SUSHMITHA P A

1NT19IS170

C2

Execrcise-3: Map Reduce Programs: Use the Hadoop framework to write a custom MapReduce program to perform word count operation on a custom data set.

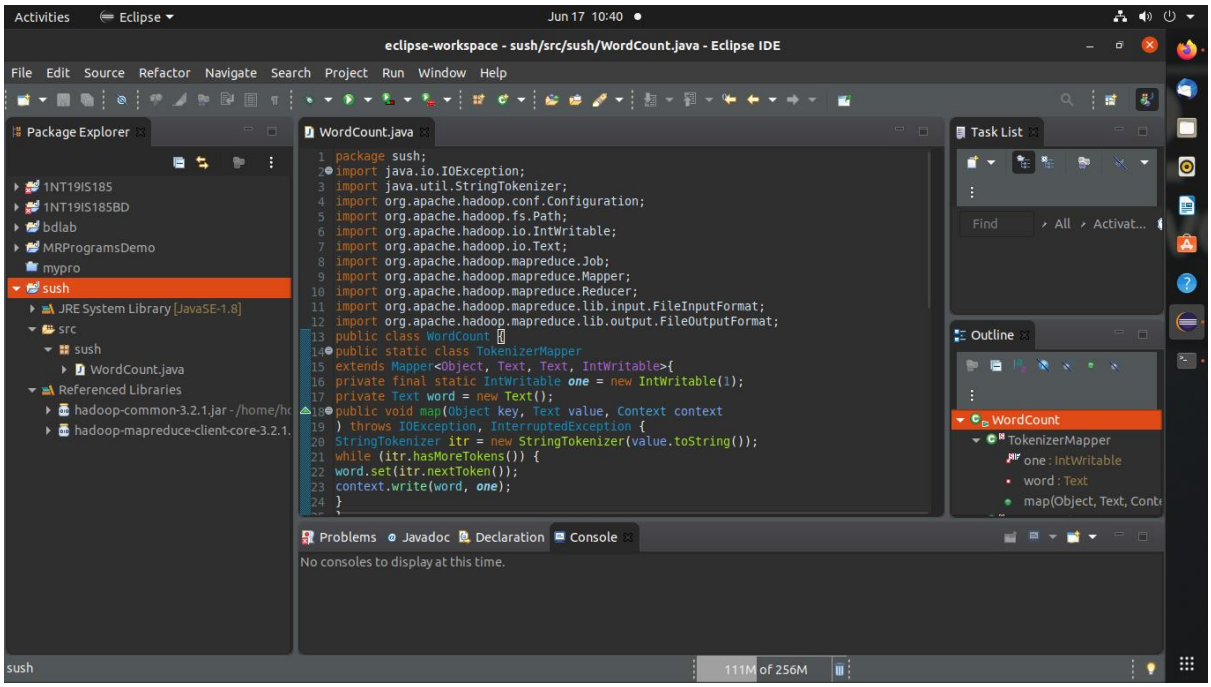
```
package sushmitha;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
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 final static IntWritable one = new IntWritable(1);
private Text word = new Text();
public void map(Object key, Text value, Context context
```

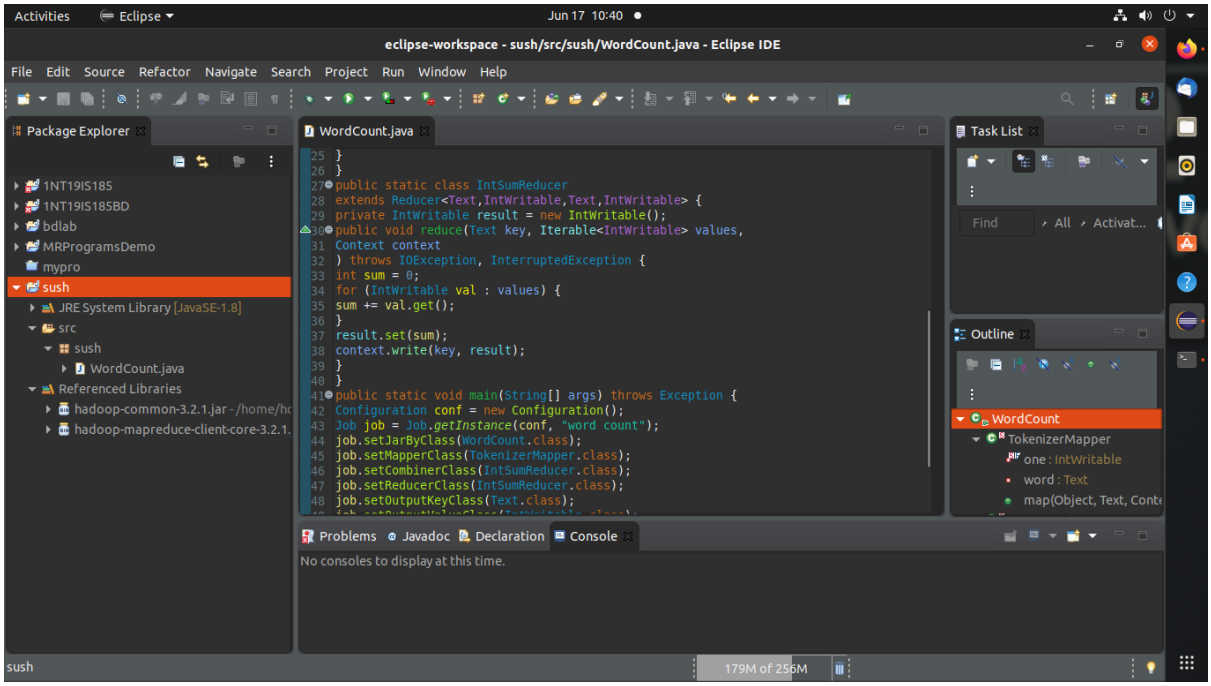
```
) throws IOException, InterruptedException {
StringTokenizer itr = new StringTokenizer(value.toString());
while (itr.hasMoreTokens()) {
word.set(itr.nextToken());
context.write(word, one);
}
}
}
public static class IntSumReducer
extends Reducer<Text,IntWritable,Text,IntWritable> {
private IntWritable 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, "word count");
```

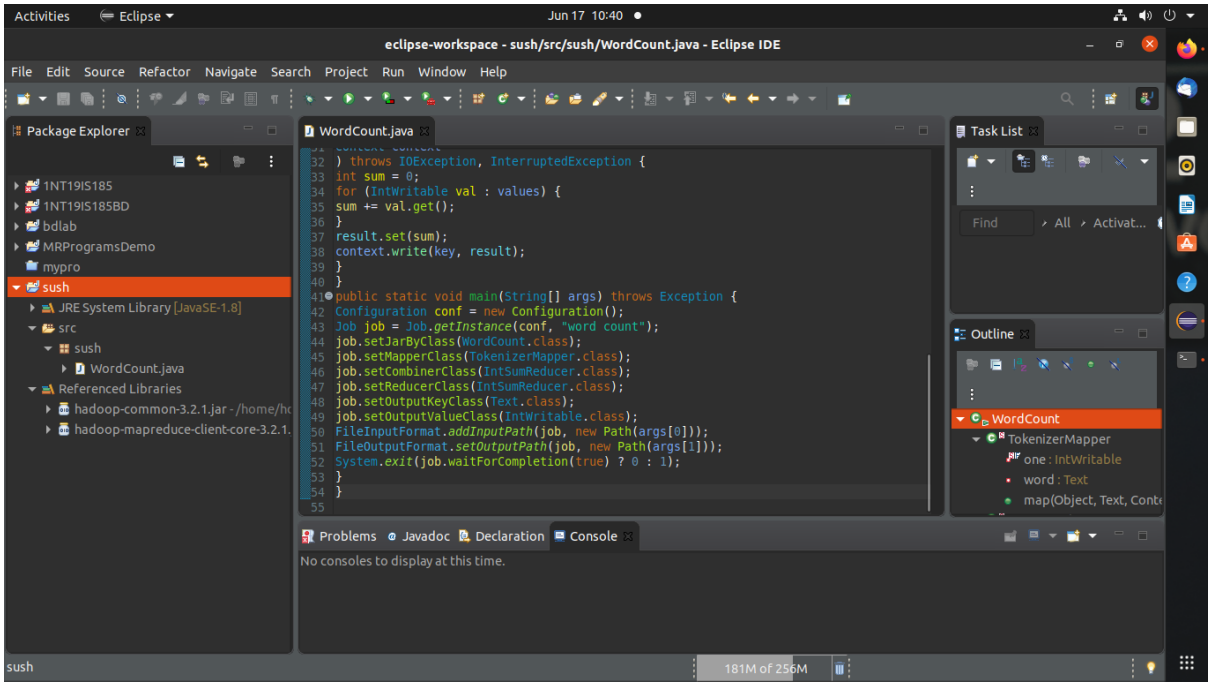
```
job.setJarByClass(WordCount.class);
job.setMapperClass(TokenizerMapper.class);
job.setCombinerClass(IntSumReducer.class);
job.setReducerClass(IntSumReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```

EXECUTION

Compiling the program in eclipse




Activities
Eclipse
Jun 17 10:40
eclipse-workspace - sush/src/sush/WordCount.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
1NT19IS185
1NT19IS185BD
bdlab
MRProgramsDemo
mypro
sush
JRE System Library [JavaSE-1.8]
src
sush
WordCount.java
Referenced Libraries
hadoop-common-3.2.1.jar - /home/hc
hadoop-mapreduce-client-core-3.2.1.
WordCount.java
25 }
26 }
27 public static class IntSumReducer
28 extends Reducer<Text,IntWritable,Text,IntWritable> {
29 private IntWritable result = new IntWritable();
30 public void reduce(Text key, Iterable<IntWritable> values,
31 Context context
32) throws IOException, InterruptedException {
33 int sum = 0;
34 for (IntWritable val : values) {
35 sum += val.get();
36 }
37 result.set(sum);
38 context.write(key, result);
39 }
40 }
41 public static void main(String[] args) throws Exception {
42 Configuration conf = new Configuration();
43 Job job = Job.getInstance(conf, "word count");
44 job.setJarByClass(WordCount.class);
45 job.setMapperClass(TokenizerMapper.class);
46 job.setCombinerClass(IntSumReducer.class);
47 job.setReducerClass(IntSumReducer.class);
48 job.setOutputKeyClass(Text.class);
Task List
Find
All
Activat...
Outline
WordCount
TokenizerMapper
one : IntWritable
word : Text
map(Object, Text, Conte
Problems
Javadoc
Declaration
Console
No consoles to display at this time.
sush
179M of 256M


Activities
Eclipse
Jun 17 10:40
eclipse-workspace - sush/src/sush/WordCount.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
1NT19IS185
1NT19IS185BD
bdlab
MRProgramsDemo
mypro
sush
JRE System Library [JavaSE-1.8]
src
sush
WordCount.java
Referenced Libraries
hadoop-common-3.2.1.jar - /home/hc
hadoop-mapreduce-client-core-3.2.1.
WordCount.java
32) throws IOException, InterruptedException {
33 int sum = 0;
34 for (IntWritable val : values) {
35 sum += val.get();
36 }
37 result.set(sum);
38 context.write(key, result);
39 }
40 }
41 public static void main(String[] args) throws Exception {
42 Configuration conf = new Configuration();
43 Job job = Job.getInstance(conf, "word count");
44 job.setJarByClass(WordCount.class);
45 job.setMapperClass(TokenizerMapper.class);
46 job.setCombinerClass(IntSumReducer.class);
47 job.setReducerClass(IntSumReducer.class);
48 job.setOutputKeyClass(Text.class);
49 job.setOutputValueClass(IntWritable.class);
50 FileInputFormat.addInputPath(job, new Path(args[0]));
51 FileOutputFormat.setOutputPath(job, new Path(args[1]));
52 System.exit(job.waitForCompletion(true) ? 0 : 1);
53 }
54 }
55
Task List
Find
All
Activat...
Outline
WordCount
TokenizerMapper
one : IntWritable
word : Text
map(Object, Text, Conte
Problems
Javadoc
Declaration
Console
No consoles to display at this time.
sush
181M of 256M

```
hdoop@admin1-HP-280-G4-MT-Business-PC: ~/hadoop-3.2.1/sbin
hdoop@admin1-HP-280-G4-MT-Business-PC:~$ cd $HADOOP_HOME
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1$ cd sbin
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ jps
6592 SecondaryNameNode
10147 Jps
6935 NodeManager
4473 org.eclipse.equinox.launcher_1.5.600.v20191014-2022.jar
6349 DataNode
6782 ResourceManager
6191 NameNode
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 6191.  Stop it first.
Starting datanodes
localhost: datanode is running as process 6349.  Stop it first.
Starting secondary namenodes [admin1-HP-280-G4-MT-Business-PC]
admin1-HP-280-G4-MT-Business-PC: secondarynamenode is running as process 6592.  Stop it first.
Starting resourcemanager
resourcemanager is running as process 6782.  Stop it first.
Starting nodemanagers
localhost: nodemanager is running as process 6935.  Stop it first.
```

Creating an input directory

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -mkdir -p ~ /input
```

Appending the contents of the file using -appendToFile

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -appendToFile - ~/input/test.txt
hello hello this is sushmitha2022-06-17 10:19:25,722 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

Running the job by passing in the input and output directories

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hadoop jar /home/hdoop/Desktop/sush.jar ~/input ~/out
2022-06-17 10:21:30,739 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-17 10:21:31,003 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-06-17 10:21:31,018 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1655439662941_0001
2022-06-17 10:21:31,106 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 10:21:31,206 INFO input.FileInputFormat: Total input files to process : 1
2022-06-17 10:21:31,285 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 10:21:31,335 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 10:21:31,359 INFO mapreduce.JobSubmitter: number of splits:1
2022-06-17 10:21:31,467 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 10:21:31,483 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1655439662941_0001
2022-06-17 10:21:31,483 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-17 10:21:31,603 INFO conf.Configuration: resource-types.xml not found
2022-06-17 10:21:31,603 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-17 10:21:31,777 INFO impl.YarnClientImpl: Submitted application application_1655439662941_0001
2022-06-17 10:21:31,810 INFO mapreduce.Job: The url to track the job: http://admin1-HP-280-G4-MT-Business-PC:8088/proxy/application_1655439662941_0001/
2022-06-17 10:21:31,810 INFO mapreduce.Job: Running job: job_1655439662941_0001
2022-06-17 10:21:37,913 INFO mapreduce.Job: Job job_1655439662941_0001 running in uber mode : false
2022-06-17 10:21:37,915 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-17 10:21:41,033 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-17 10:21:45,061 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-17 10:21:45,079 INFO mapreduce.Job: Job job_1655439662941_0001 completed successfully
2022-06-17 10:21:45,158 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=54
                FILE: Number of bytes written=450985
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=141
                HDFS: Number of bytes written=32
                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
```

```
                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)=1418
                Total time spent by all reduces in occupied slots (ms)=1627
                Total time spent by all map tasks (ms)=1418
                Total time spent by all reduce tasks (ms)=1627
                Total vcore-milliseconds taken by all map tasks=1418
                Total vcore-milliseconds taken by all reduce tasks=1627
                Total megabyte-milliseconds taken by all map tasks=1452032
                Total megabyte-milliseconds taken by all reduce tasks=1666048
        Map-Reduce Framework
                Map input records=1
                Map output records=5
                Map output bytes=50
                Map output materialized bytes=54
                Input split bytes=112
                Combine input records=5
                Combine output records=4
                Reduce input groups=4
                Reduce shuffle bytes=54
                Reduce input records=4
                Reduce output records=4
                Spilled Records=8
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=69
                CPU time spent (ms)=850
                Physical memory (bytes) snapshot=495153152
                Virtual memory (bytes) snapshot=5083226112
                Total committed heap usage (bytes)=514326528
                Peak Map Physical memory (bytes)=298160128
                Peak Map Virtual memory (bytes)=2537017344
                Peak Reduce Physical memory (bytes)=196993024
```

```
        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=29
        File Output Format Counters
                Bytes Written=32
```

Concate the part* file to get the output

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -cat ~/out/part*
2022-06-17 10:29:46,871 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hello   2
is      1
sushmitha       1
this    1
```