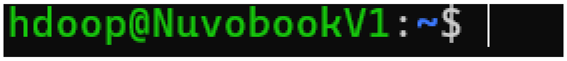
Program 3: Write a Map Reduce program that mines weather data. Hint: Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with Map Reduce, since it is semi structured and record-oriented.

: Login into Hadoop user you used while installing Hadoop , here we use hdoop user



```
hdoop@NuvobookV1:~$
```

$ su – hdoop

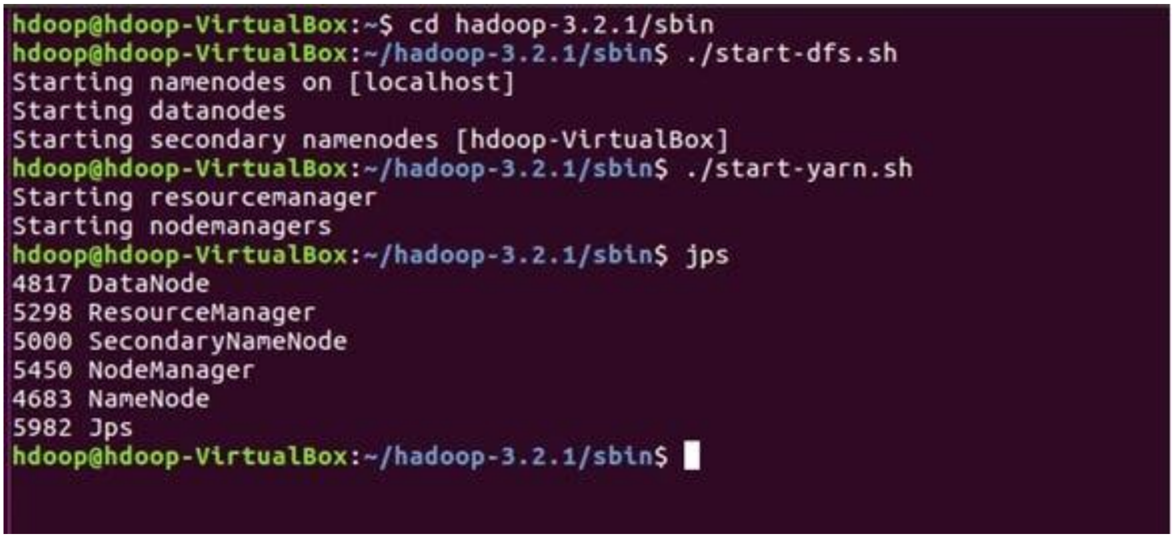
-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

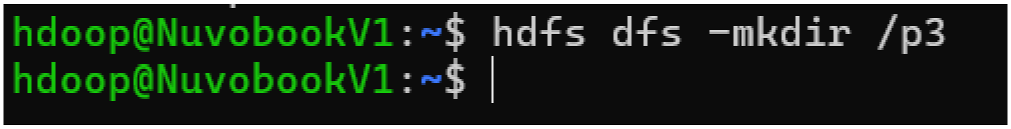
$ jps



```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

#make a dir for the program 3

$ hdfs dfs -mkdir /p3

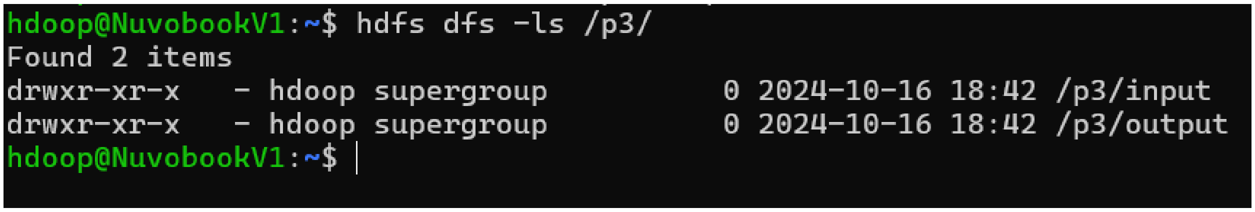


```
hdoop@NuvobookV1:~$ hdfs dfs -mkdir /p3
hdoop@NuvobookV1:~$
```

#create input and output folders

$ hdfs dfs -mkdir /p3/input

$ hdfs dfs -mkdir /p3/output



```
hdoop@NuvobookV1:~$ hdfs dfs -ls /p3/
Found 2 items
drwxr-xr-x   - hdoop supergroup          0 2024-10-16 18:42 /p3/input
drwxr-xr-x   - hdoop supergroup          0 2024-10-16 18:42 /p3/output
hdoop@NuvobookV1:~$
```

#create mapper program

$sudo mkdir program3/

$cd program3/

$sudo nano mapper.py

-insert this code

```
#!/usr/bin/env python3


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))
```

-click Ctrl+s and Ctrl+x to close it

note: make sure you have python3 in your system

#reducer program

```
$sudo nano reducer.py
```

-insert this code

```
#!/usr/bin/env python3

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
if current_month == month:
    print ('%s\t%s' % (current_month, current_max))
```

#create the dataset of temp :

Copy the content of this web page:

https://www.ncei.noaa.gov/pub/data/uscrn/products/daily01/2002/CRND0103-2002-RI_Kingston_1_NW.txt

```
$ sudo nano tempdata.txt
```

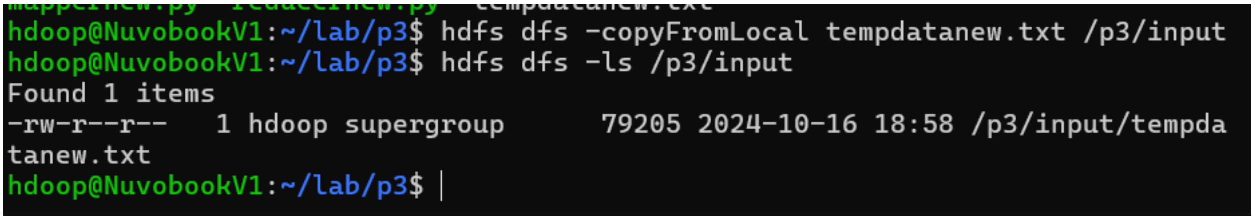
-insert the data u got from the webpage

#change the permissions of the files

$sudo chmod 777 mapper.py reducer.py

#upload the dataset to Hadoop

$ hdfs dfs -copyFromLocal tempdatanew.txt /p3/input



```
hdoop@NuvobookV1:~/lab/p3$ hdfs dfs -copyFromLocal tempdatanew.txt /p3/input
hdoop@NuvobookV1:~/lab/p3$ hdfs dfs -ls /p3/input
Found 1 items
-rw-r--r--   1 hdoop supergroup      79205 2024-10-16 18:58 /p3/input/tempda
tanew.txt
hdoop@NuvobookV1:~/lab/p3$ |
```

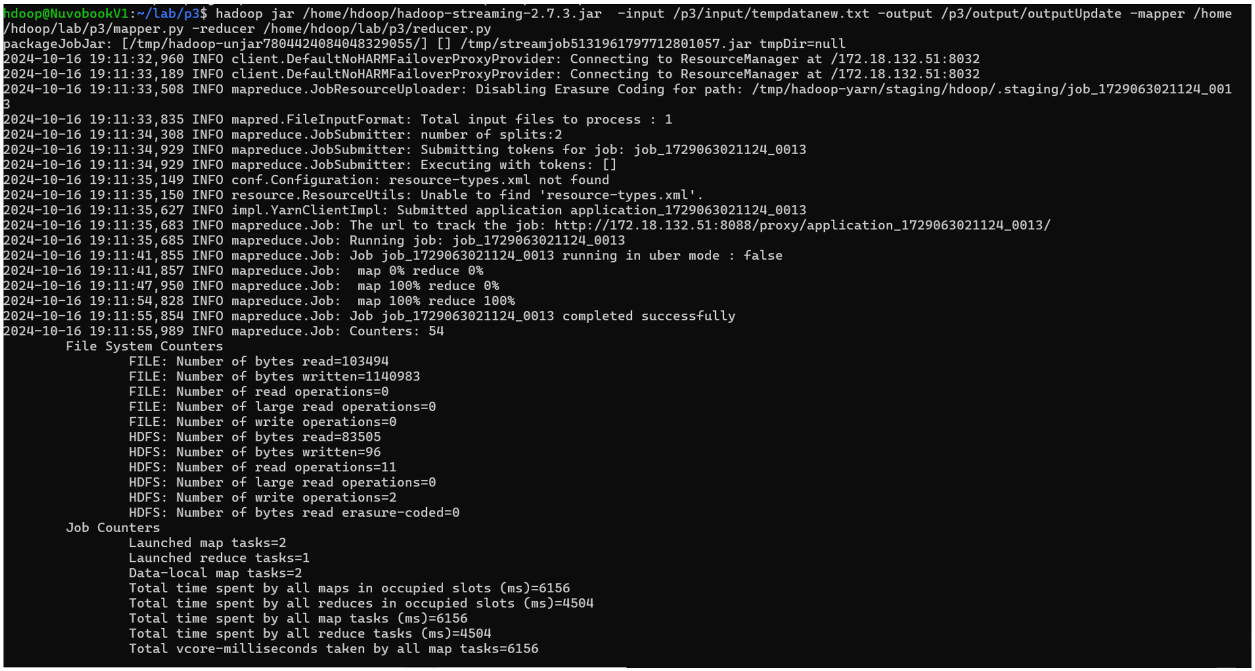
#running python program in Hadoop using Hadoop streamer

$ cd

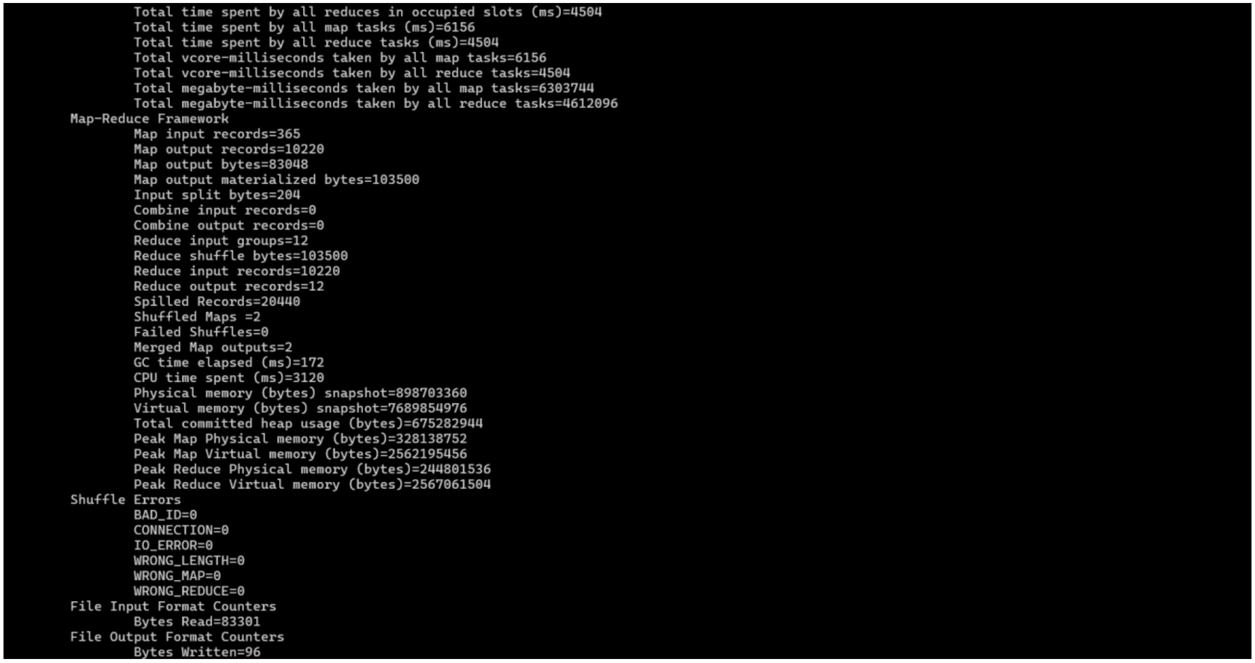
$ wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoopstreaming/2.7.3/hadoop-streaming-2.7.3.jar $ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p2/sample.txt -output /p2/output -mapper /home/hdoop/p2/mapper.py -reducer /home/hdoop/p2/reducer.py

$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p3/input/tempdatanew.txt -output /p3/output/outputUpdate -mapper /home/hdoop/p3/mapper.py -reducer /home/hdoop/p3/reducer.py

#output



```
hdoop@NuvobookV1:~/lab/p3$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar  -input /p3/input/tempdatanew.txt -output /p3/output/outputUpdate -mapper /home
/hdoop/lab/p3/mapper.py -reducer /home/hdoop/lab/p3/reducer.py
packageJobJar: [/tmp/hadoop-unjar7804424084048329055/] [] /tmp/streamjob5131961797712801057.jar tmpDir=null
2024-10-16 19:11:32,960 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 19:11:33,189 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 19:11:33,508 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1729063021124_001
3
2024-10-16 19:11:33,835 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-16 19:11:34,308 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-16 19:11:34,929 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729063021124_0013
2024-10-16 19:11:34,929 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-16 19:11:35,149 INFO conf.Configuration: resource-types.xml not found
2024-10-16 19:11:35,150 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-16 19:11:35,627 INFO impl.YarnClientImpl: Submitted application application_1729063021124_0013
2024-10-16 19:11:35,683 INFO mapreduce.Job: The url to track the job: http://172.18.132.51:8088/proxy/application_1729063021124_0013/
2024-10-16 19:11:35,685 INFO mapreduce.Job: Running job: job_1729063021124_0013
2024-10-16 19:11:41,855 INFO mapreduce.Job: Job job_1729063021124_0013 running in uber mode : false
2024-10-16 19:11:41,857 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-16 19:11:47,950 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-16 19:11:54,828 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-16 19:11:55,854 INFO mapreduce.Job: Job job_1729063021124_0013 completed successfully
2024-10-16 19:11:55,989 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=103494
                FILE: Number of bytes written=1140983
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83505
                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
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=6156
                Total time spent by all reduces in occupied slots (ms)=4504
                Total time spent by all map tasks (ms)=6156
                Total time spent by all reduce tasks (ms)=4504
                Total vcore-milliseconds taken by all map tasks=6156
```



```
                Total time spent by all reduces in occupied slots (ms)=4504
                Total time spent by all map tasks (ms)=6156
                Total time spent by all reduce tasks (ms)=4504
                Total vcore-milliseconds taken by all map tasks=6156
                Total vcore-milliseconds taken by all reduce tasks=4504
                Total megabyte-milliseconds taken by all map tasks=6303744
                Total megabyte-milliseconds taken by all reduce tasks=4612096
        Map-Reduce Framework
                Map input records=365
                Map output records=10220
                Map output bytes=83048
                Map output materialized bytes=103500
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=103500
                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)=172
                CPU time spent (ms)=3120
                Physical memory (bytes) snapshot=898703360
                Virtual memory (bytes) snapshot=7689854976
                Total committed heap usage (bytes)=675282944
                Peak Map Physical memory (bytes)=328138752
                Peak Map Virtual memory (bytes)=2562195456
                Peak Reduce Physical memory (bytes)=244801536
                Peak Reduce Virtual memory (bytes)=2567061504
        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=83301
        File Output Format Counters
                Bytes Written=96
```