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
```

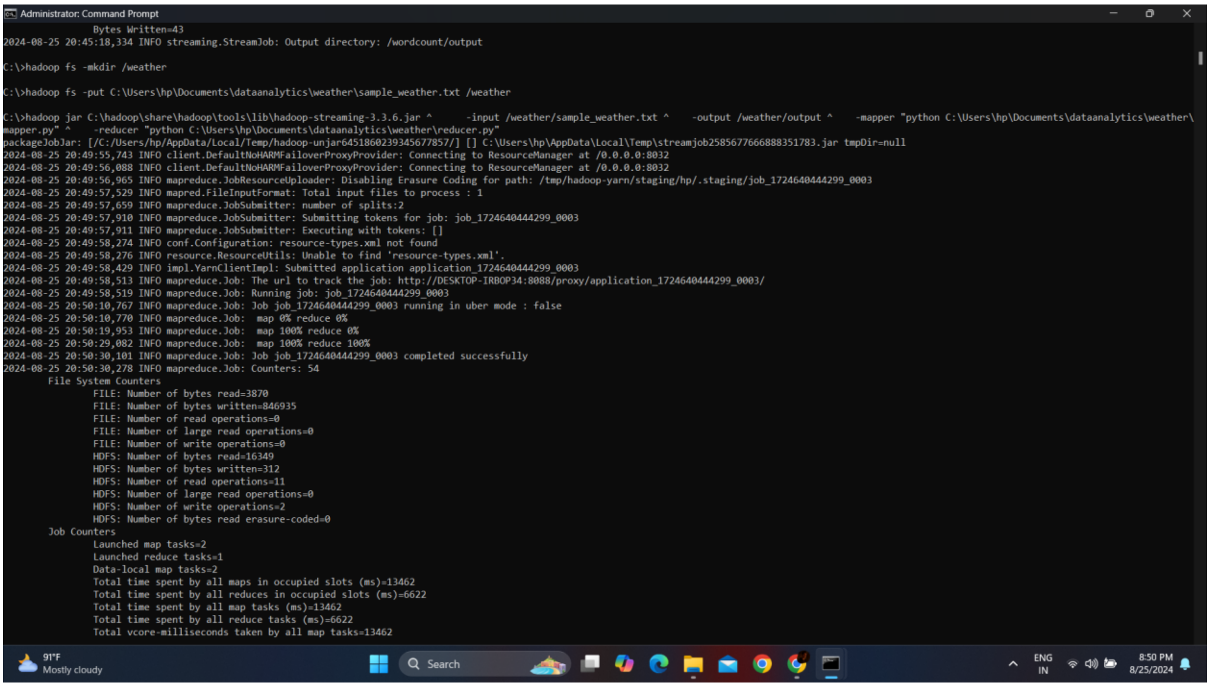
C:\hadoop>   hadoop fs -mkdir /user/

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\hp\Documents\weather\mapper.py"-reducer "C:\Users\hp\Documents\weather\reducer.py"

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

OUTPUT:

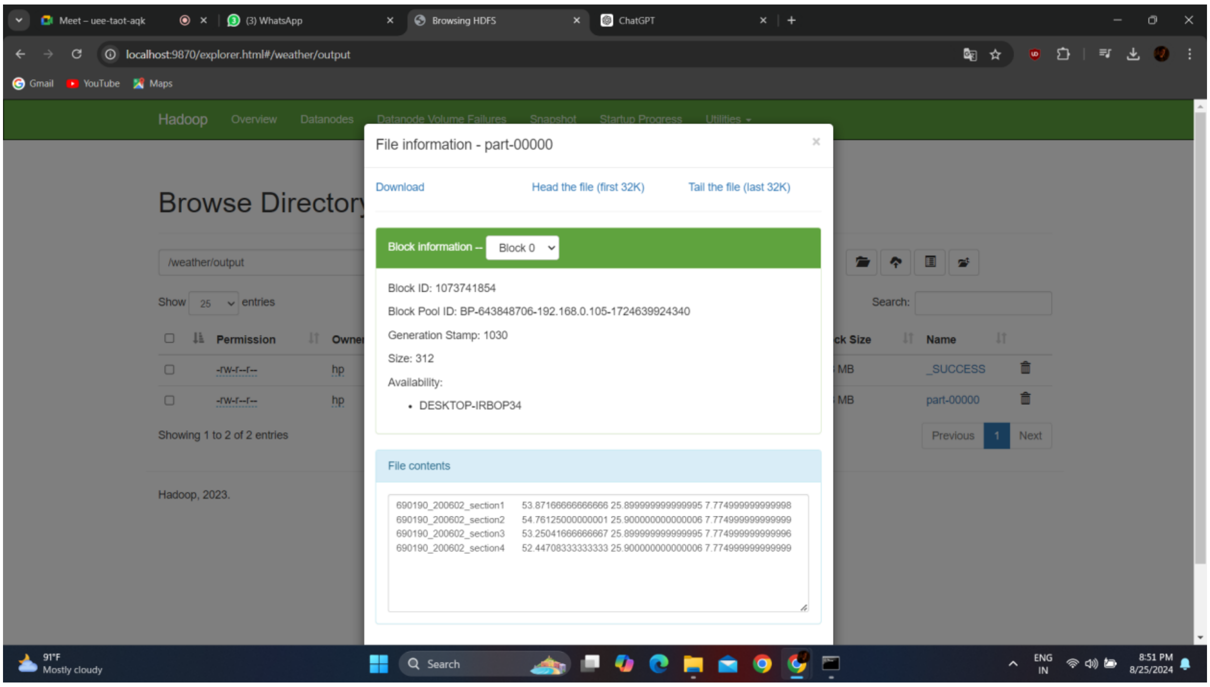


```
                Bytes Written=43
2024-08-25 20:45:18,334 INFO streaming.StreamJob: Output directory: /wordcount/output

C:\>hadoop fs -mkdir /weather

C:\>hadoop fs -put C:\Users\hp\Documents\dataanalytics\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\hp\Documents\dataanalytics\weather\
mapper.py" ^    -reducer "python C:\Users\hp\Documents\dataanalytics\weather\reducer.py"
packageJobJar: [/C:/Users/hp/AppData/Local/Temp/hadoop-unjar6451860239345677857/] [] C:\Users\hp\AppData\Local\Temp\streamjob2585677666888351783.jar tmpDir=null
2024-08-25 20:49:55,743 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-25 20:49:56,088 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-25 20:49:56,965 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hp/.staging/job_1724640444299_0003
2024-08-25 20:49:57,529 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-25 20:49:57,659 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-25 20:49:57,910 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724640444299_0003
2024-08-25 20:49:57,911 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-25 20:49:58,274 INFO conf.Configuration: resource-types.xml not found
2024-08-25 20:49:58,276 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-25 20:49:58,429 INFO impl.YarnClientImpl: Submitted application application_1724640444299_0003
2024-08-25 20:49:58,513 INFO mapreduce.Job: The url to track the job: http://DESKTOP-IRBOP34:8088/proxy/application_1724640444299_0003/
2024-08-25 20:49:58,519 INFO mapreduce.Job: Running job: job_1724640444299_0003
2024-08-25 20:50:10,767 INFO mapreduce.Job: Job job_1724640444299_0003 running in uber mode : false
2024-08-25 20:50:10,770 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-25 20:50:19,953 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-25 20:50:29,082 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-25 20:50:30,101 INFO mapreduce.Job: Job job_1724640444299_0003 completed successfully
2024-08-25 20:50:30,278 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=3870
                FILE: Number of bytes written=846935
                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
                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)=13462
                Total time spent by all reduces in occupied slots (ms)=6622
                Total time spent by all map tasks (ms)=13462
                Total time spent by all reduce tasks (ms)=6622
                Total vcore-milliseconds taken by all map tasks=13462
```



/weather/output

Show 25 entries

□	Permission	Owner	...ck Size	Name
□	-rw-r--r--	hp	MB	_SUCCESS
□	-rw-r--r--	hp	MB	part-00000

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: 1073741854

Block Pool ID: BP-643848706-192.168.0.105-1724639924340

Generation Stamp: 1030

Size: 312

Availability:

- DESKTOP-IRBOP34

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
```