EX.NO:3 **210701165**

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
```

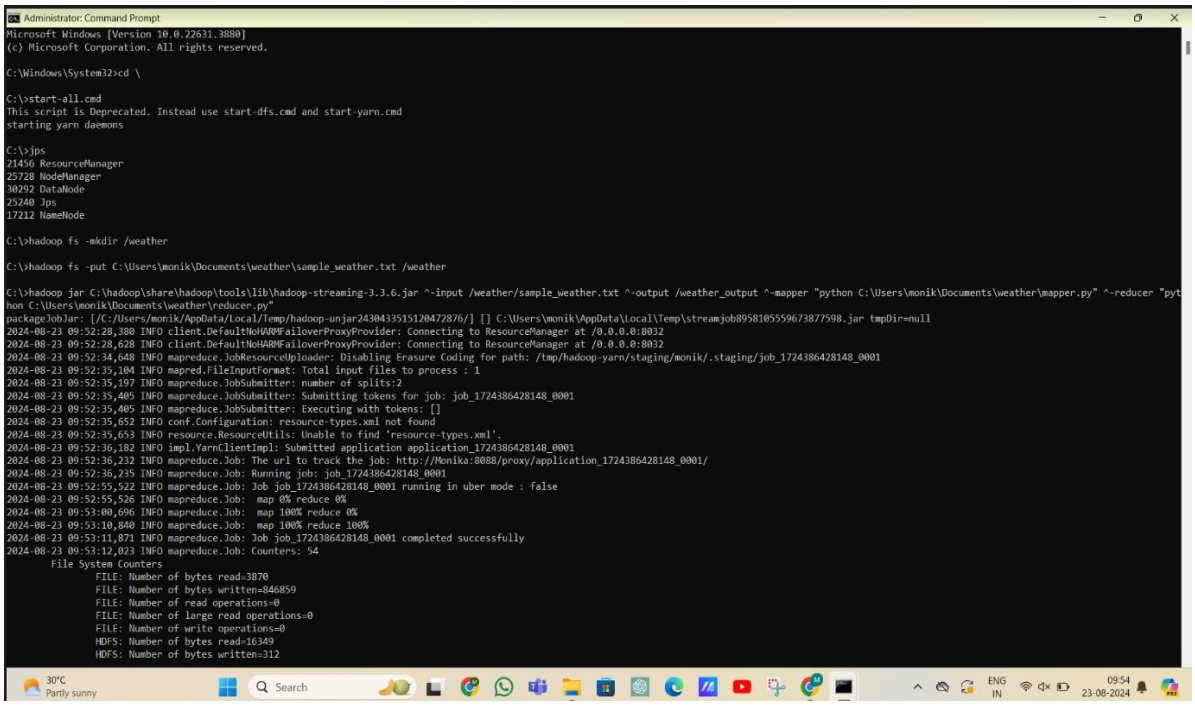
EX.NO:3 **210701165**

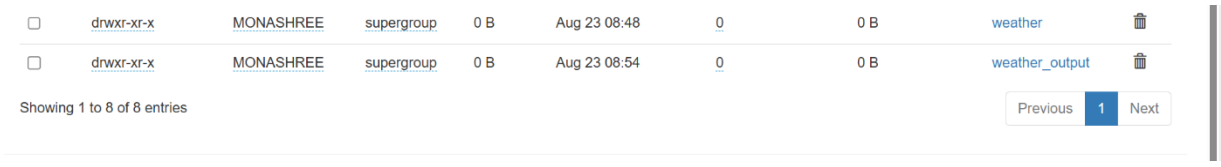
C:\hadoop> hadoop fs -mkdir /weather/

C:\hadoop> hadoop fs -put C:/weather_data/sample_weather.txt /weather

C:\hadoop> hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar^ -input /weather/sample_weather.txt^ -output /weather_output^ -mapper "C:\weather_data\mapper.py"^ -reducer "C:\weather_data\reducer.py"

OUTPUT:





Block ID: 1073741883

Block Pool ID: BP-468074218-172.16.11.91-1722913037250

Generation Stamp: 1059

Size: 12053

Availability:

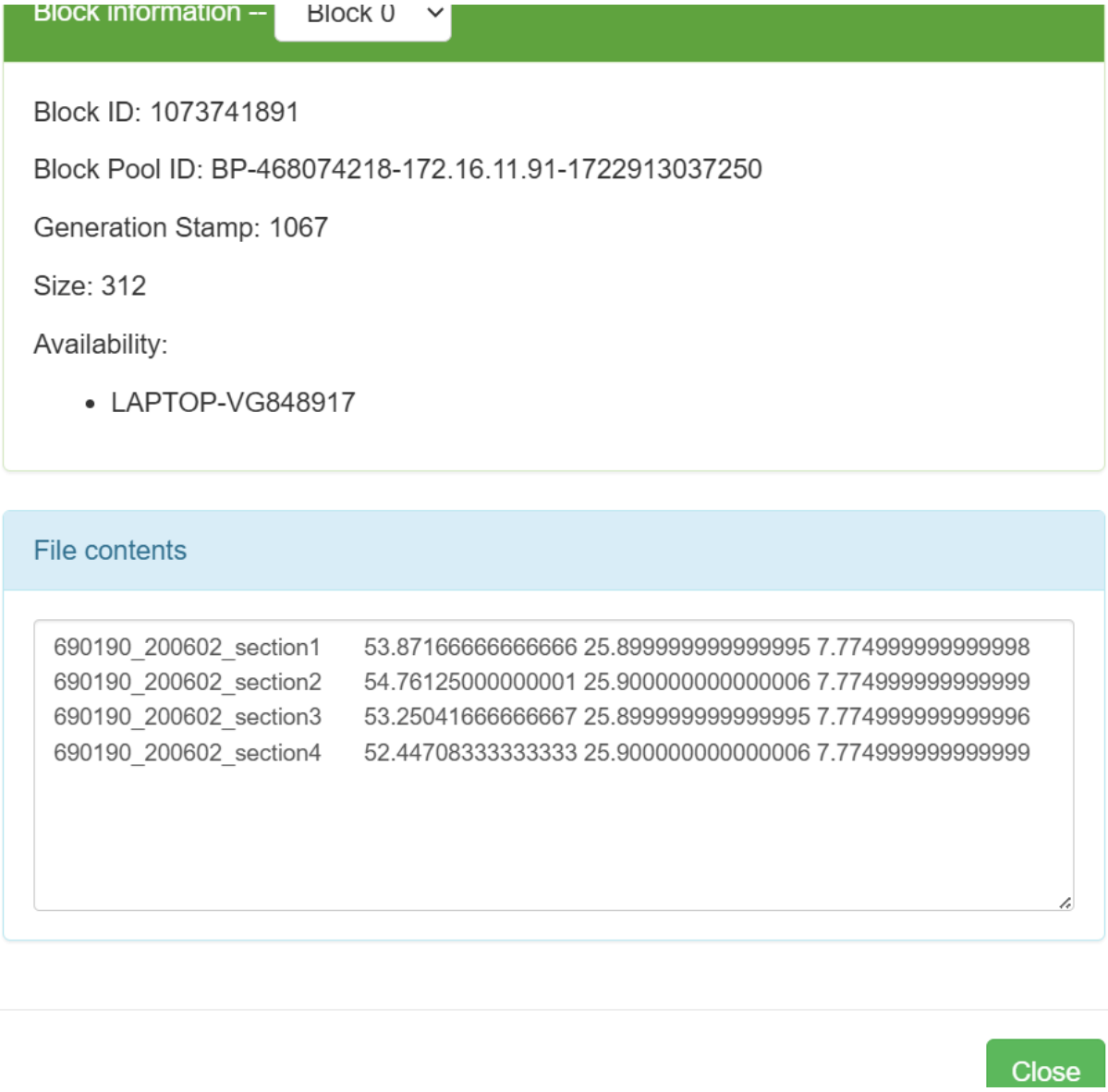
- LAPTOP-VG848917

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    0.00I 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
```

Close

EX.NO:3 **210701165**



RESULT:

Thus the implementation of the MapReduce python program a weather dataset in Hadoop is executed.