210701131
Lathika P

EX 3 IMPLEMENT A MAPREDUCE PROGRAM TO PROCESS A WEATHER DATASET

Aim:

To implement a mapreduce program to process a weather dataset using Hadoop filesystem.

Procedure:

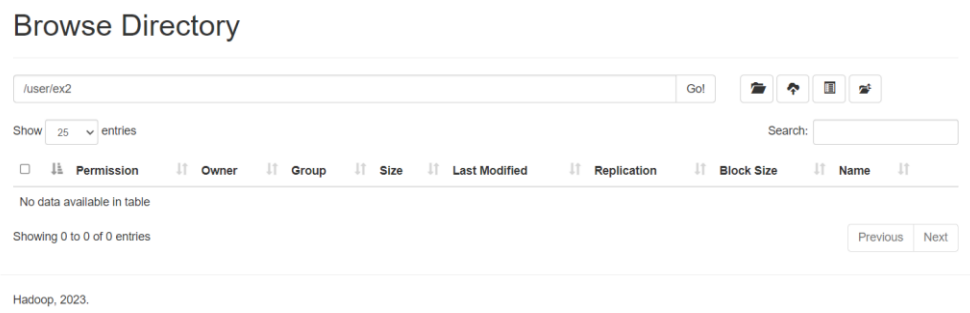
1. Start the Hadoop namenode and datanode using the command
   start-dfs.cmd
   start-yarn.cmd

2. Check if namenode and datanode are running using the command
   jps

```
C:\>jps
17248 Jps
23168 NameNode
26468 ResourceManager
22104 DataNode
16412 NodeManager
```

3. Create a directory in the Hadoop filesystem using the command

   hadoop fs -mkdir /user/ex2



   Empty directory is created.

4. Insert the input file into the directory using the command

   hadoop fs -put C:\Users\jawah\OneDrive\Desktop\LathikaDA\weather.csv /user/ex2

   //weather.csv

   date,city,temperature

   2024-08-01,New York,85

   2024-08-01,Los Angeles,90

   2024-08-01,New York,80

   2024-08-02,New York,82

   2024-08-02,Los Angeles,88

2024-08-03,Los Angeles,91

5. The MapReduce Program is written to process weather dataset.

//mapper2.py

```
#!/usr/bin/env python
import sys
import csv
def main():
    reader = csv.reader(sys.stdin)
    next(reader)  # Skip header row
    for line in reader:
        date, city, temperature = line
        try:
            temperature = float(temperature)
            print(f"{city}\t{temperature}")
        except ValueError:
            continue
if __name__ == "__main__":
    main()
```

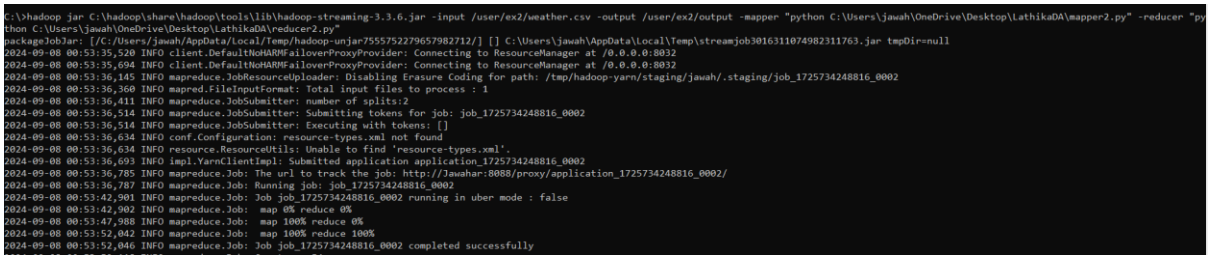
//reducer2.py

```
#!/usr/bin/env python
import sys
def main():
    current_city = None
    total_temperature = 0
    count = 0
    for line in sys.stdin:
        city, temperature = line.split('\t')
        temperature = float(temperature)
        if city == current_city:
            total_temperature += temperature
```

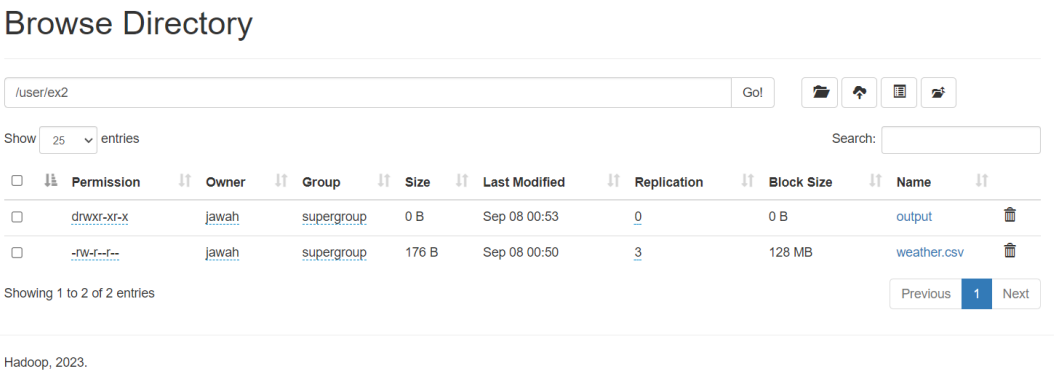
```
            count += 1
        else:
            if current_city:
                avg_temperature = total_temperature / count
                print(f"{current_city}\t{avg_temperature:.2f}")
            current_city = city
            total_temperature = temperature
            count = 1
    if current_city:
        avg_temperature = total_temperature / count
        print(f"{current_city}\t{avg_temperature:.2f}")
if __name__ == "__main__":
    main()
```

6. The mapper reducer program is executed by the following command

hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar -input /user/ex2/weather.csv -output /user/ex2/output -mapper "python C:\Users\jawah\OneDrive\Desktop\LathikaDA\mapper2.py" -reducer "python C:\Users\jawah\OneDrive\Desktop\LathikaDA\reducer2.py"



Thus the output directory is created.



7. To view the output files

```
C:\>hadoop fs -ls /user/ex2/output
Found 2 items
-rw-r--r--   3 jawah supergroup          0 2024-09-08 00:53 /user/ex2/output/_SUCCESS
-rw-r--r--   3 jawah supergroup         33 2024-09-08 00:53 /user/ex2/output/part-00000
```

hadoop fs -cat /user/ex2/output/part-00000

```
C:\>hadoop fs -cat /user/ex2/output/part-00000
Los Angeles     89.67
New York        82.50
```

8. Stop the Hadoop namenode and datanode

stop-all.cmd

Result:

Thus the mapreduce program to process a weather dataset using Hadoop filesystem is implemented successfully