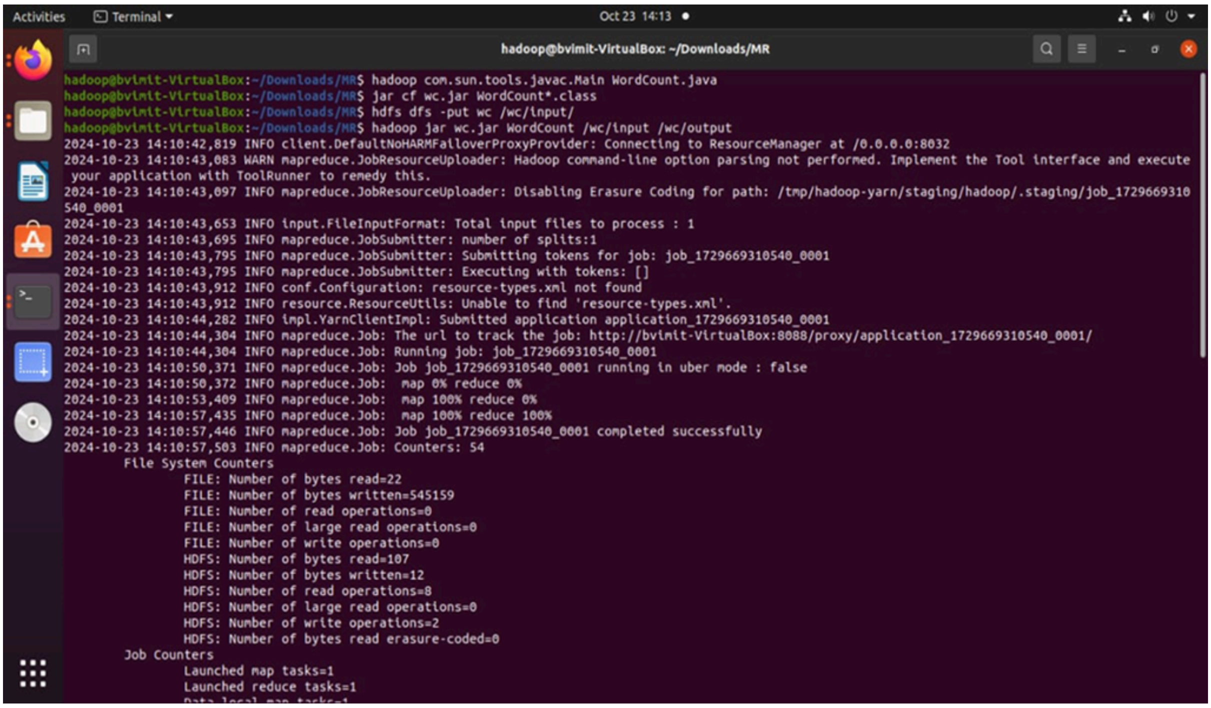
Q) Map Reduce Programming

1. **Write a program in Map Reduce for Word Count operation.**

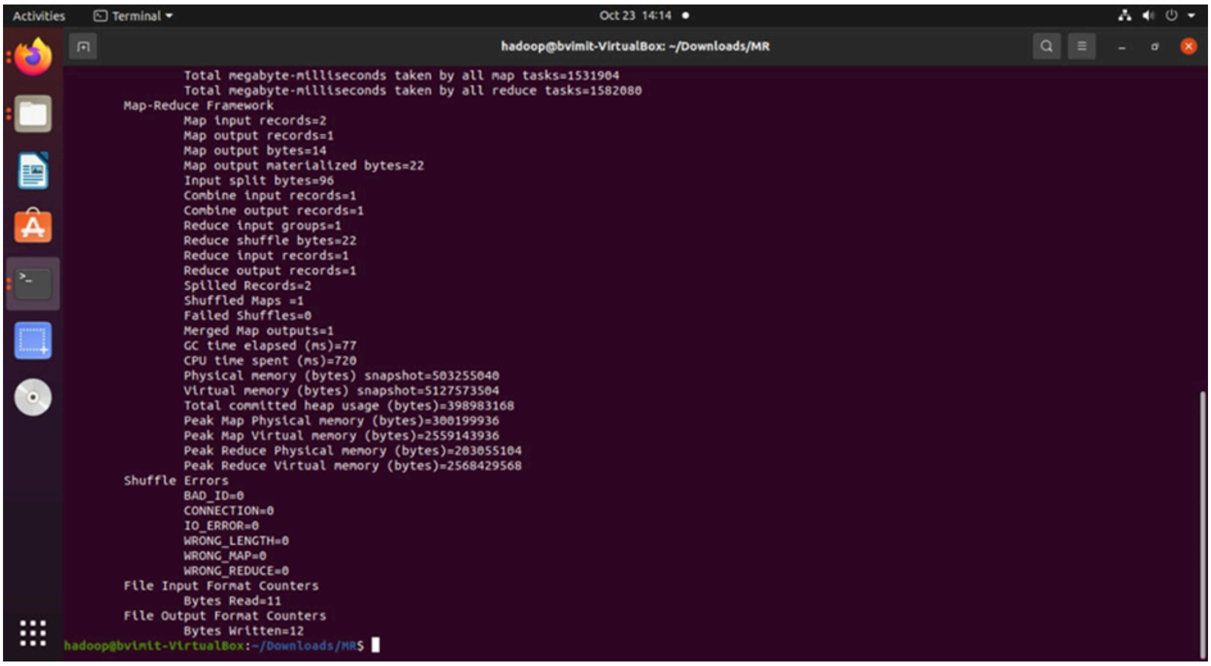
   Wordcount.java:

```
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.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 WordCount {
  public static class TokenizerMapper  extends Mapper<Object, Text,
Text, IntWritable>{
    private Text word = new Text();
    public void map(Object key, Text value,
                    Context context )
                    throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());   //word is a variable to store
words(tokens)
        context.write(word, new IntWritable(1));
}
 }
  }
  public static class IntSumReducer  extends
Reducer<Text,IntWritable,Text,IntWritable> {
    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context
                       ) throws IOException, InterruptedException {
      int sum = 0; // to calculate frequency of key or word
      for (IntWritable x : values) {
        sum += x.get();
      }
      context.write(key, new IntWritable(sum));
    }
  }
  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();   //for defining
configuration of map reducer program
    Job job = Job.getInstance(conf, "word count");   // job to be executed.
    job.setJarByClass(WordCount.class); // mentioning Main class
    job.setMapperClass(TokenizerMapper.class); // Mentioning mapper
class
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);// Mentioning Reducer
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); // exiting the job
only if the flag value becomes false
  }  }
```

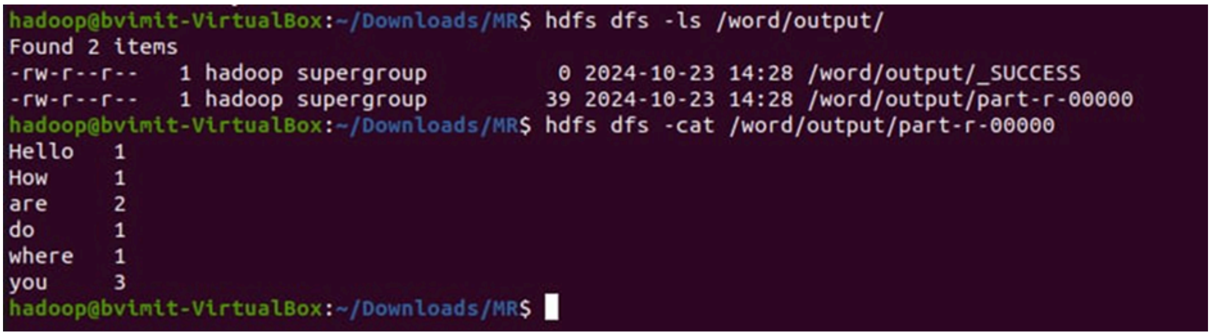
Output: -



```
hadoop@bvimit-VirtualBox:~/Downloads/MR$ hadoop com.sun.tools.javac.Main WordCount.java
hadoop@bvimit-VirtualBox:~/Downloads/MR$ jar cf wc.jar WordCount*.class
hadoop@bvimit-VirtualBox:~/Downloads/MR$ hdfs dfs -put wc /wc/input/
hadoop@bvimit-VirtualBox:~/Downloads/MR$ hadoop jar wc.jar WordCount /wc/input /wc/output
2024-10-23 14:10:42,819 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-23 14:10:43,083 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
 your application with ToolRunner to remedy this.
2024-10-23 14:10:43,097 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1729669310
540_0001
2024-10-23 14:10:43,653 INFO input.FileInputFormat: Total input files to process : 1
2024-10-23 14:10:43,695 INFO mapreduce.JobSubmitter: number of splits:1
2024-10-23 14:10:43,795 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729669310540_0001
2024-10-23 14:10:43,795 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-23 14:10:43,912 INFO conf.Configuration: resource-types.xml not found
2024-10-23 14:10:43,912 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-23 14:10:44,282 INFO impl.YarnClientImpl: Submitted application application_1729669310540_0001
2024-10-23 14:10:44,304 INFO mapreduce.Job: The url to track the job: http://bvimit-VirtualBox:8088/proxy/application_1729669310540_0001/
2024-10-23 14:10:44,304 INFO mapreduce.Job: Running job: job_1729669310540_0001
2024-10-23 14:10:50,371 INFO mapreduce.Job: Job job_1729669310540_0001 running in uber mode : false
2024-10-23 14:10:50,372 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-23 14:10:53,409 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-23 14:10:57,435 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-23 14:10:57,446 INFO mapreduce.Job: Job job_1729669310540_0001 completed successfully
2024-10-23 14:10:57,503 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=22
                FILE: Number of bytes written=545159
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=107
                HDFS: Number of bytes written=12
                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 megabyte-milliseconds taken by all map tasks=1531904
                Total megabyte-milliseconds taken by all reduce tasks=1582080
        Map-Reduce Framework
                Map input records=2
                Map output records=1
                Map output bytes=14
                Map output materialized bytes=22
                Input split bytes=96
                Combine input records=1
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=22
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=77
                CPU time spent (ms)=720
                Physical memory (bytes) snapshot=503255040
                Virtual memory (bytes) snapshot=5127573504
                Total committed heap usage (bytes)=398983168
                Peak Map Physical memory (bytes)=300199936
                Peak Map Virtual memory (bytes)=2559143936
                Peak Reduce Physical memory (bytes)=203055104
                Peak Reduce Virtual memory (bytes)=2568429568
        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=11
        File Output Format Counters
                Bytes Written=12
hadoop@bvimit-VirtualBox:~/Downloads/MR$
```



```
hadoop@bvimit-VirtualBox:~/Downloads/MR$ hdfs dfs -ls /word/output/
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2024-10-23 14:28 /word/output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         39 2024-10-23 14:28 /word/output/part-r-00000
hadoop@bvimit-VirtualBox:~/Downloads/MR$ hdfs dfs -cat /word/output/part-r-00000
Hello   1
How     1
are     2
do      1
where   1
you     3
hadoop@bvimit-VirtualBox:~/Downloads/MR$
```

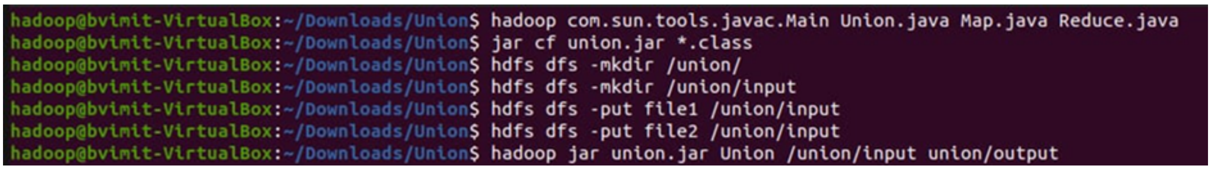
2. **Write a program in Map Reduce for Union operation.**

Union.java: -

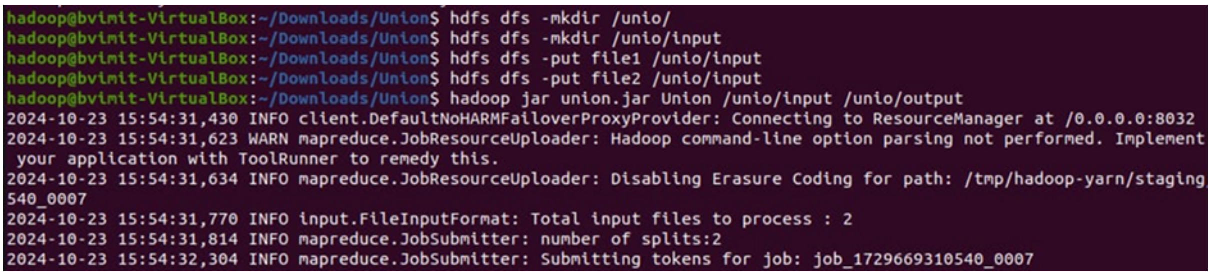
```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import java.io.IOException;
public class Union {
        private static Text Word1 = new Text("");
        public static class Mapper
extends org.apache.hadoop.mapreduce.Mapper<Object, Text, Text, Text>
{
public void map(Object key, Text value, Context context
) throws IOException, InterruptedException {
```

```
context.write(value, Word1);
}
}
public static class Reducer extends
org.apache.hadoop.mapreduce.Reducer<Text, Text, Text, Text> {
public void reduce(Text key, Iterable<Text> _values,
  Context context
) throws IOException, InterruptedException {
context.write(key,new Text(""));
}
}
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "union");
job.setJarByClass(Union.class);
job.setMapperClass(Mapper.class);
job.setCombinerClass(Reducer.class);
job.setReducerClass(Reducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
Path input = new Path( args[0]);
Path output = new Path(args[1]);
FileInputFormat.addInputPath(job, input);
FileOutputFormat.setOutputPath(job, output);
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```

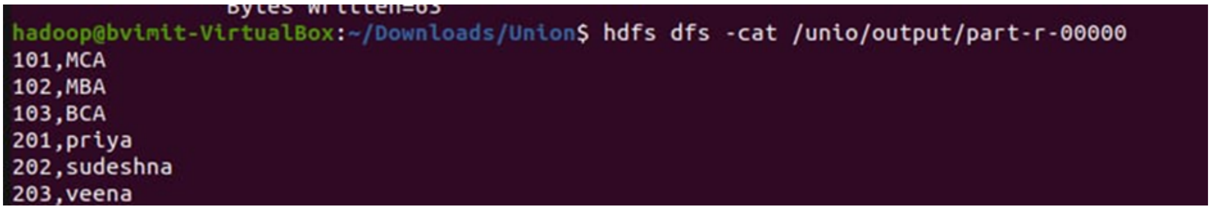
Output: -



```
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hadoop com.sun.tools.javac.Main Union.java Map.java Reduce.java
hadoop@bvimit-VirtualBox:~/Downloads/Union$ jar cf union.jar *.class
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -mkdir /union/
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -mkdir /union/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -put file1 /union/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -put file2 /union/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hadoop jar union.jar Union /union/input union/output
```



```
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -mkdir /unio/
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -mkdir /unio/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -put file1 /unio/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -put file2 /unio/input
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hadoop jar union.jar Union /unio/input /unio/output
2024-10-23 15:54:31,430 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-23 15:54:31,623 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
 your application with ToolRunner to remedy this.
2024-10-23 15:54:31,634 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
540_0007
2024-10-23 15:54:31,770 INFO input.FileInputFormat: Total input files to process : 2
2024-10-23 15:54:31,814 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-23 15:54:32,304 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729669310540_0007
```



```
                Bytes Written=63
hadoop@bvimit-VirtualBox:~/Downloads/Union$ hdfs dfs -cat /unio/output/part-r-00000
101,MCA
102,MBA
103,BCA
201,priya
202,sudeshna
203,veena
```

3. **Write a program in Map Reduce for Matrix Multiplication**
   MatrixMultiply.java: -

```
//matrix multiplication
import org.apache.hadoop.conf.*;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
public class MatrixMultiply {
  public static void main(String[] args) throws Exception {
      if (args.length != 2) {
        System.err.println("Usage: MatrixMultiply <in_dir> <out_dir>");
        System.exit(2);
     }
      Configuration conf = new Configuration();
     // M is an m-by-n matrix; N is an n-by-p matrix.
     conf.set("m", "1000");
     conf.set("n", "100");
     conf.set("p", "1000");
     @SuppressWarnings("deprecation")
          Job job = new Job(conf, "MatrixMultiply");
     job.setJarByClass(MatrixMultiply.class);
     job.setOutputKeyClass(Text.class);
```

```
        job.setOutputValueClass(Text.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
    }
}
```

Map.java: -

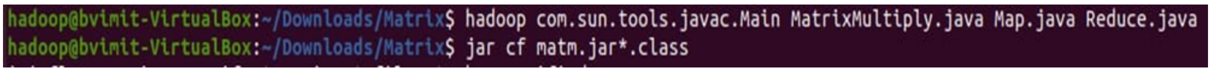
```
//matrix multiplication_Mapper
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import java.io.IOException;
public class Map
  extends org.apache.hadoop.mapreduce.Mapper<LongWritable, Text,
Text, Text> {
        @Override
        public void map(LongWritable key, Text value, Context context)
                        throws IOException, InterruptedException {
                Configuration conf = context.getConfiguration();
                int m = Integer.parseInt(conf.get("m"));
                int p = Integer.parseInt(conf.get("p"));
                String line = value.toString();
                // (M, i, j, Mij);
                String[] indicesAndValue = line.split(",");
                Text outputKey = new Text();
                Text outputValue = new Text();
                if (indicesAndValue[0].equals("M")) {
                        for (int k = 0; k < p; k++) {
                                outputKey.set(indicesAndValue[1] + "," + k);
                                // outputKey.set(i,k);
                outputValue.set(indicesAndValue[0] + "," +
indicesAndValue[2]
                        + "," + indicesAndValue[3]);
                // outputValue.set(M,j,Mij);
                context.write(outputKey, outputValue);
            }
        } else {
            // (N, j, k, Njk);
            for (int i = 0; i < m; i++) {
                outputKey.set(i + "," + indicesAndValue[2]);
                outputValue.set("N," + indicesAndValue[1] +
","
                        + indicesAndValue[3]);
                context.write(outputKey, outputValue);
            }
        }
    }
}
```

Reduce.java: -

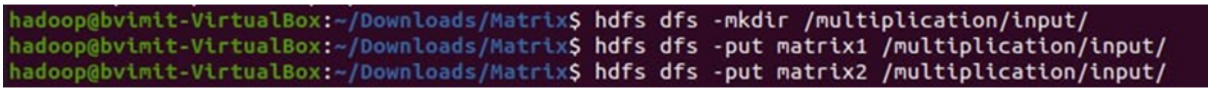
```
//matrix multiplication_Reducer
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;
import java.util.HashMap;
public class Reduce
 extends org.apache.hadoop.mapreduce.Reducer<Text, Text, Text, Text>
{
    @Override
    public void reduce(Text key, Iterable<Text> values, Context
context)
            throws IOException, InterruptedException {
        String[] value;
        //key=(i,k),
        //Values = [(M/N,j,V/W),..]
        HashMap<Integer, Float> hashA = new HashMap<Integer,
Float>();
```

```
        HashMap<Integer, Float> hashB = new HashMap<Integer,
Float>();
        for (Text val : values) {
            value = val.toString().split(",");
            if (value[0].equals("M")) {
                hashA.put(Integer.parseInt(value[1]),
Float.parseFloat(value[2]));
            } else {
                hashB.put(Integer.parseInt(value[1]),
Float.parseFloat(value[2]));
            }
        }
        int n = Integer.parseInt(context.getConfiguration().get("n"));
        float result = 0.0f;
        float m_ij;
        float n_jk;
        for (int j = 0; j < n; j++) {
            m_ij = hashA.containsKey(j) ? hashA.get(j) : 0.0f;
            n_jk = hashB.containsKey(j) ? hashB.get(j) : 0.0f;
            result += m_ij * n_jk;
        }
        if (result != 0.0f) {
            context.write(null,
                    new Text(key.toString() + "," +
Float.toString(result)));
        }
    }
}
```

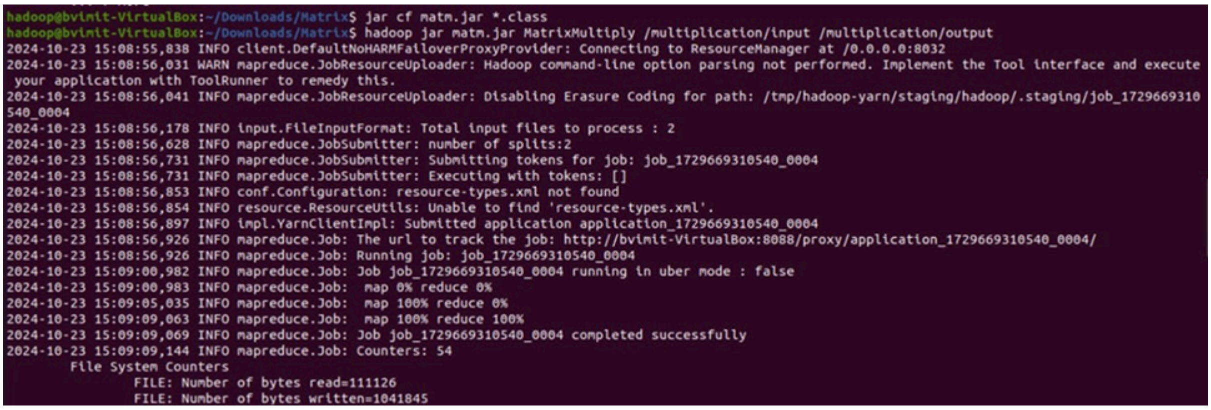
Output: -



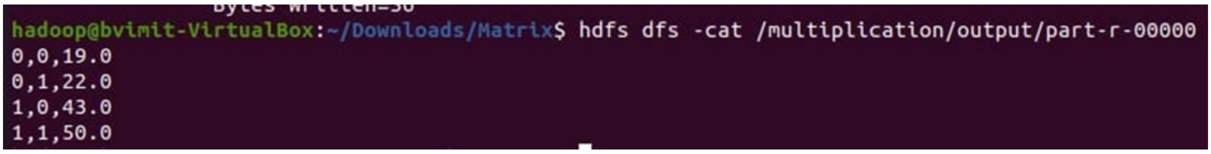
```
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hadoop com.sun.tools.javac.Main MatrixMultiply.java Map.java Reduce.java
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ jar cf matm.jar*.class
```



```
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hdfs dfs -mkdir /multiplication/input/
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hdfs dfs -put matrix1 /multiplication/input/
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hdfs dfs -put matrix2 /multiplication/input/
```



```
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ jar cf matm.jar *.class
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hadoop jar matm.jar MatrixMultiply /multiplication/input /multiplication/output
2024-10-23 15:08:55,838 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-23 15:08:56,031 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
 your application with ToolRunner to remedy this.
2024-10-23 15:08:56,041 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1729669310
540_0004
2024-10-23 15:08:56,178 INFO input.FileInputFormat: Total input files to process : 2
2024-10-23 15:08:56,628 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-23 15:08:56,731 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729669310540_0004
2024-10-23 15:08:56,731 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-23 15:08:56,853 INFO conf.Configuration: resource-types.xml not found
2024-10-23 15:08:56,854 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-23 15:08:56,897 INFO impl.YarnClientImpl: Submitted application application_1729669310540_0004
2024-10-23 15:08:56,926 INFO mapreduce.Job: The url to track the job: http://bvimit-VirtualBox:8088/proxy/application_1729669310540_0004/
2024-10-23 15:08:56,926 INFO mapreduce.Job: Running job: job_1729669310540_0004
2024-10-23 15:09:00,982 INFO mapreduce.Job: Job job_1729669310540_0004 running in uber mode : false
2024-10-23 15:09:00,983 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-23 15:09:05,035 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-23 15:09:09,063 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-23 15:09:09,069 INFO mapreduce.Job: Job job_1729669310540_0004 completed successfully
2024-10-23 15:09:09,144 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=111126
                FILE: Number of bytes written=1041845
```



```
                Bytes Written=30
hadoop@bvimit-VirtualBox:~/Downloads/Matrix$ hdfs dfs -cat /multiplication/output/part-r-00000
0,0,19.0
0,1,22.0
1,0,43.0
1,1,50.0
```