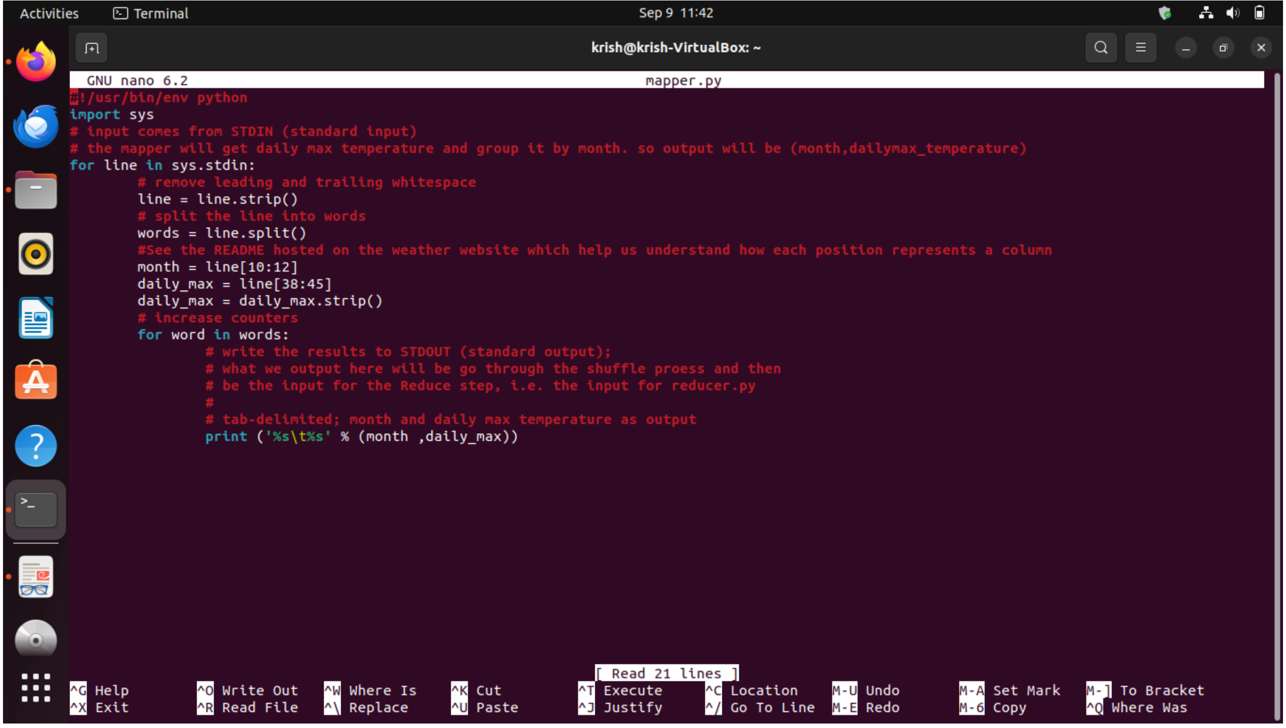
EXP NO: 3 MAP REDUCE PROGRAM TO PROCESS A WEATHER DATASET

$cd DA-Lab
$mkdir exp3
$cd exp3

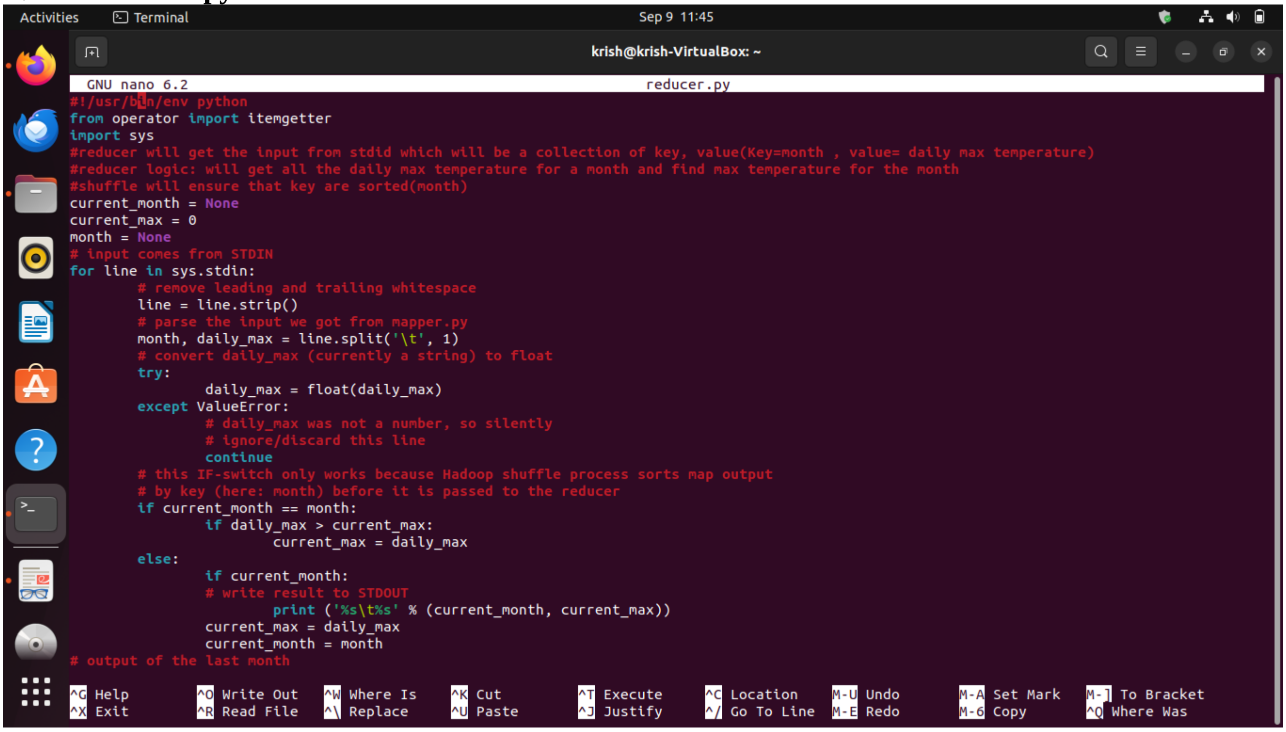
$nano dataset.txt

$nano mapper.py



```
#!/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 will be (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 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))
```

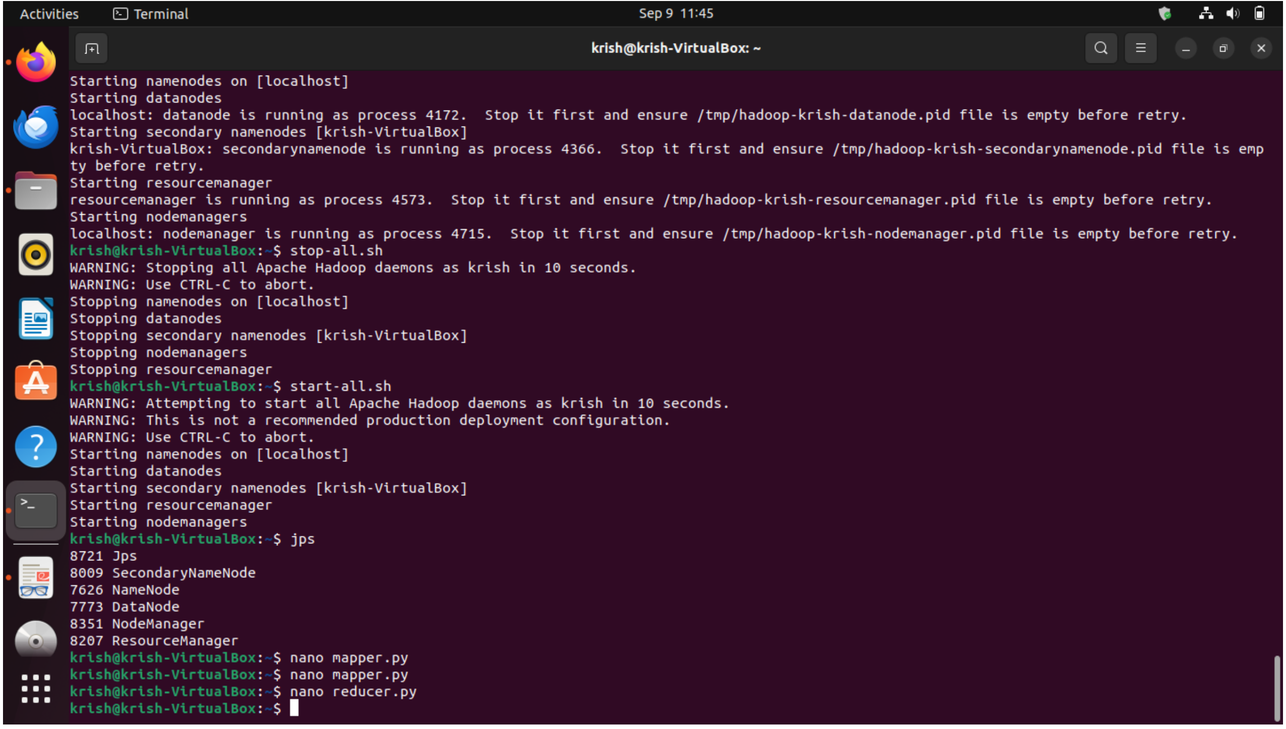
$nano reducer.py



```
#!/usr/bin/env python
from operator import itemgetter
import sys
#reducer will get the input from stdid which will be a collection of key, value(Key=month , value= daily max temperature)
#reducer logic: will get all the daily max temperature for a month and find max temperature for the month
#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
# output of the last month
```

$start-all.sh

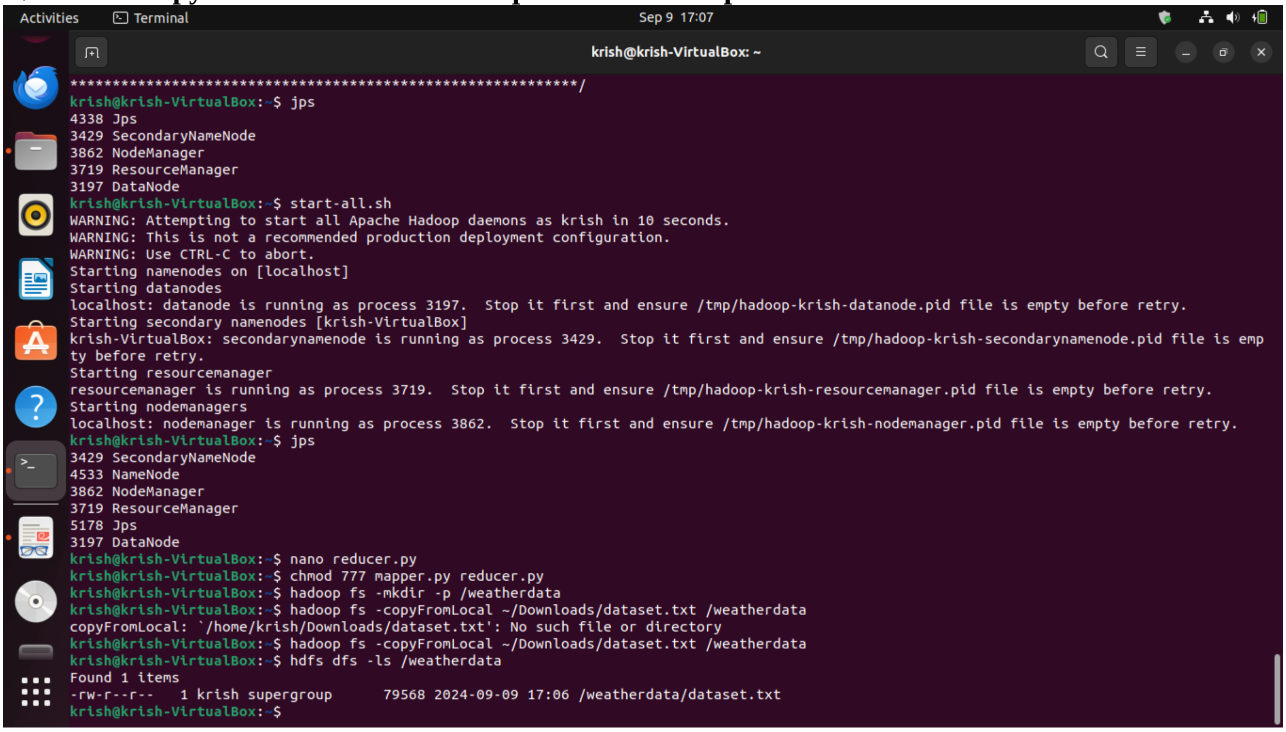
$ jps



```
Starting namenodes on [localhost]
Starting datanodes
localhost: datanode is running as process 4172.  Stop it first and ensure /tmp/hadoop-krish-datanode.pid file is empty before retry.
Starting secondary namenodes [krish-VirtualBox]
krish-VirtualBox: secondarynamenode is running as process 4366.  Stop it first and ensure /tmp/hadoop-krish-secondarynamenode.pid file is emp
ty before retry.
Starting resourcemanager
resourcemanager is running as process 4573.  Stop it first and ensure /tmp/hadoop-krish-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 4715.  Stop it first and ensure /tmp/hadoop-krish-nodemanager.pid file is empty before retry.
krish@krish-VirtualBox:~$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as krish in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [krish-VirtualBox]
Stopping nodemanagers
Stopping resourcemanager
krish@krish-VirtualBox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as krish 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 [krish-VirtualBox]
Starting resourcemanager
Starting nodemanagers
krish@krish-VirtualBox:~$ jps
8721 Jps
8009 SecondaryNameNode
7626 NameNode
7773 DataNode
8351 NodeManager
8207 ResourceManager
krish@krish-VirtualBox:~$ nano mapper.py
krish@krish-VirtualBox:~$ nano mapper.py
krish@krish-VirtualBox:~$ nano reducer.py
krish@krish-VirtualBox:~$ 
```

$hdfs dfs -mkdir /exp3

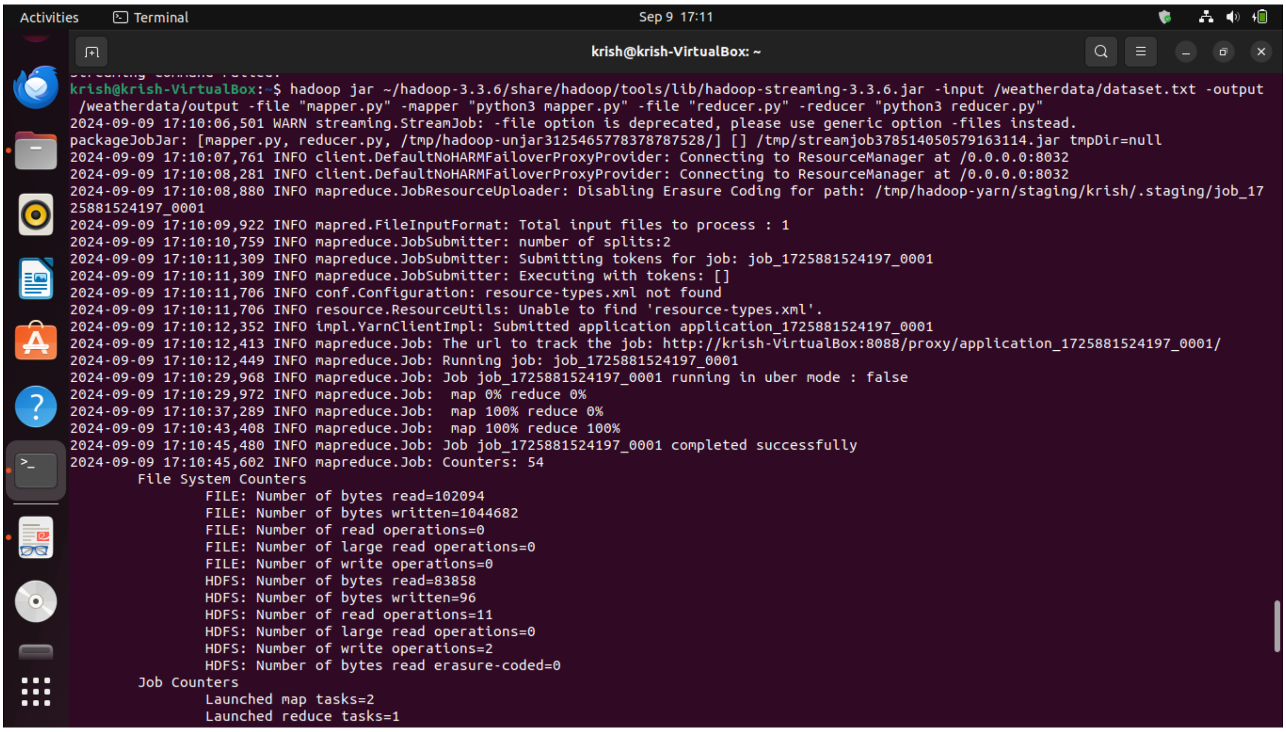
$hdfs dfs -copyFromLocal ~/DA-Lab/exp3/dataset.txt /exp2



```
*****************************************************************/
krish@krish-VirtualBox:~$ jps
4338 Jps
3429 SecondaryNameNode
3862 NodeManager
3719 ResourceManager
3197 DataNode
krish@krish-VirtualBox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as krish in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
localhost: datanode is running as process 3197.  Stop it first and ensure /tmp/hadoop-krish-datanode.pid file is empty before retry.
Starting secondary namenodes [krish-VirtualBox]
krish-VirtualBox: secondarynamenode is running as process 3429.  Stop it first and ensure /tmp/hadoop-krish-secondarynamenode.pid file is emp
ty before retry.
Starting resourcemanager
resourcemanager is running as process 3719.  Stop it first and ensure /tmp/hadoop-krish-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 3862.  Stop it first and ensure /tmp/hadoop-krish-nodemanager.pid file is empty before retry.
krish@krish-VirtualBox:~$ jps
3429 SecondaryNameNode
4533 NameNode
3862 NodeManager
3719 ResourceManager
5178 Jps
3197 DataNode
krish@krish-VirtualBox:~$ nano reducer.py
krish@krish-VirtualBox:~$ chmod 777 mapper.py reducer.py
krish@krish-VirtualBox:~$ hadoop fs -mkdir -p /weatherdata
krish@krish-VirtualBox:~$ hadoop fs -copyFromLocal ~/Downloads/dataset.txt /weatherdata
copyFromLocal: `/home/krish/Downloads/dataset.txt': No such file or directory
krish@krish-VirtualBox:~$ hadoop fs -copyFromLocal ~/Downloads/dataset.txt /weatherdata
krish@krish-VirtualBox:~$ hdfs dfs -ls /weatherdata
Found 1 items
-rw-r--r--   1 krish supergroup      79568 2024-09-09 17:06 /weatherdata/dataset.txt
krish@krish-VirtualBox:~$
```

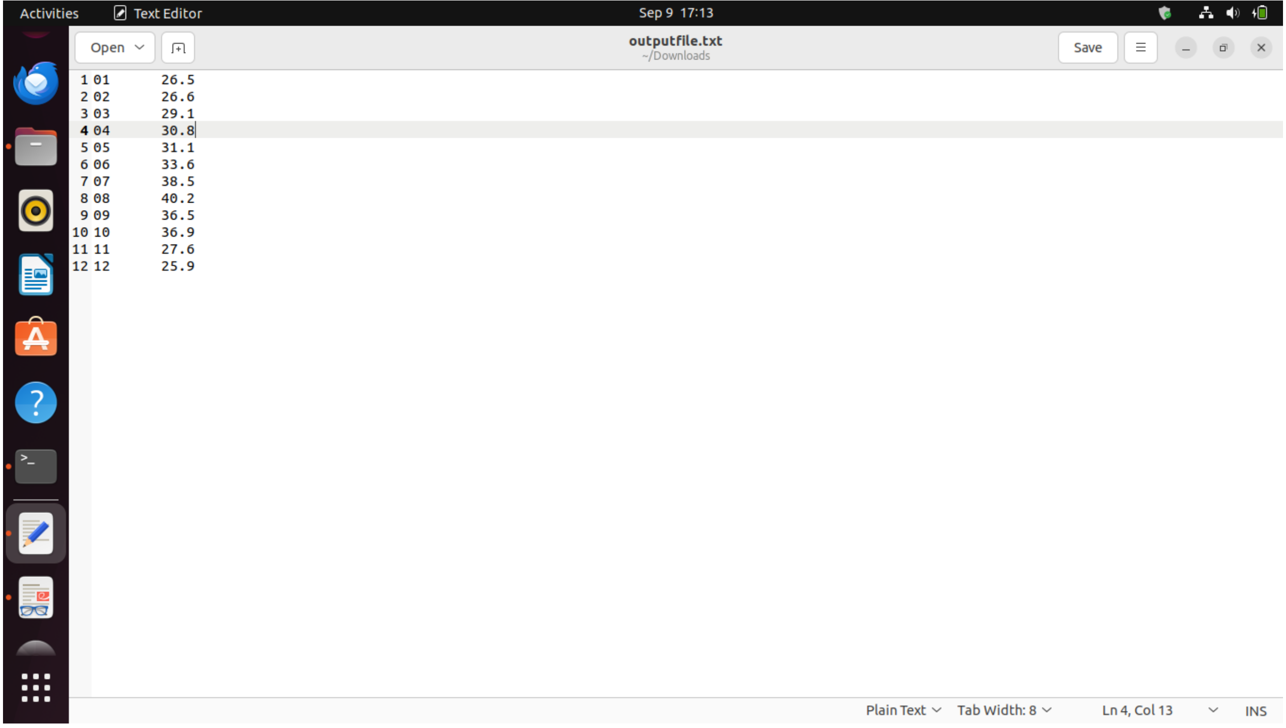
$chmod 777 mapper.py reducer.py

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



```
Streaming Command Failed!
krish@krish-VirtualBox:~$ hadoop jar ~/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /weatherdata/dataset.txt -output
 /weatherdata/output -file "mapper.py" -mapper "python3 mapper.py" -file "reducer.py" -reducer "python3 reducer.py"
2024-09-09 17:10:06,501 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-unjar3125465778378787528/] [] /tmp/streamjob3785140505791631 14.jar tmpDir=null
2024-09-09 17:10:07,761 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-09 17:10:08,281 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-09 17:10:08,880 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/krish/.staging/job_17
25881524197_0001
2024-09-09 17:10:09,922 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-09 17:10:10,759 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-09 17:10:11,309 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725881524197_0001
2024-09-09 17:10:11,309 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-09 17:10:11,706 INFO conf.Configuration: resource-types.xml not found
2024-09-09 17:10:11,706 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-09 17:10:12,352 INFO impl.YarnClientImpl: Submitted application application_1725881524197_0001
2024-09-09 17:10:12,413 INFO mapreduce.Job: The url to track the job: http://krish-VirtualBox:8088/proxy/application_1725881524197_0001/
2024-09-09 17:10:12,449 INFO mapreduce.Job: Running job: job_1725881524197_0001
2024-09-09 17:10:29,968 INFO mapreduce.Job: Job job_1725881524197_0001 running in uber mode : false
2024-09-09 17:10:29,972 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-09 17:10:37,289 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-09 17:10:43,408 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-09 17:10:45,480 INFO mapreduce.Job: Job job_1725881524197_0001 completed successfully
2024-09-09 17:10:45,602 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=102094
                FILE: Number of bytes written=1044682
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83858
                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
                Launched reduce tasks=1
```

$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
```