Exp. No : 3

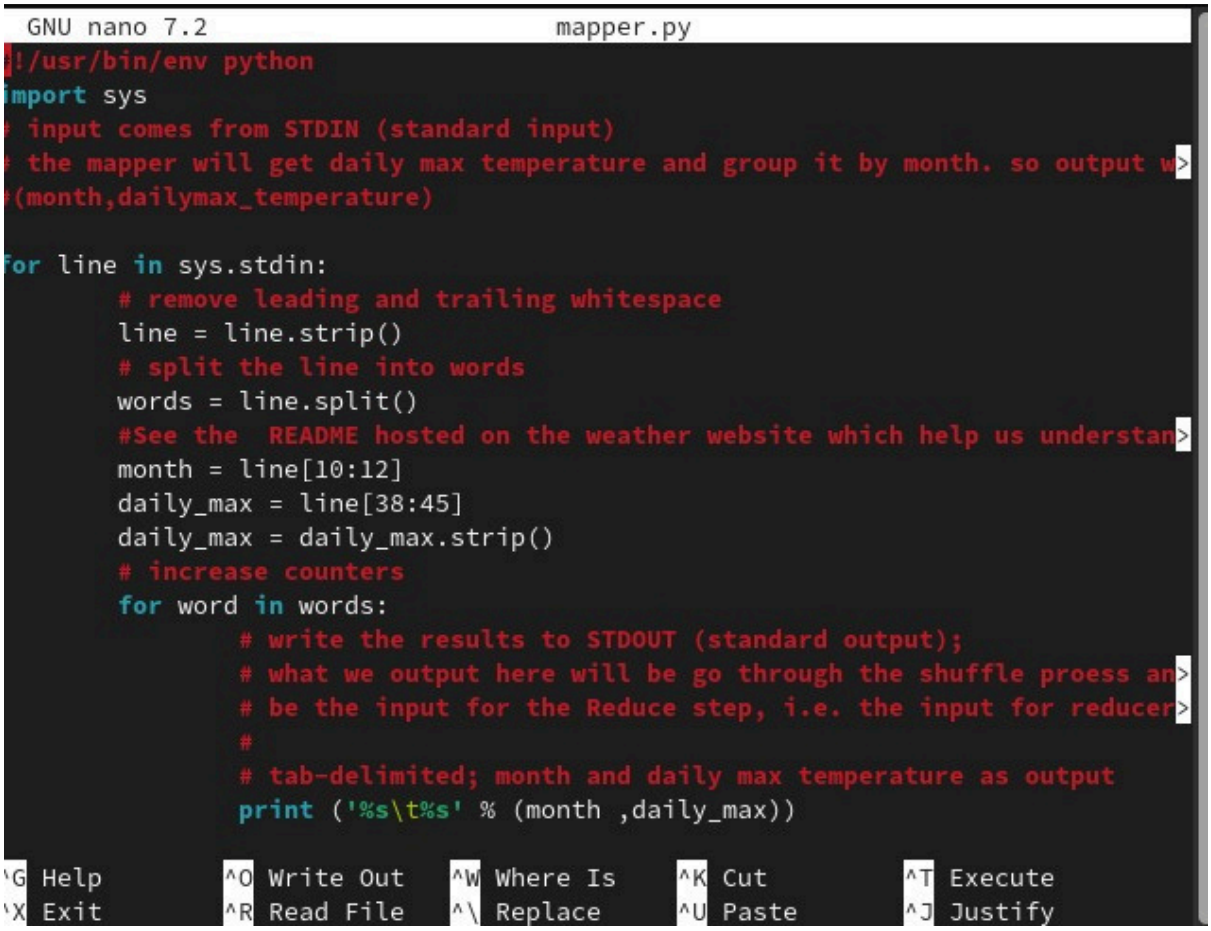
Map Reduce program to process Weather dataset

1. Download Weather dataset.

GNU nano 7.2 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   97.2  0.369  0.308 -99.000 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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 -99.00>
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   97.1  0.420  0.362 -99.000 -99.00>
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   97.8  0.412  0.350 -99.000 -99.00>
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   90.7  0.401  0.337 -99.000 -99.00>
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   97.9  0.395  0.331 -99.000 -99.00>
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   73.4  0.395  0.333 -99.000 -99.00>
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   58.2  0.391  0.330 -99.000 -99.00>
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   55.7  0.388  0.327 -99.000 -99.00>
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   31.4  0.383  0.325 -99.000 -99.00>
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   45.6  0.376  0.321 -99.000 -99.00>
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   50.2  0.373  0.320 -99.000 -99.00>
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   88.1  0.369  0.317 -99.000 -99.00>
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   98.0  0.546  0.418 -99.000 -99.00>
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   86.5  0.554  0.409 -99.000 -99.00>
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   63.0  0.494  0.381 -99.000 -99.00>
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   43.8  0.456  0.357 -99.000 -99.00>
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   40.6  0.433  0.349 -99.000 -99.00>
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   38.8  0.420  0.344 -99.000 -99.00>
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   35.5  0.410  0.339 -99.000 -99.00>
```

2. Create mapper.py program



```
 GNU nano 7.2                          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 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))

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

3. Create reducer.py

```
  GNU nano 7.2                          reducer.py                          Modified
#!/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))

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

4. Start Hadoop services.

```
arise@fedora:~/hadoop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as harithaah 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 [fedora]
jpsStarting resourcemanager
Starting nodemanagers
arise@fedora:~/hadoop$ jps
9458 SecondaryNameNode
9779 ResourceManager
9253 DataNode
9909 NodeManager
9095 NameNode
10045 Jps
```

5. Upload Weather dataset into HDFS Storage.

```
arise@fedora:ora:~/exp3$ hdfs dfs -mkdir /exp2
arise@fedora:ora:~/exp3$ hdfs dfs -put dataset.txt /exp2
```

6. Run the Map reduce program using Hadoop Streaming.

```
harithaah@fedora:~/exp3$ hadoop jar $HADOOP_STREAMING -input /exp2/dataset.txt -output /exp2/output -mapper ~/exp3/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar2895422846984926388/] [] /tmp/streamjob479605575629239246O.jar tmpDir=null
2024-10-10 20:46:38,635 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:46:40,718 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:46:42,732 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/harithaah/.staging/job_1728572703273_0002
2024-10-10 20:46:44,245 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-10 20:46:44,826 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-10 20:46:45,912 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728572703273_0002
2024-10-10 20:46:45,913 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-10 20:46:46,807 INFO conf.Configuration: resource-types.xml not found
2024-10-10 20:46:46,808 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-10 20:46:47,146 INFO impl.YarnClientImpl: Submitted application application_1728572703273_0002
2024-10-10 20:46:47,411 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1728572703273_0002/
2024-10-10 20:46:47,423 INFO mapreduce.Job: Running job: job_1728572703273_0002
```

```
in uber mode : false
2024-08-28 12:29:30,617 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-28 12:29:43,801 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-28 12:29:53,121 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-28 12:29:55,350 INFO mapreduce.Job: Job job_1724828139433_0001 complete
d successfully
2024-08-28 12:29:55,534 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=102094
                FILE: Number of bytes written=1041193
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83844
                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)=20327
                Total time spent by all reduces in occupied slots (ms)=5986
                Total time spent by all map tasks (ms)=20327
                Total time spent by all reduce tasks (ms)=5986
```

```
                Total time spent by all map tasks (ms)=20327
                Total time spent by all reduce tasks (ms)=5986
                Total vcore-milliseconds taken by all map tasks=20327
                Total vcore-milliseconds taken by all reduce tasks=5986
                Total megabyte-milliseconds taken by all map tasks=20814848
                Total megabyte-milliseconds taken by all reduce tasks=6129664
        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)=581
                CPU time spent (ms)=7020
                Physical memory (bytes) snapshot=896544768
                Virtual memory (bytes) snapshot=7764856832
                Total committed heap usage (bytes)=698875904
```

```
                Spilled Records=20440
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=581
                CPU time spent (ms)=7020
                Physical memory (bytes) snapshot=896544768
                Virtual memory (bytes) snapshot=7764856832
                Total committed heap usage (bytes)=698875904
                Peak Map Physical memory (bytes)=331964416
                Peak Map Virtual memory (bytes)=2587738112
                Peak Reduce Physical memory (bytes)=235270144
                Peak Reduce Virtual memory (bytes)=2591649792
        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=83664
        File Output Format Counters
                Bytes Written=96
2024-08-28 12:29:55,534 INFO streaming.StreamJob: Output directory: /exp3/outpu
t
```

Output :

```
arise@fedora:~/exp3$ hdfs dfs -cat /exp2/output/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
```