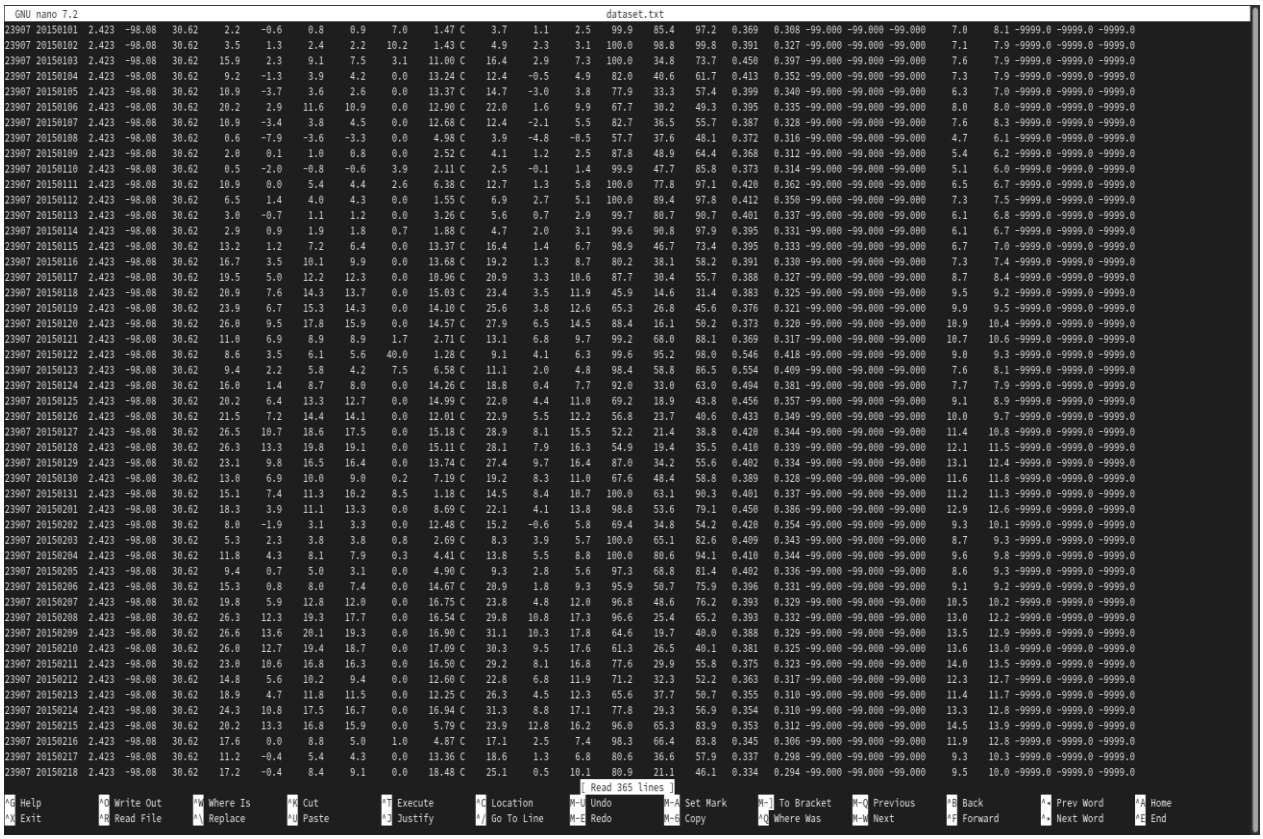
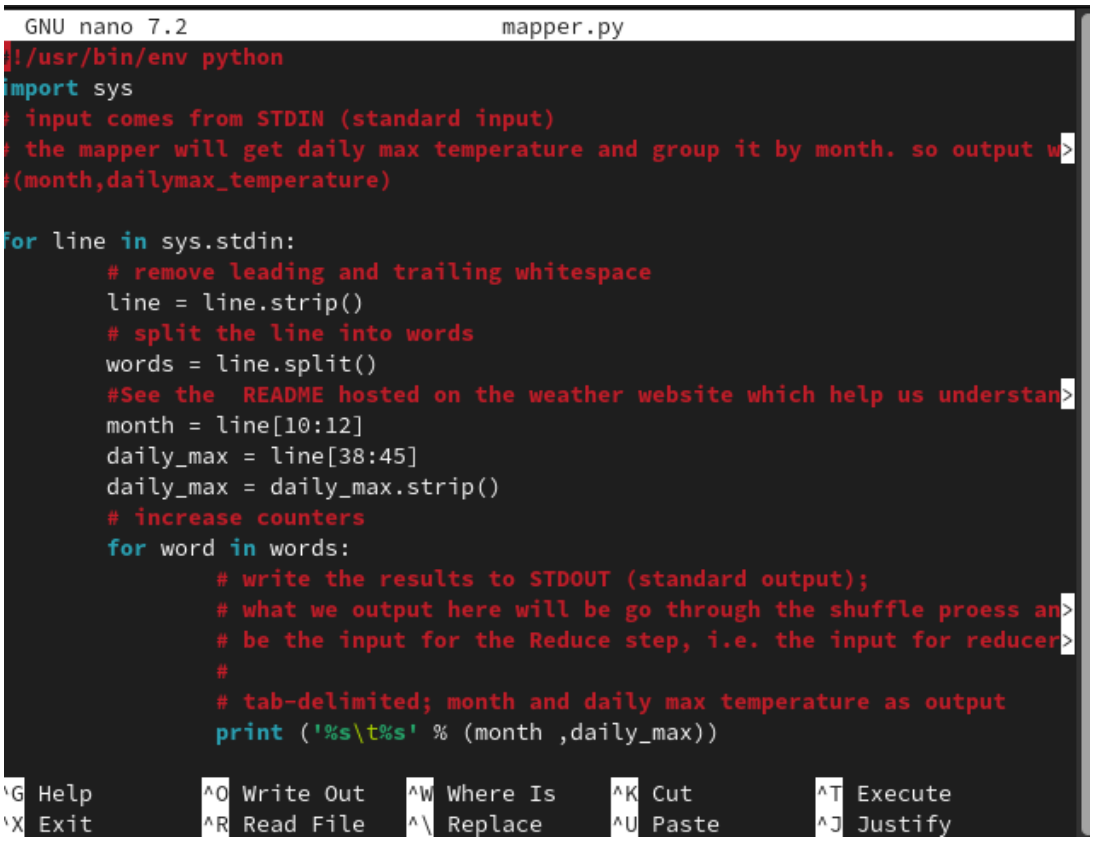
Exp. No : 3

Map Reduce program to process Weather dataset

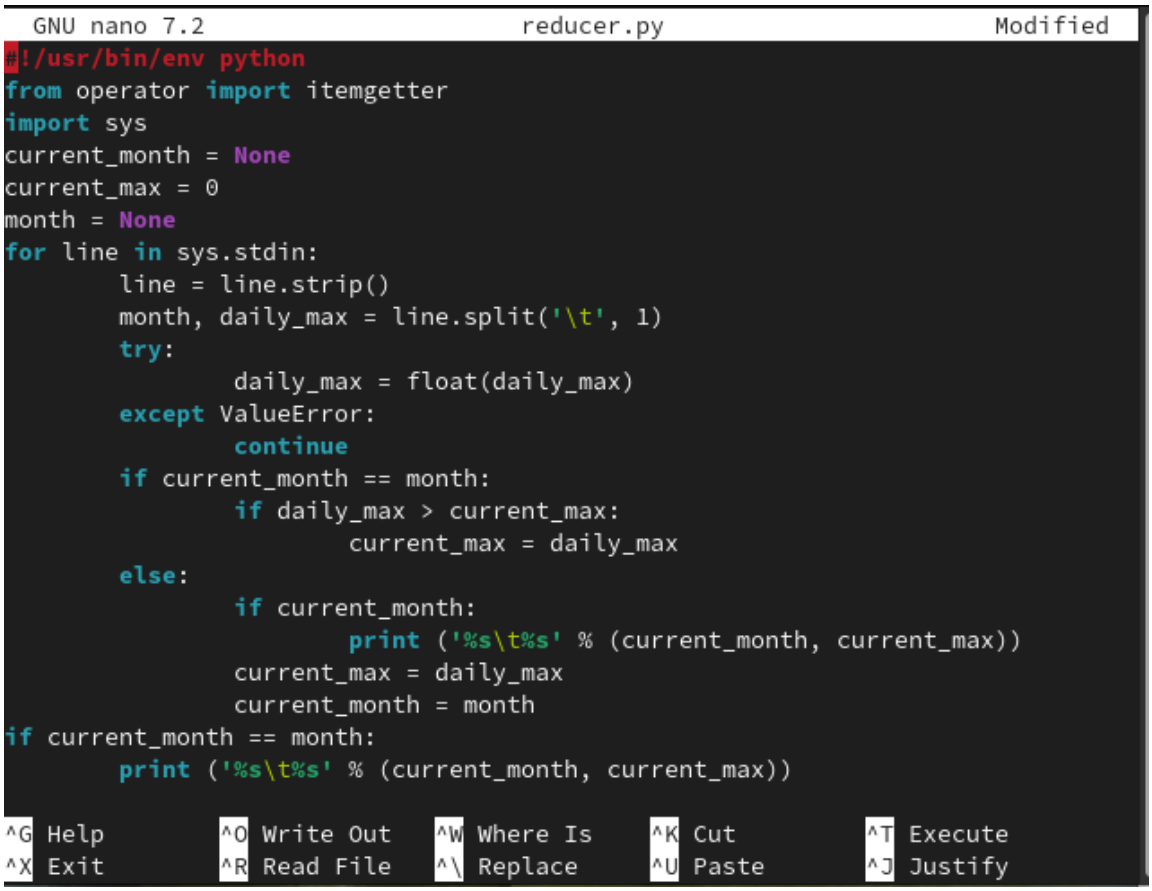
1. Download Weather dataset.



2. Create mapper.py program



3. Create reducer.py



4. Start Hadoop services.

```
lksh@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as lksh 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]
Starting resourcemanager
Starting nodemanagers
lksh@fedora:~$ jps
4643 Jps
3529 DataNode
3386 NameNode
3738 SecondaryNameNode
4108 NodeManager
3951 ResourceManager
lksh@fedora:~$
```

5. Upload Weather dataset into HDFS Storage.

```
lksh@fedora:~/exp3$ nano dataset.txt
lksh@fedora:~/exp3$ hdfs dfs -mkdir /exp2
lksh@fedora:~/exp3$ hdfs dfs -mkdir /exp3
```

6. Run the Map reduce program using Hadoop Streaming.

```
lksh@fedora:~/exp3$ hadoop jar $HADOOP_STREAMING -input /exp2/dataset.txt -output /exp2/output1 -mapper ~/exp3/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar2773513365584043905/] [] /tmp/streamjob3053124438108899539.jar tmpDir=null
2024-10-12 11:26:24,211 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 11:26:24,695 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-12 11:26:31,634 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/lksh/.staging/job_1728710244759_0001
2024-10-12 11:26:32,802 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/lksh/.staging/job_1728710244759_0001
2024-10-12 11:26:32,875 ERROR streaming.StreamJob: Error Launching job : Input path does not exist: hdfs://localhost:9000/exp2/dataset.txt
```

```
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 :

```
lksh@fedora:~$ hdfs dfs -cat /exp3/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
```