 13:36:43,425 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-02 13:36:43,943 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-02 13:36:44,520 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725263043107_0002
2024-09-02 13:36:44,520 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-02 13:36:44,824 INFO conf.Configuration: resource-types.xml not found
2024-09-02 13:36:44,824 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-02 13:36:45,108 INFO impl.YarnClientImpl: Submitted application application_1725263043107_0002
2024-09-02 13:36:45,183 INFO mapreduce.Job: The url to track the job: http://kamal-VirtualBox:8088/proxy/application_1725263043107_0
002/
2024-09-02 13:36:45,184 INFO mapreduce.Job: Running job: job_1725263043107_0002
2024-09-02 13:36:54,683 INFO mapreduce.Job: Job job_1725263043107_0002 running in uber mode : false
2024-09-02 13:36:54,722 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-02 13:37:04,242 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-02 13:37:12,376 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-02 13:37:14,427 INFO mapreduce.Job: Job job_1725263043107_0002 completed successfully
2024-09-02 13:37:15,139 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=102094
                FILE: Number of bytes written=1138411
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

$hdfs dfs -cat /exp3/output/*


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Sep 2 13:44
kamal@kamal-VirtualBox: ~/DA_LAB/exp3
Right Ctrl

```
                Reduce output records=12
                Spilled Records=20440
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=346
                CPU time spent (ms)=2710
                Physical memory (bytes) snapshot=820547584
                Virtual memory (bytes) snapshot=7602012160
                Total committed heap usage (bytes)=591396864
                Peak Map Physical memory (bytes)=320331776
                Peak Map Virtual memory (bytes)=2533212160
                Peak Reduce Physical memory (bytes)=183824384
                Peak Reduce Virtual memory (bytes)=2537431040
        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=83304
        File Output Format Counters
                Bytes Written=96
2024-09-02 13:37:15,143 INFO streaming.StreamJob: Output directory: /exp3/output
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ hdfs dfs -cat /exp3/output/*
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
kamal@kamal-VirtualBox:~/DA_LAB/exp3$
```