LAB – 4

Name: Gandevia Keval Dharmeshbhai

Sem: VII

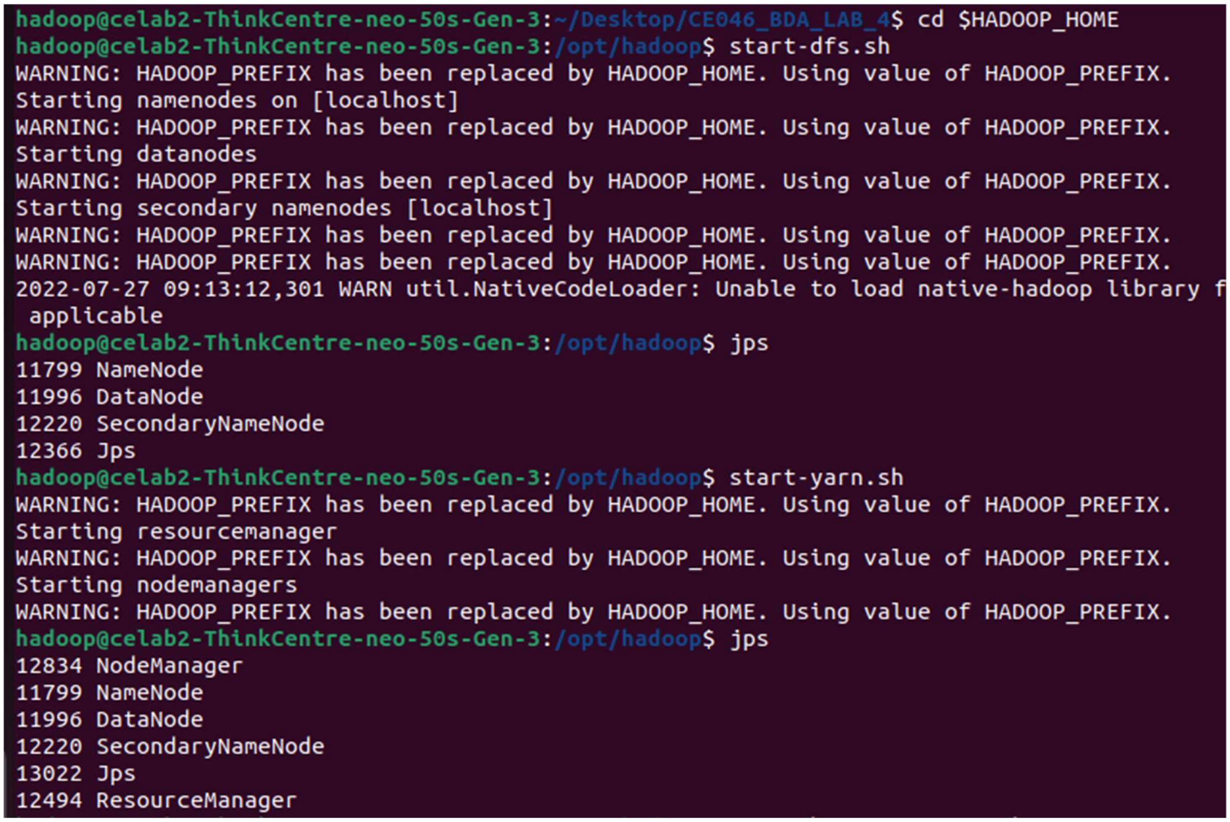
Roll No: CE046

Subject: Big Data and Analytics

Aim: Write a map-reduce program to count the frequencies of word from distributed storage source and understand the phases involved in map-reduce programming.

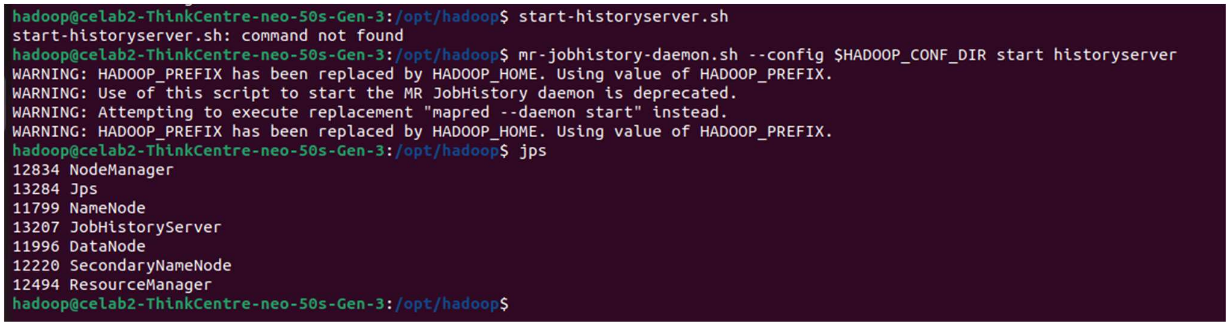
Q. 1: Wordcount program with map-reduce

(1)Starting the Hadoop server and yarn.



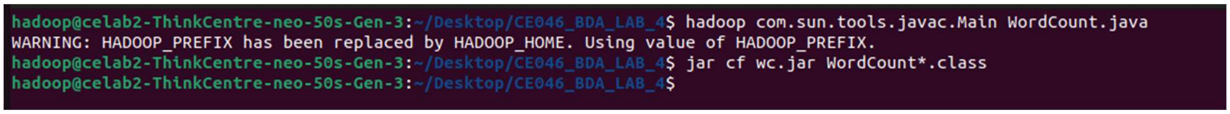
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ cd $HADOOP_HOME
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ start-dfs.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting namenodes on [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting datanodes
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting secondary namenodes [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:13:12,301 WARN util.NativeCodeLoader: Unable to load native-hadoop library f
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ jps
11799 NameNode
11996 DataNode
12220 SecondaryNameNode
12366 Jps
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ start-yarn.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting resourcemanager
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting nodemanagers
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ jps
12834 NodeManager
11799 NameNode
11996 DataNode
12220 SecondaryNameNode
13022 Jps
12494 ResourceManager
```

(2)Starting history server and verifying using jps. As we can see NameNode, DataNode, and SecodaryNameNode are the processes spawned by the Hadoop server. NodeManager and ResourceManager are the processes spawned by the yarn. Moreover, JobHistoryServer is spawned by the history server.



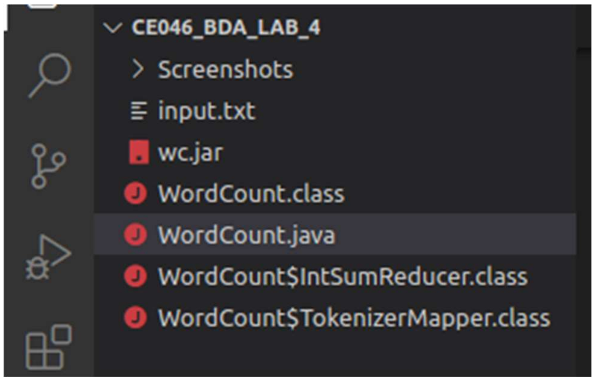
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ start-historyserver.sh
start-historyserver.sh: command not found
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ mr-jobhistory-daemon.sh --config $HADOOP_CONF_DIR start historyserver
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to start the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon start" instead.
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ jps
12834 NodeManager
13284 Jps
11799 NameNode
13207 JobHistoryServer
11996 DataNode
12220 SecondaryNameNode
12494 ResourceManager
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$
```

(3)Compiling program and creating a jar file.



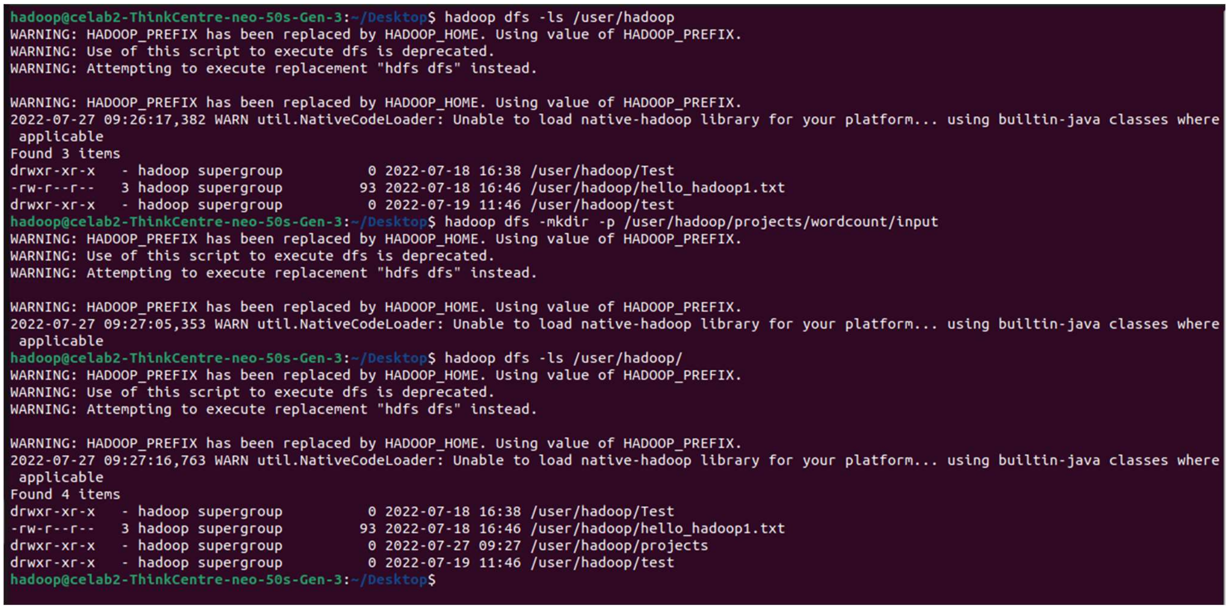
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hadoop com.sun.tools.javac.Main WordCount.java
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ jar cf wc.jar WordCount*.class
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$
```

(4)Program is complied successfully.

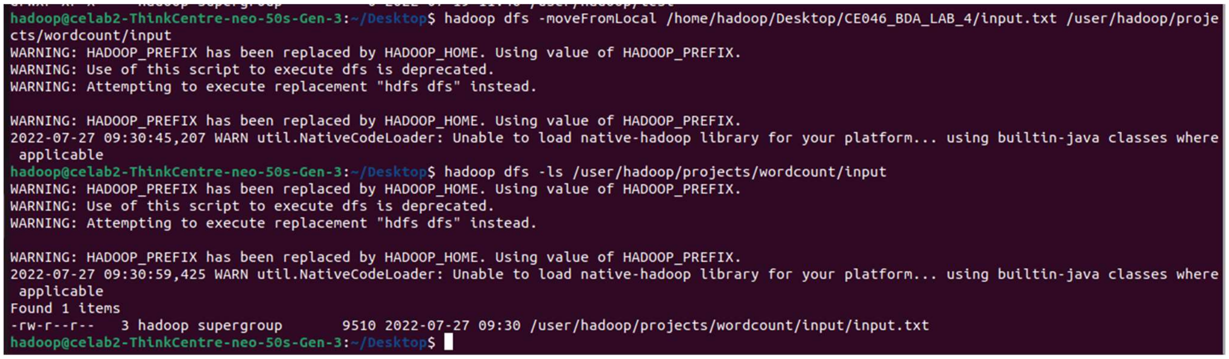


```
CE046_BDA_LAB_4
  > Screenshots
  input.txt
  wc.jar
  WordCount.class
  WordCount.java
  WordCount$IntSumReducer.class
  WordCount$TokenizerMapper.class
```

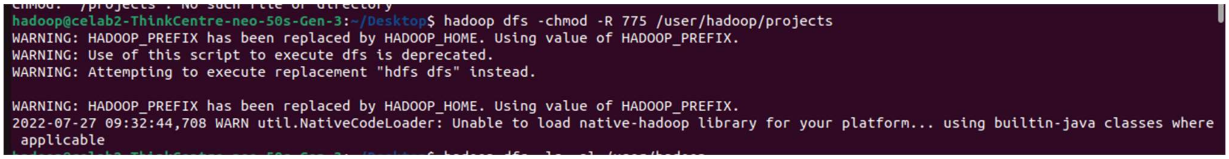
(5)Creating directory and uploading a text file into the HDFS.



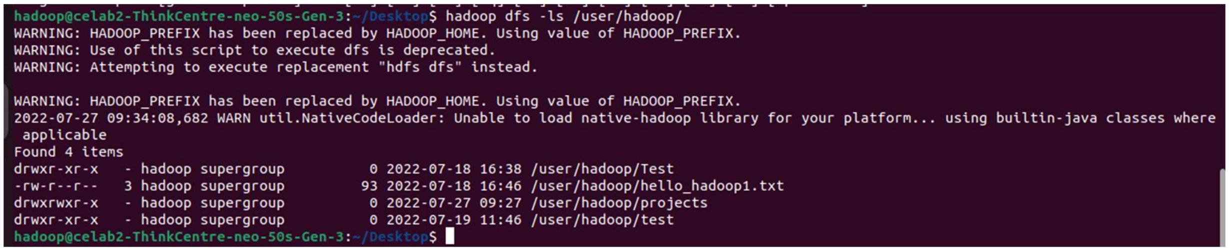
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -ls /user/hadoop
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:26:17,382 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2022-07-18 16:38 /user/hadoop/Test
-rw-r--r--   3 hadoop supergroup         93 2022-07-18 16:46 /user/hadoop/hello_hadoop1.txt
drwxr-xr-x   - hadoop supergroup          0 2022-07-19 11:46 /user/hadoop/test
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -mkdir -p /user/hadoop/projects/wordcount/input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:27:05,353 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -ls /user/hadoop/
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:27:16,763 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2022-07-18 16:38 /user/hadoop/Test
-rw-r--r--   3 hadoop supergroup         93 2022-07-18 16:46 /user/hadoop/hello_hadoop1.txt
drwxr-xr-x   - hadoop supergroup          0 2022-07-27 09:27 /user/hadoop/projects
drwxr-xr-x   - hadoop supergroup          0 2022-07-19 11:46 /user/hadoop/test
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$
```



```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -moveFromLocal /home/hadoop/Desktop/CE046_BDA_LAB_4/input.txt /user/hadoop/proje
cts/wordcount/input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:30:45,207 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -ls /user/hadoop/projects/wordcount/input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:30:59,425 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 1 items
-rw-r--r--   3 hadoop supergroup       9510 2022-07-27 09:30 /user/hadoop/projects/wordcount/input/input.txt
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$
```

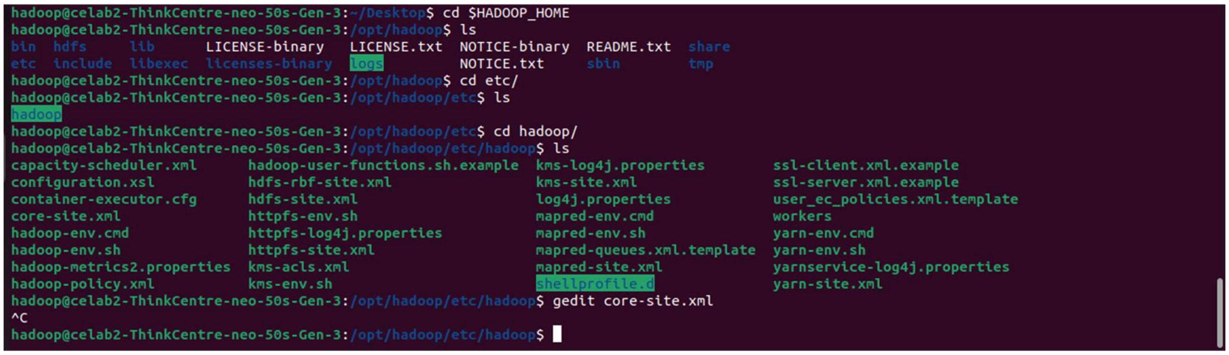


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -chmod -R 775 /user/hadoop/projects
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:32:44,708 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
```

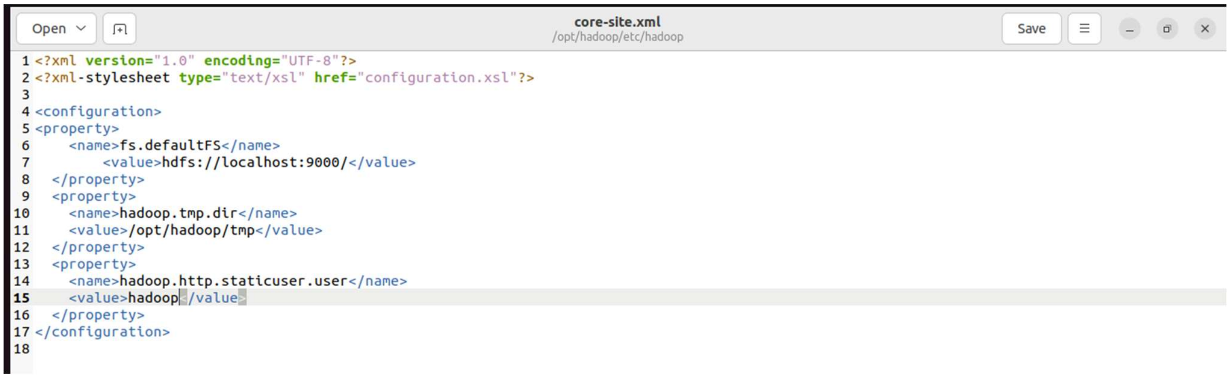


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ hadoop dfs -ls /user/hadoop/
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:34:08,682 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2022-07-18 16:38 /user/hadoop/Test
-rw-r--r--   3 hadoop supergroup         93 2022-07-18 16:46 /user/hadoop/hello_hadoop1.txt
drwxrwxr-x   - hadoop supergroup          0 2022-07-27 09:27 /user/hadoop/projects
drwxr-xr-x   - hadoop supergroup          0 2022-07-19 11:46 /user/hadoop/test
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$
```

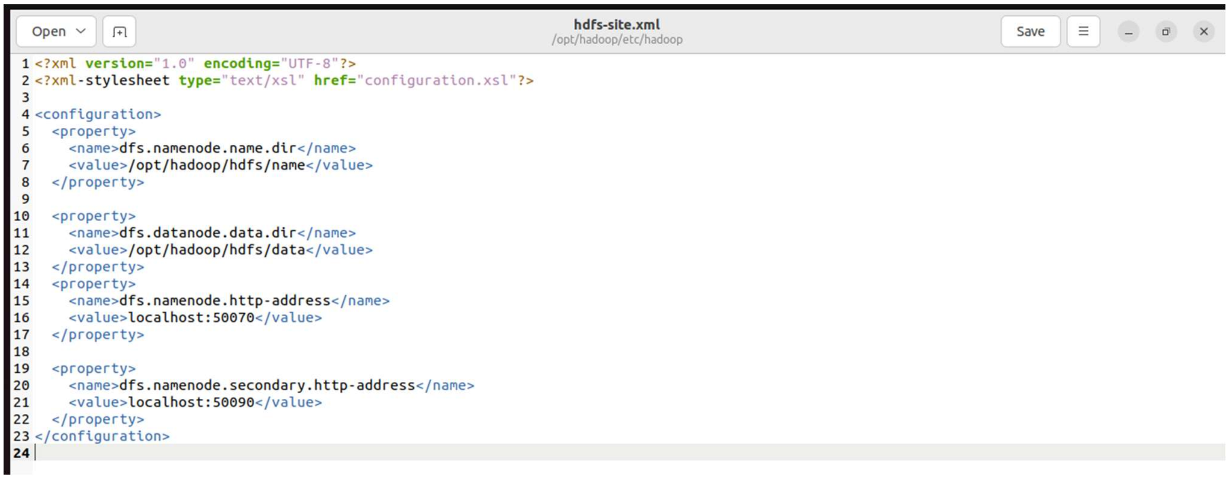
(6)Editing configuration files.



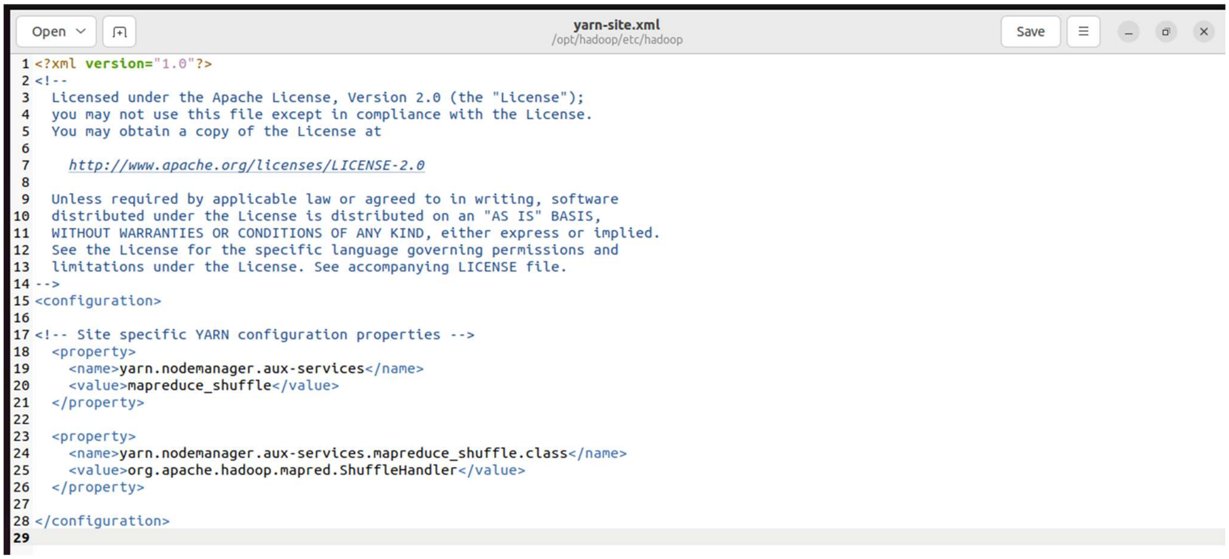
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop$ cd $HADOOP_HOME
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ ls
bin  hdfs     lib      LICENSE-binary   LICENSE.txt  NOTICE-binary  README.txt  share
etc  include  libexec  licenses-binary  logs         NOTICE.txt     sbin        tmp
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop$ cd etc/
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc$ ls
hadoop
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc$ cd hadoop/
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ ls
capacity-scheduler.xml      hadoop-user-functions.sh.example  kms-log4j.properties        ssl-client.xml.example
configuration.xsl           hdfs-rbf-site.xml                 kms-site.xml                ssl-server.xml.example
container-executor.cfg      hdfs-site.xml                     log4j.properties            user_ec_policies.xml.template
core-site.xml               httpfs-env.sh                     mapred-env.cmd              workers
hadoop-env.cmd              httpfs-log4j.properties           mapred-env.sh               yarn-env.cmd
hadoop-env.sh               httpfs-site.xml                   mapred-queues.xml.template  yarn-env.sh
hadoop-metrics2.properties  kms-acls.xml                      mapred-site.xml             yarnservice-log4j.properties
hadoop-policy.xml           kms-env.sh                        shellprofile.d              yarn-site.xml
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ gedit core-site.xml
^C
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$
```



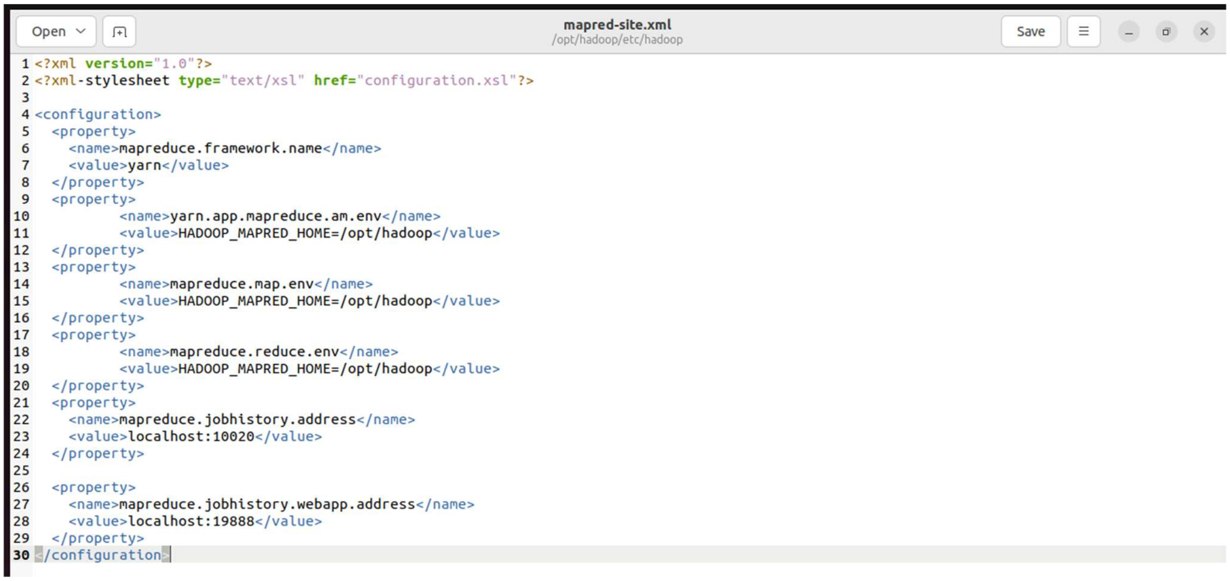
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
<property>
    <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000/</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/hadoop/tmp</value>
  </property>
  <property>
    <name>hadoop.http.staticuser.user</name>
    <value>hadoop</value>
  </property>
</configuration>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/opt/hadoop/hdfs/name</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/opt/hadoop/hdfs/data</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>localhost:50070</value>
  </property>

  <property>
    <name>dfs.namenode.secondary.http-address</name>
    <value>localhost:50090</value>
  </property>
</configuration>
```

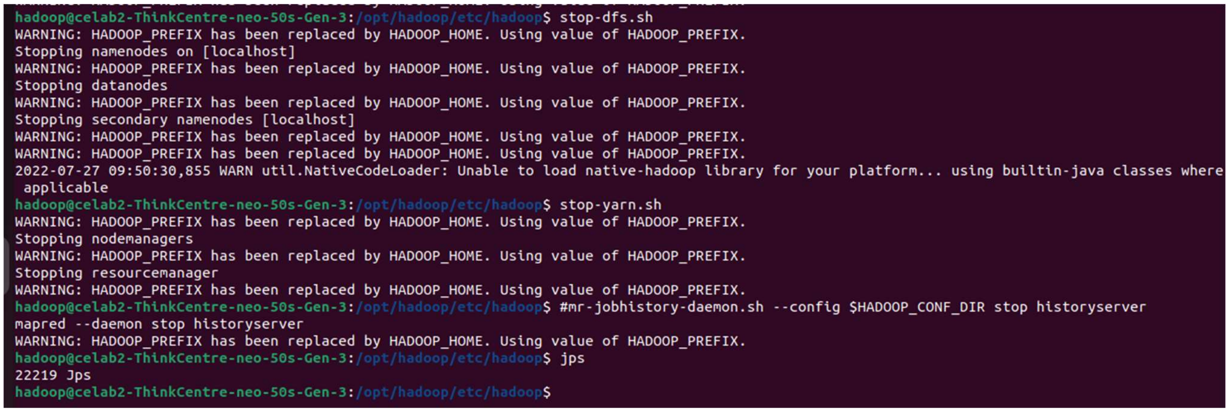


```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

  <property>
    <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>

</configuration>
```

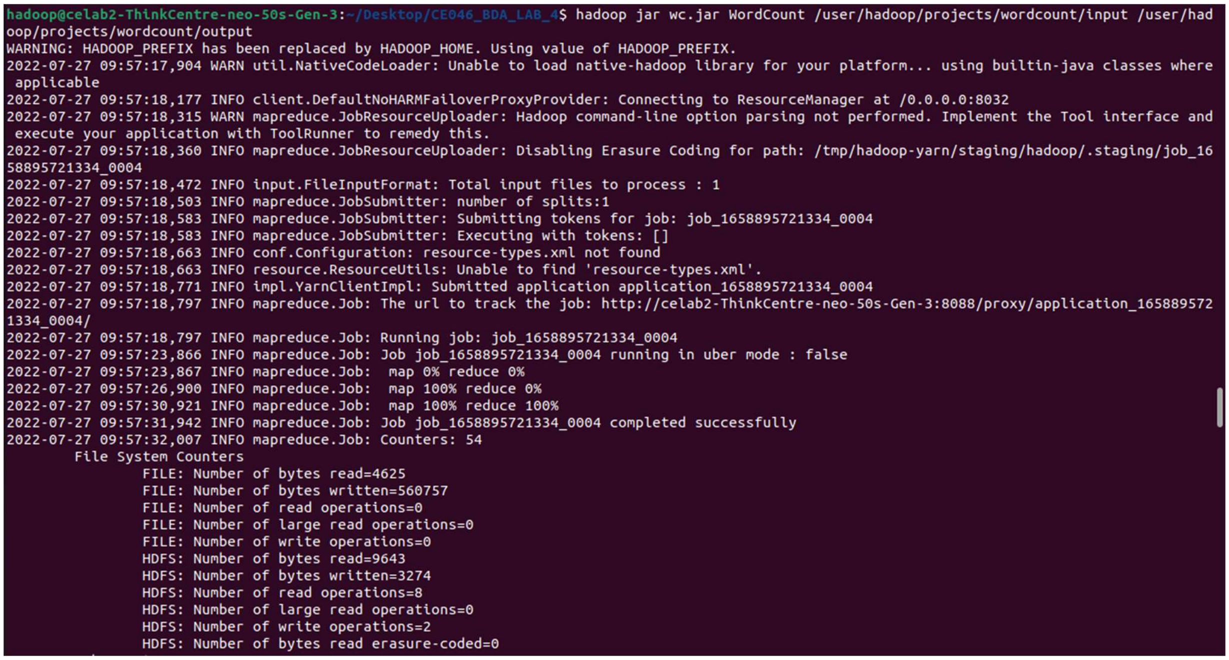


```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
          <name>yarn.app.mapreduce.am.env</name>
          <value>HADOOP_MAPRED_HOME=/opt/hadoop</value>
  </property>
  <property>
          <name>mapreduce.map.env</name>
          <value>HADOOP_MAPRED_HOME=/opt/hadoop</value>
  </property>
  <property>
          <name>mapreduce.reduce.env</name>
          <value>HADOOP_MAPRED_HOME=/opt/hadoop</value>
  </property>
  <property>
    <name>mapreduce.jobhistory.address</name>
    <value>localhost:10020</value>
  </property>

  <property>
    <name>mapreduce.jobhistory.webapp.address</name>
    <value>localhost:19888</value>
  </property>
</configuration>
```

(7)Need to stop and restart the server.

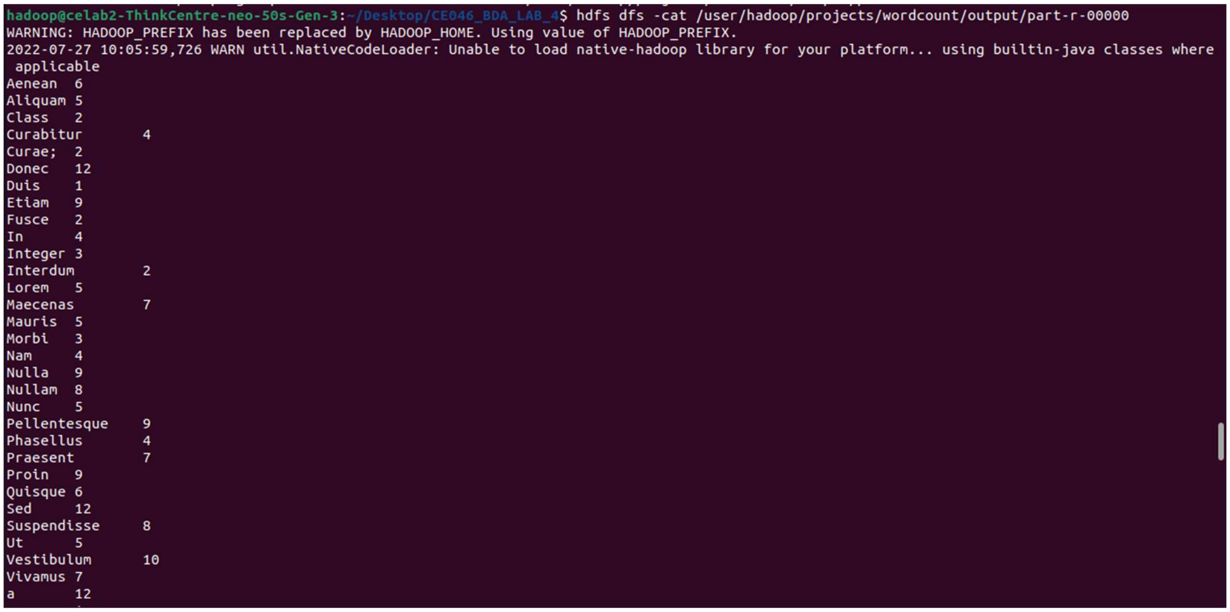


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ stop-dfs.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Stopping namenodes on [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Stopping datanodes
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Stopping secondary namenodes [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:50:30,855 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ stop-yarn.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Stopping nodemanagers
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Stopping resourcemanager
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ #mr-jobhistory-daemon.sh --config $HADOOP_CONF_DIR stop historyserver
mapred --daemon stop historyserver
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$ jps
22219 Jps
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:/opt/hadoop/etc/hadoop$
```

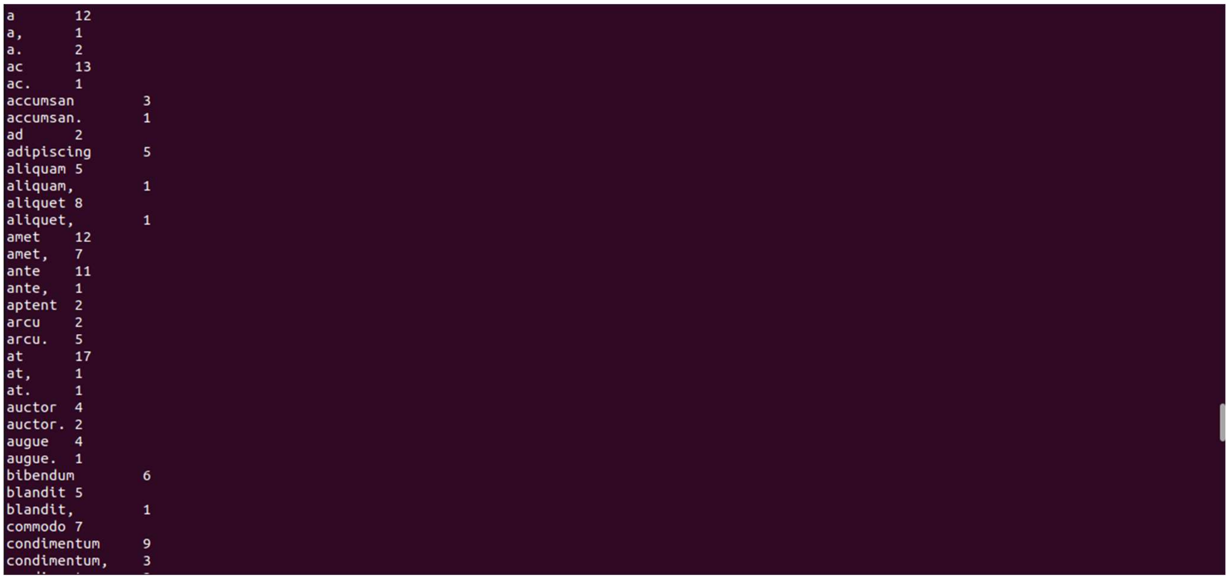


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hadoop jar wc.jar WordCount /user/hadoop/projects/wordcount/input /user/hadoop/projects/wordcount/output
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 09:57:17,904 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
2022-07-27 09:57:18,177 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-07-27 09:57:18,315 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
 execute your application with ToolRunner to remedy this.
2022-07-27 09:57:18,360 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1658895721334_0004
2022-07-27 09:57:18,472 INFO input.FileInputFormat: Total input files to process : 1
2022-07-27 09:57:18,503 INFO mapreduce.JobSubmitter: number of splits:1
2022-07-27 09:57:18,583 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1658895721334_0004
2022-07-27 09:57:18,583 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-07-27 09:57:18,663 INFO conf.Configuration: resource-types.xml not found
2022-07-27 09:57:18,663 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-07-27 09:57:18,771 INFO impl.YarnClientImpl: Submitted application application_1658895721334_0004
2022-07-27 09:57:18,797 INFO mapreduce.Job: The url to track the job: http://celab2-ThinkCentre-neo-50s-Gen-3:8088/proxy/application_1658895721334_0004/
2022-07-27 09:57:18,797 INFO mapreduce.Job: Running job: job_1658895721334_0004
2022-07-27 09:57:23,866 INFO mapreduce.Job: Job job_1658895721334_0004 running in uber mode : false
2022-07-27 09:57:23,867 INFO mapreduce.Job:  map 0% reduce 0%
2022-07-27 09:57:26,900 INFO mapreduce.Job:  map 100% reduce 0%
2022-07-27 09:57:30,921 INFO mapreduce.Job:  map 100% reduce 100%
2022-07-27 09:57:31,942 INFO mapreduce.Job: Job job_1658895721334_0004 completed successfully
2022-07-27 09:57:32,007 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=4625
                FILE: Number of bytes written=560757
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9643
                HDFS: Number of bytes written=3274
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
```

(8)Output.



```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hdfs dfs -cat /user/hadoop/projects/wordcount/output/part-r-00000
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:05:59,726 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Aenean  6
Aliquam 5
Class   2
Curabitur       4
Curae;  2
Donec   12
Duis    1
Etiam   9
Fusce   2
In      4
Integer 3
Interdum        2
Lorem   5
Maecenas        7
Mauris  5
Morbi   3
Nam     4
Nulla   9
Nullam  8
Nunc    5
Pellentesque    9
Phasellus       4
Praesent        7
Proin   9
Quisque 6
Sed     12
Suspendisse     8
Ut      5
Vestibulum      10
Vivamus 7
a       12
```

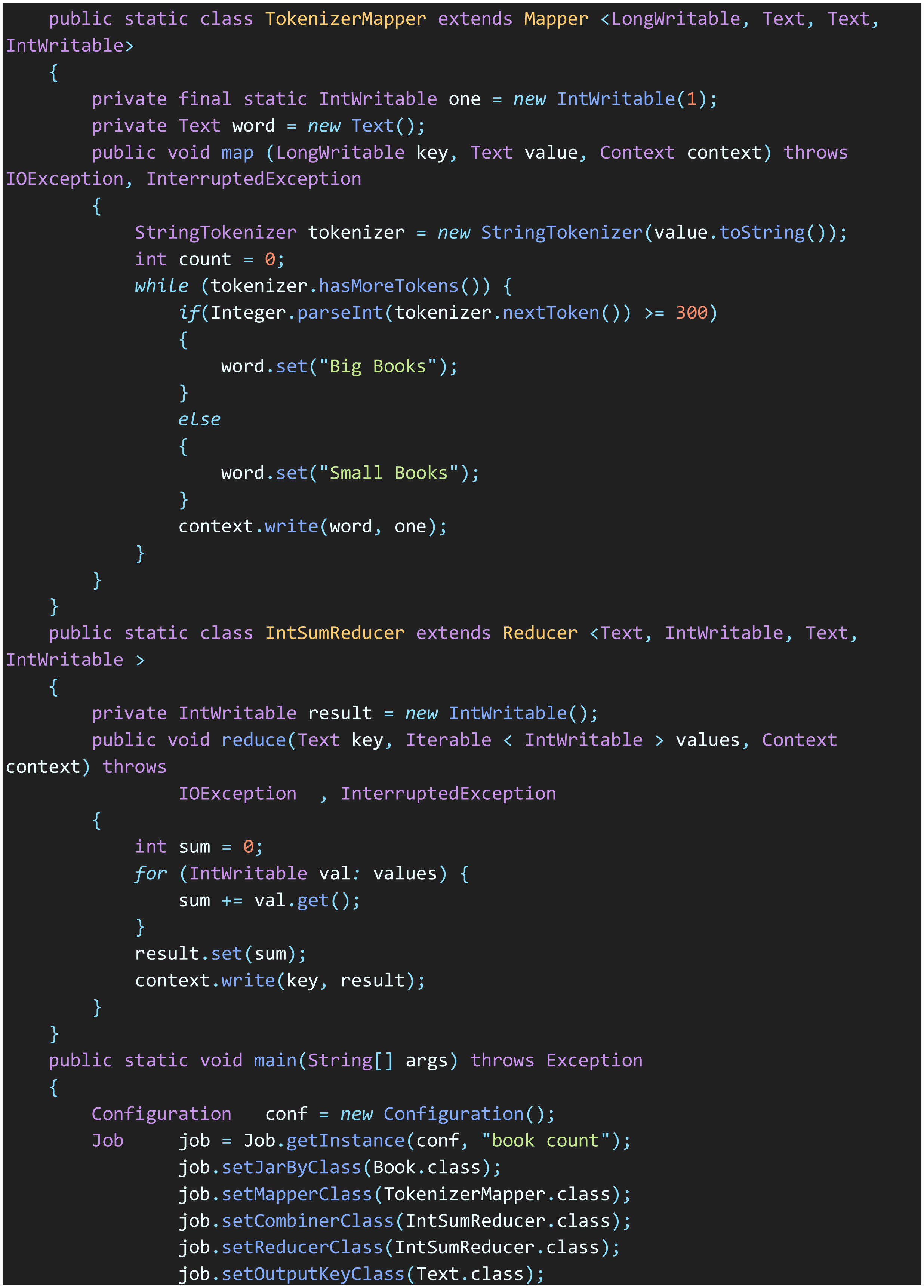


```
a       12
a,      1
a.      2
ac      13
ac.     1
accumsan        3
accumsan.       1
ad      2
adipiscing      5
aliquam 5
aliquam,        1
aliquet 8
aliquet,        1
amet    12
amet,   7
ante    11
ante,   1
aptent  2
arcu    2
arcu.   5
at      17
at,     1
at.     1
auctor  4
auctor. 2
augue   4
augue.  1
bibendum        6
blandit 5
blandit,        1
commodo 7
condimentum     9
condimentum,    3
```

Q. 2: Write map and reduce functions to split the books into the following two categories: (a) Big Books, (b) Small Books. Books which have more than 300 pages should be in the big book category. Books which have less than 300 pages should be in the small book category.

❖ **Code:**

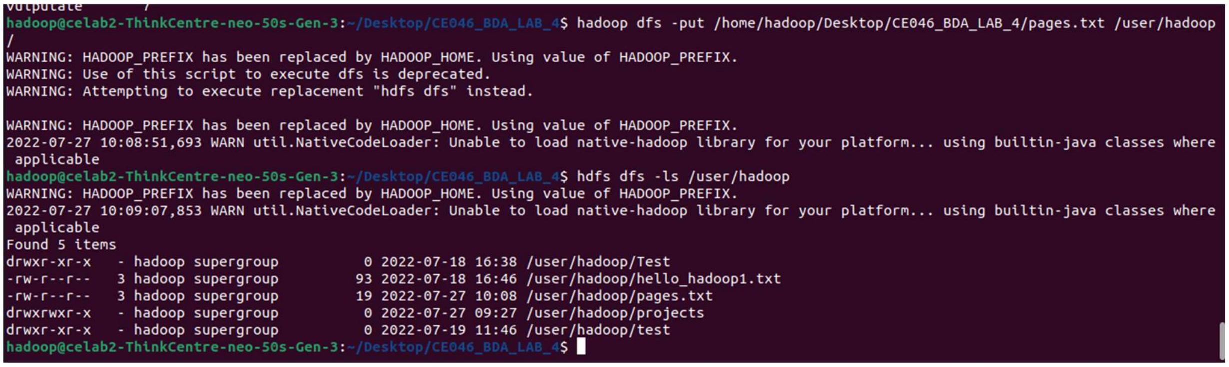
```
import      java.   io.IOException;
import      java.   util.StringTokenizer;
import      org.    apache.hadoop.conf.Configuration;
import      org.    apache.hadoop.fs.Path;
import      org.    apache.hadoop.io.*;
import      org.    apache.hadoop.io.Text;
import      org.    apache.hadoop.mapreduce.Job;
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    Book
{
```



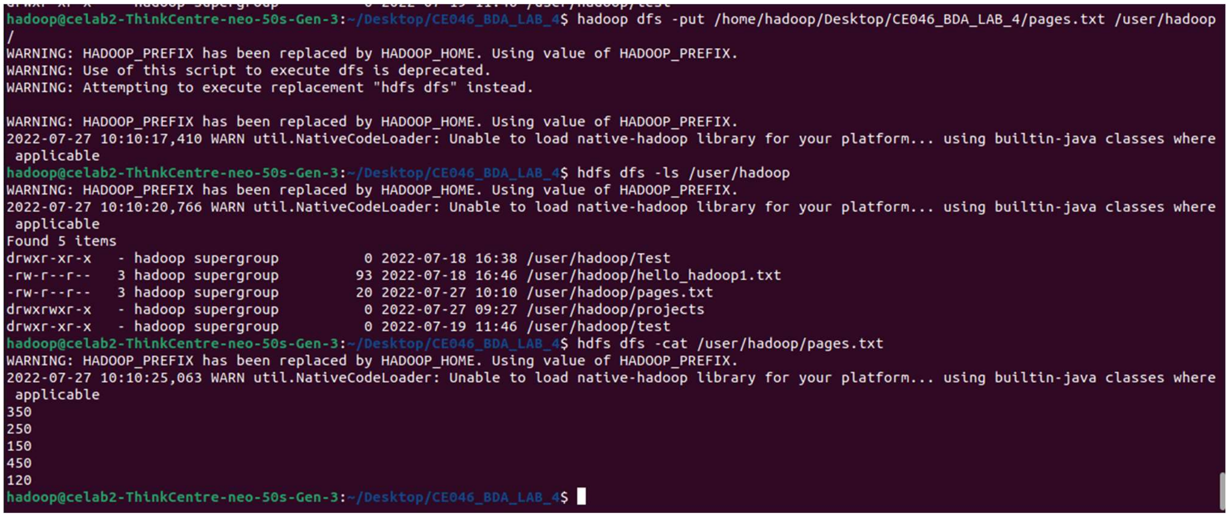
```
    public static class TokenizerMapper extends Mapper <LongWritable, Text, Text,
IntWritable>
    {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        public void map (LongWritable key, Text value, Context context) throws
IOException, InterruptedException
        {
            StringTokenizer tokenizer = new StringTokenizer(value.toString());
            int count = 0;
            while (tokenizer.hasMoreTokens()) {
                if(Integer.parseInt(tokenizer.nextToken()) >= 300)
                {
                    word.set("Big Books");
                }
                else
                {
                    word.set("Small Books");
                }
                context.write(word, one);
            }
        }
    }
    public static class IntSumReducer extends Reducer <Text, IntWritable, Text,
IntWritable >
    {
        private IntWritable result = new IntWritable();
        public void reduce(Text key, Iterable < IntWritable > values, Context
context) throws
                IOException  , InterruptedException
        {
            int sum = 0;
            for (IntWritable val: values) {
                sum += val.get();
            }
            result.set(sum);
            context.write(key, result);
        }
    }
    public static void main(String[] args) throws Exception
    {
        Configuration   conf = new Configuration();
        Job     job = Job.getInstance(conf, "book count");
                job.setJarByClass(Book.class);
                job.setMapperClass(TokenizerMapper.class);
                job.setCombinerClass(IntSumReducer.class);
                job.setReducerClass(IntSumReducer.class);
                job.setOutputKeyClass(Text.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);
    }
}
```

❖ Steps:

(1)Putting pages.txt file onto the Hadoop server.

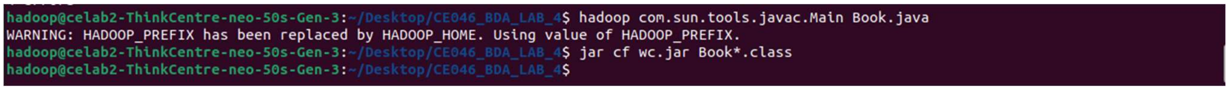


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hadoop dfs -put /home/hadoop/Desktop/CE046_BDA_LAB_4/pages.txt /user/hadoop
/
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:08:51,693 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hdfs dfs -ls /user/hadoop
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:09:07,853 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2022-07-18 16:38 /user/hadoop/Test
-rw-r--r--   3 hadoop supergroup         93 2022-07-18 16:46 /user/hadoop/hello_hadoop1.txt
-rw-r--r--   3 hadoop supergroup         19 2022-07-27 10:08 /user/hadoop/pages.txt
drwxrwxr-x   - hadoop supergroup          0 2022-07-27 09:27 /user/hadoop/projects
drwxr-xr-x   - hadoop supergroup          0 2022-07-19 11:46 /user/hadoop/test
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ 
```



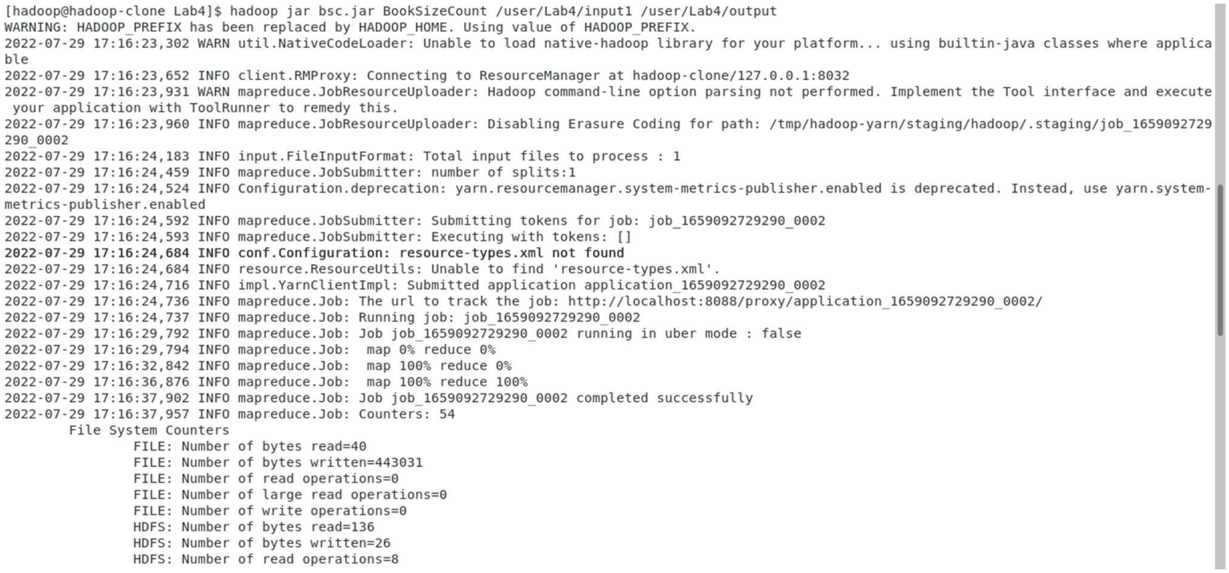
```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hadoop dfs -put /home/hadoop/Desktop/CE046_BDA_LAB_4/pages.txt /user/hadoop
/
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:10:17,410 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hdfs dfs -ls /user/hadoop
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:10:20,766 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2022-07-18 16:38 /user/hadoop/Test
-rw-r--r--   3 hadoop supergroup         93 2022-07-18 16:46 /user/hadoop/hello_hadoop1.txt
-rw-r--r--   3 hadoop supergroup         20 2022-07-27 10:10 /user/hadoop/pages.txt
drwxrwxr-x   - hadoop supergroup          0 2022-07-27 09:27 /user/hadoop/projects
drwxr-xr-x   - hadoop supergroup          0 2022-07-19 11:46 /user/hadoop/test
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hdfs dfs -cat /user/hadoop/pages.txt
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-27 10:10:25,063 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
350
250
150
450
120
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ 
```

(2)Compiling program and creating a jar file.

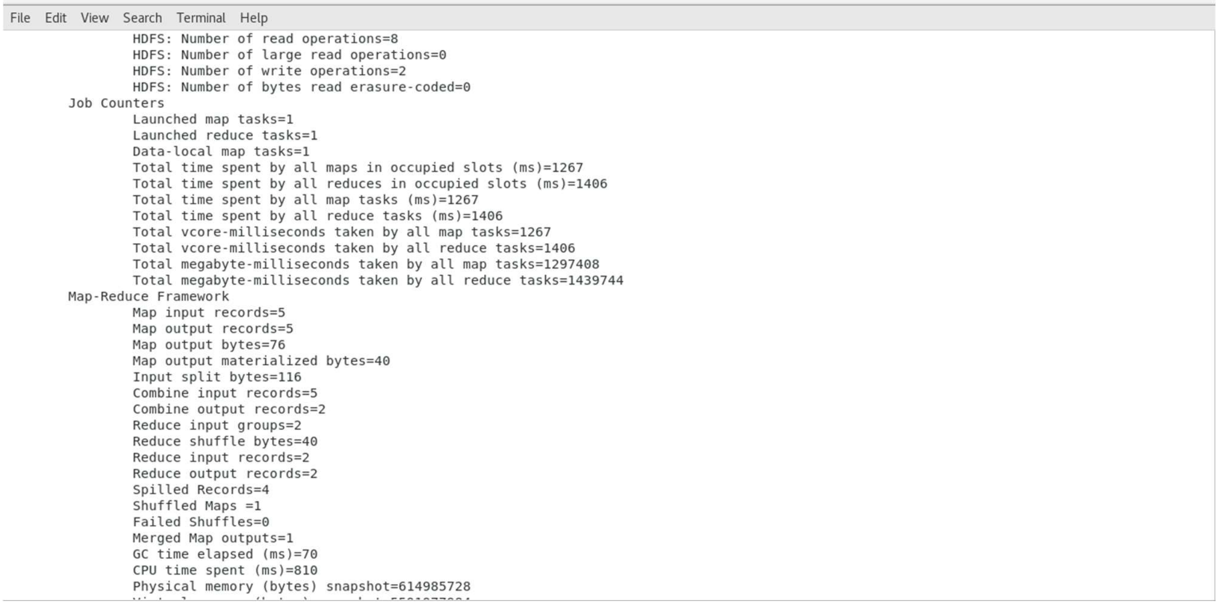


```
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ hadoop com.sun.tools.javac.Main Book.java
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$ jar cf wc.jar Book*.class
hadoop@celab2-ThinkCentre-neo-50s-Gen-3:~/Desktop/CE046_BDA_LAB_4$
```

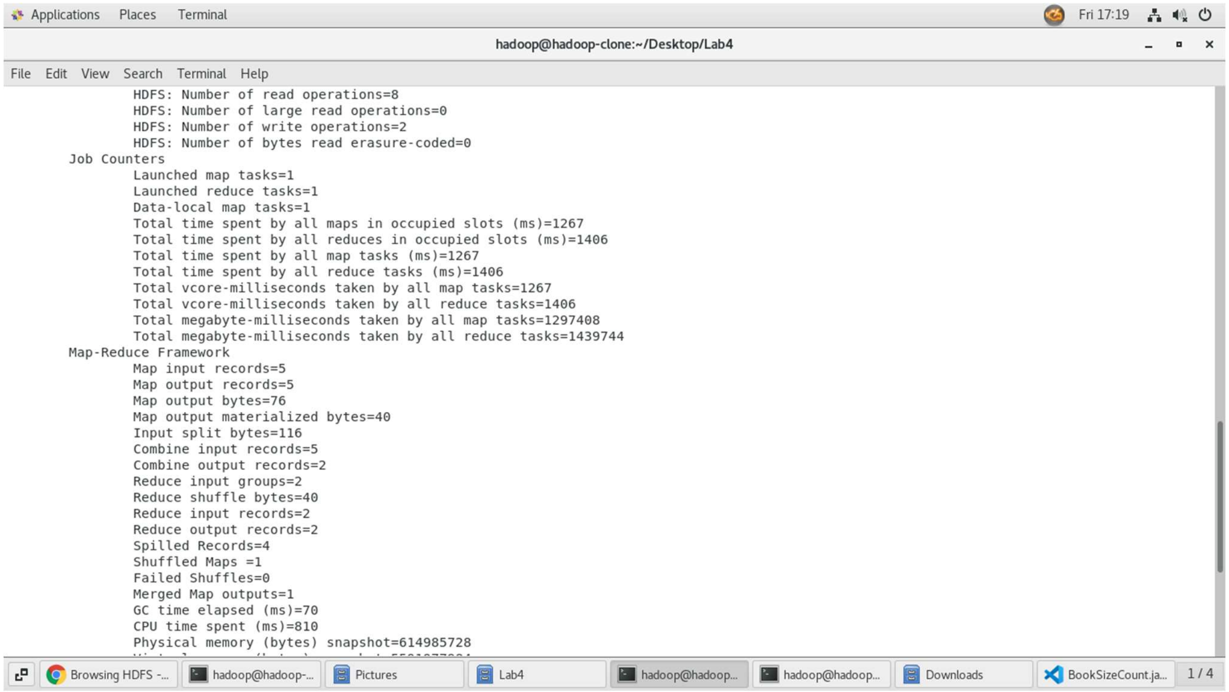
(3)Output.



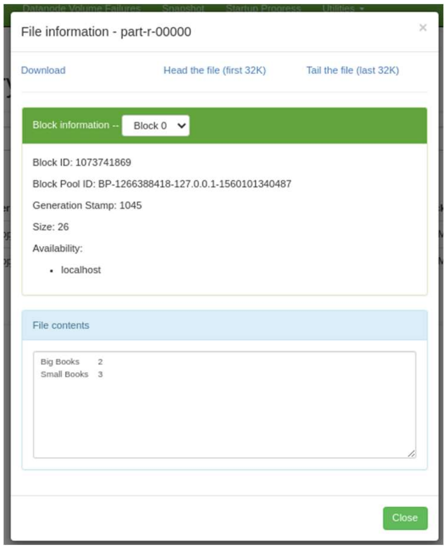
```
[hadoop@hadoop-clone Lab4]$ hadoop jar bsc.jar BookSizeCount /user/Lab4/input1 /user/Lab4/output
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2022-07-29 17:16:23,302 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
2022-07-29 17:16:23,652 INFO client.RMProxy: Connecting to ResourceManager at hadoop-clone/127.0.0.1:8032
2022-07-29 17:16:23,931 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
 your application with ToolRunner to remedy this.
2022-07-29 17:16:23,960 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1659092729
290_0002
2022-07-29 17:16:24,183 INFO input.FileInputFormat: Total input files to process : 1
2022-07-29 17:16:24,459 INFO mapreduce.JobSubmitter: number of splits:1
2022-07-29 17:16:24,524 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-
metrics-publisher.enabled
2022-07-29 17:16:24,592 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1659092729290_0002
2022-07-29 17:16:24,593 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-07-29 17:16:24,684 INFO conf.Configuration: resource-types.xml not found
2022-07-29 17:16:24,684 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-07-29 17:16:24,716 INFO impl.YarnClientImpl: Submitted application application_1659092729290_0002
2022-07-29 17:16:24,736 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1659092729290_0002/
2022-07-29 17:16:24,737 INFO mapreduce.Job: Running job: job_1659092729290_0002
2022-07-29 17:16:29,792 INFO mapreduce.Job: Job job_1659092729290_0002 running in uber mode : false
2022-07-29 17:16:29,794 INFO mapreduce.Job:  map 0% reduce 0%
2022-07-29 17:16:32,842 INFO mapreduce.Job:  map 100% reduce 0%
2022-07-29 17:16:36,876 INFO mapreduce.Job:  map 100% reduce 100%
2022-07-29 17:16:37,902 INFO mapreduce.Job: Job job_1659092729290_0002 completed successfully
2022-07-29 17:16:37,957 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=40
                FILE: Number of bytes written=443031
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=136
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=8
```



```
                HDFS: Number of read operations=8
                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=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1267
                Total time spent by all reduces in occupied slots (ms)=1406
                Total time spent by all map tasks (ms)=1267
                Total time spent by all reduce tasks (ms)=1406
                Total vcore-milliseconds taken by all map tasks=1267
                Total vcore-milliseconds taken by all reduce tasks=1406
                Total megabyte-milliseconds taken by all map tasks=1297408
                Total megabyte-milliseconds taken by all reduce tasks=1439744
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Map output bytes=76
                Map output materialized bytes=40
                Input split bytes=116
                Combine input records=5
                Combine output records=2
                Reduce input groups=2
                Reduce shuffle bytes=40
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=70
                CPU time spent (ms)=810
                Physical memory (bytes) snapshot=614985728
```



```
                HDFS: Number of read operations=8
                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=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1267
                Total time spent by all reduces in occupied slots (ms)=1406
                Total time spent by all map tasks (ms)=1267
                Total time spent by all reduce tasks (ms)=1406
                Total vcore-milliseconds taken by all map tasks=1267
                Total vcore-milliseconds taken by all reduce tasks=1406
                Total megabyte-milliseconds taken by all map tasks=1297408
                Total megabyte-milliseconds taken by all reduce tasks=1439744
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Map output bytes=76
                Map output materialized bytes=40
                Input split bytes=116
                Combine input records=5
                Combine output records=2
                Reduce input groups=2
                Reduce shuffle bytes=40
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=70
                CPU time spent (ms)=810
                Physical memory (bytes) snapshot=614985728
```



File information - part-r-00000

Block information -- Block 0

Block ID: 1073741869

Block Pool ID: BP-1266388418-127.0.0.1-1560101340487

Generation Stamp: 1045

Size: 26

Availability:

- localhost

File contents

```
Big Books     2
Small Books   3
```