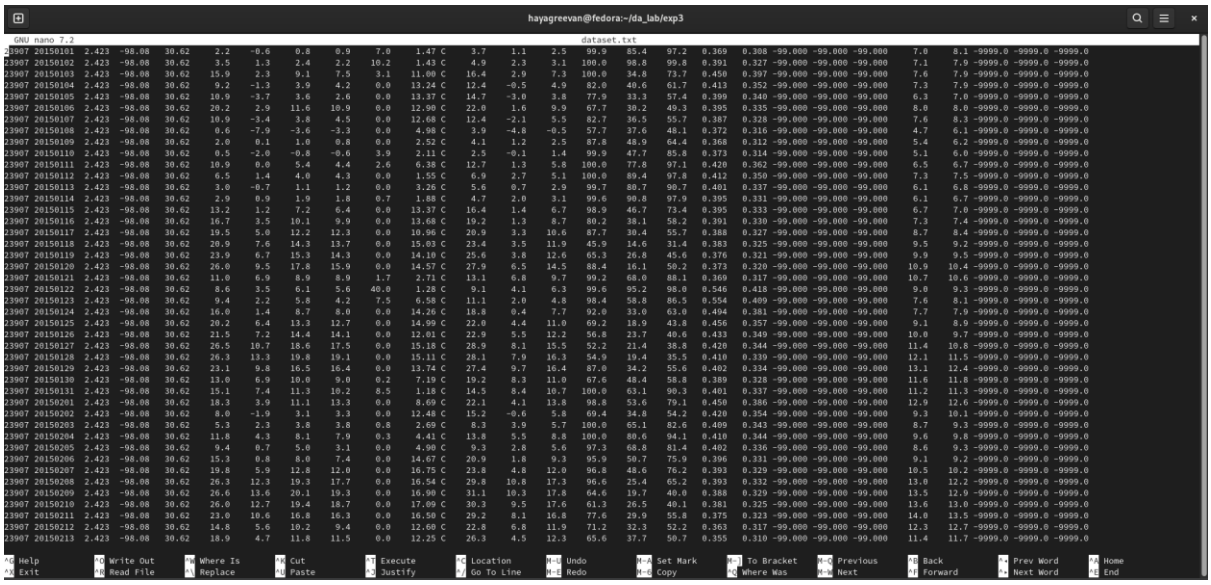
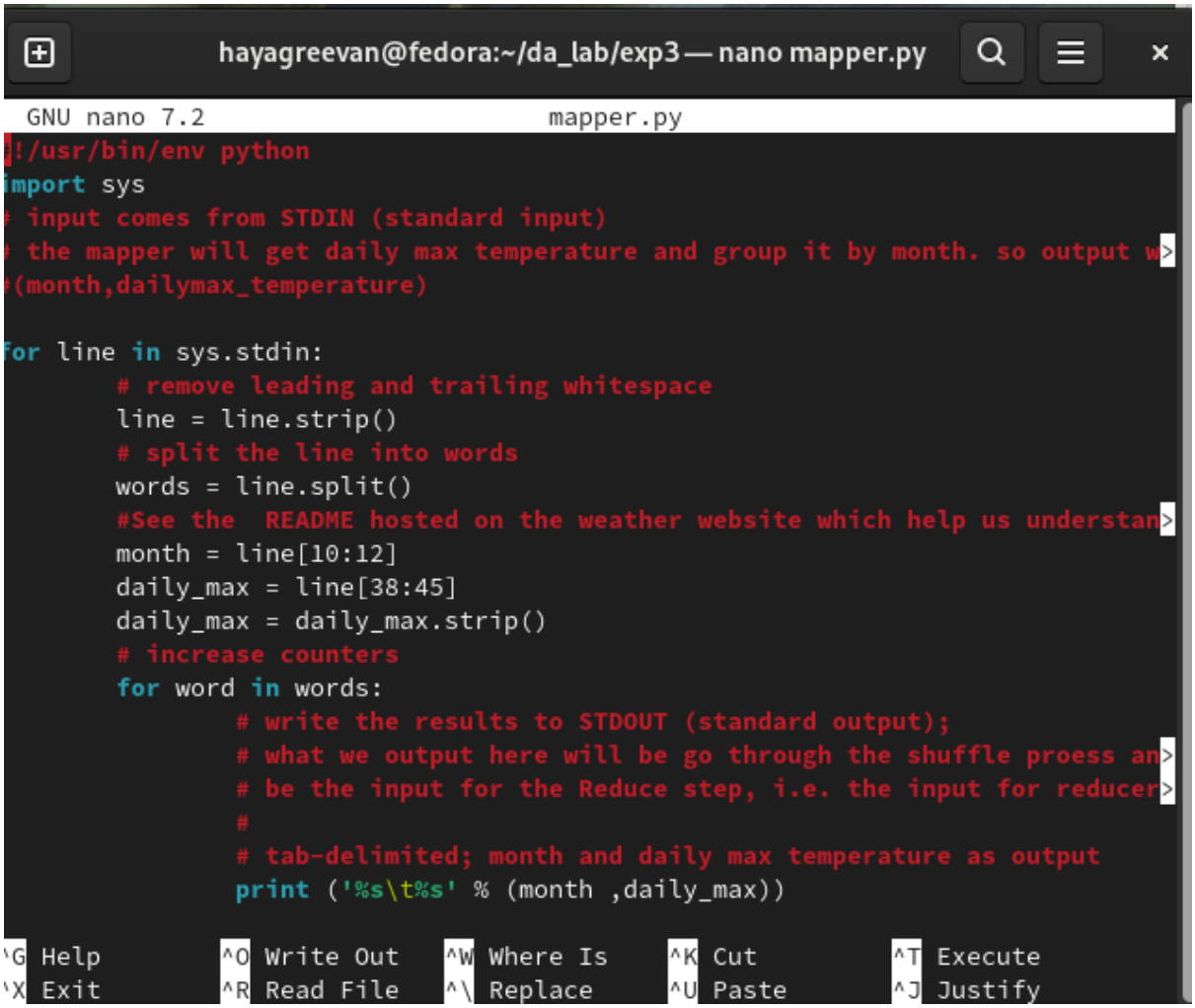
Exp. No : 3

Map Reduce program to process Weather dataset

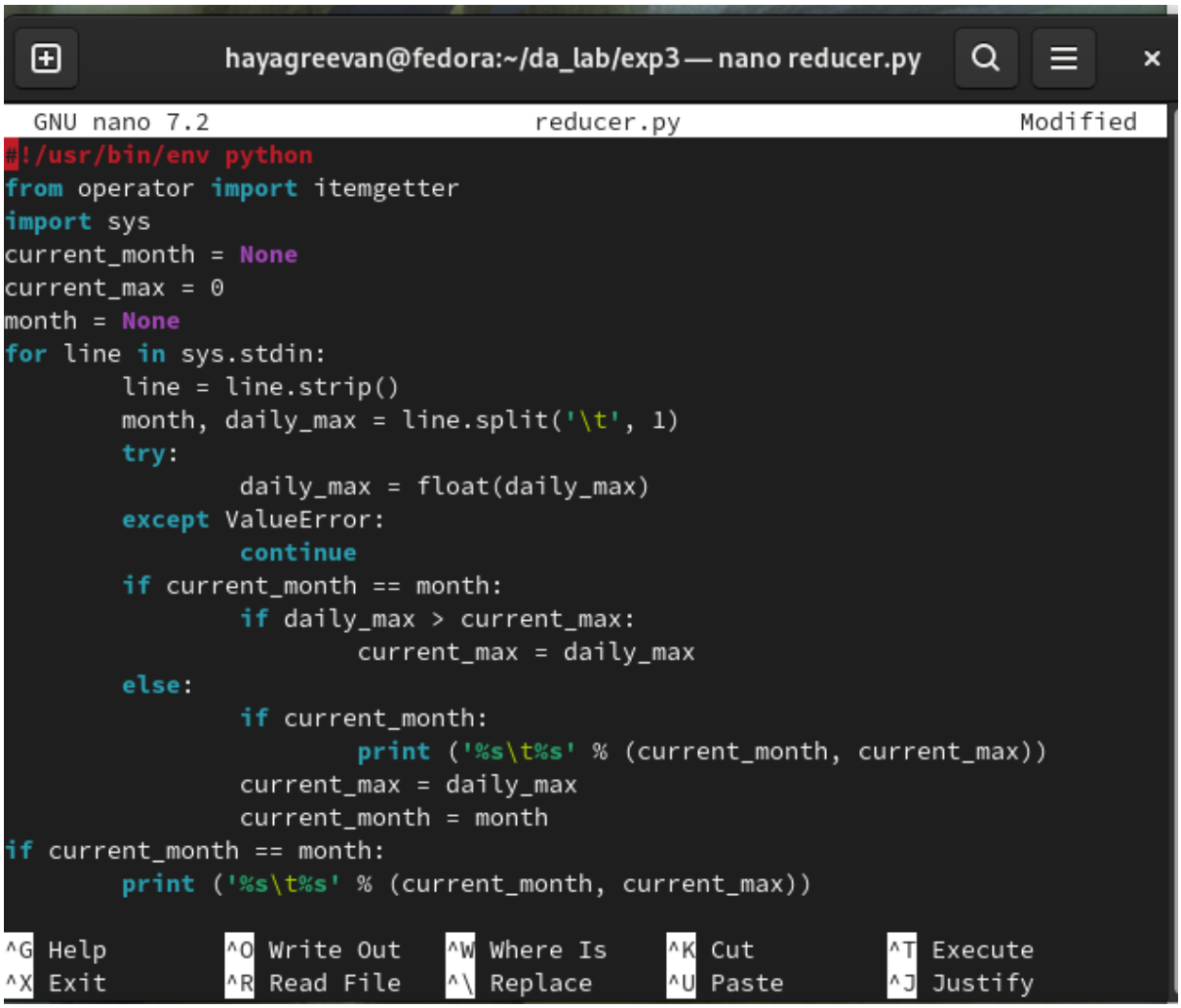
1. Download Weather dataset.



2. Create mapper.py program



3. Create reducer.py



4. Start Hadoop services.

```
hayagreevan@fedora:~/da_lab/exp3$ service sshd start
Redirecting to /bin/systemctl start sshd.service
hayagreevan@fedora:~/da_lab/exp3$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hayagreevan in 10 sec
onds.
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
hayagreevan@fedora:~/da_lab/exp3$ jps
4355 ResourceManager
4899 Jps
3814 DataNode
4058 SecondaryNameNode
4494 NodeManager
3631 NameNode
```

5. Upload Weather dataset into HDFS Storage.

```
hayagreevan@fedora:~/da_lab/exp3$ hdfs dfs -mkdir /exp3/
hayagreevan@fedora:~/da_lab/exp3$ hdfs dfs -copyFromLocal dataset.txt /exp3/
hayagreevan@fedora:~/da_lab/exp3$ hdfs dfs -ls /exp3/
Found 1 items
-rw-r--r--   1 hayagreevan supergroup      79568 2024-08-28 12:27 /exp3/dataset
.txt
hayagreevan@fedora:~/da_lab/exp3$
```

6. Run the Map reduce program using Hadoop Streaming.

```
hayagreevan@fedora:~/da_lab/exp3$ hadoop jar ~/da_lab/exp3/hadoop-streaming-3.3
.6.jar -input /exp3/dataset.txt -output /exp3/output -mapper ~/da_lab/exp3/mapp
er.py -reducer ~/da_lab/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar514626091708043449/] [] /tmp/streamjob81084674
87130491776.jar tmpDir=null
2024-08-28 12:29:06,156 INFO client.DefaultNoHARMFailoverProxyProvider: Connect
ing to ResourceManager at /0.0.0.0:8032
2024-08-28 12:29:06,878 INFO client.DefaultNoHARMFailoverProxyProvider: Connect
ing to ResourceManager at /0.0.0.0:8032
2024-08-28 12:29:08,553 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/hayagreevan/.staging/job_1724828139433
_0001
2024-08-28 12:29:09,732 INFO mapred.FileInputFormat: Total input files to proce
ss : 1
2024-08-28 12:29:10,083 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-28 12:29:11,063 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1724828139433_0001
2024-08-28 12:29:11,063 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-28 12:29:11,464 INFO conf.Configuration: resource-types.xml not found
2024-08-28 12:29:11,465 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2024-08-28 12:29:12,414 INFO impl.YarnClientImpl: Submitted application applica
tion_1724828139433_0001
2024-08-28 12:29:12,514 INFO mapreduce.Job: The url to track the job: http://fe
dora:8088/proxy/application_1724828139433_0001/
2024-08-28 12:29:12,517 INFO mapreduce.Job: Running job: job_1724828139433_0001
2024-08-28 12:29:30,610 INFO mapreduce.Job: Job job_1724828139433_0001 running
```

```
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
hayagreevan@fedora:~/da_lab/exp3$
```

Output :

```
hayagreevan@fedora:~/da_lab/exp3$ 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
hayagreevan@fedora:~/da_lab/exp3$
```