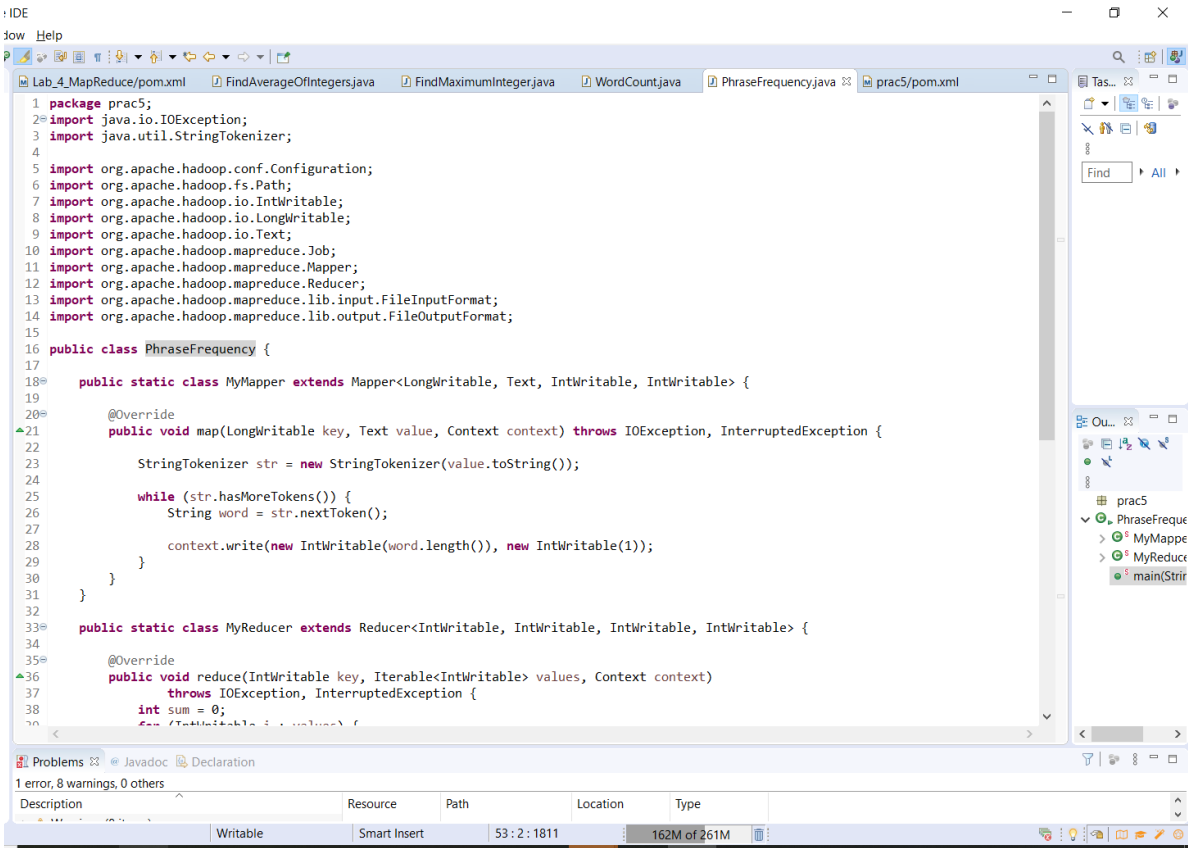
2CS702 Big Data Analytics

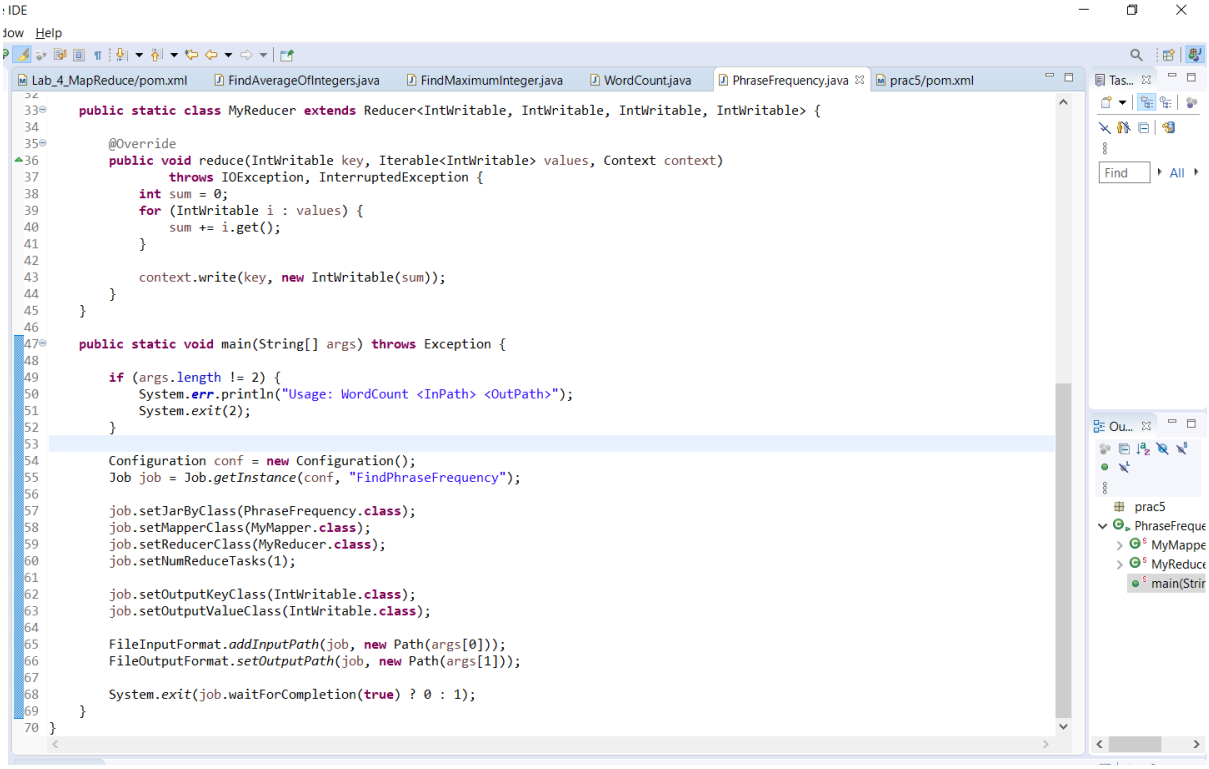
Lab-5 Task

Submitted by: Labdhi Sheth 18BCE101

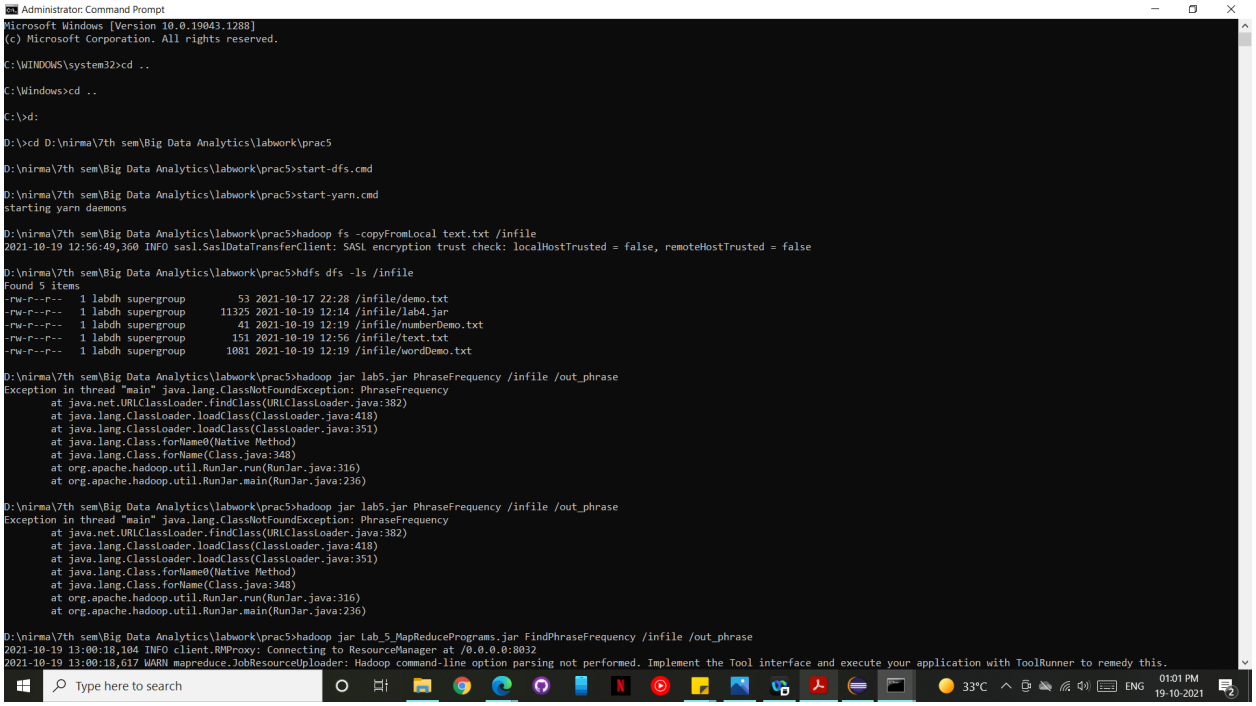
Aim: **Apply MapReduce algorithms to find phrase frequency from a given dataset.**

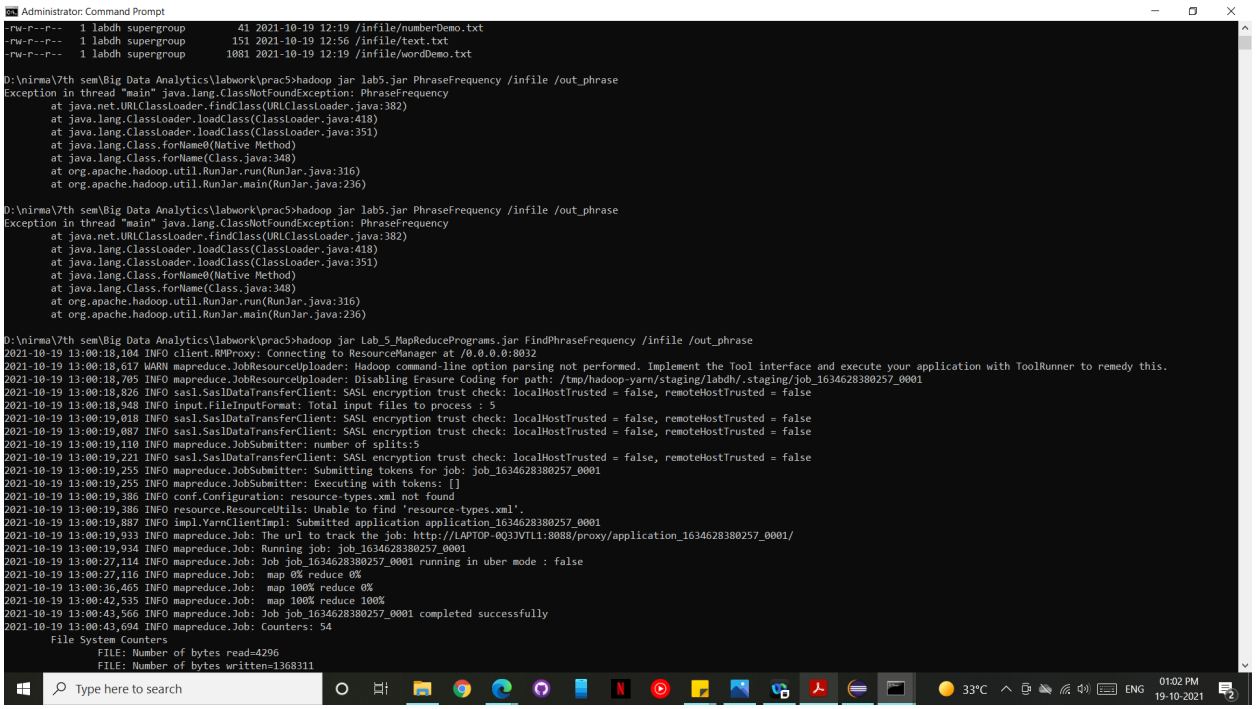
Code:

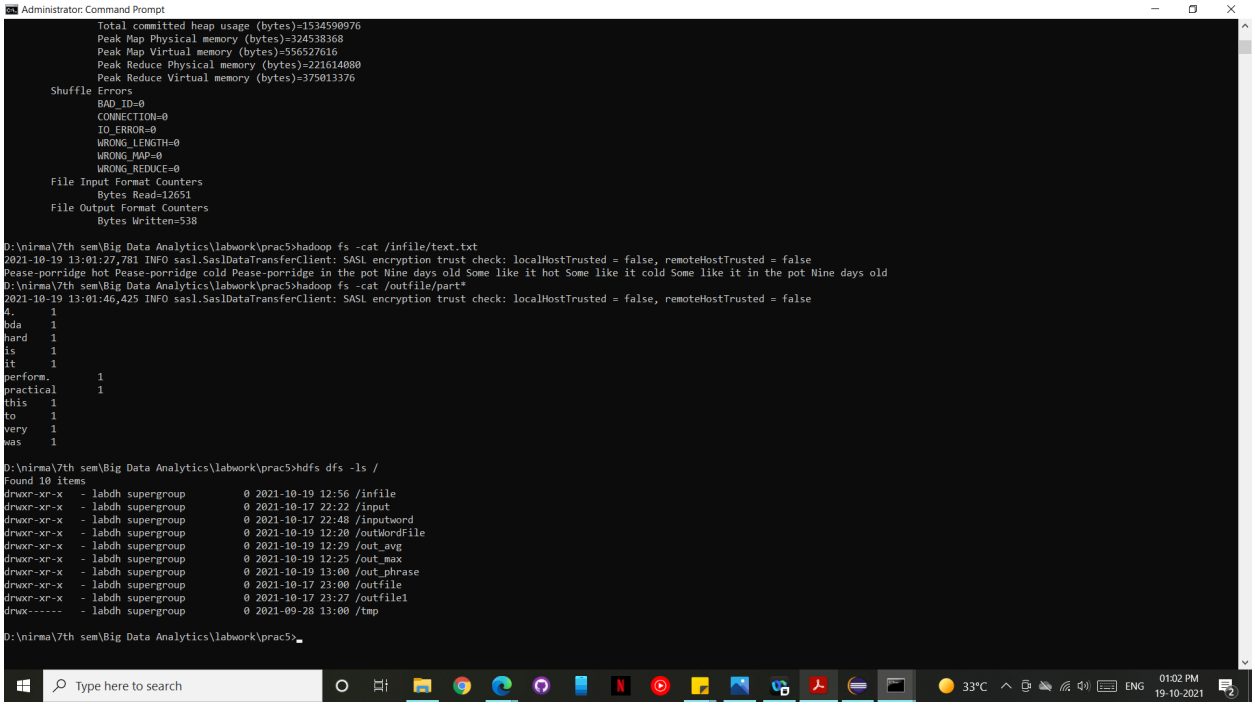

IDE
dow Help
Lab_4_MapReduce/pom.xml
FindAverageOfIntegers.java
FindMaximumInteger.java
WordCount.java
PhraseFrequency.java
prac5/pom.xml
package prac5;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class PhraseFrequency {
public static class MyMapper extends Mapper<LongWritable, Text, IntWritable, IntWritable> {
@Override
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
StringTokenizer str = new StringTokenizer(value.toString());
while (str.hasMoreTokens()) {
String word = str.nextToken();
context.write(new IntWritable(word.length()), new IntWritable(1));
}
}
}
public static class MyReducer extends Reducer<IntWritable, IntWritable, IntWritable, IntWritable> {
@Override
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
int sum = 0;
Problems
Javadoc
Declaration
1 error, 8 warnings, 0 others
Description
Resource
Path
Location
Type
Writable
Smart Insert
53 : 2 : 1811
162M of 261M


IDE
dow Help
Lab_4_MapReduce/pom.xml
FindAverageOfIntegers.java
FindMaximumInteger.java
WordCount.java
PhraseFrequency.java
prac5/pom.xml
public static class MyReducer extends Reducer<IntWritable, IntWritable, IntWritable, IntWritable> {
@Override
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
int sum = 0;
for (IntWritable i : values) {
sum += i.get();
}
context.write(key, new IntWritable(sum));
}
}
public static void main(String[] args) throws Exception {
if (args.length != 2) {
System.err.println("Usage: WordCount <InPath> <OutPath>");
System.exit(2);
}
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "FindPhraseFrequency");
job.setJarByClass(PhraseFrequency.class);
job.setMapperClass(MyMapper.class);
job.setReducerClass(MyReducer.class);
job.setNumReduceTasks(1);
job.setOutputKeyClass(IntWritable.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}

Output:


Administrator: Command Prompt
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\prac5
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>start-dfs.cmd
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>start-yarn.cmd
starting yarn daemons
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop fs -copyFromLocal text.txt /infile
2021-10-19 12:56:49,360 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hdfs dfs -ls /infile
Found 5 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   151 2021-10-19 12:56 /infile/text.txt
-rw-r--r--   1 labdh supergroup   1081 2021-10-19 12:19 /infile/wordDemo.txt
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar lab5.jar PhraseFrequency /infile /out_phrase
Exception in thread "main" java.lang.ClassNotFoundException: PhraseFrequency
at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
at java.lang.Class.forName0(Native Method)
at java.lang.Class.forName(Class.java:348)
at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar lab5.jar PhraseFrequency /infile /out_phrase
Exception in thread "main" java.lang.ClassNotFoundException: PhraseFrequency
at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
at java.lang.Class.forName0(Native Method)
at java.lang.Class.forName(Class.java:348)
at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar Lab_5_MapReducePrograms.jar FindPhraseFrequency /infile /out_phrase
2021-10-19 13:00:18,104 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-19 13:00:18,617 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
33°C
ENG
01:01 PM
19-10-2021


Administrator: Command Prompt
-rw-r--r--   1 labdh supergroup   41 2021-10-19 12:19 /infile/numberDemo.txt
-rw-r--r--   1 labdh supergroup   151 2021-10-19 12:56 /infile/text.txt
-rw-r--r--   1 labdh supergroup   1081 2021-10-19 12:19 /infile/wordDemo.txt
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar lab5.jar PhraseFrequency /infile /out_phrase
Exception in thread "main" java.lang.ClassNotFoundException: PhraseFrequency
at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
at java.lang.Class.forName0(Native Method)
at java.lang.Class.forName(Class.java:348)
at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar lab5.jar PhraseFrequency /infile /out_phrase
Exception in thread "main" java.lang.ClassNotFoundException: PhraseFrequency
at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
at java.lang.Class.forName0(Native Method)
at java.lang.Class.forName(Class.java:348)
at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop jar Lab_5_MapReducePrograms.jar FindPhraseFrequency /infile /out_phrase
2021-10-19 13:00:18,104 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-19 13:00:18,617 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 13:00:18,705 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1634628380257_0001
2021-10-19 13:00:18,826 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 13:00:18,948 INFO input.FileInputFormat: Total input files to process : 5
2021-10-19 13:00:19,018 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 13:00:19,087 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 13:00:19,110 INFO mapreduce.JobSubmitter: number of splits:5
2021-10-19 13:00:19,221 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-10-19 13:00:19,255 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634628380257_0001
2021-10-19 13:00:19,255 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-19 13:00:19,386 INFO conf.Configuration: resource-types.xml not found
2021-10-19 13:00:19,386 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-19 13:00:19,887 INFO impl.YarnClientImpl: Submitted application application_1634628380257_0001
2021-10-19 13:00:19,933 INFO mapreduce.Job: The url to track the job: http://LAPTOP-0Q3JVTL1:8088/proxy/application_1634628380257_0001/
2021-10-19 13:00:19,934 INFO mapreduce.Job: Running job: job_1634628380257_0001
2021-10-19 13:00:27,114 INFO mapreduce.Job: Job job_1634628380257_0001 running in uber mode : false
2021-10-19 13:00:27,116 INFO mapreduce.Job: map 0% reduce 0%
2021-10-19 13:00:36,465 INFO mapreduce.Job: map 100% reduce 0%
2021-10-19 13:00:42,535 INFO mapreduce.Job: map 100% reduce 100%
2021-10-19 13:00:43,566 INFO mapreduce.Job: Job job_1634628380257_0001 completed successfully
2021-10-19 13:00:43,694 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=4296
FILE: Number of bytes written=1368311
Type here to search
33°C
ENG
01:02 PM
19-10-2021


Administrator: Command Prompt
Total committed heap usage (bytes)=1534590976
Peak Map Physical memory (bytes)=324538368
Peak Map Virtual memory (bytes)=556527616
Peak Reduce Physical memory (bytes)=221614080
Peak Reduce Virtual memory (bytes)=375013376
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=12651
File Output Format Counters
Bytes Written=538
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop fs -cat /infile/text.txt
2021-10-19 13:01:27,781 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Pease-porridge hot Pease-porridge cold Pease-porridge in the pot Nine days old Some like it hot Some like it cold Some like it in the pot Nine days old
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hadoop fs -cat /outfile/part*
2021-10-19 13:01:46,425 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
4. 1
bda 1
hard 1
is 1
it 1
perform. 1
practical 1
this 1
to 1
very 1
was 1
D:\nirma\7th sem\Big Data Analytics\labwork\prac5>hdfs dfs -ls /
Found 10 items
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:56 /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 12:29 /out_avg
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:25 /out_max
drwxr-xr-x - labdh supergroup 0 2021-10-19 13:00 /out_phrase
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\prac5>
Type here to search
33°C
ENG
01:02 PM
19-10-2021