

Implement a MapReduce program to process a weather dataset

Steps:

1. Open command prompt and run as administrator

Go to hadoop sbin directory

```
C:\Windows\system32>cd C:\Hadoop\sbin  
C:\Hadoop\sbin>_
```

Note:

1. Check hadoop/data/datanode and hadoop/data/namenode and if both folders are empty, type “hdfs namenode -format”.
2. Check python version with “python --version”.
3. Check “C:\Python39\” is added in Environment variables > System variables > Path, if not add your python path.
4. Check Environment variables > System variables > HADOOP_HOME is set as “C:\Hadoop”.

```
C:\Hadoop\sbin>echo %HADOOP_HOME%  
C:\Hadoop  
  
C:\Hadoop\sbin>python --version  
Python 3.11.4
```

2. Start Hadoop Services

start-dfs.cmd

start-yarn.cmd

```
C:\Hadoop\sbin>start-dfs.cmd  
  
C:\Hadoop\sbin>start-yarn.cmd  
starting yarn daemons  
  
C:\Hadoop\sbin>jps  
13120 NameNode  
2384 NodeManager  
4100 DataNode  
7956 ResourceManager  
9124 Jps
```

3. Open the browser and go to the URL “localhost:9870”

Overview 'localhost:9000' (✓active)

Started:	Sun Aug 18 18:45:16 +0530 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f06fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-a23ce25d-ee9d-4000-ac1f-044f436c4c8a
Block Pool ID:	BP-934656018-192.168.56.1-1723971050909

Summary

Security is off.
Safemode is off.

19 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 24 total filesystem object(s).

Heap Memory used 74.86 MB of 193 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 61.65 MB of 63.11 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	118.63 GB
-----------------------------	-----------

4. Create a Directory in HDFS

```
hdfs dfs -mkdir -p /weather/hadoop/input
```

```
C:\Hadoop\sbin>hdfs dfs -mkdir -p /weather/hadoop/input
C:\Hadoop\sbin>
```

5. Copy the Input File to HDFS

```
hdfs dfs -put C:/Users/Admin/mapreduce_weather/sample_weather.txt /weather/hadoop/input
```

```
C:\Hadoop\sbin>hdfs dfs -put C:/Users/Admin/mapreduce_weather/sample_weather.txt /weather/hadoop/input
C:\Hadoop\sbin>hdfs dfs -ls /weather/hadoop/input
Found 1 items
-rw-r--r-- 1 Admin supergroup 12053 2024-08-18 18:52 /weather/hadoop/input/sample_weather.txt
C:\Hadoop\sbin>hdfs dfs -cat /weather/hadoop/input/sample_weather.txt
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
```

Note:**mapper.py:**

```
#!/usr/bin/env python
```

```
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/env python
```

```
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()
```

6. Run the Hadoop Streaming Job

```
hadoop jar %HADOOP_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
```

```
-mapper "python C:\\Users\\Admin\\mapreduce_weather\\mapper.py" -reducer "python  
C:\\Users\\Admin\\mapreduce_weather\\reducer.py" ^
```

```
-input /weather/hadoop/input/sample_weather.txt -output /weather/hadoop/output
```

```
C:\Hadoop\sbin>hadoop jar %HADOOP_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
More? -mapper "python C:\\Users\\Admin\\mapreduce_weather\\mapper.py" -reducer "python C:\\Users\\Admin\\mapreduce_weather\\reducer.py" ^
More? -input /weather/hadoop/input/sample_weather.txt -output /weather/hadoop/output
packageJobJar: [/C:/Users/Admin/AppData/Local/Temp/hadoop-unjar754311025374819372/] [] C:\Users\Admin\AppData\Local\Temp\streamjob1857461994
2024-08-18 19:02:14,577 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-18 19:02:14,970 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-18 19:02:16,543 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_17
2024-08-18 19:02:17,513 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-18 19:02:17,703 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-18 19:02:18,141 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1723986937631_0001
2024-08-18 19:02:18,142 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-18 19:02:18,524 INFO conf.Configuration: resource-types.xml not found
2024-08-18 19:02:18,525 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-18 19:02:19,829 INFO impl.YarnClientImpl: Submitted application application_1723986937631_0001
2024-08-18 19:02:19,971 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TF65P79:8088/proxy/application_1723986937631_0001/
2024-08-18 19:02:19,981 INFO mapreduce.Job: Running job: job_1723986937631_0001
2024-08-18 19:02:44,954 INFO mapreduce.Job: Job job_1723986937631_0001 running in uber mode : false
2024-08-18 19:02:44,961 INFO mapreduce.Job: map 0% reduce 0%
2024-08-18 19:03:01,944 INFO mapreduce.Job: map 100% reduce 0%
2024-08-18 19:03:12,124 INFO mapreduce.Job: map 100% reduce 100%
2024-08-18 19:03:13,149 INFO mapreduce.Job: Job job_1723986937631_0001 completed successfully
2024-08-18 19:03:13,480 INFO mapreduce.Job: Counters: 54
```

```
File Input Format Counters
  Bytes Read=16149
File Output Format Counters
  Bytes Written=312
2024-08-18 19:03:13,482 INFO streaming.StreamJob: Output directory: /weather/hadoop/output
C:\Hadoop\sbin>
```

7. View the Output

```
hdfs dfs -cat /weather/hadoop/output/part-00000
```

```
C:\Windows\System32>hdfs dfs -cat /input/hadoop/output1/weatherset_result1/part-00000
000    15.533333333333333
001    15.433333333333332
002    16.433333333333334
003    15.600000000000001
```

8. Once the map reduce operations are performed successfully, the output will be present in the specified directory.

“/weather/hadoop/output/part-00000”

Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	jazil	supergroup	0 B	Aug 26 23:50	1	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-r--r--	jazil	supergroup	96 B	Aug 26 23:50	1	128 MB	part-00000	

```
2000    15.533333333333333
2001    15.433333333333332
2002    16.433333333333334
2003    15.600000000000001
```

9. Stop Hadoop Services

`stop-dfs.cmd`

`stop-yarn.cmd`

```
C:\Hadoop\sbin>stop-dfs.cmd
SUCCESS: Sent termination signal to the process with PID 7964.
SUCCESS: Sent termination signal to the process with PID 13580.

C:\Hadoop\sbin>stop-yarn.cmd
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 14412.
SUCCESS: Sent termination signal to the process with PID 7092.

INFO: No tasks running with the specified criteria.

C:\Hadoop\sbin>
```