Ex no: 3

Implement a MapReduce program to process a weather dataset

Aim:

To implement a MapReduce program to process a weather dataset.

Procedure:

1. Open Command prompt and run as administrator and start the Hadoop service.
2. Check if the namenode is empty, type “hdfs namenode -format” if it is empty else skip this step.
3. Create a new directory in the Hadoop file system using the command “hdfs dfs -mkdir /directory_name”.
4. Upload the input text file to the new directory using the command “hdfs dfs -put C:\DA_EX_3\sample_weather.txt /weather_data”
5. Create the mapper and reducer python programs separately for this program.
6. For Hadoop Streaming execute the following command
   hadoop jar C:\hadoop-3.3.6\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar ^
   -mapper “python C:/DA_EX_3/mapper.py” ^ -reducer “python C:/DA_EX_3/reducer.py” ^ -input /word_count/input.txt ^ -output /weather_data/hadoop_output/output

Program:

mapper.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()
```

reducer.py

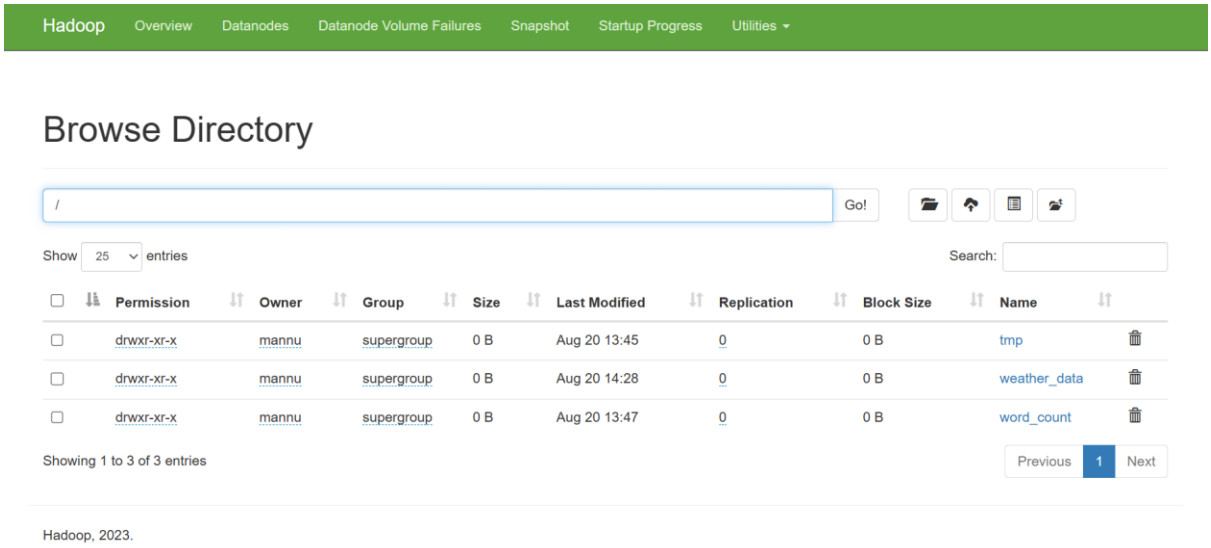
```
#!/usr/bin/python3
import sys
def reduce1():
    current_key = None
    sum_temp, sum_dew, sum_wind = 0, 0, 0
    count = 0
    for line in sys.stdin:
        key, value = line.strip().split("\t")
```

```
            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()
```

Output Screenshots:

```
C:\Hadoop_programs>start-dfs.cmd

C:\Hadoop_programs>start-yarn.cmd
starting yarn daemons
```



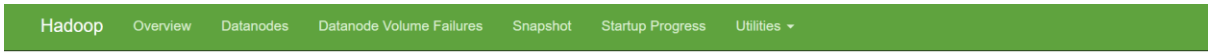

```
C:\hadoop-3.3.6\sbin>hadoop fs -mkdir /weather_data

C:\hadoop-3.3.6\sbin>hdfs dfs -put C:\DA_EX_3\sample_weather.txt /weather_data
```

```
C:\hadoop-3.3.6\sbin>hadoop jar C:\hadoop-3.3.6\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar ^
More? -mapper "python C:\\DA_EX_3\\mapper.py" -reducer "python C:\\DA_EX_3\\reducer.py" ^
More? -input /weather_data/sample_weather.txt -output /weather_data/hadoop_output/output
packageJobJar: [/C:/Users/mannu/AppData/Local/Temp/hadoop-unjar14206619621890066657/] [] C:\Users\mannu\AppData\Local\Temp\streamjob77725541
07362293945.jar tmpDir=null
2024-08-20 14:28:21,207 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-20 14:28:21,375 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-20 14:28:27,043 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/mannu/.staging/job_
1724141229673_0003
2024-08-20 14:28:27,283 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-20 14:28:27,351 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-20 14:28:27,481 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724141229673_0003
2024-08-20 14:28:27,482 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-20 14:28:27,619 INFO conf.Configuration: resource-types.xml not found
2024-08-20 14:28:27,620 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-20 14:28:27,691 INFO impl.YarnClientImpl: Submitted application application_1724141229673_0003
2024-08-20 14:28:27,728 INFO mapreduce.Job: The url to track the job: http://Shreeya:8088/proxy/application_1724141229673_0003/
2024-08-20 14:28:27,729 INFO mapreduce.Job: Running job: job_1724141229673_0003
2024-08-20 14:28:40,989 INFO mapreduce.Job: Job job_1724141229673_0003 running in uber mode : false
2024-08-20 14:28:40,992 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-20 14:28:48,141 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-20 14:28:48,155 INFO mapreduce.Job: Job job_1724141229673_0003 completed successfully
2024-08-20 14:28:48,279 INFO mapreduce.Job: Counters: 54
```

```
        File Input Format Counters
                Bytes Read=16149
        File Output Format Counters
                Bytes Written=312
2024-08-20 14:28:48,279 INFO streaming.StreamJob: Output directory: /weather_data/hadoop_output/output

C:\hadoop-3.3.6\sbin>
```



Browse Directory

/weather_data Go!

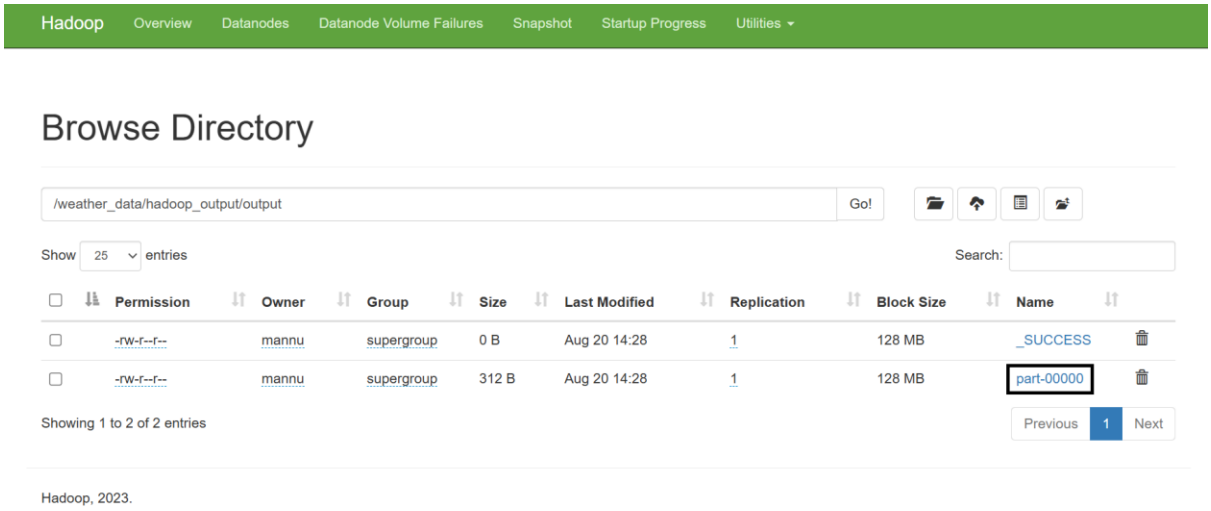
Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐	drwxr-xr-x	mannu	supergroup	0 B	Aug 20 14:28	0	0 B	hadoop_output	🗑
☐	-rw-r--r--	mannu	supergroup	11.77 KB	Aug 20 14:24	1	128 MB	sample_weather.txt	🗑

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2023.



File information - part-00000

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

Block information -- Block 0

Block ID: 1073741844

Block Pool ID: BP-174739359-192.168.182.1-1723880367120

Generation Stamp: 1020

Size: 312

Availability:

- Shreeya

File contents

```
690190_200602_section1    53.87166666666666 25.899999999999995 7.774999999999998
690190_200602_section2    54.76125000000001 25.900000000000006 7.774999999999999
690190_200602_section3    53.25041666666667 25.899999999999995 7.774999999999996
690190_200602_section4    52.44708333333333 25.900000000000006 7.774999999999999
```

```
C:\hadoop-3.3.6\sbin>hdfs dfs -cat /weather_data/hadoop_output/output/part-00000
690190_200602_section1  53.87166666666666 25.899999999999995 7.774999999999998
690190_200602_section2  54.76125000000001 25.900000000000006 7.774999999999999
690190_200602_section3  53.25041666666667 25.899999999999995 7.774999999999996
690190_200602_section4  52.44708333333333 25.900000000000006 7.774999999999999
```

Result:

Thus the implementation of MapReducer program to process weather data has been executed successfully.