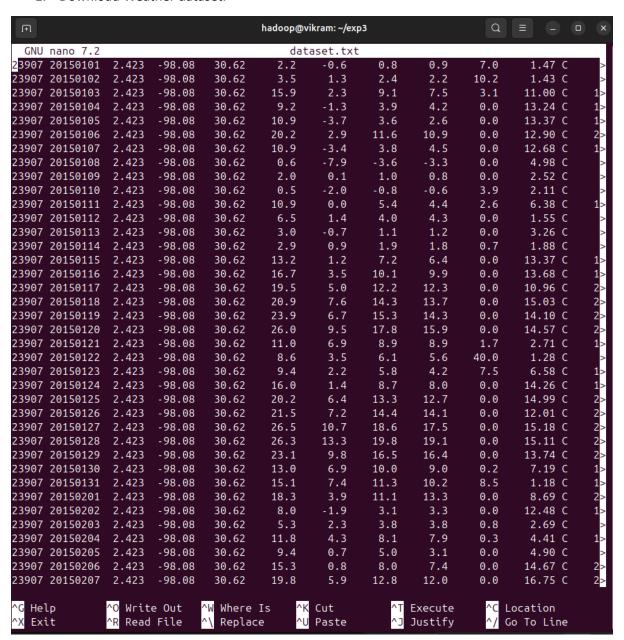
Exp. No: 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.



2. Create mapper.py program

```
Q ≡
                                              hadoop@vikram: ~/exp3
  GNU nano 7.2
                                                     mapper.py
   on t sys

nput comes from STDIN (standard input)

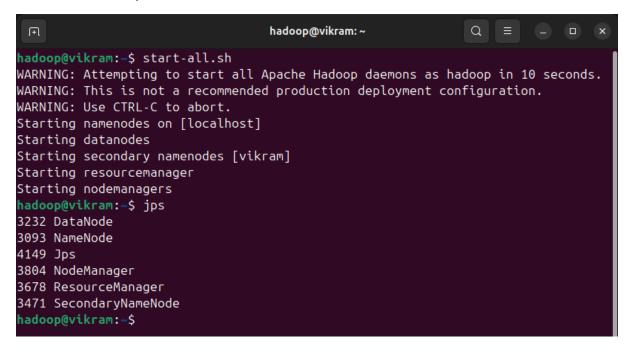
he mapper will get daily max temperature and group it by month. so output will be

onth,dailymax_temperature)
import sys
for line in sys.stdin:
         line = line.strip()
         words = line.split()
#See the README hosted on the weather website which help us understand how each pos
         month = line[10:12]
         daily_max = line[38:45]
         daily_max = daily_max.strip()
         for word in words:
                   print ('%s\t%s' % (month ,daily_max))
                 ^O Write Out
                                   ^W Where Is
                                                     ^K Cut
                                                                                          ^C Location
 G Help
                                                                          Execute
                  NR Read File
                                   ^\ Replace
                                                      'U Paste
                                                                           Justify
                                                                                            Go To Line
```

3. Create reducer.py

```
hadoop@vikram: ~/exp3
                                                                                                                                                                                                                                                                                    Q =
       GNU nano 7.2
                                                                                                                                                               reducer.py
 from operator import itemgetter
 import sys
                                daily max temperature)
logic: will get all the daily max temperature for a month and find max temperature for a month and 
current_month = None
 current_max = 0
month = None
for line in sys.stdin:
                             line = line.strip()
                             month, daily_max = line.split('\t', 1)
                                                         daily_max = float(daily_max)
                             except ValueError:
                            # 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
                                                          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))
                                                       ^O Write Out
                                                                                                             ^W Where Is
                                                                                                                                                                    ^K Cut
                                                                                                                                                                                                                            ^T Execute
                                                                                                                                                                                                                                                                                  ^C Location
          Exit
                                                       ^R Read File
                                                                                                               ^\ Replace
                                                                                                                                                                              Paste
                                                                                                                                                                                                                                     Justify
                                                                                                                                                                                                                                                                                            Go To Line
```

4. Start Hadoop services.



5. Upload Weather dataset into HDFS Storage.

```
hadoop@vikram:~/exp3$ hdfs dfs -ls /exp3
Found 2 items
-rw-r--r- 1 hadoop supergroup 79204 2024-09-13 09:55 /exp3/dataset.txt
drwxr-xr-x - hadoop supergroup 0 2024-09-13 09:56 /exp3/output
hadoop@vikram:~/exp3$
```

6. Run the Map reduce program using Hadoop Streaming.

```
hadoop@vikram: ~/exp3
hadoop@vikram:~/exp3$ hadoop jar $HADOOP STREAMING -input /exp3/dataset.txt -output /exp3/outp
ut -mapper ~/exp3/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar3248343435949788667/] [] /tmp/streamjob1848067600838242943.ja
r tmpDir=null
2024-10-13 17:25:34,591 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to Resource
Manager at /0.0.0.0:8032
2024-10-13 17:25:34,800 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to Resource
Manager at /0.0.0.0:8032
2024-10-13 17:25:35,109 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/hadoop/.staging/job_1728819809156_0002
2024-10-13 17:25:35,421 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-13 17:25:35,490 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-13 17:25:35,628 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728819809
156 0002
2024-10-13 17:25:35,628 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-13 17:25:35,819 INFO conf.Configuration: resource-types.xml not found
.
2024-10-13 17:25:35,820 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-13 17:25:35,892 INFO impl.YarnClientImpl: Submitted application application_1728819809
156 0002
2024-10-13 17:25:35,942 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy
/application_1728819809156_0002/
2024-10-13 17:25:35,945 INFO mapreduce.Job: Running job: job_1728819809156_0002
2024-10-13 17:25:43,086 INFO mapreduce.Job: Job job 1728819809156 0002 running in uber mode :
false
2024-10-13 17:25:43,088 INFO mapreduce.Job: map 0% reduce 0%
2024-10-13 17:25:48,211 INFO mapreduce.Job: map 50% reduce 0%
2024-10-13 17:25:49,257 INFO mapreduce.Job: map 100% reduce 0%
2024-10-13 17:25:53,301 INFO mapreduce.Job: map 100% reduce 100% 2024-10-13 17:25:54,322 INFO mapreduce.Job: Job job_1728819809156_0002 completed successfully
2024-10-13 17:25:54,443 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=102094
                FILE: Number of bytes written=1040452
                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 bytes read erasure-coded=0
```

```
hadoop@vikram: ~/exp3
                Total vcore-milliseconds taken by all map tasks=7437
                Total vcore-milliseconds taken by all reduce tasks=2821
                Total megabyte-milliseconds taken by all map tasks=7615488
                Total megabyte-milliseconds taken by all reduce tasks=2888704
        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)=227
                CPU time spent (ms)=2870
                Physical memory (bytes) snapshot=906362880
Virtual memory (bytes) snapshot=7637671936
                Total committed heap usage (bytes)=939524096
                Peak Map Physical memory (bytes)=325271552
                Peak Map Virtual memory (bytes)=2544807936
                Peak Reduce Physical memory (bytes)=256819200
                Peak Reduce Virtual memory (bytes)=2548711424
        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
2024-10-13 17:25:54,443 INFO streaming.StreamJob: Output directory: /exp3/output
hadoop@vikram:~/exp3$
```

Output:

```
hadoop@vikram: ~/exp3
hadoop@vikram:~/exp3$ hdfs dfs -cat /exp3/output/*
01
        26.5
        26.6
02
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
hadoop@vikram:~/exp3$
```