

### **EXP 3: Map Reduce program to process a weather dataset**

#### **AIM:**

To implement MapReduce program to process a weather dataset.

#### **PROCEDURE:**

**Step 1: Create Data File:** Create a file named "sample\_weather.txt" and populate it with text data that you wish to analyse.

**Step 2: Mapper Logic - mapper.py:** Create a file named "mapper.py" to implement the logic for the mapper. The mapper will read input data from STDIN, split lines into words, and output each word with its count.

#### **mapper1.py**

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

```
def map1():
    for line in sys.stdin:
        tokens = line.strip().split()
        if len(tokens) < 13:
            continue

        station = tokens[0]
        if "STN" in station:
            continue

        date_hour = tokens[2]
        temp = tokens[3]
        dew = tokens[4]
        wind = tokens[12]

        if temp == "9999.9" or dew == "9999.9" or wind == "999.9":
            continue
```

```
hour = int(date_hour.split("_")[-1])
date = date_hour[:date_hour.rfind("_")-2]

if 4 < hour <= 10:
    section = "section1"
elif 10 < hour <= 16:
    section = "section2"
elif 16 < hour <= 22:
    section = "section3"
else:
    section = "section4"

key_out = f"{station}_{date}_{section}"
value_out = f"{temp} {dew} {wind}"
print(f"{key_out}\t{value_out}")

if __name__ == "__main__":
    map1()
```

**Step 3: Reducer Logic - reducer.py:** Create a file named "reducer.py" to implement the logic for the reducer. The reducer will aggregate the occurrences of each word and generate the final output.

### **reducer1.py**

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

def reduce1():
    current_key = None
    sum_temp, sum_dew, sum_wind = 0, 0, 0
    count = 0

    for line in sys.stdin:
        key, value = line.strip().split("\t")
        temp, dew, wind = map(float, value.split())

        if current_key is None:
```

```
current_key = key

if key == current_key:
    sum_temp += temp
    sum_dew += dew
    sum_wind += wind
    count += 1
else:
    avg_temp = sum_temp / count
    avg_dew = sum_dew / count
    avg_wind = sum_wind / count
    print(f'{current_key}\t{avg_temp} {avg_dew} {avg_wind}')

    current_key = key
    sum_temp, sum_dew, sum_wind = temp, dew, wind
    count = 1

if current_key is not None:
    avg_temp = sum_temp / count
    avg_dew = sum_dew / count
    avg_wind = sum_wind / count
    print(f'{current_key}\t{avg_temp} {avg_dew} {avg_wind}')

if __name__ == "__main__":
    reduce1()
```

**Step 4: Prepare Hadoop Environment:** Start the Hadoop daemons and create a directory in HDFS to store your data. Run the following commands to store the data in the WeatherData Directory.

```
Administrator: Command Prompt

Directory of D:\madhumitha\DA\weather

09/17/2024 05:19 PM <DIR> .
09/17/2024 05:19 PM <DIR> ..
09/17/2024 05:19 PM      12,053 input1.txt
09/17/2024 05:19 PM      934 mapper1.py
09/17/2024 05:19 PM      1,012 reducer1.py
                3 File(s)      13,999 bytes
                2 Dir(s)      335,340,986,368 bytes free

D:\madhumitha\DA\weather>hdfs dfs -put input1.txt /madhu

D:\madhumitha\DA\weather>hdfs dfs -put mapper1.py /madhu

D:\madhumitha\DA\weather>hdfs dfs -chmod 777 /madhu/mapper1.py

D:\madhumitha\DA\weather>hdfs dfs -put reducer1.py /madhu

D:\madhumitha\DA\weather>hdfs dfs -chmod 777 /madhu/reducer1.py
```

```
D:\madhumitha\DA\weather>hadoop jar
D:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -files
"hdfs:///madhu/mapper1.py,hdfs:///madhu/reducer1.py" -input
/madhu/input1.txt -output /madhu/output4 -mapper "python mapper1.py"
-reducer "python reducer1.py"
```

## OUTPUT:

```
Administrator: Command Prompt
D:\madhumitha\DA\weather>hdfs dfs -chmod 777 /madhu/reducer1.py

D:\madhumitha\DA\weather>hadoop jar D:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -files "hdfs:///madhu/mapper1.py,hdfs:///madhu/reducer1.py" -input /madhu/input1.txt -output /madhu/output4 -
mapper "python mapper1.py" -reducer "python reducer1.py"
packageJobJar: [/C:/Users/ADMIN/AppData/Local/Temp/hadoop-unjar1687931298959750318/] [] C:\Users\ADMIN\AppData\Local\Temp\streamjob2350782234208546254.jar tmpDir=null
2024-09-17 17:21:57,875 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-17 17:21:58,059 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-17 17:22:03,436 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ADMIN/.staging/job_1726572362226_0002
2024-09-17 17:22:03,676 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-17 17:22:03,723 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-17 17:22:03,878 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1726572362226_0002
2024-09-17 17:22:03,878 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-17 17:22:04,021 INFO conf.Configuration: resource-types.xml not found
2024-09-17 17:22:04,021 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-17 17:22:04,080 INFO impl.YarnClientImpl: Submitted application application_1726572362226_0002
2024-09-17 17:22:04,113 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TAPF3T5:8088/proxy/application_1726572362226_0002/
2024-09-17 17:22:04,114 INFO mapreduce.Job: Running job: job_1726572362226_0002
2024-09-17 17:22:35,535 INFO mapreduce.Job: Job job_1726572362226_0002 running in uber mode : false
2024-09-17 17:22:35,537 INFO mapreduce.Job: map 0% reduce 0%
2024-09-17 17:22:55,798 INFO mapreduce.Job: map 50% reduce 0%
2024-09-17 17:23:00,844 INFO mapreduce.Job: map 100% reduce 0%
2024-09-17 17:23:21,085 INFO mapreduce.Job: map 100% reduce 100%
2024-09-17 17:23:26,150 INFO mapreduce.Job: Job job_1726572362226_0002 completed successfully
2024-09-17 17:23:26,231 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=3870
    FILE: Number of bytes written=851819
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=16329
    HDFS: Number of bytes written=312
    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)=35526
    Total time spent by all reduces in occupied slots (ms)=17581
    Total time spent by all map tasks (ms)=35526
    Total time spent by all reduce tasks (ms)=17581
    Total vcore-milliseconds taken by all map tasks=35526
    Total vcore-milliseconds taken by all reduce tasks=17581
    Total megabyte-milliseconds taken by all map tasks=36378624
    Total megabyte-milliseconds taken by all reduce tasks=18002944
  Map-Reduce Framework
```

```
Administrator: Command Prompt

Map input records=96
Map output records=96
Map output bytes=3672
Map output materialized bytes=3876
Input split bytes=180
Combine input records=0
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=3876
Reduce input records=96
Reduce output records=4
Spilled Records=192
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=87
CPU time spent (ms)=406
Physical memory (bytes) snapshot=830640128
Virtual memory (bytes) snapshot=1212493824
Total committed heap usage (bytes)=624427008
Peak Map Physical memory (bytes)=300793856
Peak Map Virtual memory (bytes)=428761088
Peak Reduce Physical memory (bytes)=230240256
Peak Reduce Virtual memory (bytes)=406401024

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=16149
File Output Format Counters
  Bytes Written=312
2024-09-17 17:23:26,232 INFO streaming.StreamJob: Output directory: /madhu/output4

D:\madhumitha\DA\weather>hdfs dfs -ls /madhu/output4
Found 2 items
-rw-r--r-- 1 ADMIN supergroup          0 2024-09-17 17:23 /madhu/output4/_SUCCESS
-rw-r--r-- 1 ADMIN supergroup      312 2024-09-17 17:23 /madhu/output4/part-00000

D:\madhumitha\DA\weather>hdfs dfs -cat /madhu/output4/part-00000
690190_200602_section1 53.87166666666666 25.899999999999995 7.7749999999999998
690190_200602_section2 54.76125000000001 25.900000000000006 7.7749999999999999
690190_200602_section3 53.25041666666667 25.899999999999995 7.7749999999999996
690190_200602_section4 52.44708333333333 25.900000000000006 7.7749999999999999

D:\madhumitha\DA\weather>_

Browse HDFS

localhost:9870/explorer.html#/madhu/output4

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory
/madhu/output4

Show 25 entries

Permission Owner
-rw-r--r-- ADMIN
-rw-r--r-- ADMIN

Showing 1 to 2 of 2 entries

Hadoop, 2023.

File information - part-00000

Download Head the file (first 32K) Tail the file (last 32K)

Block information - Block 0

Block ID: 1073742012
Block Pool ID: BP-2145839966-192.168.72.1-1726306606534
Generation Stamp: 1190
Size: 312
Availability:
  • DESKTOP-TAPFJT5

File contents

690190_200602_section1 53.87166666666666 25.899999999999995 7.7749999999999998
690190_200602_section2 54.76125000000001 25.900000000000006 7.7749999999999999
690190_200602_section3 53.25041666666667 25.899999999999995 7.7749999999999996
690190_200602_section4 52.44708333333333 25.900000000000006 7.7749999999999999
```

## RESULT:

Thus, the program for weather dataset using Map Reduce has been executed successfully.