BDA LAB ASSIGNMENT NO.4
MAP REDUCE 1

NAME : KETAKI PATIL
ROLL NO : PA-17
BATCH : A1

CODE :

WORD MAPPER

```java
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;


public class Word_Mapper extends
Mapper<LongWritable,Text,Text,IntWritable>
{

    public  void map(LongWritable key,Text value,Context context)
throws IOException,InterruptedException {
        String line = value.toString();
        StringTokenizer tokenizer=new StringTokenizer(line);
        while(tokenizer.hasMoreTokens()) {
            value.set(tokenizer.nextToken());
            context.write(value,new IntWritable(1));
        }
    }
}
```

WORD DRIVER
```java
import java.io.IOException;

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;
```

```java
//import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
//import java.util.Map;

public class Word_Driver {
    public static void main(String[] args) throws Exception{
        Configuration conf=new Configuration();
        Job job =new Job(conf,"My Word Count Program");
        job.setJarByClass(Word_Driver.class);
        job.setMapperClass(Word_Mapper.class);
        job.setReducerClass(Word_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);
    }
}
```

WORD REDUCER

```java
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;


public class Word_Reducer extends
Reducer<Text,IntWritable,Text,IntWritable> {

    public void reduce(Text key, Iterable <IntWritable> values,
Context context)
        throws IOException,InterruptedException{
        int sum=0;
        for(IntWritable x: values)
        {
            sum+=x.get();
        }
        context.write(key, new IntWritable(sum));
    }
}
```

OUTPUT :

Starting datanode, namenode, node manager and resource manager through start dfs and yarn commands.

```
Administrator: Command Prompt

Microsoft Windows [Version 10.0.19042.985]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>D:

D:\>cd D:\hadoop-3.1.4\hadoop-3.1.4\sbin

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>.\start-dfs.cmd

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>jps
4608
16756 Jps
21044 NameNode
8868 DataNode

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>.\start-yarn.cmd
starting yarn daemons

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>jps
23392 NodeManager
4608
11172 ResourceManager
21044 NameNode
8868 DataNode
11176 Jps

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>
```

```
Apache Hadoop Distribution - hadoop  datanode

2021-06-01 18:45:42,524 INFO impl.FsDatasetImpl: Time
n D:\Hadoop\data\datanode: 106ms
2021-06-01 18:45:42,524 INFO impl.FsDatasetImpl: Tota
7-1620916509044: 108ms
2021-06-01 18:45:42,528 INFO impl.FsDatasetImpl: Addi
509044 on volume D:\Hadoop\data\datanode...
2021-06-01 18:45:42,529 INFO impl.BlockPoolSlice: Rep
68.1.27-1620916509044\current\replicas doesn't exist
2021-06-01 18:45:42,565 INFO impl.FsDatasetImpl: Time
20916509044 on volume D:\Hadoop\data\datanode: 36ms
2021-06-01 18:45:42,568 INFO impl.FsDatasetImpl: Tota
68.1.27-1620916509044: 41ms
2021-06-01 18:45:42,568 INFO checker.ThrottledAsyncCh
2021-06-01 18:45:42,582 INFO checker.DatasetVolumeChe
2021-06-01 18:45:42,636 INFO datanode.VolumeScanner:
d3eddad220d): no suitable block pools found to scan.
2021-06-01 18:45:42,646 INFO datanode.DirectoryScanne
8 PM with interval of 21600000ms
2021-06-01 18:45:42,654 INFO datanode.DataNode: Block
54-3ea8-445a-bc4a-4961dbd7cb13) service to localhost/
2021-06-01 18:45:42,738 INFO datanode.DataNode: Block
Uuid fea80054-3ea8-445a-bc4a-4961dbd7cb13) service to
2021-06-01 18:45:42,739 INFO datanode.DataNode: For n
0000msec CACHEREPORT_INTERVAL of 10000msec Initial de
2021-06-01 18:45:42,980 INFO datanode.DataNode: Succe
 report(s), of which we sent 1. The reports had 27 to
msecs for RPC and NN processing. Got back one command
2021-06-01 18:45:42,981 INFO datanode.DataNode: Got f
044
```

```
Apache Hadoop Distribution - hadoop  namenode

00 milliseconds
2021-06-01 18:45:42,720 INFO hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(127.0.0.1:9866, datano
deUuid=fea80054-3ea8-445a-bc4a-4961dbd7cb13, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-3
9ee263a-33a1-4e03-94a7-0d9824b7d599;nsid=1653617406;c=1620916509044) storage fea80054-3ea8-445a-bc4a-4961dbd7cb13
2021-06-01 18:45:42,722 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2021-06-01 18:45:42,722 INFO blockmanagement.BlockReportLeaseManager: Registered DN fea80054-3ea8-445a-bc4a-4961dbd7cb13
(127.0.0.1:9866).
2021-06-01 18:45:42,818 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-c5e150a7-f692-4180-9d7b-ad3edd
ad220d for DN 127.0.0.1:9866
2021-06-01 18:45:42,922 INFO BlockStateChange: BLOCK* processReport 0x6b97955f735f4a8b: Processing first storage report
for DS-c5e150a7-f692-4180-9d7b-ad3eddad220d from datanode fea80054-3ea8-445a-bc4a-4961dbd7cb13
2021-06-01 18:45:42,930 INFO blockmanagement.BlockManager: initializing replication queues
2021-06-01 18:45:42,931 INFO hdfs.StateChange: STATE* Safe mode extension entered.
The reported blocks 26 has reached the threshold 0.9990 of total blocks 27. The minimum number of live datanodes is not
required. In safe mode extension. Safe mode will be turned off automatically in 29 seconds.
2021-06-01 18:45:42,933 INFO BlockStateChange: BLOCK* processReport 0x6b97955f735f4a8b: from storage DS-c5e150a7-f692-41
80-9d7b-ad3eddad220d node DatanodeRegistration(127.0.0.1:9866, datanodeUuid=fea80054-3ea8-445a-bc4a-4961dbd7cb13, infoPo
rt=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-39ee263a-33a1-4e03-94a7-0d9824b7d599;nsid=1653617406
c=1620916509044), blocks: 27, hasStaleStorage: false, processing time: 10 msecs, invalidatedBlocks: 0
2021-06-01 18:45:42,935 INFO blockmanagement.BlockManager: Total number of blocks            = 27
2021-06-01 18:45:42,935 INFO blockmanagement.BlockManager: Number of invalid blocks          = 0
2021-06-01 18:45:42,936 INFO blockmanagement.BlockManager: Number of under-replicated blocks = 8
2021-06-01 18:45:42,936 INFO blockmanagement.BlockManager: Number of  over-replicated blocks = 0
2021-06-01 18:45:42,937 INFO blockmanagement.BlockManager: Number of blocks being written    = 0
2021-06-01 18:45:42,937 INFO hdfs.StateChange: STATE* Replication Queue initialization scan for invalid, over- and under
-replicated blocks completed in 6 msec
2021-06-01 18:46:03,055 INFO hdfs.StateChange: STATE* Safe mode ON, in safe mode extension.
The reported blocks 27 has reached the threshold 0.9990 of total blocks 27. The minimum number of live datanodes is not
required. In safe mode extension. Safe mode will be turned off automatically in 9 seconds.
```

Making a new directory in hadoop

```
D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hdfs dfs -mkdir /WordCount

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hdfs dfs -ls
Found 1 items
drwxr-xr-x   - ketak supergroup          0 2021-05-20 11:19 wordmr_out

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hdfs dfs -ls /
Found 9 items
drwxr-xr-x   - ketak supergroup          0 2021-05-24 12:12 /Weather
drwxr-xr-x   - ketak supergroup          0 2021-06-01 18:56 /WordCount
drwxr-xr-x   - ketak supergroup          0 2021-05-18 22:20 /WordMap
-rw-r--r--   1 ketak supergroup         60 2021-05-18 22:06 /hdfs
-rw-r--r--   1 ketak supergroup         60 2021-05-18 22:07 /ketaki
-rw-r--r--   1 ketak supergroup         60 2021-05-18 22:13 /mapper
drwx------   - ketak supergroup          0 2021-05-21 13:02 /tmp
drwxr-xr-x   - ketak supergroup          0 2021-05-20 11:19 /user
-rw-r--r--   1 ketak supergroup        123 2021-05-18 22:16 /wordmapper

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hadoop dfs -copyFromLocal D:\input1.txt /WordCount/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hadoop dfs -ls /WordCount/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Found 1 items
-rw-r--r--   1 ketak supergroup        429 2021-06-01 18:58 /WordCount/input1.txt

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>cd ..

D:\hadoop-3.1.4\hadoop-3.1.4>cd bin
```
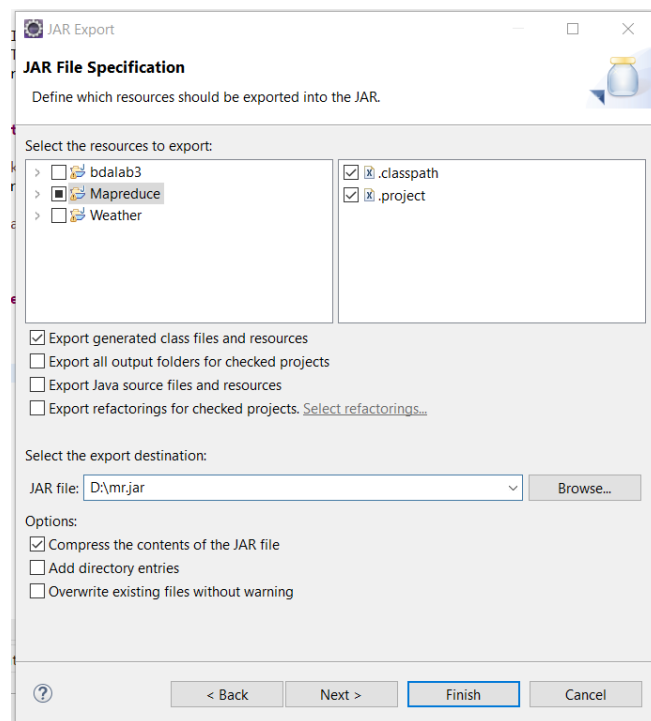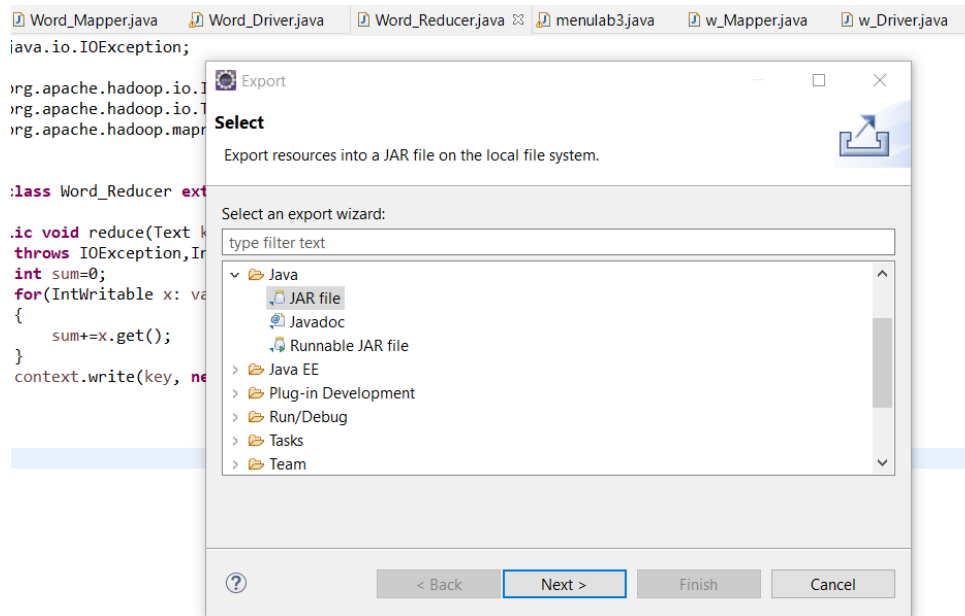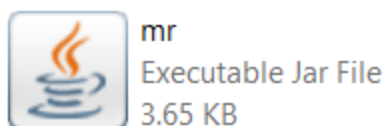
Jar file of project is exported from java program and stored :

Jar file stored :

Execution of project :

```
D:\hadoop-3.1.4\hadoop-3.1.4\bin>hadoop jar D:\mr.jar Word_Driver /WordCount/input1.txt wmapred_out
2021-06-01 19:02:24,808 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-06-01 19:02:27,311 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execut
2021-06-01 19:02:28,081 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ketak/.staging/job_1622553419
2021-06-01 19:02:28,524 INFO input.FileInputFormat: Total input files to process : 1
2021-06-01 19:02:28,935 INFO mapreduce.JobSubmitter: number of splits:1
2021-06-01 19:02:29,370 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622553419474_0001
2021-06-01 19:02:29,373 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-06-01 19:02:29,702 INFO conf.Configuration: resource-types.xml not found
2021-06-01 19:02:29,702 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-06-01 19:02:30,203 INFO impl.YarnClientImpl: Submitted application application_1622553419474_0001
2021-06-01 19:02:30,296 INFO mapreduce.Job: The url to track the job: http://LAPTOP-CGT0UDGT:8088/proxy/application_1622553419474_0001/
2021-06-01 19:02:30,297 INFO mapreduce.Job: Running job: job_1622553419474_0001
2021-06-01 19:02:53,930 INFO mapreduce.Job: Job job_1622553419474_0001 running in uber mode : false
2021-06-01 19:02:53,933 INFO mapreduce.Job:  map 0% reduce 0%
2021-06-01 19:03:02,239 INFO mapreduce.Job:  map 100% reduce 0%
2021-06-01 19:03:11,365 INFO mapreduce.Job:  map 100% reduce 100%
2021-06-01 19:03:12,395 INFO mapreduce.Job: Job job_1622553419474_0001 completed successfully
2021-06-01 19:03:12,951 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=846
                FILE: Number of bytes written=446519
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=536
                HDFS: Number of bytes written=441
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=6252
                Total time spent by all reduces in occupied slots (ms)=5823
```

Active Nodes in Hadoop Web interface at localhost:8088 :

Word Count on cmd :

```
                  Bytes Written=441

D:\hadoop-3.1.4\hadoop-3.1.4\bin>hadoop fs -cat /user/ketak/wmapred_out/part-r-00000
Although        1
Far      1
If       1
Piranhas        1
Their    1
When     1
and      1
animals 1
aquatic 1
are      4
attack. 1
be       1
bite     1
by       2
confronted      1
consider        1
dangerous       1
eat      1
eaten   2
entirely        1
fear     1
feed     1
first    1
fish     2
flee     1
for      1
harmless.       1
humans  2
humans. 1
instinct        1
is       1
large    1
makes    1
more     1
most     2
not      1
of       1
on       1
part     1
people  3
piranhas        3
piranhas.       1
plants. 1
quite    1
rarely  1
```

Downloading part-r-00000 from localhost:9870 :

localhost:9870/explorer.html#/user/ketak/wmapred_out

Hadoop   Overview   Datanodes   Datanode Volume Failures   Snapshot   Startup Progress   Utilities

Browse Directory

/user/ketak/wmapred_out   Go!

Show 25 entries   Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐	-rw-r--r--	ketak	supergroup	0 B	Jun 01 19:03	1	128 MB	_SUCCESS	🗑
☐	-rw-r--r--	ketak	supergroup	441 B	Jun 01 19:03	1	128 MB	part-r-00000	🗑

Showing 1 to 2 of 2 entries   Previous 1 Next

Hadoop, 2020.

Word Count file :

```
part-r-00000 (1) - Notepad

File  Edit  Format  View  Help
Although      1
Far      1
If       1
Piranhas      1
Their    1
When     1
and      1
animals 1
aquatic 1
are      4
attack. 1
be       1
bite     1
by       2
confronted    1
consider      1
dangerous     1
eat      1
eaten    2
entirely      1
fear     1
feed     1
first    1
fish     2
flee     1
for      1
harmless.     1
humans   2
humans.  1
instinct      1
is       1

            Ln 1, Col 1          100%   Unix (LF)
```