2CS702 Big Data Analytics

Lab-4 Task

Submitted by: Labdhi Sheth 18BCE101

Aim: Design MapReduce algorithms to take a very large file of integers and produce as output:

a) **The largest integer**
b) **The average of all the integers**
c) **Word count**

Codes:

1. **The largest integer**

Lab_4_MapReduce/pom.xml | FindAverageOfIntegers.java | FindMaximumInteger.java | WordCount.java

```java
public class FindMaximumInteger {

    public static class MyMapper extends Mapper<Object, Text, Text, IntWritable>{

        private Text word = new Text("Local Maximum Integer");

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

            StringTokenizer i = new StringTokenizer(value.toString());
            int maximum_value = Integer.MIN_VALUE;

            while (i.hasMoreTokens()) {
                int current_value = Integer.parseInt(i.nextToken());
                if(current_value>maximum_value)
                    maximum_value=current_value;
            }

            context.write(word, new IntWritable(maximum_value));
        }
    }

    public static class MyReducer extends Reducer<Text,IntWritable,Text,IntWritable> {

        private IntWritable final_result = new IntWritable();
        private Text word = new Text("Global Maximum Integer");

        public void reduce(Text key, Iterable<IntWritable> values,Context context) throws IOException, InterruptedException {

            int maximum_integer = Integer.MIN_VALUE;
            for (IntWritable val : values) {
                if(val.get()>maximum_integer)
                    maximum_integer=val.get();
            }

            final_result.set(maximum_integer);
            context.write(word, final_result);
        }
    }
```

2. The average of the integers:

Lab_4_MapReduce/pom.xml | FindAverageOfIntegers.java | FindMaximumInteger.java | WordCount.java

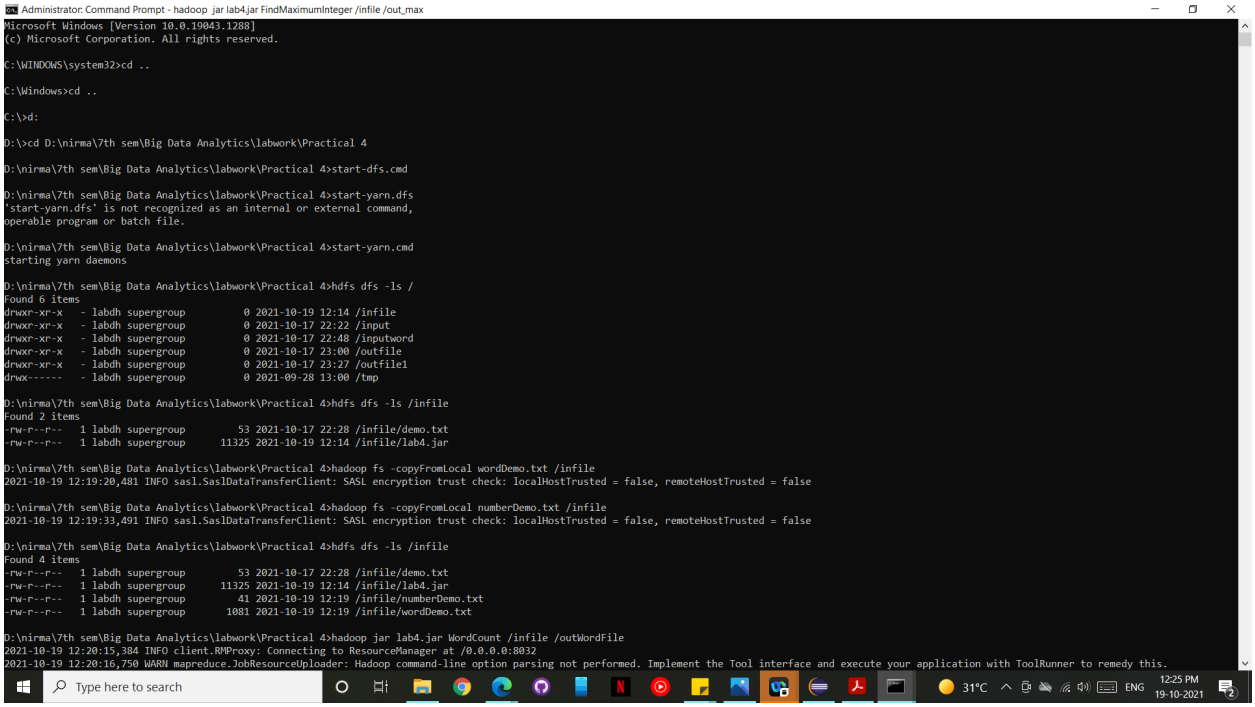
```
18    public static class MyMapper extends Mapper<Object, Text, Text, FloatWritable>{
19
20      private Text word = new Text("Average Of Integer");
21
22      public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
23          StringTokenizer i = new StringTokenizer(value.toString());
24          float local_total = 0;
25          int count=0;
26
27          while (i.hasMoreTokens()) {
28              float current_value = Float.parseFloat(i.nextToken());
29              local_total+= current_value;
30              count++;
31          }
32
33          float local_average =local_total/count;
34          context.write(word, new FloatWritable(local_average));
35      }
36    }
37
38    public static class MyReducer extends Reducer<Text,FloatWritable,Text,FloatWritable> {
39
40        private FloatWritable final_result = new FloatWritable();
41        private Text word = new Text("Global Avergae Of Integer");
42
43        public void reduce(Text key, Iterable<FloatWritable> values,Context context) throws IOException, InterruptedException {
44
45            float total = 0;
46            int count=0;
47
48            for (FloatWritable val : values) {
49
50                total+=val.get();
51                count++;
52            }
53
54           float global_average = total/count;
55           final_result.set(global_average);
56           context.write(word, final_result);
```

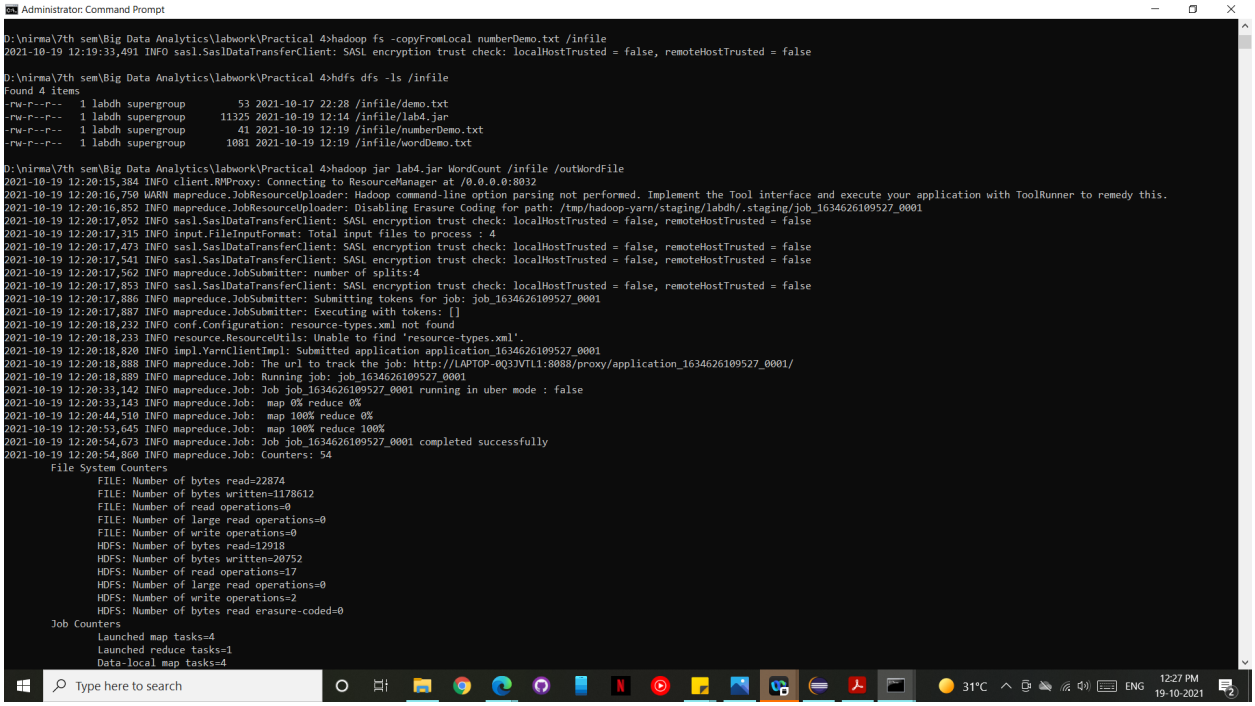
3. Word count:

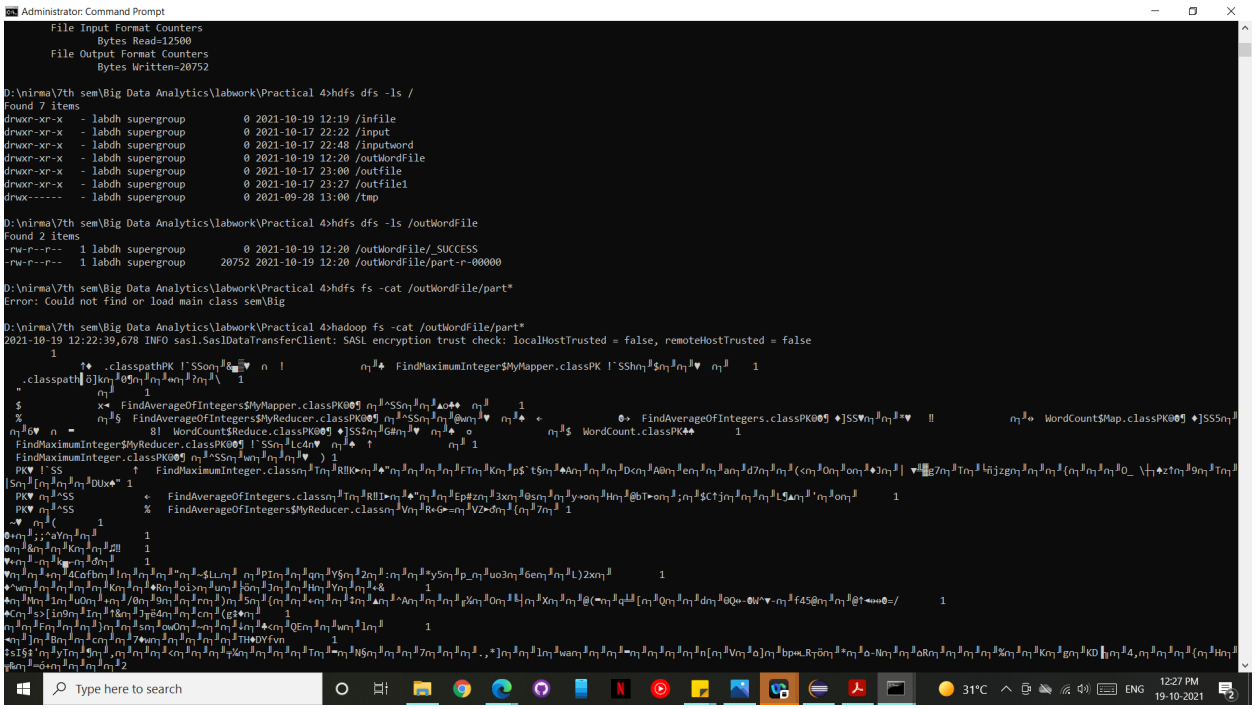
Lab_4_MapReduce/pom.xml | FindAverageOfIntegers.java | FindMaximumInteger.java | WordCount.java

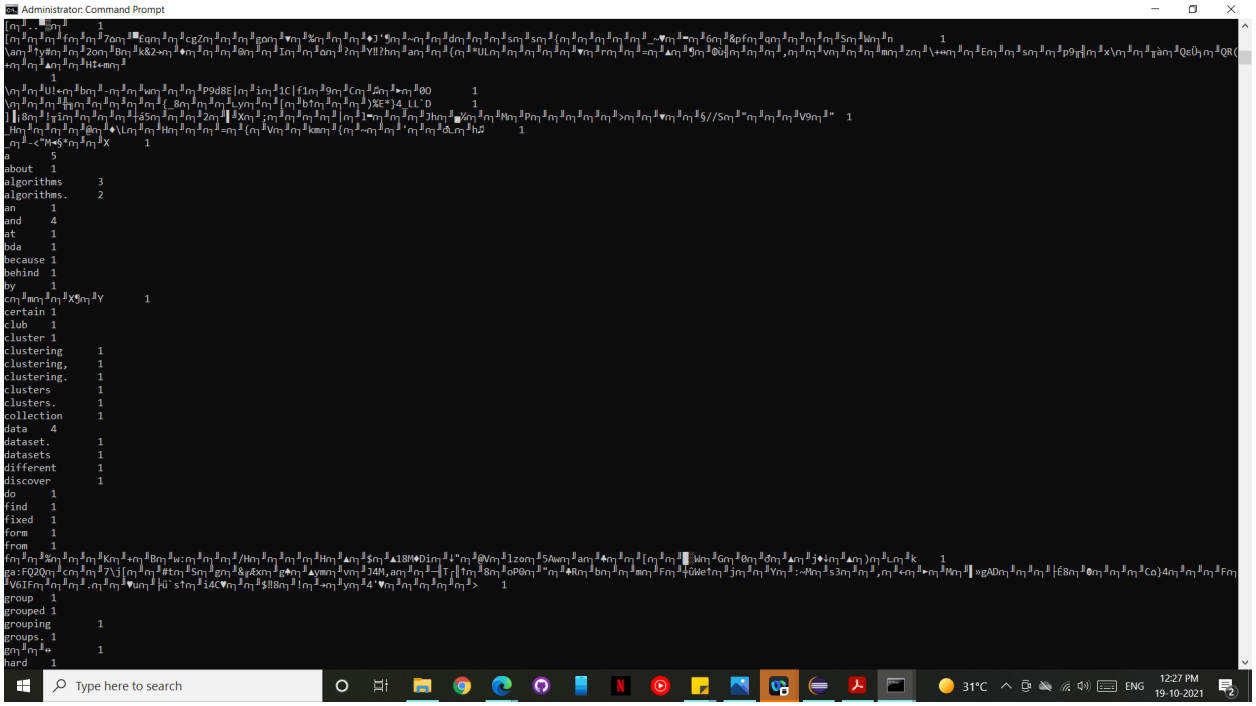
```
1  import java.io.IOException;
14
15 public class WordCount {
16
17     public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
18
19         @Override
20         public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
21
22             StringTokenizer str = new StringTokenizer(value.toString());
23
24             while (str.hasMoreTokens()) {
25                 String word = str.nextToken();
26
27                 context.write(new Text(word), new IntWritable(1));
28             }
29         }
30     }
31
32     public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
33
34         @Override
35         public void reduce(Text key, Iterable<IntWritable> values, Context context)
36                 throws IOException, InterruptedException {
37             int sum = 0;
38             for (IntWritable i : values) {
39                 sum += i.get();
40             }
41
42             context.write(key, new IntWritable(sum));
43         }
44     }
45
46     public static void main(String[] args) throws Exception {
47
48         if (args.length != 2) {
49             System.err.println("Usage: WordCount <InPath> <OutPath>");
50             System.exit(2);
51         }
```

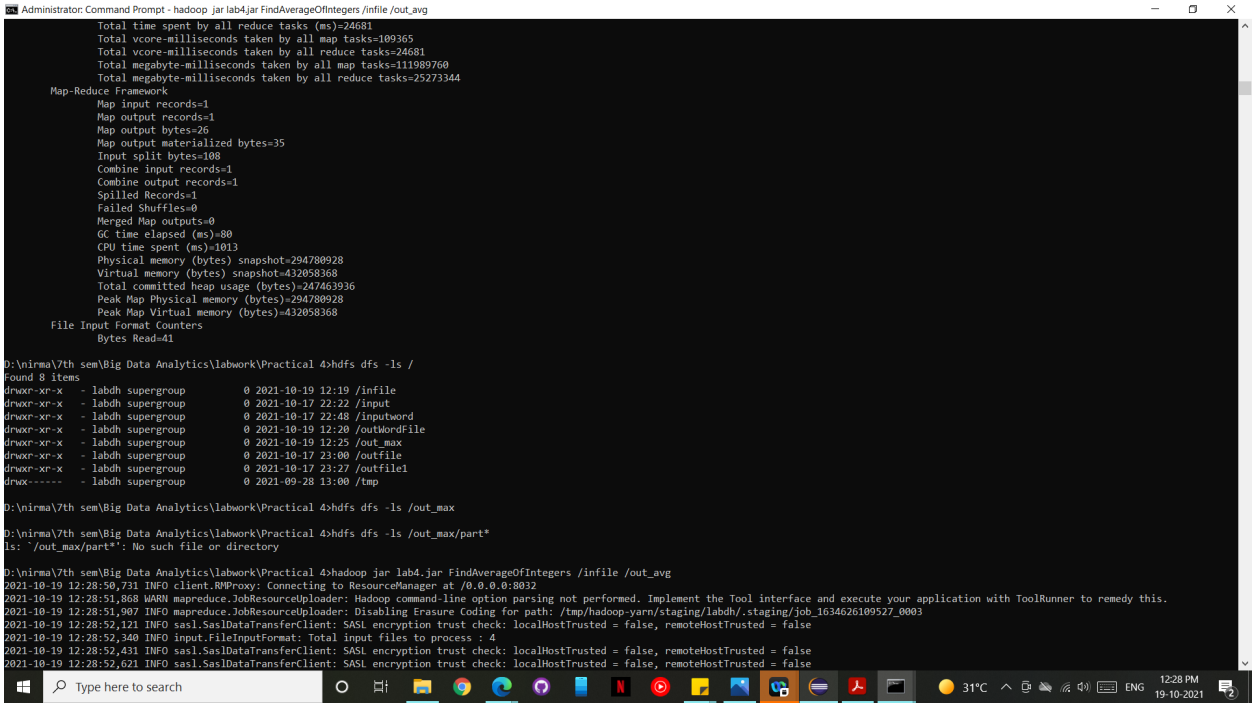
Output:


Administrator: Command Prompt - hadoop jar lab4.jar FindMaximumInteger /infile /out_max
Microsoft Windows [Version 10.0.19043.1288]
(c) Microsoft Corporation. All rights reserved.
C:\WINDOWS\system32>cd ..
C:\Windows>cd ..
C:\>d:
D:\>cd D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>start-dfs.cmd
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>start-yarn.dfs
'start-yarn.dfs' is not recognized as an internal or external command,
operable program or batch file.
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>start-yarn.cmd
starting yarn daemons
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /
Found 6 items
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:14 /infile
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:22 /input
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:48 /inputword
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:00 /outfile
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:27 /outfile1
drwx------ - labdh supergroup 0 2021-09-28 13:00 /tmp
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /infile
Found 2 items
-rw-r--r-- 1 labdh supergroup 53 2021-10-17 22:28 /infile/demo.txt
-rw-r--r-- 1 labdh supergroup 11325 2021-10-19 12:14 /infile/lab4.jar
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop fs -copyFromLocal wordDemo.txt /infile
2021-10-19 12:19:20,481 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop fs -copyFromLocal numberDemo.txt /infile
2021-10-19 12:19:33,491 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /infile
Found 4 items
-rw-r--r-- 1 labdh supergroup 53 2021-10-17 22:28 /infile/demo.txt
-rw-r--r-- 1 labdh supergroup 11325 2021-10-19 12:14 /infile/lab4.jar
-rw-r--r-- 1 labdh supergroup 41 2021-10-19 12:19 /infile/numberDemo.txt
-rw-r--r-- 1 labdh supergroup 1081 2021-10-19 12:19 /infile/wordDemo.txt
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop jar lab4.jar WordCount /infile /outWordFile
2021-10-19 12:20:15,384 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-19 12:20:16,750 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Type here to search
31°C
ENG
12:25 PM
19-10-2021


Administrator: Command Prompt
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop fs -copyFromLocal numberDemo.txt /infile
2021-10-19 12:19:33,491 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /infile
Found 4 items
-rw-r--r--   1 labdh supergroup   53 2021-10-17 22:28 /infile/demo.txt
-rw-r--r--   1 labdh supergroup   11325 2021-10-19 12:14 /infile/lab4.jar
-rw-r--r--   1 labdh supergroup   41 2021-10-19 12:19 /infile/numberDemo.txt
-rw-r--r--   1 labdh supergroup   1081 2021-10-19 12:19 /infile/wordDemo.txt
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop jar lab4.jar WordCount /infile /outWordFile
2021-10-19 12:20:15,384 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-19 12:20:16,750 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-10-19 12:20:16,852 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1634626109527_0001
2021-10-19 12:20:17,052 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:20:17,315 INFO input.FileInputFormat: Total input files to process : 4
2021-10-19 12:20:17,473 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:20:17,541 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:20:17,562 INFO mapreduce.JobSubmitter: number of splits:4
2021-10-19 12:20:17,853 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:20:17,886 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634626109527_0001
2021-10-19 12:20:17,887 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-19 12:20:18,232 INFO conf.Configuration: resource-types.xml not found
2021-10-19 12:20:18,233 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-19 12:20:18,820 INFO impl.YarnClientImpl: Submitted application application_1634626109527_0001
2021-10-19 12:20:18,888 INFO mapreduce.Job: The url to track the job: http://LAPTOP-0Q3JVTL1:8088/proxy/application_1634626109527_0001/
2021-10-19 12:20:18,889 INFO mapreduce.Job: Running job: job_1634626109527_0001
2021-10-19 12:20:33,142 INFO mapreduce.Job: Job job_1634626109527_0001 running in uber mode : false
2021-10-19 12:20:33,143 INFO mapreduce.Job: map 0% reduce 0%
2021-10-19 12:20:44,510 INFO mapreduce.Job: map 100% reduce 0%
2021-10-19 12:20:53,645 INFO mapreduce.Job: map 100% reduce 100%
2021-10-19 12:20:54,673 INFO mapreduce.Job: Job job_1634626109527_0001 completed successfully
2021-10-19 12:20:54,860 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=22874
FILE: Number of bytes written=1178612
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=12918
HDFS: Number of bytes written=20752
HDFS: Number of read operations=17
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=4
Launched reduce tasks=1
Data-local map tasks=4
Type here to search
31°C
ENG
12:27 PM
19-10-2021


Administrator: Command Prompt
File Input Format Counters
Bytes Read=12500
File Output Format Counters
Bytes Written=20752
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - labdh supergroup   0 2021-10-19 12:19 /infile
drwxr-xr-x   - labdh supergroup   0 2021-10-17 22:22 /input
drwxr-xr-x   - labdh supergroup   0 2021-10-17 22:48 /inputword
drwxr-xr-x   - labdh supergroup   0 2021-10-19 12:20 /outWordFile
drwxr-xr-x   - labdh supergroup   0 2021-10-17 23:00 /outfile
drwxr-xr-x   - labdh supergroup   0 2021-10-17 23:27 /outfile1
drwx------   - labdh supergroup   0 2021-09-28 13:00 /tmp
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /outWordFile
Found 2 items
-rw-r--r--   1 labdh supergroup   0 2021-10-19 12:20 /outWordFile/_SUCCESS
-rw-r--r--   1 labdh supergroup   20752 2021-10-19 12:20 /outWordFile/part-r-00000
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs fs -cat /outWordFile/part*
Error: Could not find or load main class sem\Big
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop fs -cat /outWordFile/part*
2021-10-19 12:22:39,678 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Type here to search
31°C
ENG
12:27 PM
19-10-2021


Administrator: Command Prompt
algorithms 3
algorithms. 2
an 1
and 4
at 1
bda 1
because 1
behind 1
by 1
certain 1
club 1
cluster 1
clustering 1
clustering, 1
clustering. 1
clusters 1
clusters. 1
collection 1
data 4
dataset. 1
datasets 1
different 1
discover 1
do 1
find 1
fixed 1
form 1
from 1
group 1
grouped 1
grouping 1
groups. 1
hard 1
Type here to search
31°C
ENG
12:27 PM
19-10-2021


Administrator: Command Prompt - hadoop jar lab4.jar FindAverageOfIntegers /infile /out_avg
Total time spent by all reduce tasks (ms)=24681
Total vcore-milliseconds taken by all map tasks=109365
Total vcore-milliseconds taken by all reduce tasks=24681
Total megabyte-milliseconds taken by all map tasks=111989760
Total megabyte-milliseconds taken by all reduce tasks=25273344
Map-Reduce Framework
Map input records=1
Map output records=1
Map output bytes=26
Map output materialized bytes=35
Input split bytes=108
Combine input records=1
Combine output records=1
Spilled Records=1
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=80
CPU time spent (ms)=1013
Physical memory (bytes) snapshot=294780928
Virtual memory (bytes) snapshot=432058368
Total committed heap usage (bytes)=247463936
Peak Map Physical memory (bytes)=294780928
Peak Map Virtual memory (bytes)=432058368
File Input Format Counters
Bytes Read=41
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /
Found 8 items
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:19 /infile
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:22 /input
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:48 /inputword
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:20 /outWordFile
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:25 /out_max
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:00 /outfile
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:27 /outfile1
drwx------ - labdh supergroup 0 2021-09-28 13:00 /tmp
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /out_max
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hdfs dfs -ls /out_max/part*
ls: `/out_max/part*': No such file or directory
D:\nirma\7th sem\Big Data Analytics\labwork\Practical 4>hadoop jar lab4.jar FindAverageOfIntegers /infile /out_avg
2021-10-19 12:28:50,731 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-19 12:28:51,868 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-10-19 12:28:51,907 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1634626109527_0003
2021-10-19 12:28:52,121 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:28:52,340 INFO input.FileInputFormat: Total input files to process : 4
2021-10-19 12:28:52,431 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 12:28:52,621 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Type here to search
31°C
ENG
12:28 PM
19-10-2021