DIGITAL ASSIGNMENT- 2

Semester: ***Winter Semester 2023-24***

Course Code: BITE411L

Course Title: ***Big Data Analytics***

Faculty Name: ***RANICHANDRA C – SCORE***

NAME: ***POLI VARDHINI REDDY***

REGISTER NUMBER: 21BIT0382

BUSINESS INTELLIGENCE USING HADOOP EXAMPLE

Check Hadoop is installed or not

```
Command Prompt

Microsoft Windows [Version 10.0.22631.3447]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Lenovo>hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da
9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/hadoop/share/hadoop/common/hadoop-common-3.3.
6.jar

C:\Users\Lenovo>
```

```
Windows PowerShell

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Lenovo> start-all.sh
PS C:\Users\Lenovo>
[main 2024-04-26T13:36:30.155Z] update#setState idle
[main 2024-04-26T13:37:00.165Z] update#setState checking for updates
[main 2024-04-26T13:37:01.023Z] update#setState idle
```

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka
.ms/PSWindows

PS C:\Users\Lenovo> hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da
9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/hadoop/share/hadoop/common/hadoop-common-3.3.
6.jar
PS C:\Users\Lenovo> start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
PS C:\Users\Lenovo> jps
19152 ResourceManager
8672 DataNode
3828 NodeManager
14984 Jps
PS C:\Users\Lenovo>
```

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka
.ms/PSWindows

PS C:\Users\Lenovo> hdfs namenode -format
2024-04-26 20:47:13,859 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DESKTOP-2M6SFOU/192.168.138.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.6
STARTUP_MSG:   classpath = C:\hadoop\etc\hadoop;C:\hadoop\share\hadoop\commo
n;C:\hadoop\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;C:\h
adoop\share\hadoop\common\lib\audience-annotations-0.5.0.jar;C:\hadoop\share
\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop\share\hadoop\common\lib\checker-
qual-2.5.2.jar;C:\hadoop\share\hadoop\common\lib\commons-beanutils-1.9.4.jar
;C:\hadoop\share\hadoop\common\lib\commons-cli-1.2.jar;C:\hadoop\share\hadoo
p\common\lib\commons-codec-1.15.jar;C:\hadoop\share\hadoop\common\lib\common
s-collections-3.2.2.jar;C:\hadoop\share\hadoop\common\lib\commons-compress-1
.21.jar;C:\hadoop\share\hadoop\common\lib\commons-configuration2-2.8.0.jar;C
:\hadoop\share\hadoop\common\lib\commons-daemon-1.0.13.jar;C:\hadoop\share\h
adoop\common\lib\commons-io-2.8.0.jar;C:\hadoop\share\hadoop\common\lib\comm
ons-lang3-3.12.0.jar;C:\hadoop\share\hadoop\common\lib\commons-logging-1.1.3
.jar;C:\hadoop\share\hadoop\common\lib\commons-math3-3.1.1.jar;C:\hadoop\sha
re\hadoop\common\lib\commons-net-3.9.0.jar;C:\hadoop\share\hadoop\common\lib
\commons-text-1.10.0.jar;C:\hadoop\share\hadoop\common\lib\curator-client-5.
2.0.jar;C:\hadoop\share\hadoop\common\lib\curator-framework-5.2.0.jar;C:\had
oop\share\hadoop\common\lib\curator-recipes-5.2.0.jar;C:\hadoop\share\hadoop
\common\lib\dnsjava-2.1.7.jar;C:\hadoop\share\hadoop\common\lib\failureacces
s-1.0.jar;C:\hadoop\share\hadoop\common\lib\gson-2.9.0.jar;C:\hadoop\share\h
adoop\common\lib\guava-27.0-jre.jar;C:\hadoop\share\hadoop\common\lib\hadoop
-annotations-3.3.6.jar;C:\hadoop\share\hadoop\common\lib\hadoop-auth-3.3.6.j
ar;C:\hadoop\share\hadoop\common\lib\hadoop-shaded-guava-1.1.1.jar;C:\hadoop
\share\hadoop\common\lib\hadoop-shaded-protobuf_3_7-1.1.1.jar;C:\hadoop\shar
e\hadoop\common\lib\httpclient-4.5.13.jar;C:\hadoop\share\hadoop\common\lib\
httpcore-4.4.13.jar;C:\hadoop\share\hadoop\common\lib\j2objc-annotations-1.1
.jar;C:\hadoop\share\hadoop\common\lib\jackson-annotations-2.12.7.jar;C:\had
oop\share\hadoop\common\lib\jackson-core-2.12.7.jar;C:\hadoop\share\hadoop\c
ommon\lib\jackson-core-asl-1.9.13.jar;C:\hadoop\share\hadoop\common\lib\jack
```

```
PS C:\Users\Lenovo> cd \hadoop
PS C:\hadoop> cd sbin
PS C:\hadoop\sbin> start.dfs
start.dfs : The term 'start.dfs' is not recognized as the name of a
cmdlet, function, script file, or operable program. Check the spelling of
the name, or if a path was included, verify that the path is correct and
try again.
At line:1 char:1
+ start.dfs
+ ~~~~~~~~~
    + CategoryInfo          : ObjectNotFound: (start.dfs:String) [], Comma
   ndNotFoundException
    + FullyQualifiedErrorId : CommandNotFoundException

PS C:\hadoop\sbin> start-dfs
PS C:\hadoop\sbin>
```

```
Apache Hadoop Distribution
-3.3.6.jar
STARTUP_MSG:   build = https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c; compiled by 'ub
untu' on 2023-06-18T08:22Z
STARTUP_MSG:   java = 1.8.0_411
************************************************************/
2024-04-26 20:51:46,290 INFO namenode.NameNode: createNameNode []
2024-04-26 20:51:46,554 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-26 20:51:46,742 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-04-26 20:51:46,742 INFO impl.MetricsSystemImpl: NameNode metrics system started
2024-04-26 20:51:46,787 INFO namenode.NameNodeUtils: fs.defaultFS is file:///
2024-04-26 20:51:47,051 ERROR namenode.NameNode: Failed to start namenode.
java.lang.IllegalArgumentException: Invalid URI for NameNode address (check fs.defaultFS): file:/// has no authority.
        at org.apache.hadoop.hdfs.DFSUtilClient.getNNAddress(DFSUtilClient.java:781)
        at org.apache.hadoop.hdfs.DFSUtilClient.getNNAddressCheckLogical(DFSUtilClient.java:810)
        at org.apache.hadoop.hdfs.DFSUtilClient.getNNAddress(DFSUtilClient.java:772)
```

```
Apache Hadoop Distribution

024-04-26 20:51:51,620 INFO server.session: node0 Stopped scavenging
024-04-26 20:51:51,620 INFO handler.ContextHandler: Stopped o.e.j.s.ServletContextHandler@42bc14c1{static,/static,file:
//C:/hadoop/share/hadoop/hdfs/webapps/static/,STOPPED}
024-04-26 20:51:51,620 INFO handler.ContextHandler: Stopped o.e.j.s.ServletContextHandler@7b94089b{logs,/logs,file:///C
/hadoop/logs/,STOPPED}
024-04-26 20:51:51,620 INFO datanode.DataNode: Waiting up to 30 seconds for transfer threads to complete
024-04-26 20:51:51,620 INFO ipc.Server: Stopping server on 9867
024-04-26 20:51:51,636 INFO impl.MetricsSystemImpl: Stopping DataNode metrics system...
024-04-26 20:51:51,636 INFO impl.MetricsSystemImpl: DataNode metrics system stopped.
024-04-26 20:51:51,637 INFO impl.MetricsSystemImpl: DataNode metrics system shutdown complete.
024-04-26 20:51:51,637 INFO datanode.DataNode: Shutdown complete.
024-04-26 20:51:51,637 ERROR datanode.DataNode: Exception in secureMain
ava.io.IOException: No services to connect, missing NameNode address.
      at org.apache.hadoop.hdfs.server.datanode.BlockPoolManager.refreshNamenodes(BlockPoolManager.java:165)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.startDataNode(DataNode.java:1755)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.<init>(DataNode.java:564)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.makeInstance(DataNode.java:3148)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.instantiateDataNode(DataNode.java:3054)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.createDataNode(DataNode.java:3098)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.secureMain(DataNode.java:3242)
      at org.apache.hadoop.hdfs.server.datanode.DataNode.main(DataNode.java:3266)
```

```
Apache Hadoop Distribution
24-04-26 20:52:49,686 INFO resourcemanager.ResourceManager: STARTUP_MSG:
/************************************************************
RTUP_MSG: Starting ResourceManager
RTUP_MSG:   host = DESKTOP-2M6SFOU/192.168.56.1
RTUP_MSG:   args = []
RTUP_MSG:   version = 3.3.6
RTUP_MSG:   classpath = C:\hadoop\etc\hadoop;C:\hadoop\etc\hadoop;C:\hadoop\etc\
hadoop\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;C:\hadoop\sha
s-0.5.0.jar;C:\hadoop\share\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop\share\had
\hadoop\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\hadoop\share\had
oop\share\hadoop\common\lib\commons-codec-1.15.jar;C:\hadoop\share\hadoop\commo
doop\share\hadoop\common\lib\commons-compress-1.21.jar;C:\hadoop\share\hadoop\c
r;C:\hadoop\share\hadoop\common\lib\commons-daemon-1.0.13.jar;C:\hadoop\share\ha
hadoop\share\hadoop\common\lib\commons-lang3-3.12.0.jar;C:\hadoop\share\hadoop\c
adoop\share\hadoop\common\lib\commons-math3-3.1.1.jar;C:\hadoop\share\hadoop\com
share\hadoop\common\lib\commons-text-1.10.0.jar;C:\hadoop\share\hadoop\common\li

Apache Hadoop Distribution
024-04-26 20:52:49,974 INFO nodemanager.NodeManager: STARTUP_MSG:
/************************************************************
TARTUP_MSG: Starting NodeManager
TARTUP_MSG:   host = DESKTOP-2M6SFOU/192.168.56.1
TARTUP_MSG:   args = []
TARTUP_MSG:   version = 3.3.6
TARTUP_MSG:   classpath = C:\hadoop\etc\hadoop;C:\hadoop\etc\hadoop;C:\hadoop\et
:\hadoop\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;C:\hadoop\s
ons-0.5.0.jar;C:\hadoop\share\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop\share\h
:\hadoop\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\hadoop\share\had
adoop\share\hadoop\common\lib\commons-codec-1.15.jar;C:\hadoop\share\hadoop\comm
hadoop\share\hadoop\common\lib\commons-compress-1.21.jar;C:\hadoop\share\hadoop\
jar;C:\hadoop\share\hadoop\common\lib\commons-daemon-1.0.13.jar;C:\hadoop\share\
:\hadoop\share\hadoop\common\lib\commons-lang3-3.12.0.jar;C:\hadoop\share\hadoop
\hadoop\share\hadoop\common\lib\commons-math3-3.1.1.jar;C:\hadoop\share\hadoop\c
\share\hadoop\common\lib\commons-text-1.10.0.jar;C:\hadoop\share\hadoop\common\
are\hadoop\common\lib\curator-framework-5.2.0.jar;C:\hadoop\share\hadoop\common
share\hadoop\common\lib\dnsjava-2.1.7.jar;C:\hadoop\share\hadoop\common\lib\fai
\common\lib\gson-2.9.0.jar;C:\hadoop\share\hadoop\common\lib\guava-27.0-jre.jar
-annotations-3.3.6.jar;C:\hadoop\share\hadoop\common\lib\hadoop-auth-3.3.6.jar;
shaded-guava-1.1.1.jar;C:\hadoop\share\hadoop\common\lib\hadoop-shaded-protobuf
mon\lib\httpclient-4.5.13.jar;C:\hadoop\share\hadoop\common\lib\httpcore-4.4.13
objc-annotations-1.1.jar;C:\hadoop\share\hadoop\common\lib\jackson-annotations-

Windows PowerShell
.3.6\data\namenode\current\fsimage_0000000000000000000, C:\hadoop-3.3.6\data
\namenode\current\fsimage_0000000000000000000.md5, C:\hadoop-3.3.6\data\name
node\current\seen_txid, C:\hadoop-3.3.6\data\namenode\current\VERSION]
2024-04-26 20:48:25,363 INFO common.Storage: Storage directory C:\hadoop-3.3
.6\data\namenode has been successfully formatted.
2024-04-26 20:48:25,413 INFO namenode.FSImageFormatProtobuf: Saving image fi
le C:\hadoop-3.3.6\data\namenode\current\fsimage.ckpt_0000000000000000000 us
ing no compression
2024-04-26 20:48:25,563 INFO namenode.FSImageFormatProtobuf: Image file C:\h
adoop-3.3.6\data\namenode\current\fsimage.ckpt_0000000000000000000 of size 4
01 bytes saved in 0 seconds .
2024-04-26 20:48:25,585 INFO namenode.NNStorageRetentionManager: Going to re
tain 1 images with txid >= 0
2024-04-26 20:48:25,609 INFO namenode.FSNamesystem: Stopping services starte
d for active state
2024-04-26 20:48:25,609 INFO namenode.FSNamesystem: Stopping services starte
d for standby state
2024-04-26 20:48:25,615 INFO namenode.FSImage: FSImageSaver clean checkpoint
: txid=0 when meet shutdown.
2024-04-26 20:48:25,615 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-2M6SFOU/192.168.138.1
************************************************************/
PS C:\Users\Lenovo> cd \hadoop
PS C:\hadoop> cd sbin
PS C:\hadoop\sbin> start.dfs
start.dfs : The term 'start.dfs' is not recognized as the name of a
cmdlet, function, script file, or operable program. Check the spelling of
the name, or if a path was included, verify that the path is correct and
try again.
At line:1 char:1
+ start.dfs
+ ~~~~~~~~~
    + CategoryInfo          : ObjectNotFound: (start.dfs:String) [], Comma
  ndNotFoundException
    + FullyQualifiedErrorId : CommandNotFoundException

PS C:\hadoop\sbin> start-dfs
PS C:\hadoop\sbin> start-yarn
starting yarn daemons
PS C:\hadoop\sbin>
```

```
PS C:\hadoop\sbin> start-dfs
PS C:\hadoop\sbin> start-yarn
starting yarn daemons
PS C:\hadoop\sbin>
```

8:53 PM
26/04/2024

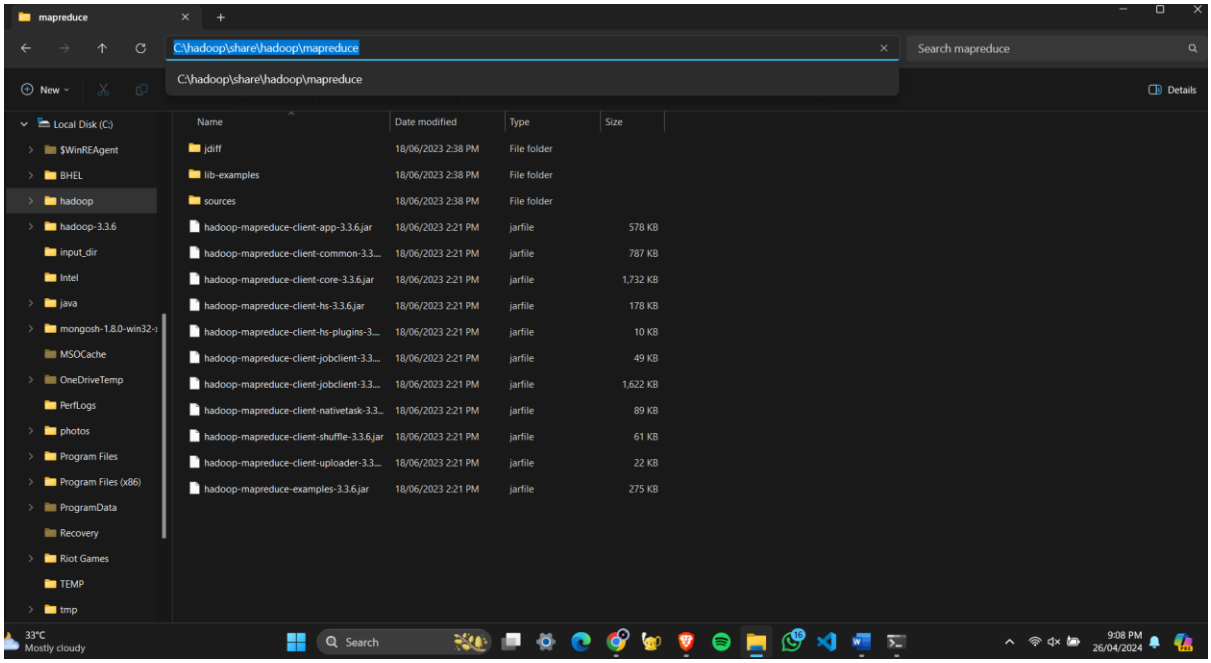
```
PS C:\hadoop\sbin> start-dfs
PS C:\hadoop\sbin> start-yarn
starting yarn daemons
PS C:\hadoop\sbin> hadoop fs -mkdir /input_dir
PS C:\hadoop\sbin>
```

8:58 PM
26/04/2024

```
PS C:\hadoop\sbin> start-dfs
PS C:\hadoop\sbin> start-yarn
starting yarn daemons
PS C:\hadoop\sbin> hadoop fs -mkdir /input_dir
PS C:\hadoop\sbin> hadoop fs -put C:\hadoop\data.txt\input_dir
put: `C:/hadoop/data.txt/input_dir': No such file or directory
PS C:\hadoop\sbin> hadoop fs -put C:/hadoop/data.txt/input_dir
put: `/hadoop/data.txt/input_dir': No such file or directory
PS C:\hadoop\sbin> hadoop fs -put C:/hadoop/data.txt
PS C:\hadoop\sbin> hadoop fs -ls /input_dir/
PS C:\hadoop\sbin>
```

```
C:\hadoop\sbin>hadoop fs -ls /input_dir/
Found 1 items
-rw-r--r--   1 JAWAD supergroup        111 2022-06-07 00:27 /input_dir/data.txt
```

```
C:\hadoop\sbin>hadoop fs -cat /input_dir/data.txt
Pakistan
India
China
Bangladesh
Pakistan
Iran
America
India
Iran
America
China
Pakistan
Iraq
China
C:\hadoop\sbin>
```




```
C:\hadoop\sbin>hadoop jar C:/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.3.jar wordcount /input_dir /out
put_dir
2022-06-07 00:31:04,238 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-06-07 00:31:08,149 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
JAWAD/.staging/job_1654543374869_0001
2022-06-07 00:31:09,337 INFO input.FileInputFormat: Total input files to process : 1
2022-06-07 00:31:09,868 INFO mapreduce.JobSubmitter: number of splits:1
2022-06-07 00:31:10,963 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1654543374869_0001
2022-06-07 00:31:10,967 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-07 00:31:11,918 INFO conf.Configuration: resource-types.xml not found
2022-06-07 00:31:11,920 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-07 00:31:13,412 INFO impl.YarnClientImpl: Submitted application application_1654543374869_0001
2022-06-07 00:31:14,091 INFO mapreduce.Job: The url to track the job: http://DESKTOP-MGLR0BE:8088/proxy/application_1654
543374869_0001/
2022-06-07 00:31:14,094 INFO mapreduce.Job: Running job: job_1654543374869_0001
```

```
C:\hadoop\sbin>hadoop fs -cat /output_dir/*
America 2
Bangladesh      1
China   3
India   2
Iran    2
Iraq    1
Pakistan        3

C:\hadoop\sbin>_
```

```
PS C:\hadoop\sbin> stop-all.cmd
This script is Deprecated. Instead use stop-dfs.cmd and stop-yarn.cmd

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.
stopping yarn daemons

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.
PS C:\hadoop\sbin>
```

BUSINESS INTELLIGENCE USING HADOOP EXAMPLE

Business intelligence (BI) using Hadoop involves processing and analyzing large volumes of data to extract meaningful insights and make informed business decisions. Hadoop provides a scalable and distributed framework for storing, processing, and analyzing big data. In Java, you can develop BI applications using Hadoop's MapReduce paradigm or higher-level frameworks like Apache Hive or Apache Spark.

CODE

MAPPER CLASS:

```
import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;


public class SalesMapper extends Mapper<LongWritable, Text, Text, DoubleWritable> {


   @Override

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

      // Assuming each line in the input represents a sales transaction in the format:
product_name,sales_amount

      String[] parts = value.toString().split(",");

      if (parts.length == 2) {

         String product = parts[0];
```

```
            double salesAmount = Double.parseDouble(parts[1]);
            context.write(new Text(product), new DoubleWritable(salesAmount));
        }
    }
}
```

REDUCER CLASS

```
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;

public class SalesReducer extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {

    @Override
    public void reduce(Text key, Iterable<DoubleWritable> values, Context context) throws
IOException, InterruptedException {
        double totalSales = 0;
        for (DoubleWritable value : values) {
            totalSales += value.get();
        }
        context.write(key, new DoubleWritable(totalSales));
    }
}
```

MAIN CLASS

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
```

```
public class SalesAnalysis {

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "sales analysis");

        job.setJarByClass(SalesAnalysis.class);
        job.setMapperClass(SalesMapper.class);
        job.setReducerClass(SalesReducer.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(DoubleWritable.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        TextInputFormat.addInputPath(job, new Path(args[0]));
        TextOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

INPUT:

```
ProductA,100.00
ProductB,50.00
ProductA,75.00
ProductC,120.00
ProductB,30.00
```

OUTPUT:

```
ProductA    175.0
ProductB    80.0
ProductC    120.0
```

INPUT:

```
Laptop,1200.00
Phone,800.00
Tablet,500.00
Phone,700.00
Laptop,1500.00
Tablet,600.00
```

OUTPUT:

```
Laptop    2700.0
Phone     1500.0
Tablet    1100.0
```