

Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

The screenshot shows a text editor window titled 'dataset.txt' with a file size of 1.1 MB. The editor displays a large table of weather data. The first few rows of the table are as follows:

Date	Time	Location	Temp (C)	Humidity (%)	Wind Speed (km/h)	Cloudiness (%)	Pressure (hPa)	Visibility (km)	UV Index	Condition
23/09/2015	10:01	2, 423	-98.08	30.62	2.2	-0.6	0.8	0.9	7.0	1.47 C
23/09/2015	10:02	2, 423	-98.08	30.62	3.5	1.3	2.4	2.2	10.2	1.43 C
23/09/2015	10:03	2, 423	-98.08	30.62	15.9	2.3	9.1	7.5	3.1	11.00 C

The table continues with many more rows, each representing a different time and location. The data is formatted as a CSV file with commas separating the columns.

2. Create mapper.py program

```

GNU nano 6.2
#!/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

```

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. Start Hadoop services.

```

hadoop@kiran:~$ nano mapper.py
hadoop@kiran:~$ nano reducer.py
hadoop@kiran:~$ jps
5009 SecondaryNameNode
5410 NodeManager
4723 NameNode
4835 DataNode
5293 ResourceManager
26303 Jps
hadoop@kiran:~$ ch
chacl          chardetect          chcon          cheese          chgrp          chown          chroot
chage          chat          chcpu          chfn            chmem          chown          chrt
chardet        chattr         check-language-support  chgpsswd       chnod          chpasswd       chsh
hadoop@kiran:~$ chmod 777 mapper.py reducer.py
hadoop@kiran:~$ hadoop fs -copyFromLocal ~/Downloads/dataset.txt /weatherdata
copyFromLocal: '/weatherdata/dataset.txt': File exists
hadoop@kiran:~$ cd weatherdata
bash: cd: /weatherdata: No such file or directory
hadoop@kiran:~$ hadoop fs -rm -r /weatherdata/output
rm: '/weatherdata/output': No such file or directory
hadoop@kiran:~$ hadoop fs -mkdir -p /weatherdata
hadoop@kiran:~$ hadoop fs -copyFromLocal ~/Downloads/dataset.txt /weatherdata
copyFromLocal: '/weatherdata/dataset.txt': File exists
hadoop@kiran:~$ ^C
hadoop@kiran:~$ hadoop fs -ls /weatherdata/
Found 1 items
-rw-r--r-- 1 hadoop supergroup 79568 2024-09-14 12:07 /weatherdata/dataset.txt
hadoop@kiran:~$ ^C

```

5. Upload Weather dataset into HDFS Storage.

```

hadoop@kiran:~$ nano mapper.py
hadoop@kiran:~$ nano reducer.py
hadoop@kiran:~$ hdfs dfs -text /weatherdata/output/* > /home/Downloads/output/
/part-000000
bash: /home/Downloads/output/: No such file or directory
bash: /part-000000: No such file or directory
hadoop@kiran:~$ hdfs dfs -cat /weatherdata/output/part-000000 > /home/hadoop/Downloads/output.txt
hadoop@kiran:~$

```

6. Run the Map reduce program using Hadoop Streaming.

```

hadoop@kiran:~$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
-input /weatherdata/dataset.txt \
-output /weatherdata/output \
-file /mapper.py \
-reducer /reducer.py \
-napper "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/streamjob26185863505703859.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 12:39:58,180 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-14 12:39:59,192 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-ffb723f3826/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-0df1ab4dc169/reducer.py
2024-09-14 12:39:58,843 INFO mapreduce.Job: The url to track the job: http://localhost:8888/
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 12:39:59,072 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2024-09-14 12:39:59,884 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; bufvold = 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 12:39:59,207 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]

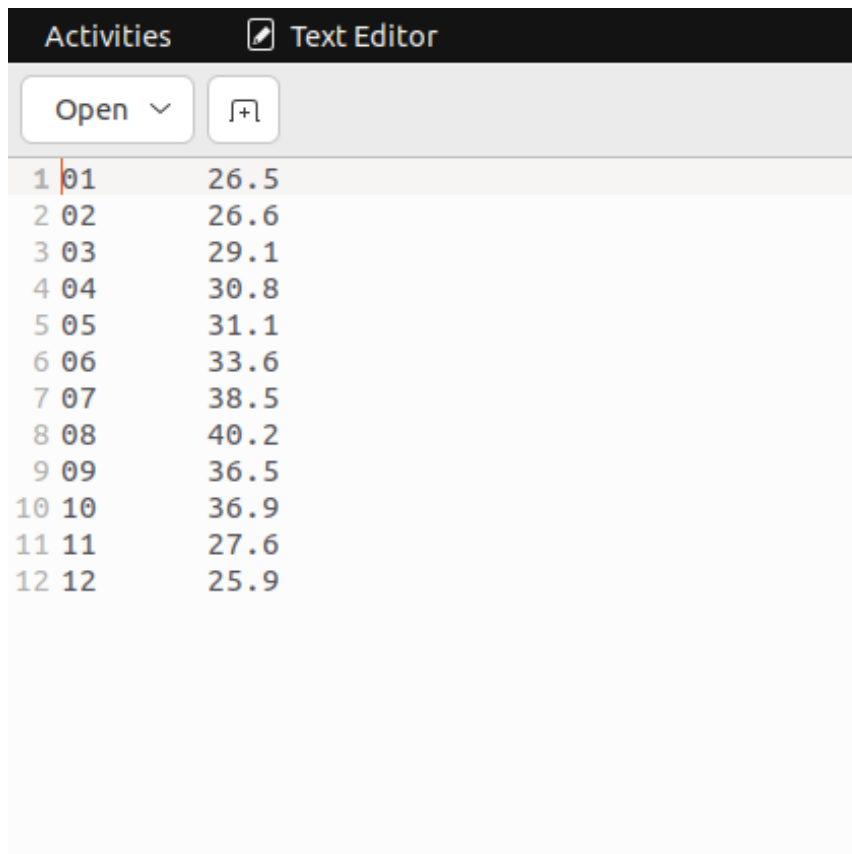
```




```

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=81048
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=2040
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  		
1	01	26.5
2	02	26.6
3	03	29.1
4	04	30.8
5	05	31.1
6	06	33.6
7	07	38.5
8	08	40.2
9	09	36.5
10	10	36.9
11	11	27.6
12	12	25.9