EXP NO: 3   MAP REDUCE PROGRAM TO PROCESS A WEATHER DATASET

$nano dataset.txt

$nano mapper.py


```python
#!/usr/bin/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,dailymax_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 help 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 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))
```

$nano reducer.py

```
GNU nano 7.2                                                          reducer.py
#!/usr/bin/python

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
#for the month
#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))
```

```
  ┌──(hadoop㉿kali)-[~/hadoop/bin]
  └─$ ./hdfs dfs -ls /exp3
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-21 00:11:13,818 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Found 3 items
-rw-r--r--   1 hadoop supergroup      79205 2024-08-29 10:50 /exp3/dataset.txt
drwxr-xr-x   - hadoop supergroup          0 2024-08-29 10:52 /exp3/new_output
drwxr-xr-x   - hadoop supergroup          0 2024-09-13 01:00 /exp3/output
```

$chmod 777 mapper.py reducer.py

$hadoop jar $HADOOP_STREAMING -input /dataset.txt -output /output -mapper

mapper.py -reducer reducer.py

```
hadoop@fedora:~/hadoop/experiments$ hadoop jar ../share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar -input /dataset.txt -output /out -mapper $HOME/hadoop/experiments/mapper.py -reducer $HOME/hadoop/experiments/reducer.py
packageJobJar: [/tmp/hadoop-unjar18004748121278480987/] [] /tmp/streamjob17676603820645521190.jar tmpDir=null
2024-08-30 00:52:09,047 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-30 00:52:09,240 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-30 00:52:09,500 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1724955478554_0006
2024-08-30 00:52:09,809 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-30 00:52:10,710 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-30 00:52:11,295 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724955478554_0006
2024-08-30 00:52:11,295 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-30 00:52:11,476 INFO conf.Configuration: resource-types.xml not found
2024-08-30 00:52:11,476 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-30 00:52:11,552 INFO impl.YarnClientImpl: Submitted application application_1724955478554_0006
2024-08-30 00:52:11,583 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1724955478554_0006/
2024-08-30 00:52:11,585 INFO mapreduce.Job: Running job: job_1724955478554_0006
2024-08-30 00:52:17,711 INFO mapreduce.Job: Job job_1724955478554_0006 running in uber mode : false
2024-08-30 00:52:17,712 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-30 00:52:23,827 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-30 00:52:29,891 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-30 00:52:29,898 INFO mapreduce.Job: Job job_1724955478554_0006 completed successfully
2024-08-30 00:52:30,033 INFO mapreduce.Job: Counters: 54
```

$hdfs dfs -cat /out/*

```
hadoop@fedora:~/hadoop/experiments$ hdfs dfs -cat /out/*
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
```