

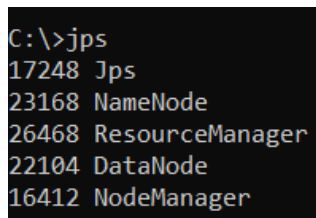
EX 3 IMPLEMENT A MAPREDUCE PROGRAM TO PROCESS A WEATHER DATASET

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

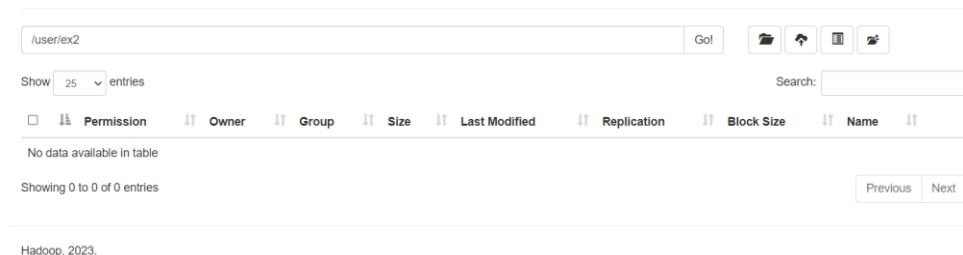


```
C:\>jps
17248 Jps
23168 NameNode
26468 ResourceManager
22104 DataNode
16412 NodeManager
```

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

hadoop fs -mkdir /user/ex2

Browse Directory



/user/ex2

Go

Show 25 entries

Search:

Permission Owner Group Size Last Modified Replication Block Size Name

No data available in table

Showing 0 to 0 of 0 entries

Previous Next

Hadoop, 2023.

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"

```

```

C:\>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"
packageJobJar: [/C:/Users/jawah/AppData/Local/Temp/hadoop-unjar7555752279657982712/] [] C:\Users\jawah\AppData\Local\Temp\streamjob3016311074982311763.jar tmpDir=null
2024-09-08 00:53:35,520 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-08 00:53:35,694 INFO client.DefaultHARMPFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-08 00:53:36,145 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/jawah/.staging/job_1725734248816_0002
2024-09-08 00:53:36,360 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-08 00:53:36,411 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-08 00:53:36,514 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725734248816_0002
2024-09-08 00:53:36,514 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-08 00:53:36,634 INFO conf.Configuration: resource-types.xml not found
2024-09-08 00:53:36,634 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
2024-09-08 00:53:36,693 INFO impl.YarnClientImpl: Submitted application application_1725734248816_0002
2024-09-08 00:53:36,785 INFO mapreduce.Job: The url to track the job: http://jawahar:8088/proxy/application_1725734248816_0002/
2024-09-08 00:53:36,787 INFO mapreduce.Job: Running job: job_1725734248816_0002
2024-09-08 00:53:42,901 INFO mapreduce.Job: Job job_1725734248816_0002 running in uber mode : false
2024-09-08 00:53:42,902 INFO mapreduce.Job: map 0% reduce 0%
2024-09-08 00:53:47,988 INFO mapreduce.Job: map 100% reduce 0%
2024-09-08 00:53:52,042 INFO mapreduce.Job: map 100% reduce 100%
2024-09-08 00:53:52,046 INFO mapreduce.Job: Job job_1725734248816_0002 completed successfully

```

Thus the output directory is created.

Browse Directory

/user/ex2

Go!

Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div>Permission</div>	<div><div></div>Owner</div>	<div><div></div>Group</div>	<div><div></div>Size</div>	<div><div></div>Last Modified</div>	<div><div></div>Replication</div>	<div><div></div>Block Size</div>	<div><div></div>Name</div>	<div><div></div></div>
<input type="checkbox"/>	drwxr-xr-x	jawah	supergroup	0 B	Sep 08 00:53	0	0 B	output	<div><div></div></div>
<input type="checkbox"/>	-rw-r--r--	jawah	supergroup	176 B	Sep 08 00:50	3	128 MB	weather.csv	<div><div></div></div>

Showing 1 to 2 of 2 entries

Previous

1

Next

Hadoop, 2023.

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