ASSIGNMENT 5

PROBLEM STATEMENT: Design and develop a distributed application to find the coolest/hottest year from the available weather data. Use weather data from the Internet and process it using MapReduce.

CODE

mapper.py

```
#!/usr/bin/env python3
import sys

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue

    parts = line.split()
    if len(parts) == 3:
        date_str, min_temp, max_temp = parts
        if "-" in date_str:
            year = date_str.split("-")[0]
            print(f"{year} {min_temp} {max_temp}")
```

reducer.py

```
#!/usr/bin/env python3
import sys
from collections import defaultdict

temps_by_year = defaultdict(list)

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue
    parts = line.split()
    if len(parts) != 3:
        continue
    year, min_temp, max_temp = parts
    try:

temps_by_year[year].append((int(min_temp), int(max_temp)))
    except:
        continue

for year in sorted(temps_by_year.keys()):
    mins, maxs = zip(*temps_by_year[year])
    print(f"{year}\t{min(mins)}\t{max(maxs)}")
```

OUTPUT

```
hduser24@HDUSER-24865:~/hadoop_practicals$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [HDUSER-24865]
hduser24@HDUSER-24865:~/hadoop_practicals$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hduser24@HDUSER-24865:~/hadoop_practicals$ jp
Command 'jp' not found, but can be installed with:
sudo apt install jp
hduser24@HDUSER-24865:~/hadoop_practicals$ jps
13952 NodeManager
13379 DataNode
14375 Jps
13255 NameNode
13832 ResourceManager
13615 SecondaryNameNode
hduser24@HDUSER-24865:~/hadoop_practicals$ cd input/
hduser24@HDUSER-24865:~/hadoop_practicals/input$ head weather_data.txt
1958-09-24 6 14
1951-10-09 6 31
1956-10-15 8 26
1955-12-04 19 31
1959-01-06 -11 -4
1957-02-25 15 21
1957-11-30 7 29
1956-09-20 -14 7
1952-10-17 -8 11
1958-12-06 -16 -8
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -put weather_data.txt /input/
hduser24@HDUSER-24865:~/hadoop_practicals/input$ hdfs dfs -ls /input/
Found 3 items
```

-rw-r--r--   1 hduser24 supergroup   15 2025-04-19 07:50 /input/character_count.txt

-rw-r--r--   1 hduser24 supergroup   1684 2025-04-19 08:53 /input/weather_data.txt

-rw-r--r--   1 hduser24 supergroup   53 2025-04-19 07:43 /input/wordcount_input.txt

hduser24@HDUSER-24865:~/hadoop_practicals/input$ cd ..

hduser24@HDUSER-24865:~/hadoop_practicals$ nano ./mapper/mapper.py

hduser24@HDUSER-24865:~/hadoop_practicals$ nano ./reducer/reducer.py

hduser24@HDUSER-24865:~/hadoop_practicals$ chmod +x ./mapper/mapper.py

hduser24@HDUSER-24865:~/hadoop_practicals$ chmod +x ./reducer/reducer.py

hduser24@HDUSER-24865:~/hadoop_practicals$ hadoop jar
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \

> -input /input/weather_data.txt \

> -output /output/weather_output \

> -mapper mapper.py \

> -reducer reducer.py \

> -file ./mapper/mapper.py \

> -file ./reducer/reducer.py

2025-04-19 08:56:18,539 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [./mapper/mapper.py, ./reducer/reducer.py, /tmp/hadoop-unjar6941302822368605056/] []
/tmp/streamjob3300983547996344782.jar tmpDir=null

2025-04-19 08:56:19,446 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-04-19 08:56:19,630 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-04-19 08:56:19,959 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/hduser24/.staging/job_1745052761951_0001

2025-04-19 08:56:21,206 INFO mapred.FileInputFormat: Total input files to process : 1

2025-04-19 08:56:22,162 INFO mapreduce.JobSubmitter: number of splits:2

2025-04-19 08:56:22,396 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745052761951_0001

2025-04-19 08:56:22,396 INFO mapreduce.JobSubmitter: Executing with tokens: []

2025-04-19 08:56:22,593 INFO conf.Configuration: resource-types.xml not found

2025-04-19 08:56:22,593 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2025-04-19 08:56:22,930 INFO impl.YarnClientImpl: Submitted application application_1745052761951_0001

2025-04-19 08:56:22,978 INFO mapreduce.Job: The url to track the job: http://HDUSER-
24865.:8088/proxy/application_1745052761951_0001/

2025-04-19 08:56:22,980 INFO mapreduce.Job: Running job: job_1745052761951_0001

2025-04-19 08:56:31,134 INFO mapreduce.Job: Job job_1745052761951_0001 running in uber mode : false

2025-04-19 08:56:31,136 INFO mapreduce.Job: map 0% reduce 0%

2025-04-19 08:56:36,241 INFO mapreduce.Job: map 100% reduce 0%

2025-04-19 08:56:40,530 INFO mapreduce.Job: map 100% reduce 100%

2025-04-19 08:56:41,564 INFO mapreduce.Job: Job job_1745052761951_0001 completed successfully

2025-04-19 08:56:41,678 INFO mapreduce.Job: Counters: 54

File System Counters

FILE: Number of bytes read=1391

FILE: Number of bytes written=845334

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=2718

HDFS: Number of bytes written=120

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)=5513

Total time spent by all reduces in occupied slots (ms)=2669

Total time spent by all map tasks (ms)=5513

Total time spent by all reduce tasks (ms)=2669

Total vcore-milliseconds taken by all map tasks=5513

Total vcore-milliseconds taken by all reduce tasks=2669

Total megabyte-milliseconds taken by all map tasks=2822656

Total megabyte-milliseconds taken by all reduce tasks=1366528

Map-Reduce Framework

Map input records=100

Map output records=100

Map output bytes=1185

Map output materialized bytes=1397

Input split bytes=192

Combine input records=0

Combine output records=0

Reduce input groups=100

Reduce shuffle bytes=1397

Reduce input records=100

Reduce output records=10

Spilled Records=200

Shuffled Maps =2

Failed Shuffles=0

Merged Map outputs=2

GC time elapsed (ms)=226

CPU time spent (ms)=1720

Physical memory (bytes) snapshot=944644096

Virtual memory (bytes) snapshot=6361174016

Total committed heap usage (bytes)=897056768

Peak Map Physical memory (bytes)=364916736

Peak Map Virtual memory (bytes)=2118967296

Peak Reduce Physical memory (bytes)=260411392

Peak Reduce Virtual memory (bytes)=2124247040

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=2526

File Output Format Counters

Bytes Written=120

```
2025-04-19 08:56:41,678 INFO streaming.StreamJob: Output directory: /output/weather_output
hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -ls /output/weather_output/
Found 2 items
-rw-r--r--   1 hduser24 supergroup          0 2025-04-19 08:56 /output/weather_output/_SUCCESS
-rw-r--r--   1 hduser24 supergroup        120 2025-04-19 08:56 /output/weather_output/part-00000
hduser24@HDUSER-24865:~/hadoop_practicals$ hdfs dfs -cat /output/weather_output/part-00000
1950    -18     43
1951    -17     44
1952    -12     32
1953    -20     41
1954    -13     40
1955    -16     45
1956    -14     33
1957    -19     38
1958    -20     28
1959    -19     40
hduser24@HDUSER-24865:~/hadoop_practicals$
```