Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

hadoop@vikram: ~/exp3

```
  GNU nano 7.2                                dataset.txt
23907 20150101  2.423  -98.08   30.62     2.2    -0.6     0.8     0.9     7.0     1.47 C     >
23907 20150102  2.423  -98.08   30.62     3.5     1.3     2.4     2.2    10.2     1.43 C     >
23907 20150103  2.423  -98.08   30.62    15.9     2.3     9.1     7.5     3.1    11.00 C    1>
23907 20150104  2.423  -98.08   30.62     9.2    -1.3     3.9     4.2     0.0    13.24 C    1>
23907 20150105  2.423  -98.08   30.62    10.9    -3.7     3.6     2.6     0.0    13.37 C    1>
23907 20150106  2.423  -98.08   30.62    20.2     2.9    11.6    10.9     0.0    12.90 C    2>
23907 20150107  2.423  -98.08   30.62    10.9    -3.4     3.8     4.5     0.0    12.68 C    1>
23907 20150108  2.423  -98.08   30.62     0.6    -7.9    -3.6    -3.3     0.0     4.98 C     >
23907 20150109  2.423  -98.08   30.62     2.0     0.1     1.0     0.8     0.0     2.52 C     >
23907 20150110  2.423  -98.08   30.62     0.5    -2.0    -0.8    -0.6     3.9     2.11 C     >
23907 20150111  2.423  -98.08   30.62    10.9     0.0     5.4     4.4     2.6     6.38 C    1>
23907 20150112  2.423  -98.08   30.62     6.5     1.4     4.0     4.3     0.0     1.55 C     >
23907 20150113  2.423  -98.08   30.62     3.0    -0.7     1.1     1.2     0.0     3.26 C     >
23907 20150114  2.423  -98.08   30.62     2.9     0.9     1.9     1.8     0.7     1.88 C     >
23907 20150115  2.423  -98.08   30.62    13.2     1.2     7.2     6.4     0.0    13.37 C    1>
23907 20150116  2.423  -98.08   30.62    16.7     3.5    10.1     9.9     0.0    13.68 C    1>
23907 20150117  2.423  -98.08   30.62    19.5     5.0    12.2    12.3     0.0    10.96 C    2>
23907 20150118  2.423  -98.08   30.62    20.9     7.6    14.3    13.7     0.0    15.03 C    2>
23907 20150119  2.423  -98.08   30.62    23.9     6.7    15.3    14.3     0.0    14.10 C    2>
23907 20150120  2.423  -98.08   30.62    26.0     9.5    17.8    15.9     0.0    14.57 C    2>
23907 20150121  2.423  -98.08   30.62    11.0     6.9     8.9     8.9     1.7     2.71 C    1>
23907 20150122  2.423  -98.08   30.62     8.6     3.5     6.1     5.6    40.0     1.28 C     >
23907 20150123  2.423  -98.08   30.62     9.4     2.2     5.8     4.2     7.5     6.58 C    1>
23907 20150124  2.423  -98.08   30.62    16.0     1.4     8.7     8.0     0.0    14.26 C    1>
23907 20150125  2.423  -98.08   30.62    20.2     6.4    13.3    12.7     0.0    14.99 C    2>
23907 20150126  2.423  -98.08   30.62    21.5     7.2    14.4    14.1     0.0    12.01 C    2>
23907 20150127  2.423  -98.08   30.62    26.5    10.7    18.6    17.5     0.0    15.18 C    2>
23907 20150128  2.423  -98.08   30.62    26.3    13.3    19.8    19.1     0.0    15.11 C    2>
23907 20150129  2.423  -98.08   30.62    23.1     9.8    16.5    16.4     0.0    13.74 C    2>
23907 20150130  2.423  -98.08   30.62    13.0     6.9    10.0     9.0     0.2     7.19 C    1>
23907 20150131  2.423  -98.08   30.62    15.1     7.4    11.3    10.2     8.5     1.18 C    1>
23907 20150201  2.423  -98.08   30.62    18.3     3.9    11.1    13.3     0.0     8.69 C    2>
23907 20150202  2.423  -98.08   30.62     8.0    -1.9     3.1     3.3     0.0    12.48 C    1>
23907 20150203  2.423  -98.08   30.62     5.3     2.3     3.8     3.8     0.8     2.69 C     >
23907 20150204  2.423  -98.08   30.62    11.8     4.3     8.1     7.9     0.3     4.41 C    1>
23907 20150205  2.423  -98.08   30.62     9.4     0.7     5.0     3.1     0.0     4.90 C     >
23907 20150206  2.423  -98.08   30.62    15.3     0.8     8.0     7.4     0.0    14.67 C    2>
23907 20150207  2.423  -98.08   30.62    19.8     5.9    12.8    12.0     0.0    16.75 C    2>

^G Help       ^O Write Out  ^W Where Is   ^K Cut       ^T Execute    ^C Location
^X Exit       ^R Read File  ^\ Replace    ^U Paste     ^J Justify    ^/ Go To Line
```

2. Create mapper.py program

```
hadoop@vikram: ~/exp3
  GNU nano 7.2                                   mapper.py
#!/usr/bin/env python3
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,dailymax_temperature)

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 help us understand how each posi>
        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 be go through the shuffle proess 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))

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line
```

3. Create reducer.py

```
hadoop@vikram: ~/exp3
  GNU nano 7.2                          reducer.py
#!/usr/bin/env python3

from operator import itemgetter
import sys
#reducer will get the input from stdid which will be a collection of key, value(Key=month ,
#value= daily max temperature)
#reducer logic: will get all the daily max temperature for a month and find max temperature f>
#shuffle will ensure that key are sorted(month)
current_month = None
current_max = 0
month = None
# input comes from STDIN
for line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()
        # parse the input we got from mapper.py
        month, daily_max = line.split('\t', 1)
# convert daily_max (currently a string) to float
        try:
                daily_max = float(daily_max)
        except ValueError:
                # daily_max was not a number, so silently
                # ignore/discard this line
                continue
        # 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
        else:
                if current_month:
                        # write result to STDOUT
                        print ('%s\t%s' % (current_month, current_max))
                current_max = daily_max
                current_month = month
# output of the last month
if current_month == month:
        print ('%s\t%s' % (current_month, current_max))

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line
```

4. Start Hadoop services.

```
hadoop@vikram: ~
hadoop@vikram:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop 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 [vikram]
Starting resourcemanager
Starting nodemanagers
hadoop@vikram:~$ jps
3232 DataNode
3093 NameNode
4149 Jps
3804 NodeManager
3678 ResourceManager
3471 SecondaryNameNode
hadoop@vikram:~$
```

5. Upload Weather dataset into HDFS Storage.

```
hadoop@vikram: ~/exp3
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
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
hadoop@vikram:~/exp3$
```