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.

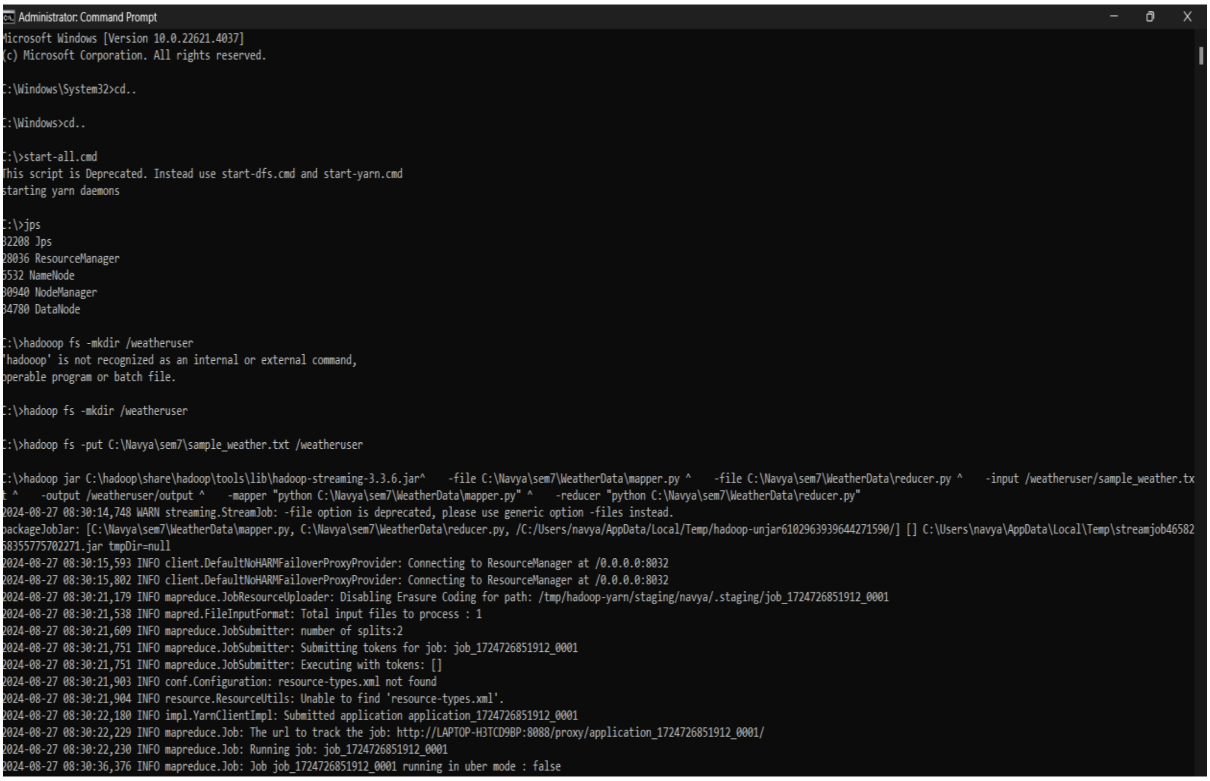
Step 8: Check Output
1. View the results of the MapReduce job in the HDFS output directory.
2. If needed, you can copy the results to your local machine for further analysis.

Commands:

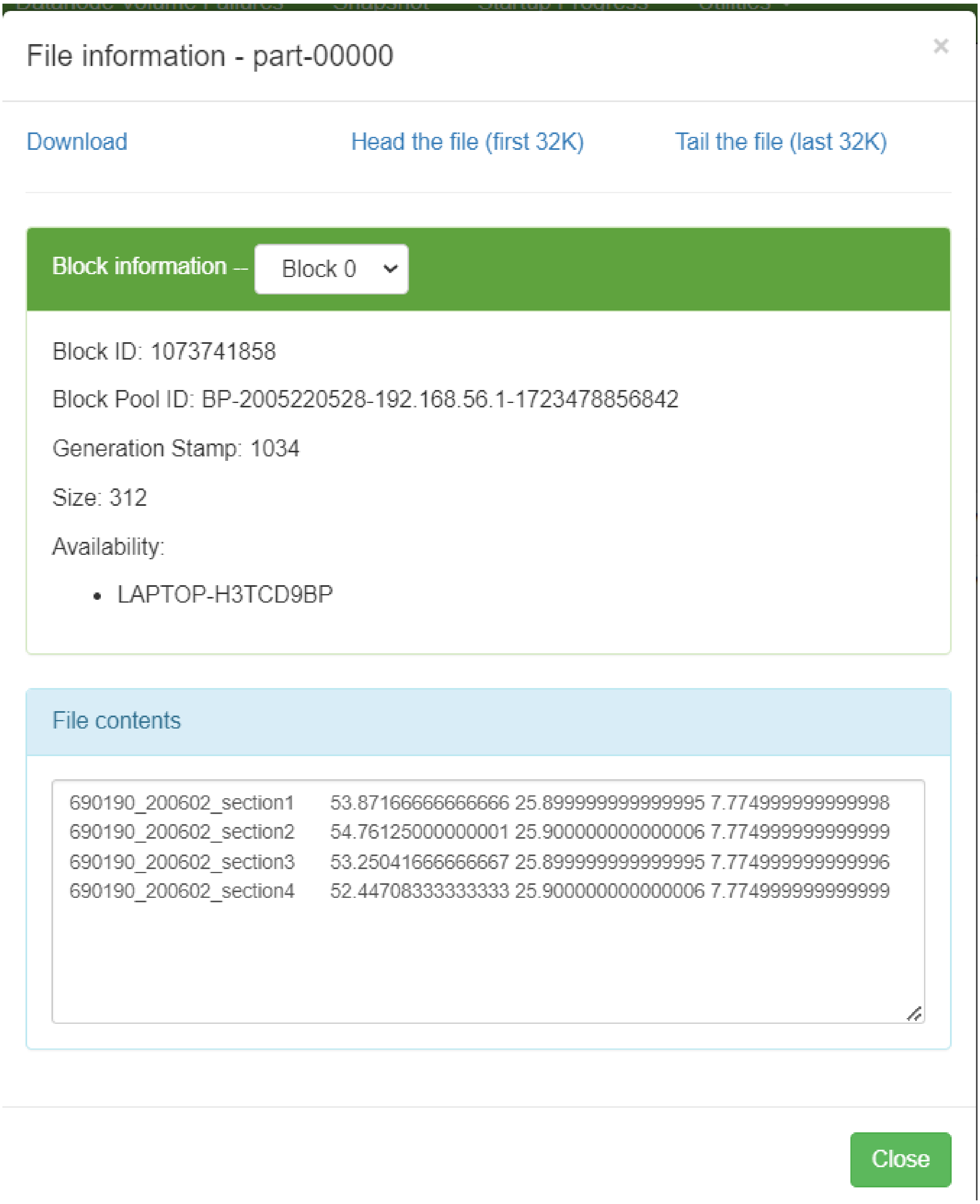
```
C:\hadoop\sbin>      start-all.cmd

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:



```
icrosoft Windows [Version 10.0.22621.4037]
(c) Microsoft Corporation. All rights reserved.

:\Windows\System32>cd..

:\Windows>cd..

:\>start-all.cmd
his script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
tarting yarn daemons

:\>jps
2208 Jps
8036 ResourceManager
532 NameNode
0940 NodeManager
4780 DataNode

:\>hadooop fs -mkdir /weatheruser
hadooop' is not recognized as an internal or external command,
perable program or batch file.

:\>hadoop fs -mkdir /weatheruser

:\>hadoop fs -put C:\Navya\sem7\sample_weather.txt /weatheruser

:\>hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar^    -file C:\Navya\sem7\WeatherData\mapper.py ^    -file C:\Navya\sem7\WeatherData\reducer.py ^    -input /weatheruser/sample_weather.tx
 ^    -output /weatheruser/output ^    -mapper "python C:\Navya\sem7\WeatherData\mapper.py" ^    -reducer "python C:\Navya\sem7\WeatherData\reducer.py"
024-08-27 08:30:14,748 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
ackageJobJar: [C:\Navya\sem7\WeatherData\mapper.py, C:\Navya\sem7\WeatherData\reducer.py, /C:/Users/navya/AppData/Local/Temp/hadoop-unjar6102963939644271590/] [] C:\Users\navya\AppData\Local\Temp\streamjob46582
8355775702271.jar tmpDir=null
024-08-27 08:30:15,593 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
024-08-27 08:30:15,802 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
024-08-27 08:30:21,179 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/navya/.staging/job_1724726851912_0001
024-08-27 08:30:21,538 INFO mapred.FileInputFormat: Total input files to process : 1
024-08-27 08:30:21,609 INFO mapreduce.JobSubmitter: number of splits:2
024-08-27 08:30:21,751 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724726851912_0001
024-08-27 08:30:21,751 INFO mapreduce.JobSubmitter: Executing with tokens: []
024-08-27 08:30:21,903 INFO conf.Configuration: resource-types.xml not found
024-08-27 08:30:21,904 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
024-08-27 08:30:22,180 INFO impl.YarnClientImpl: Submitted application application_1724726851912_0001
024-08-27 08:30:22,229 INFO mapreduce.Job: The url to track the job: http://LAPTOP-H3TCD9BP:8088/proxy/application_1724726851912_0001/
024-08-27 08:30:22,230 INFO mapreduce.Job: Running job: job_1724726851912_0001
024-08-27 08:30:36,376 INFO mapreduce.Job: Job job_1724726851912_0001 running in uber mode : false
```



File information - part-00000

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

Block information -- Block 0

Block ID: 1073741858

Block Pool ID: BP-2005220528-192.168.56.1-1723478856842

Generation Stamp: 1034

Size: 312

Availability:

- LAPTOP-H3TCD9BP

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
```