Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

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.

```
harithaah@fedora:~/hadoop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as harithaah 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 [fedora]
jpsStarting resourcemanager
Starting nodemanagers
harithaah@fedora:~/hadoop$ jps
9458 SecondaryNameNode
9779 ResourceManager
9253 DataNode
9909 NodeManager
9095 NameNode
10045 Jps
harithaah@fedora:~/hadoop$
```

5. Upload Weather dataset into HDFS Storage.

```
harithaah@fedora:~$ cd exp3
harithaah@fedora:~/exp3$ hdfs dfs -mkdir /exp2
harithaah@fedora:~/exp3$ hdfs dfs -put dataset.txt /exp2
```

6. Run the Map reduce program using Hadoop Streaming.

```
harithaah@fedora:~/exp3$ hadoop jar $HADOOP_STREAMING -input /exp2/dataset.txt -output /exp2/output -mapper ~/exp3/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar2895422846984926388/] [] /tmp/streamjob4796055756292392460.jar tmpDir=null
2024-10-10 20:46:38,635 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:46:40,718 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:46:42,732 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/harithaah/.staging/job_1728572703273_0002
2024-10-10 20:46:44,245 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-10 20:46:44,826 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-10 20:46:45,912 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728572703273_0002
2024-10-10 20:46:45,913 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-10 20:46:46,807 INFO conf.Configuration: resource-types.xml not found
2024-10-10 20:46:46,808 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-10 20:46:47,146 INFO impl.YarnClientImpl: Submitted application application_1728572703273_0002
2024-10-10 20:46:47,411 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1728572703273_0002/
2024-10-10 20:46:47,423 INFO mapreduce.Job: Running job: job_1728572703273_0002
```

```
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
```

Output :

```
harithaah@fedora:~/exp3$ hdfs dfs -cat /exp2/output/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
harithaah@fedora:~/exp3$
harithaah@fedora:~/exp3$
```