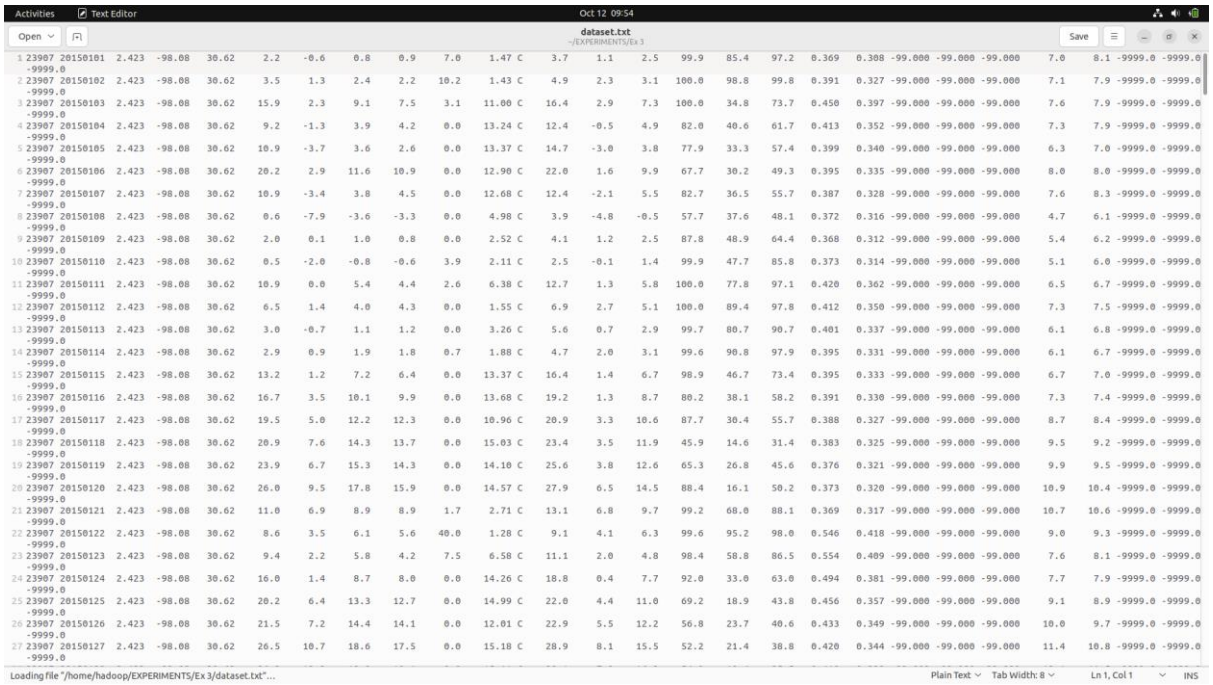
Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.



2. Create mapper.py program

```
GNU nano 6.2                                                    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 will be (month, daily_max_temperature)

# Download the dataset (weather data)
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 helps 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 go through the shuffle process 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))
```

3. Create reducer.py

```
Activities    Terminal

  GNU nano 6.2
#!/usr/bin/env python
from operator import itemgetter
import sys

current_month = None
current_max = float('-inf')
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))
```

4. Run the Map reduce program using Hadoop Streaming.

```
-input /weatherdata/dataset.txt \
-output /weatherdata/output \
-file ~/mapper.py \
-file ~/reducer.py \
-mapper "python3 mapper.py" \
-reducer "python3 reducer.py"
2024-09-14 12:39:56,313 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/mapper.py, /home/hadoop/reducer.py] [] /tmp/streamjob8261850635057036 59.jar tmpDir=null
2024-09-14 12:39:57,567 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-14 12:39:57,764 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-09-14 12:39:57,764 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-09-14 12:39:57,788 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-09-14 12:39:58,108 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-14 12:39:58,192 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-14 12:39:58,395 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local458174712_0001
2024-09-14 12:39:58,395 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-14 12:39:58,670 INFO mapred.LocalDistributedCacheManager: Localized file:/home/hadoop/mapper.py as file:/tmp/hadoop-hadoop/mapred/local/job_local458174712_0001_2bba3c2d-2b34-4464-be51-ffba723f3826/mapper
.py
2024-09-14 12:39:58,722 INFO mapred.LocalDistributedCacheManager: Localized file:/home/hadoop/reducer.py as file:/tmp/hadoop-hadoop/mapred/local/job_local458174712_0001_722ffb92-ba2a-447b-b6fa-6df1ab4dc169/reduc
er.py
2024-09-14 12:39:58,843 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-09-14 12:39:58,845 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-09-14 12:39:58,845 INFO mapreduce.Job: Running job: job_local458174712_0001
2024-09-14 12:39:58,848 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-09-14 12:39:58,852 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 12:39:58,853 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 12:39:58,912 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-09-14 12:39:58,916 INFO mapred.LocalJobRunner: Starting task: attempt_local458174712_0001_m_000000_0
2024-09-14 12:39:58,955 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 12:39:58,955 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 12:39:58,973 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-09-14 12:39:58,984 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/weatherdata/dataset.txt:0+79568
2024-09-14 12:39:59,011 INFO mapred.MapTask: numReduceTasks: 1
2024-09-14 12:39:59,045 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-09-14 12:39:59,045 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-09-14 12:39:59,045 INFO mapred.MapTask: soft limit at 83886080
2024-09-14 12:39:59,045 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-09-14 12:39:59,045 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-09-14 12:39:59,049 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-09-14 12:39:59,063 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3, mapper.py]
2024-09-14 12:39:59,067 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2024-09-14 12:39:59,067 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2024-09-14 12:39:59,068 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2024-09-14 12:39:59,068 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2024-09-14 12:39:59,068 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2024-09-14 12:39:59,068 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2024-09-14 12:39:59,069 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2024-09-14 12:39:59,071 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2024-09-14 12:39:59,071 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2024-09-14 12:39:59,207 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2024-09-14 12:39:59,207 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2024-09-14 12:39:59,209 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=96
2024-09-14 12:40:00,302 INFO mapred.LocalJobRunner: Finishing task: attempt_local458174712_0001_r_000000_0
2024-09-14 12:40:00,302 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-14 12:40:00,913 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-14 12:40:00,914 INFO mapreduce.Job: Job job_local458174712_0001 completed successfully
2024-09-14 12:40:00,927 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=209834
                FILE: Number of bytes written=1603444
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=159136
                HDFS: Number of bytes written=96
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=365
                Map output records=10220
                Map output bytes=81648
                Map output materialized bytes=102094
                Input split bytes=97
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=102094
                Reduce input records=10220
                Reduce output records=12
                Spilled Records=20440
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=13
                Total committed heap usage (bytes)=547356672
        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=79568
        File Output Format Counters
                Bytes Written=96
2024-09-14 12:40:00,927 INFO streaming.StreamJob: Output directory: /weatherdata/output
```

Output :

Activities Text Editor

Open

```
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
```