Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

```
helen@fedora:~/exp2
  GNU nano 7.2                          data.txt
23907 20150101  2.423  -98.08   30.62     2.2    -0.6     0.8     0.9     7.0     1.>
23907 20150102  2.423  -98.08   30.62     3.5     1.3     2.4     2.2    10.2     1.>
23907 20150103  2.423  -98.08   30.62    15.9     2.3     9.1     7.5     3.1    11.>
23907 20150104  2.423  -98.08   30.62     9.2    -1.3     3.9     4.2     0.0    13.>
23907 20150105  2.423  -98.08   30.62    10.9    -3.7     3.6     2.6     0.0    13.>
23907 20150106  2.423  -98.08   30.62    20.2     2.9    11.6    10.9     0.0    12.>
23907 20150107  2.423  -98.08   30.62    10.9    -3.4     3.8     4.5     0.0    12.>
23907 20150108  2.423  -98.08   30.62     0.6    -7.9    -3.6    -3.3     0.0     4.>
23907 20150109  2.423  -98.08   30.62     2.0     0.1     1.0     0.8     0.0     2.>
23907 20150110  2.423  -98.08   30.62     0.5    -2.0    -0.8    -0.6     3.9     2.>
23907 20150111  2.423  -98.08   30.62    10.9     0.0     5.4     4.4     2.6     6.>
23907 20150112  2.423  -98.08   30.62     6.5     1.4     4.0     4.3     0.0     1.>
23907 20150113  2.423  -98.08   30.62     3.0    -0.7     1.1     1.2     0.0     3.>
23907 20150114  2.423  -98.08   30.62     2.9     0.9     1.9     1.8     0.7     1.>
23907 20150115  2.423  -98.08   30.62    13.2     1.2     7.2     6.4     0.0    13.>
23907 20150116  2.423  -98.08   30.62    16.7     3.5    10.1     9.9     0.0    13.>
23907 20150117  2.423  -98.08   30.62    19.5     5.0    12.2    12.3     0.0    10.>
23907 20150118  2.423  -98.08   30.62    20.9     7.6    14.3    13.7     0.0    15.>
23907 20150119  2.423  -98.08   30.62    23.9     6.7    15.3    14.3     0.0    14.>
23907 20150120  2.423  -98.08   30.62    26.0     9.5    17.8    15.9     0.0    14.>
                                [ Read 365 lines ]
^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

```
  GNU nano 7.2                          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 w>
#(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 understan>
        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 an>
                # be the input for the Reduce step, i.e. the input for reducer>
                #
                # 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
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

3. Create reducer.py

```
  GNU nano 7.2                    reducer.py                    Modified
#!/usr/bin/env python
from operator import itemgetter
import sys
current_month = None
current_max = 0
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))

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

4. Start Hadoop services.

```
helen@fedora:~
helen@fedora:~/exp2        helen@fedora:~
helen@fedora:~/exp2$ cd
helen@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as helen in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 3449.  Stop it first and ensure /tmp/hadoop
-helen-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 3638.  Stop it first and ensure /tmp/hadoop
-helen-datanode.pid file is empty before retry.
Starting secondary namenodes [fedora]
fedora: secondarynamenode is running as process 3924.  Stop it first and ensure /tmp/
hadoop-helen-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 4213.  Stop it first and ensure /tmp/hadoop-hel
en-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 4348.  Stop it first and ensure /tmp/had
oop-helen-nodemanager.pid file is empty before retry.
helen@fedora:~$ jps
3924 SecondaryNameNode
4213 ResourceManager
3638 DataNode
3449 NameNode
4348 NodeManager
8061 Jps
helen@fedora:~$
```

5. Run the Map reduce program using Hadoop Streaming.

```
helen@fedora:~
helen@fedora:~/exp2        helen@fedora:~
helen@fedora:~/exp2$ cd
helen@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp2/data.txt -output /exp2/outp
ut3 -mapper ~/exp2/mapper.py -reducer ~/exp2/reducer.py
packageJobJar: [/tmp/hadoop-unjar6285464154385550498/] [] /tmp/streamjob1479607642965
193751.jar tmpDir=null
2024-10-12 07:29:18,538 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
 ResourceManager at /0.0.0.0:8032
2024-10-12 07:29:18,750 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
 ResourceManager at /0.0.0.0:8032
2024-10-12 07:29:19,201 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding
for path: /tmp/hadoop-yarn/staging/helen/.staging/job_1728729630364_0002
2024-10-12 07:29:19,717 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-12 07:29:19,994 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-12 07:29:20,746 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1
728729630364_0002
2024-10-12 07:29:20,746 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-12 07:29:21,006 INFO conf.Configuration: resource-types.xml not found
2024-10-12 07:29:21,006 INFO resource.ResourceUtils: Unable to find 'resource-types.x
ml'.
2024-10-12 07:29:21,187 INFO impl.YarnClientImpl: Submitted application application_1
728729630364_0002
2024-10-12 07:29:21,216 INFO mapreduce.Job: The url to track the job: http://fedora:8
088/proxy/application_1728729630364_0002/
2024-10-12 07:29:21,218 INFO mapreduce.Job: Running job: job_1728729630364_0002
2024-10-12 07:29:29,626 INFO mapreduce.Job: Job job_1728729630364_0002 running in ube
r mode : false
2024-10-12 07:29:29,627 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-12 07:29:39,796 INFO mapreduce.Job:  map 100% reduce 0%
```

helen@fedora:~

helen@fedora:~/exp2 | helen@fedora:~

```
2024-10-12 07:29:48,233 INFO mapreduce.Job: Job job_1728729630364_0002 completed successfully
2024-10-12 07:29:48,337 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=3652
                FILE: Number of bytes written=842563
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83475
                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
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=15870
                Total time spent by all reduces in occupied slots (ms)=4177
                Total time spent by all map tasks (ms)=15870
                Total time spent by all reduce tasks (ms)=4177
                Total vcore-milliseconds taken by all map tasks=15870
                Total vcore-milliseconds taken by all reduce tasks=4177
                Total megabyte-milliseconds taken by all map tasks=16250880
                Total megabyte-milliseconds taken by all reduce tasks=4277248
        Map-Reduce Framework
```

helen@fedora:~

helen@fedora:~/exp2 | helen@fedora:~

```
                Reduce input records=365
                Reduce output records=12
                Spilled Records=730
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=285
                CPU time spent (ms)=5150
                Physical memory (bytes) snapshot=778772480
                Virtual memory (bytes) snapshot=9581989888
                Total committed heap usage (bytes)=488636416
                Peak Map Physical memory (bytes)=320217088
                Peak Map Virtual memory (bytes)=3203178496
                Peak Reduce Physical memory (bytes)=180899840
                Peak Reduce Virtual memory (bytes)=3193049088
        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=83301
        File Output Format Counters
                Bytes Written=96
2024-10-12 07:29:48,343 INFO streaming.StreamJob: Output directory: /exp2/output3
helen@fedora:~$
```

Output :

```
helen@fedora:~$ hdfs dfs -cat /exp2/output3/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
helen@fedora:~$
```