

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

1. Log in with your Hadoop user.
2. Download the weather dataset and save it locally, for example, as `dataset.txt`.

Step 2: Mapper Logic

1. Create a file named `mapper.py`.
2. Implement the mapper logic:
 - The mapper processes each line of the dataset.
 - Extract the month and daily maximum temperature from each record and output them.

Step 3: Reducer Logic

1. Create a file named `reducer.py`.
2. Implement the reducer logic:
 - The reducer receives the output from the mapper, which contains the month and temperature data.
 - Aggregate the daily maximum temperatures by month and find the highest temperature for each month.

Step 4: Prepare Hadoop Environment

1. Start the necessary Hadoop services (daemons).
2. Create a directory in HDFS for storing the weather dataset.

Step 5: Upload Data to HDFS

1. Upload the dataset file to the HDFS directory created in the previous step.

Step 6: Make Python Files Executable

1. Provide executable permissions to the `mapper.py` and `reducer.py` files.

Step 7: Run the MapReduce Program Using Hadoop Streaming

1. Download the Hadoop Streaming JAR file if not already available.
2. Run the MapReduce job by specifying the input data (dataset), the output directory, and the mapper and reducer Python files using Hadoop Streaming.

Step 8: Check Output

1. View the results of the MapReduce job in the HDFS output directory.
2. If needed, you can copy the results to your local machine for further analysis.

Commands:

```
C:\hadoop\sbin> start-all.cmd  
C:\hadoop\sbin> jps  
C:\hadoop\sbin> cd /  
C:\> cd hadoop
```

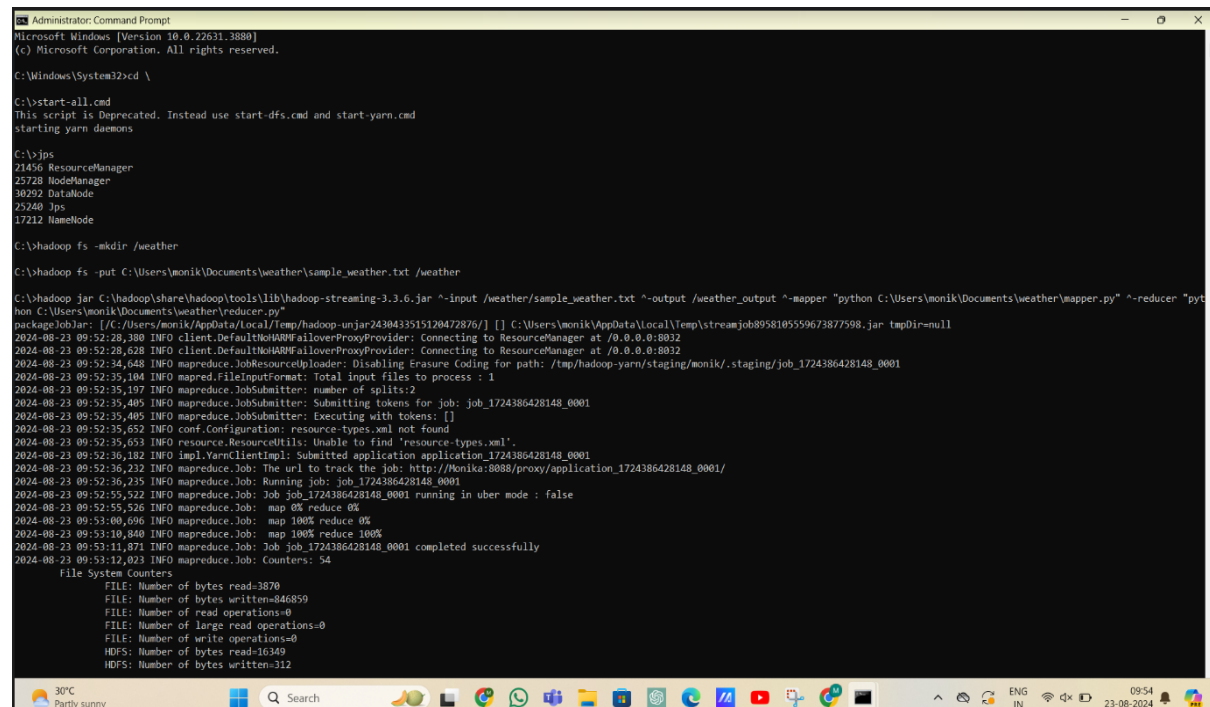
```
C:\hadoop> hadoop fs -mkdir /user/

C:\hadoop> hadoop fs -put C:/DataAnalytics/sample_weather.csv /input

C:\hadoop> hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar -input /user/sample_weather.csv -output /user/output-data -mapper "C:\Users\monik\Documents\weather\mapper.py" -reducer "C:\Users\monik\Documents\weather\reducer.py"

hadoop fs -cat /user/jayas/output/part-00000
```

OUTPUT:



```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.3880]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd \

C:\>start-all.cmd
This script is deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\>jps
21456 ResourceManager
25728 NodeManager
30292 DataNode
25240 Jps
17212 NameNode

C:\>hadoop fs -mkdir /weather

C:\>hadoop fs -put C:\Users\monik\Documents\weather\sample_weather.txt /weather

C:\>hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar -input /weather/sample_weather.txt -output /weather_output -mapper "python C:\Users\monik\Documents\weather\mapper.py" -reducer "python C:\Users\monik\Documents\weather\reducer.py"
packageJobJar: [/C:/Users/monik/AppData/Local/Temp/hadoop-unjar2438433515120472876/] [] C:\Users\monik\AppData\Local\Temp\streamjob8958185559673877598.jar tmpDir=null
2024-08-23 09:52:28,380 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-23 09:52:28,628 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-23 09:52:34,648 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/monik/.staging/job_1724386428148_0001
2024-08-23 09:52:35,197 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-23 09:52:35,405 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724386428148_0001
2024-08-23 09:52:35,652 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-23 09:52:35,652 INFO conf.Configuration: resource-types.xml not found
2024-08-23 09:52:35,653 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-23 09:52:36,182 INFO impl.YarnClientImpl: Submitted application application_1724386428148_0001
2024-08-23 09:52:36,232 INFO mapreduce.Job: The url to track the job: http://monika:8088/proxy/application_1724386428148_0001/
2024-08-23 09:52:36,232 INFO mapreduce.Job: Running job: job_1724386428148_0001
2024-08-23 09:52:55,522 INFO mapreduce.Job: Job job_1724386428148_0001 running in uber mode : false
2024-08-23 09:52:55,526 INFO mapreduce.Job: map 0% reduce 0%
2024-08-23 09:53:00,696 INFO mapreduce.Job: map 100% reduce 0%
2024-08-23 09:53:10,840 INFO mapreduce.Job: map 100% reduce 100%
2024-08-23 09:53:11,871 INFO mapreduce.Job: Job job_1724386428148_0001 completed successfully
2024-08-23 09:53:12,023 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=3870
  FILE: Number of bytes written=846859
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=16349
  HDFS: Number of bytes written=312
```

localhost:9870/explorer.html#/weather

Browse Directory

/weather

Show 25 entries

Permission Owner

-rw-r--r-- monik

Showing 1 to 1 of 1 entries

Hadoop, 2023.

File information - sample_weather.txt

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

Block Information -- Block 0

Block ID: 1073741861
Block Pool ID: BP-1760020507-192.168.56.1-1724161408547
Generation Stamp: 1037
Size: 12053
Availability:
• Monika

File contents

```
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 0.001 999.9 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 0.001 999.9 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 0.001 999.9 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 0.001 999.9 000000
```

Close

localhost:9870/explorer.html#/weather_output

Browse Directory

/weather_output

Show 25 entries

Permission Owner

-rw-r--r-- monik

-rw-r--r-- monik

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: 1073741868
Block Pool ID: BP-1760020507-192.168.56.1-1724161408547
Generation Stamp: 1044
Size: 312
Availability:
• Monika

File contents

```
690190_200602_section1 53.87166666666666 25.899999999999995 7.774999999999999
690190_200602_section2 54.78125000000001 25.900000000000006 7.774999999999999
690190_200602_section3 53.25041666666667 25.899999999999995 7.774999999999999
690190_200602_section4 52.44708333333333 25.900000000000006 7.774999999999999
```