EX NO: 3 Map Reduce program to process a weather dataset

Aim:

To implement MapReduce program to process a weather dataset.

Procedure:

Step 1:

Create a file named “sample_weather.txt” and populate it with the data.

```
690190 13910  20060201_0  51.75     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_1  54.74     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_2  50.59     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_3  51.67     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_4  65.67     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_5  55.37     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_6  49.26     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_7  55.44     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_8  64.05     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_9  68.77     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_10  48.93     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_11  65.37     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_12  69.45     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_13  52.91     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_14  53.69     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_15  53.30     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_16  66.17     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_17  53.83     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_18  50.54     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_19  50.27     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_20  59.08     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_21  53.05     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_22  57.97     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_23  48.23     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060202_0  47.16     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_1  69.72     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_2  62.71     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_3  46.34     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_4  53.15     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_5  64.59     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_6  58.26     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_7  53.27     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202_8  43.68     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
690190 13910  20060202 9  65.70     28.5 24  1003.2 24   940.6 24   15.0 24    5.7 24   12.0  999.9    0.00I 999.9  000000
```

Step 2:

Create program **mapper.py**

```
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()
```

Create program **reducer.py**

```
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 3:

Start the hadoop deamons and create a directory in HDFS to store data.

Make a directory, put the text file inside it.

hdfs dfs –mkir –p /user/hadoop/input

```
C:\hadoop\hadoop\sbin>start-dfs.cmd

C:\hadoop\hadoop\sbin>start-yarn.cmd
starting yarn daemons

C:\hadoop\hadoop\sbin>hdfs dfs -mkdir -p /user/hadoop/input

C:\hadoop\hadoop\sbin>hdfs dfs -put C:/text/sample_weather.txt /user/hadoop/input
```

Step 4:

Run the mapper.py and reducer.py in hadoop environment.

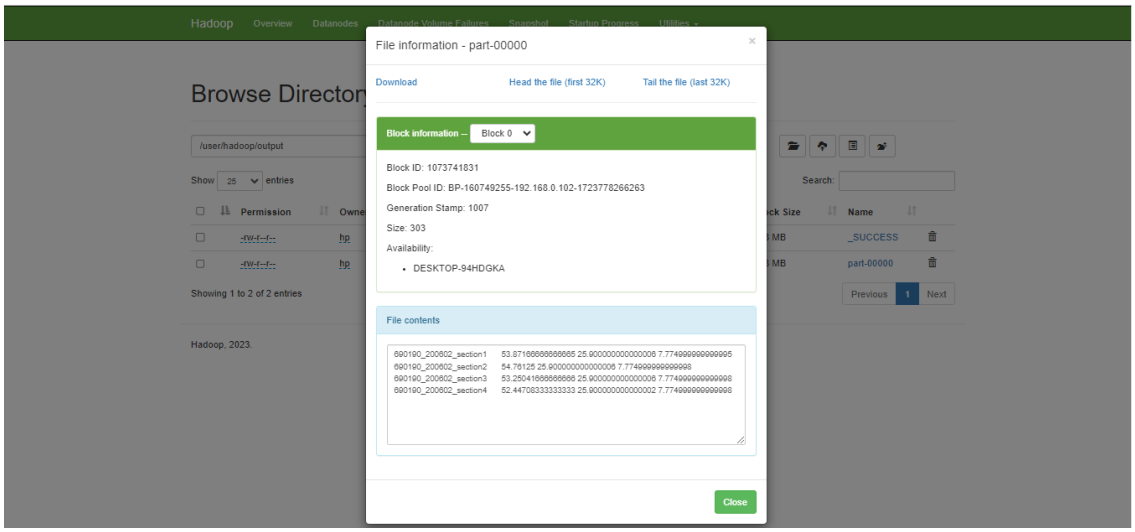
```
C:\hadoop\hadoop\sbin>hadoop jar C:\hadoop\hadoop\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
More? -mapper "python C:\text\mapperw.py" -reducer "python C:\text\reducerw.py" ^
More? -input /user/hadoop/input/sample_weather.txt -output /user/hadoop/output
2024-08-19 08:48:06,520 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-08-19 08:48:06,731 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-08-19 08:48:06,731 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-08-19 08:48:06,772 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-08-19 08:48:08,141 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-19 08:48:08,382 INFO mapreduce.JobSubmitter: number of splits:1
2024-08-19 08:48:08,691 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1457147166_0001
2024-08-19 08:48:08,694 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-19 08:48:09,013 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-08-19 08:48:09,019 INFO mapreduce.Job: Running job: job_local1457147166_0001
2024-08-19 08:48:09,022 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-08-19 08:48:09,029 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-08-19 08:48:09,113 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-08-19 08:48:09,113 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignor
e
2024-08-19 08:48:09,211 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-08-19 08:48:09,220 INFO mapred.LocalJobRunner: Starting task: attempt_local1457147166_0001_m_000000_0
2024-08-19 08:48:09,279 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-08-19 08:48:09,279 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignor
e
2024-08-19 08:48:09,296 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2024-08-19 08:48:09,380 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@7e1a7ae0
2024-08-19 08:48:09,450 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hadoop/input/sample_weather.txt:0+12053
2024-08-19 08:48:09,490 INFO mapred.MapTask: numReduceTasks: 1
2024-08-19 08:48:09,638 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-08-19 08:48:09,639 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-08-19 08:48:09,640 INFO mapred.MapTask: soft limit at 83886080
2024-08-19 08:48:09,641 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-08-19 08:48:09,642 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-08-19 08:48:09,653 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-08-19 08:48:09,796 INFO streaming.PipeMapRed: PipeMapRed exec [python, C:\text\mapperw.py]
2024-08-19 08:48:09,804 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2024-08-19 08:48:09,808 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
```

```
                Merged Map outputs=1
                GC time elapsed (ms)=23
                Total committed heap usage (bytes)=524812288
        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=12053
        File Output Format Counters
                Bytes Written=303
024-08-19 08:48:13,116 INFO streaming.StreamJob: Output directory: /user/hadoop/output
```

Step 5:

Check the output in the output directory

/user/hadoop/output



Result:

Thus the program to process weather dataset using map reduce was executed successfully.