

# MapReduce - Lập trình chương trình WordCount

Apr 24, 2022

## I. Thử nghiệm với hàm mẫu WordCount của Hadoop

Trong thư mục **C:\hadoop-3.3.0\share\hadoop\mapreduce** Hadoop đã có sẵn chương trình MapReduce **hadoop-mapreduce-examples-3.3.0.jar**. Ta sẽ thử nghiệm bài toán đếm từ bằng cách tạo ra file text chứa dữ liệu và đầu ra mong muốn là các cặp **[từ: số lượng xuất hiện]**

Bước 1: Tạo file data.txt

Nội dung của file data.txt là:

```
Bus Car bus  
car train car  
bus car train  
bus TRAIN BUS  
buS caR CAR  
car BUS TRAIN
```

Bước 2: Tạo thư mục input tại hdfs và lưu file data.txt

Tạo thư mục **input** trong hdfs với câu lệnh:

```
hdfs dfs -mkdir /input
```

Đẩy file **data.txt** vào folder **input** vừa tạo:

```
hdfs dfs -put "C:\input\data.txt" /input
```

*Lưu ý: Thay "C:\input\data.txt" bằng nơi lưu trữ file trong máy*

Vào trang quản lý NameNode <http://localhost:9870/> để kiểm tra file

← → ↻

localhost:9870/dfshealth.html#tab-overview

🔖 ☆ 🗖 👤 ⋮

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

Browse the file system

Logs

Log Level

Metrics

Configuration

Process Thread Dump

## Overview 'localhost:9000' (✓active)

Started:	Sun Apr 24 17:37:06 +0700 2022
Version:	3.3.0, raa9ff1b71b1d350f5bac59cf2a01ec470da649af
Compiled:	Tue Jul 07 01:44:00 +0700 2020 by brahma from branch-3.3.0
Cluster ID:	CID-002a012d-a224-4ec1-be63-f7060f7004ee
Block Pool ID:	BP-469905230-192-160-23-120-1650051200140

## Summary

Security is off.

Setmode is off.

3 files and directories, 1 blocks (1 replicated blocks: 0 erasure coded block groups) = 4 total filesystem object(s).

← → ↻

localhost:9870/dfshealth.html#tab-overview

🔖 ☆ 🗖 👤 ⋮

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

## Browse Directory

/

Go!

🗖 ⬆ 🗖 🗑

Show 25 entries

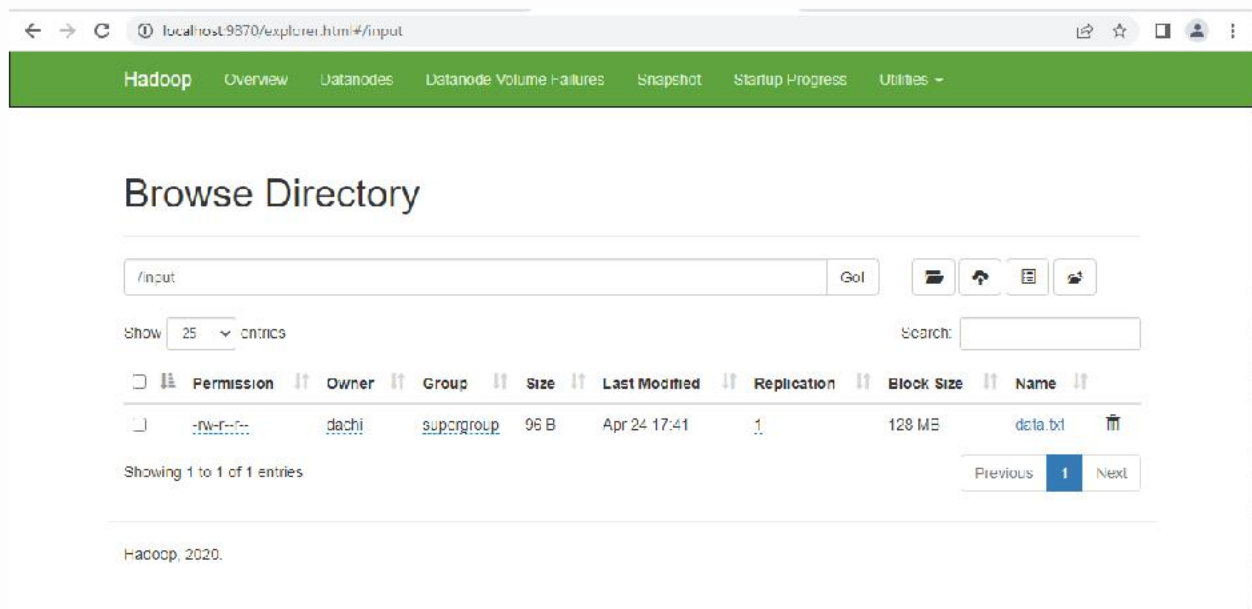
Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 17:41	0	0 B	input	🗑

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2020



### Bước 3: Chạy chương trình MapReduce và xem kết quả

Chương trình mẫu MapReduce của Hadoop nằm tại **C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar**. Ta sẽ thử nghiệm đầu vào chương trình là file **data.txt** và kết quả sẽ được lưu tại folder **/output**, lệnh thực hiện"

```
hadoop jar "C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar" wordcount /input/data.txt /output
```

```

C:\Users\dachi>hadoop jar "C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar" wordcount /input/
data.txt /output
2022-04-24 17:50:04,417 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-04-24 17:50:04,990 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
dachi/.staging/job_1650796630442_0001
2022-04-24 17:50:05,336 INFO input.FileInputFormat: Total input files to process : 1
2022-04-24 17:50:05,448 INFO mapreduce.JobSubmitter: number of splits:1
2022-04-24 17:50:05,595 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1650796630442_0001
2022-04-24 17:50:05,597 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-24 17:50:05,810 INFO conf.Configuration: resource-types.xml not found
2022-04-24 17:50:05,811 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-24 17:50:06,323 INFO impl.YarnClientImpl: Submitted application application_1650796630442_0001
2022-04-24 17:50:06,368 INFO mapreduce.Job: The url to track the job: http://DESKTOP-QBKQL72:8080/proxy/application_1650
796630442_0001/
2022-04-24 17:50:06,369 INFO mapreduce.Job: Running job: job_1650796630442_0001
2022-04-24 17:50:15,650 INFO mapreduce.Job: Job job_1650796630442_0001 running in uber mode : false
2022-04-24 17:50:15,651 INFO mapreduce.Job: map 0% reduce 0%
2022-04-24 17:50:20,779 INFO mapreduce.Job: map 100% reduce 0%
2022-04-24 17:50:25,855 INFO mapreduce.Job: map 100% reduce 100%
2022-04-24 17:50:25,852 INFO mapreduce.Job: Job job_1650796630442_0001 completed successfully
2022-04-24 17:50:25,963 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=173
  FILE: Number of bytes written=531069
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  
```

Xem kết quả thu được:

← → ↻ localhost:9870/explore.html#

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

## Browse Directory

/

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 17:41	0	0 B	input	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 17:50	0	0 B	output	
<input type="checkbox"/>	drwx-----	dachi	supergroup	0 B	Apr 24 17:50	0	0 B	tmp	

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2020.

← → ↻ localhost:9870/explore.html#/output

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

## Browse Directory

/output

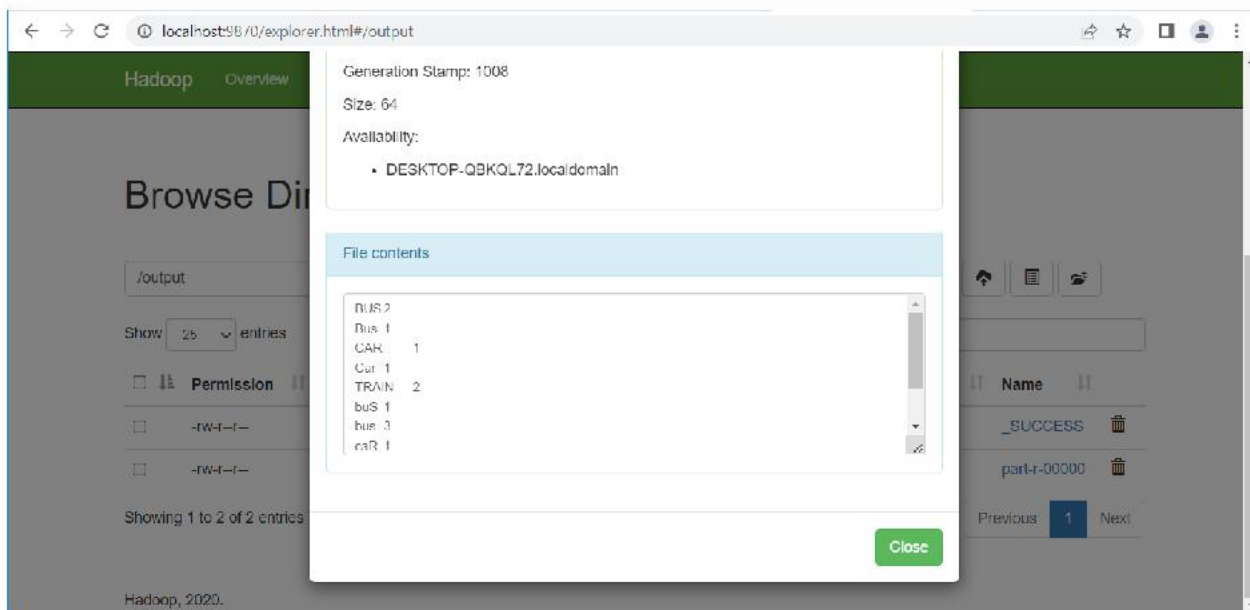
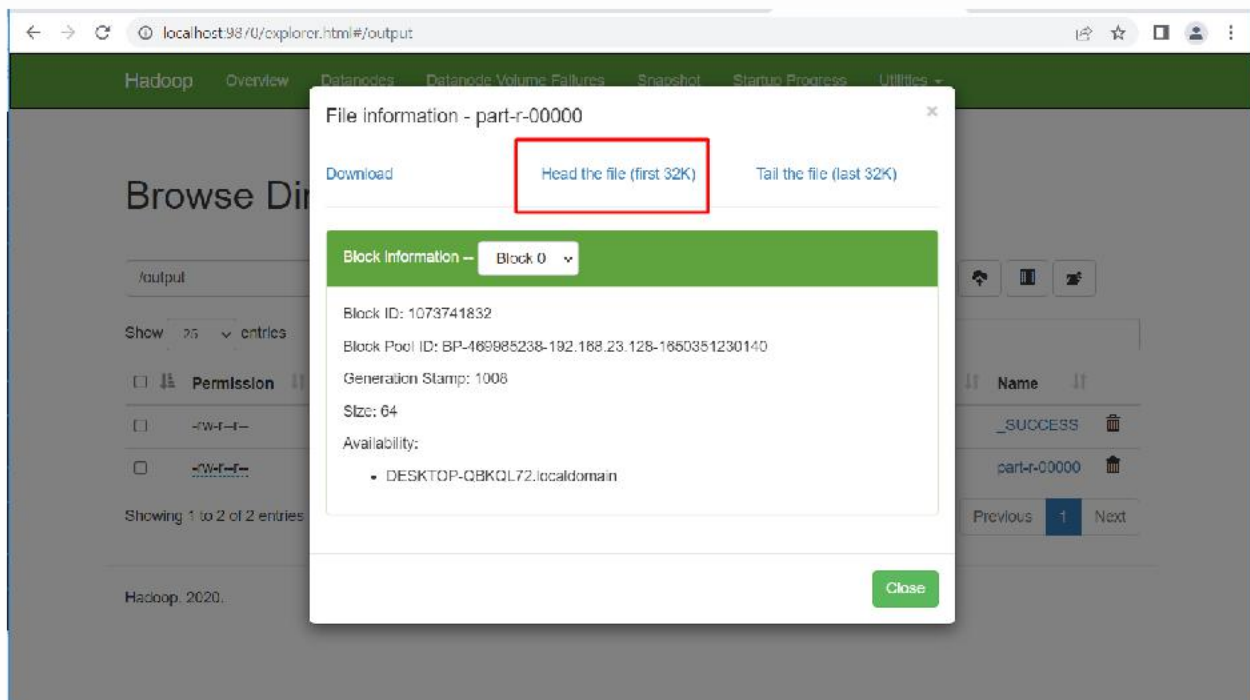
Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	dachi	supergroup	0 B	Apr 24 17:50	1	128 MB	SUCCESS	
<input type="checkbox"/>	-rw-r--r--	dachi	supergroup	107 B	Apr 24 17:50	1	128 MB	part-r-00000	

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2020.



Hoặc dùng lệnh cmd:

```
hdfs dfs -cat /output/part-r-00000
```

Lưu ý: thay /output/part-r-00000 thành đường dẫn file muốn xem



```
Command Prompt
Microsoft Windows [Version 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

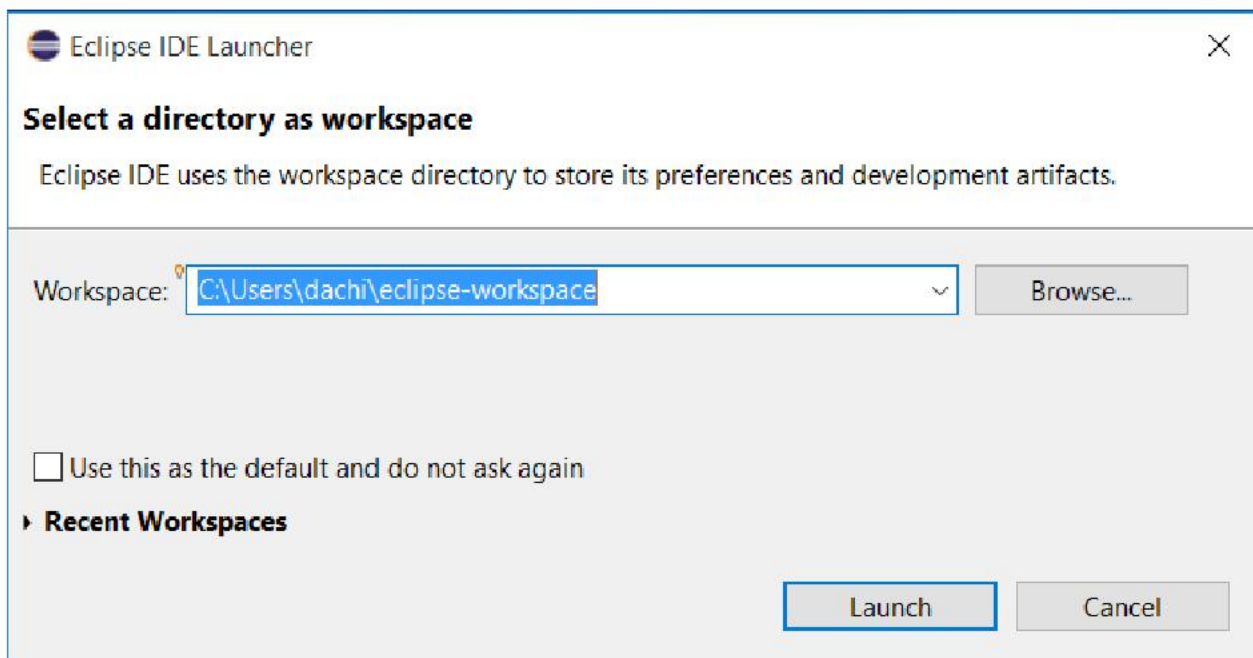
C:\Users\dachi>hdfs dfs -cat /output/part-r-00000
BUS      2
Bus      1
CAR      1
Car      1
TRAIN    2
buS      1
bus      3
caR      1
car      4
train    2
```

Như vậy ta đã chạy thành công chương trình mẫu MapReduce của Hadoop cung cấp.

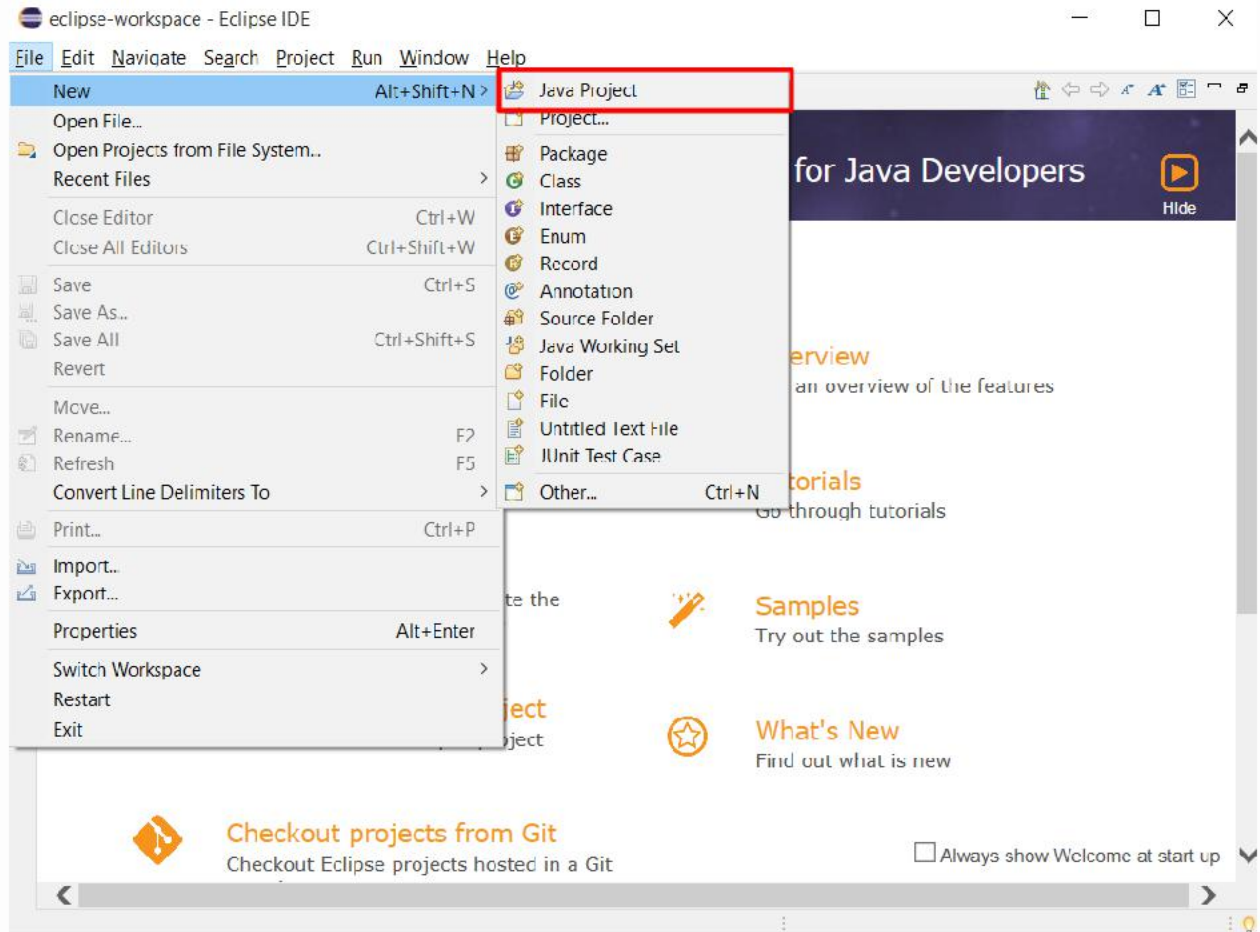
## II. Lập trình chương trình WordCount bằng Eclipse

Bước 1: Tạo project

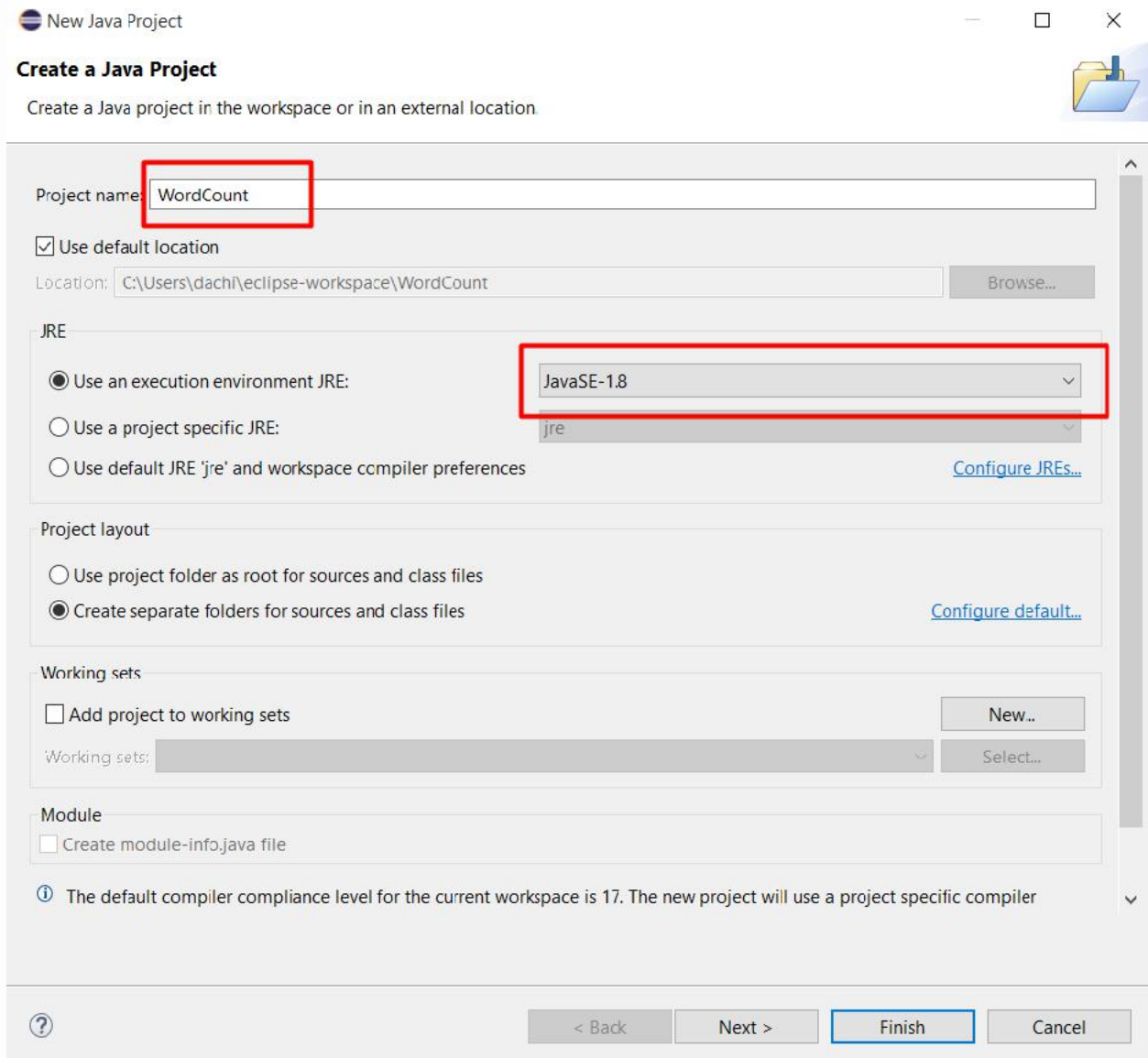
Mở chương trình Eclipse. Chọn **workspace** (nên để mặc định)



Tạo project Java, chọn **File > New > Java Project**



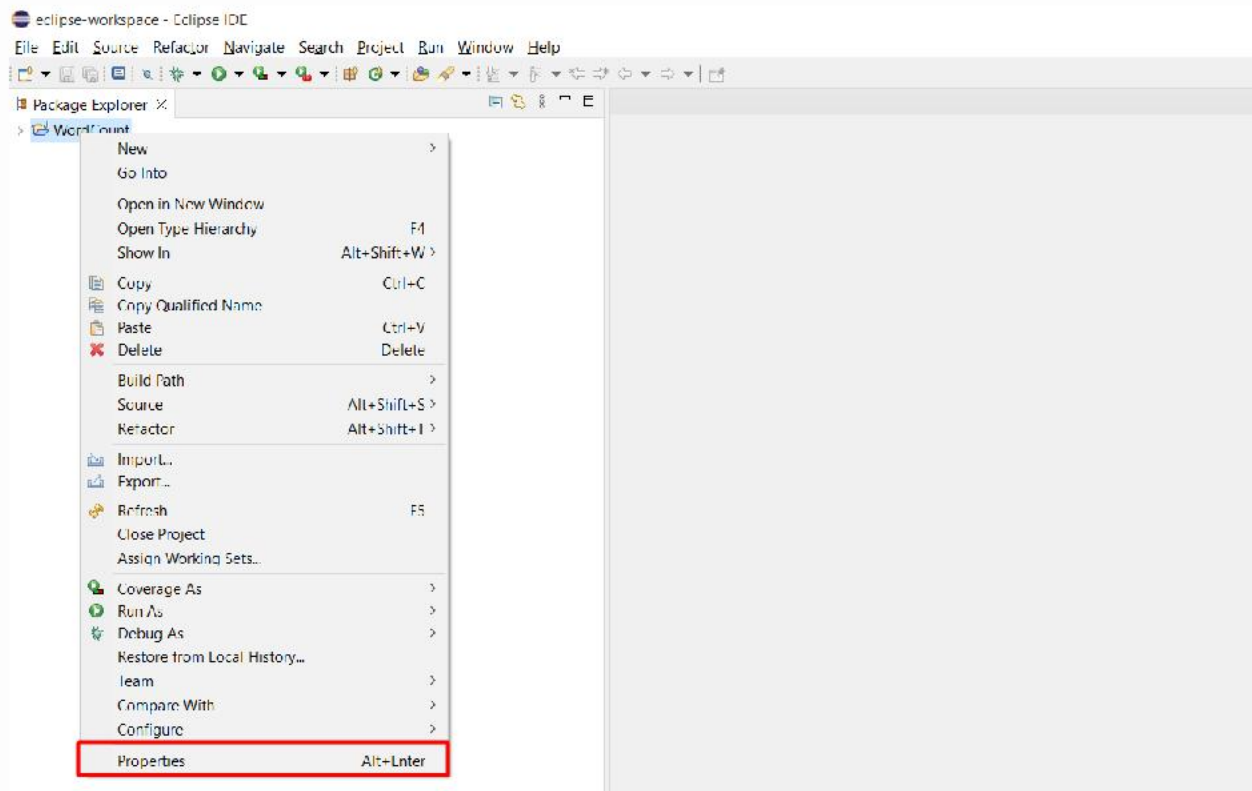
Đặt tên project là **WordCount** và chọn môi trường là **JavaSE-1.8**. Xong ấn **Finish**.



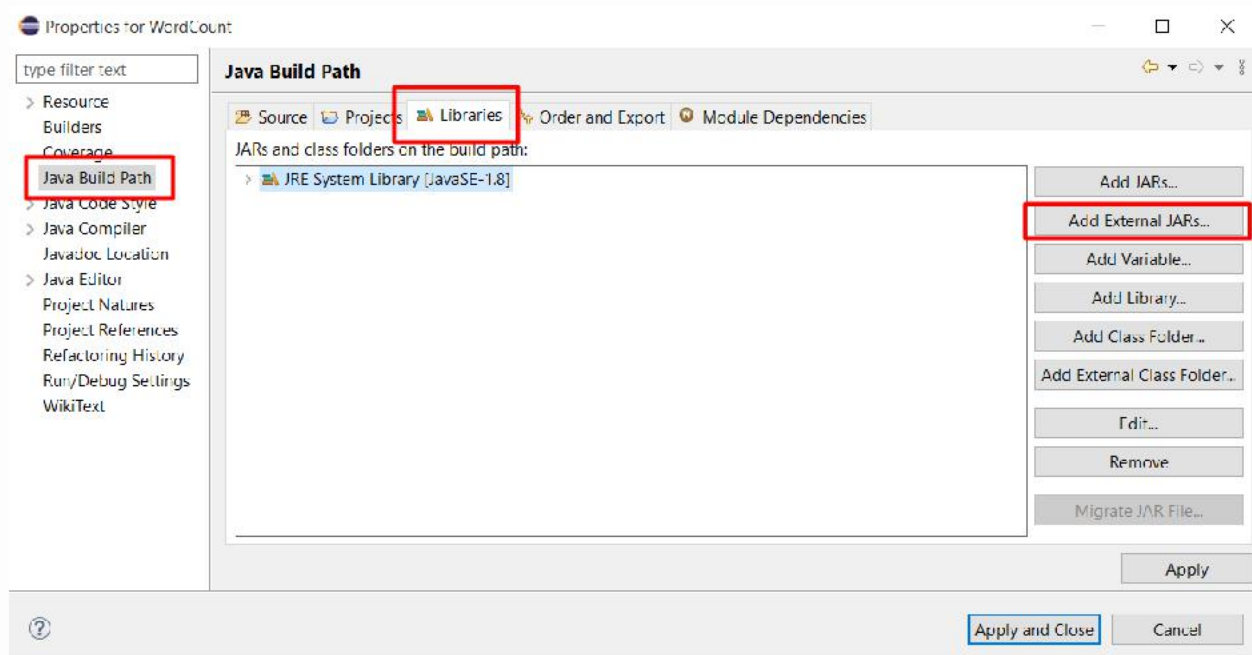
Bước 2: Thêm thư viện cần thiết để chạy MapReduce

Chuột phải vào project **WordCount** chọn **Properties**

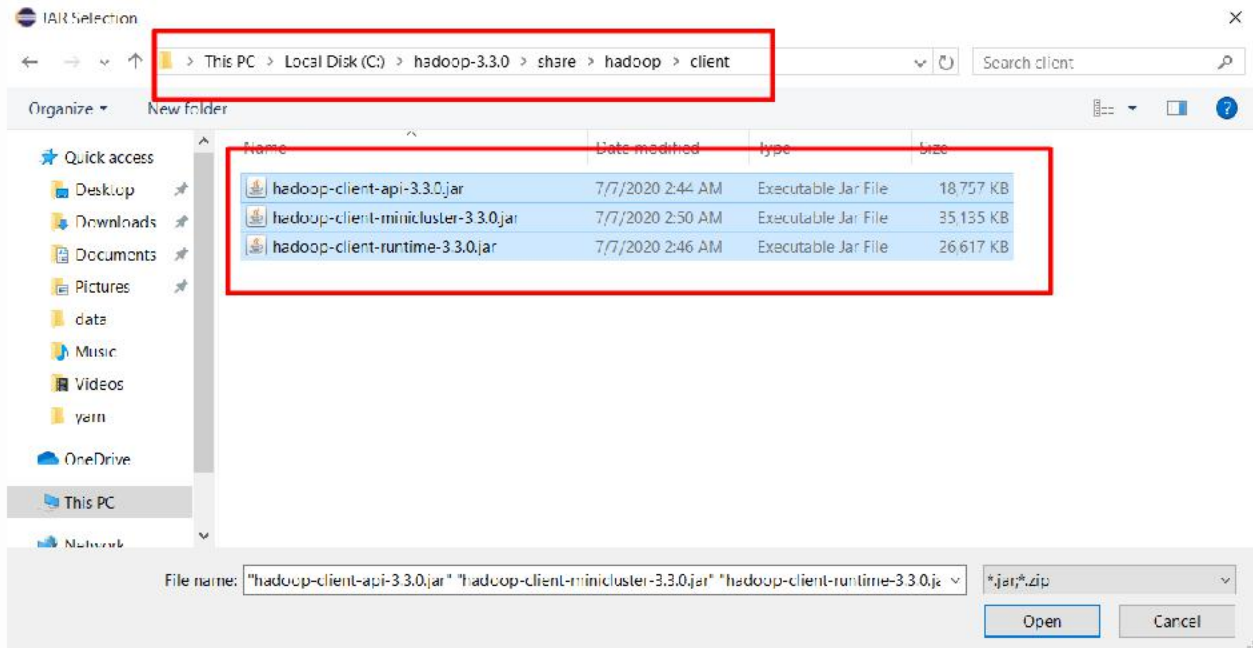




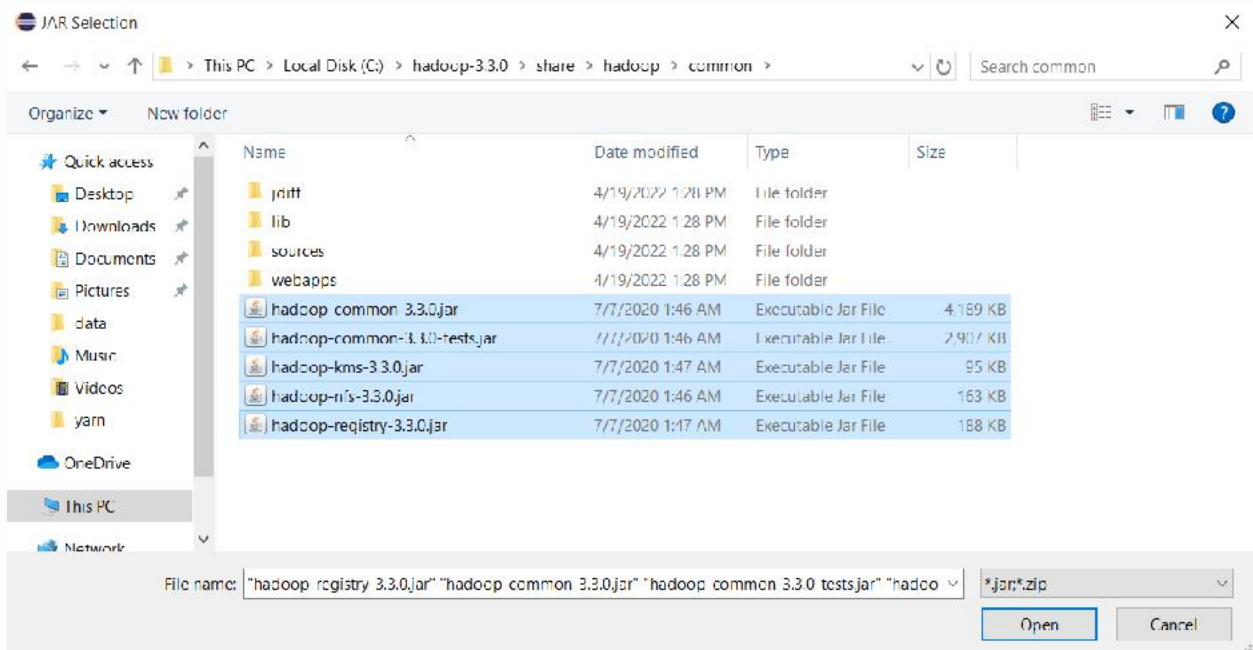
Chọn **Java Build Path**, chọn tab **Libraries** và bấm **Add External JARs**



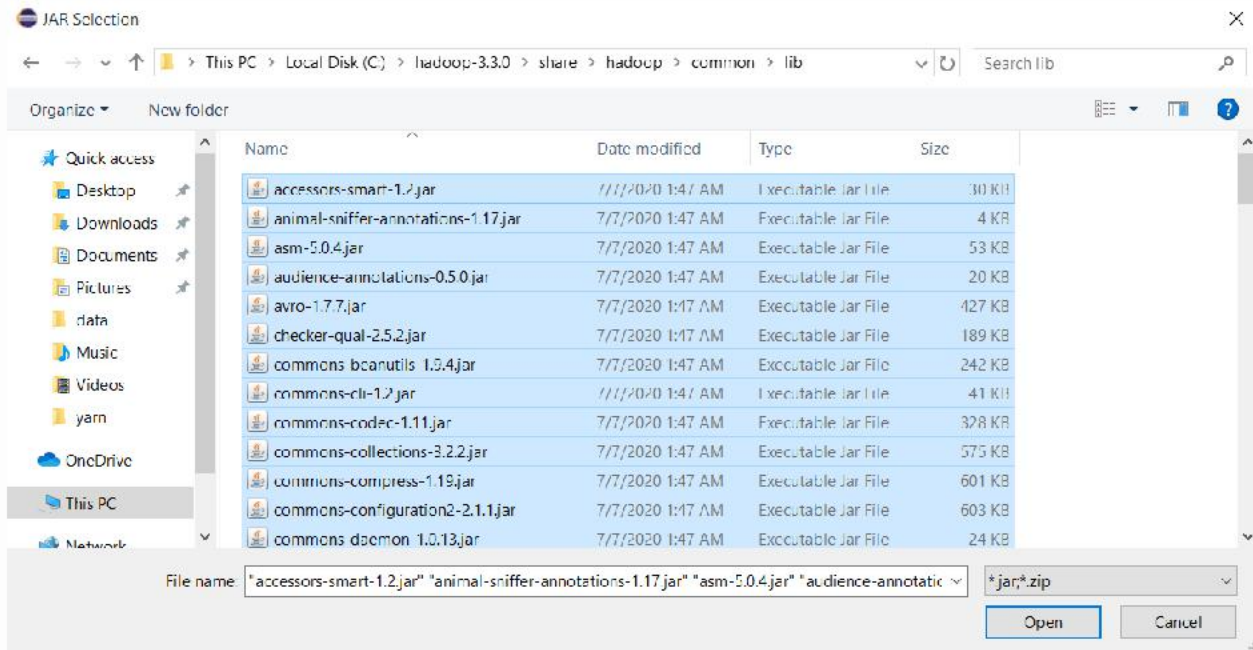
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\client** và ấn **Open**



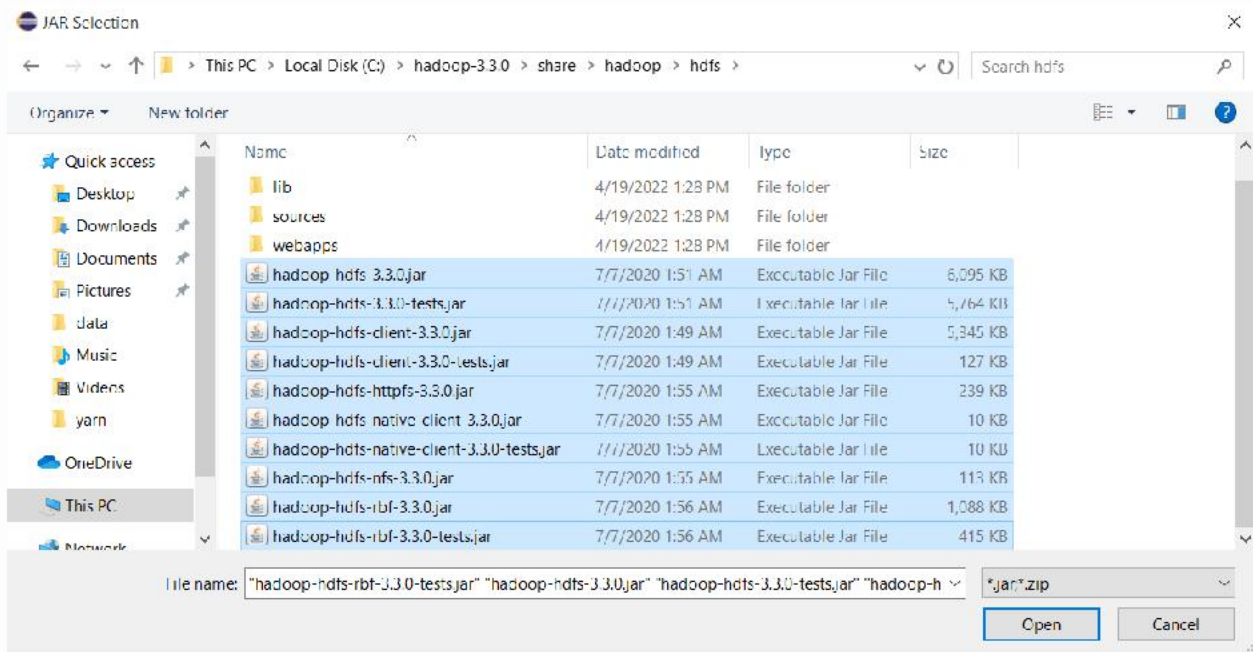
Tương tự chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common** và ấn **Open**



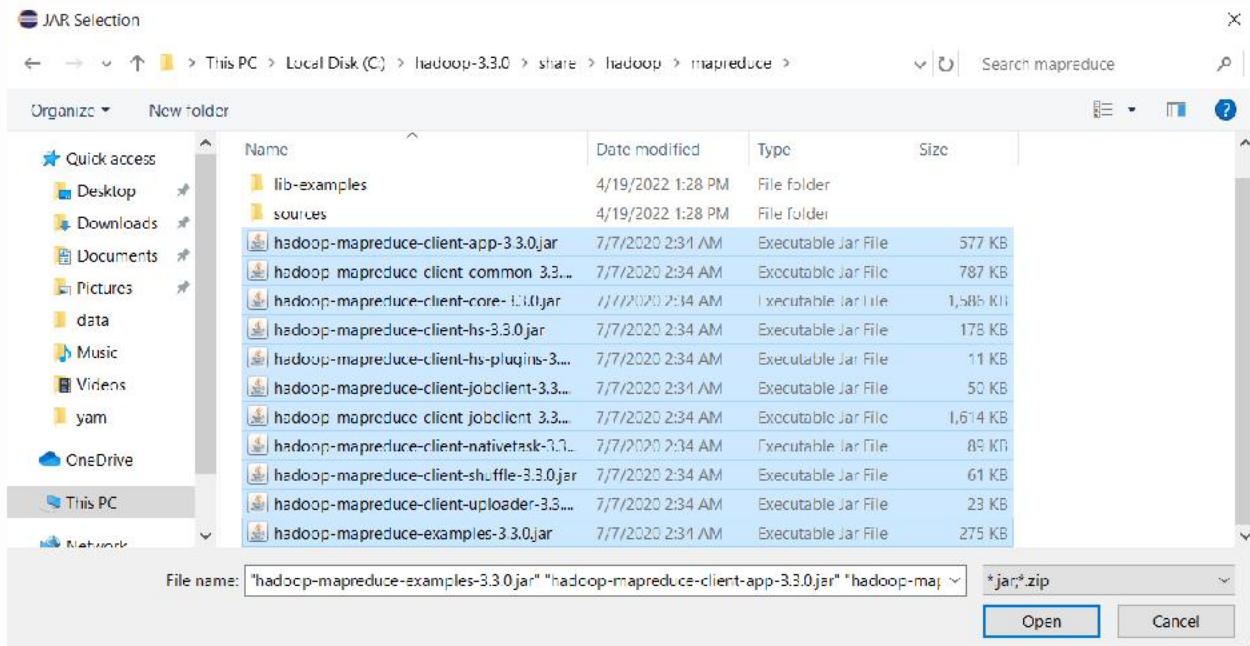
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common\lib** và ấn **Open**



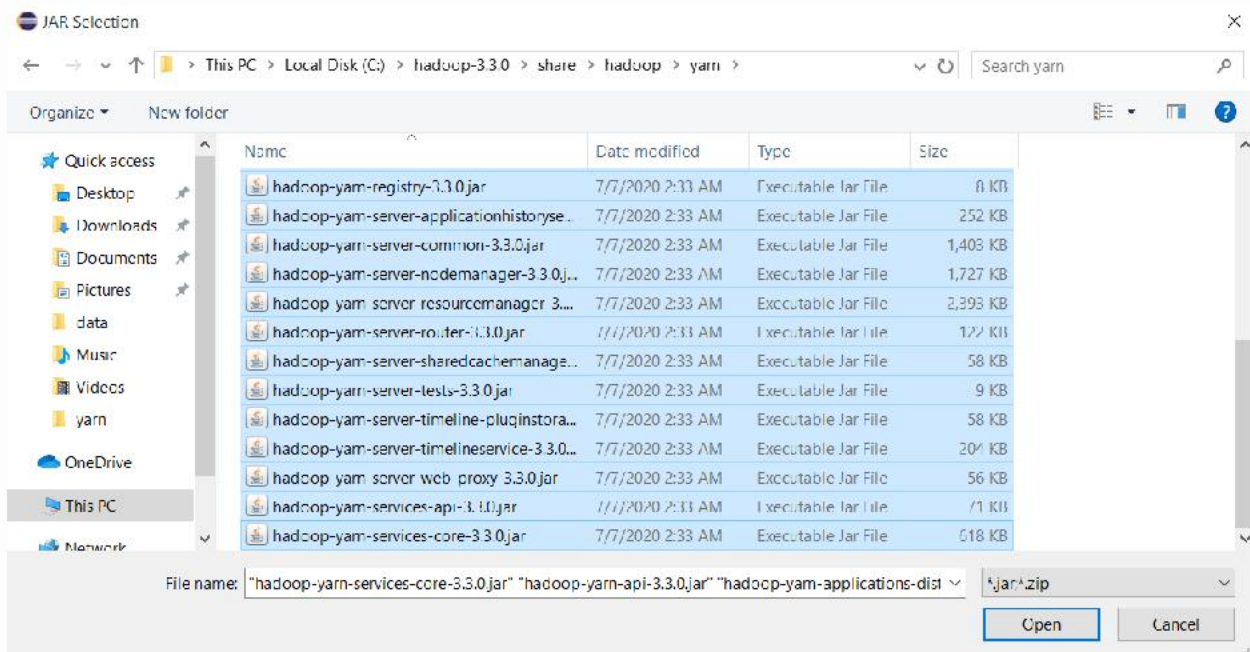
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\hdfs** và ấn **Open**



Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\mapreduce** và ấn **Open**

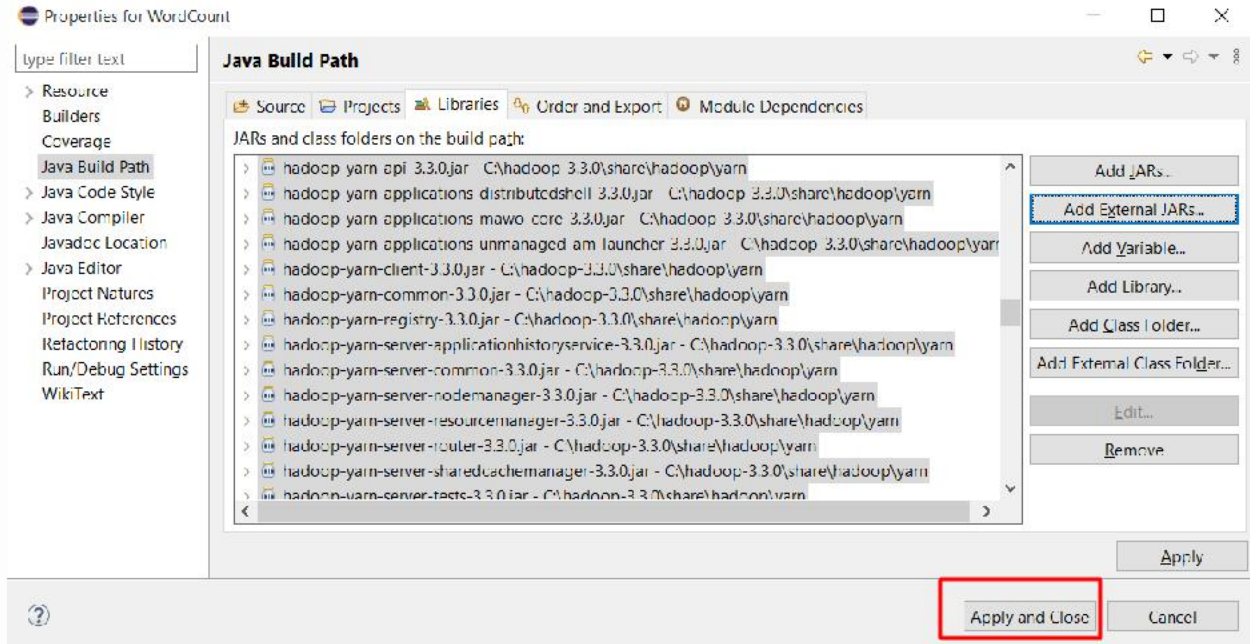


Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\yarn** và ấn **Open**



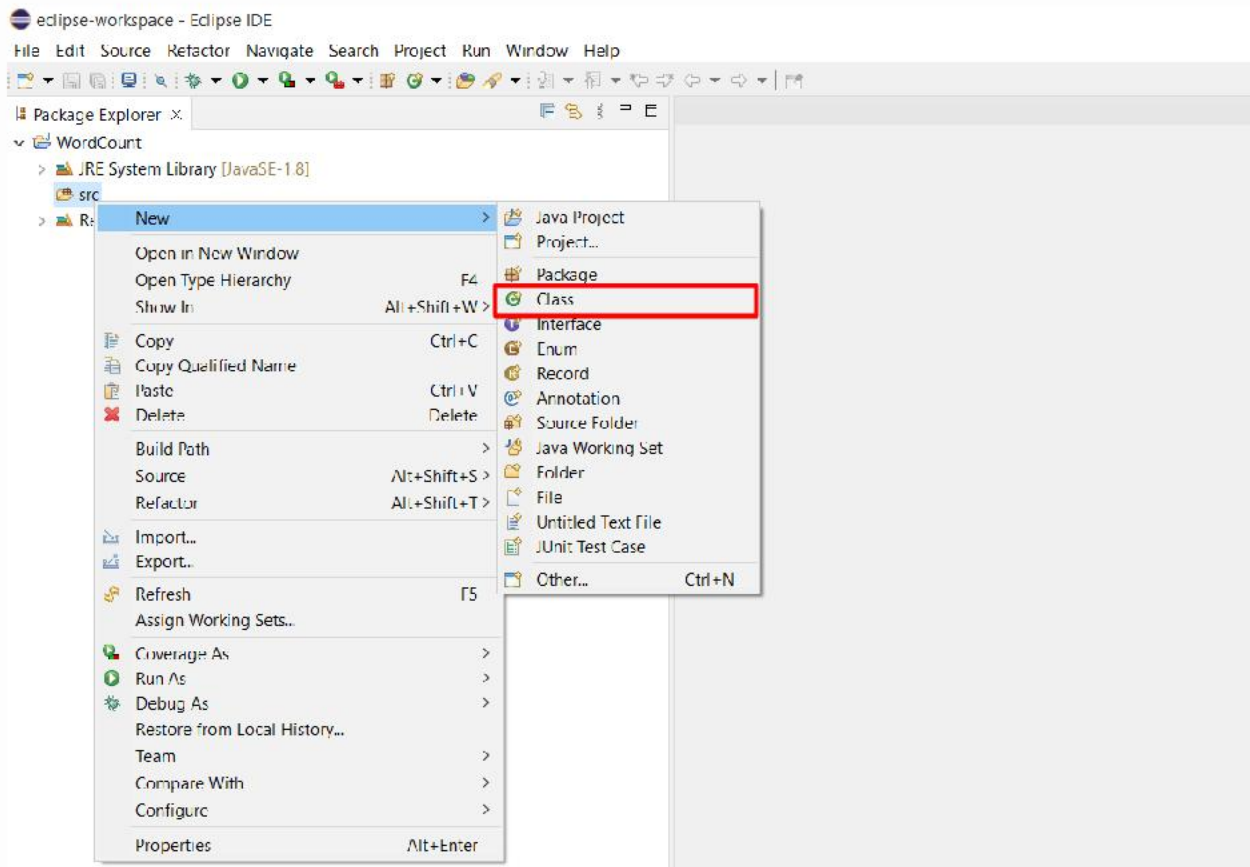
Ấn **Apply** and **Close**





### Bước 3: Tạo class xử lý tác vụ MapReduce

Double click vào project **WordCount**, chuột phải vào **src** và chọn **New > Class**



Tạo class để xử lý nhiệm vụ **Map**, đặt tên là **WordCountMapper**

New Java Class

**Java Class**

⚠ The use of the default package is discouraged.

Source folder: WordCount/src Browse...

Package: (default) Browse...

☐ Enclosing type: Browse...

---

Name: WordCountMapper

Modifiers: ☒ public ☐ package ☐ private ☐ protected  
☐ abstract ☐ final ☐ static

Superclass: java.lang.Object Browse...

Interfaces: Add... Remove

Which method stubs would you like to create?

☐ public static void main(String[] args)

☐ Constructors from superclass

☒ Inherited abstract methods

Do you want to add comments? (Configure templates and default value [here](#))

☐ Generate comments

Finish Cancel

Nội dung bên trong file **WordCountMapper.java**:

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
```



```

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

public class WordCountMapper extends Mapper <LongWritable, Text, Text, IntWritable>
{
    private Text wordToken = new Text();
    public void map(LongWritable key, Text value, Context context) throws IOException,
    InterruptedException
    {
        StringTokenizer tokens = new StringTokenizer(value.toString()); //Dividing
String into tokens
        while (tokens.hasMoreTokens())
        {
            wordToken.set(tokens.nextToken());
            context.write(wordToken, new IntWritable(1));
        }
    }
}

```

Tương tự tạo class xử lý nhiệm vụ **Reduce**, đặt tên là **WordCountReducer**:

```

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

public class WordCountReducer extends Reducer <Text, IntWritable, Text, IntWritable>
{
    private IntWritable count = new IntWritable();
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws
    IOException, InterruptedException
    {
        int valueSum = 0;
        for (IntWritable val : values)
        {
            valueSum += val.get();
        }
        count.set(valueSum);
        context.write(key, count);
    }
}

```

Và tạo class **WordCount** chứa hàm **main** để khởi chạy chương trình:

```

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;

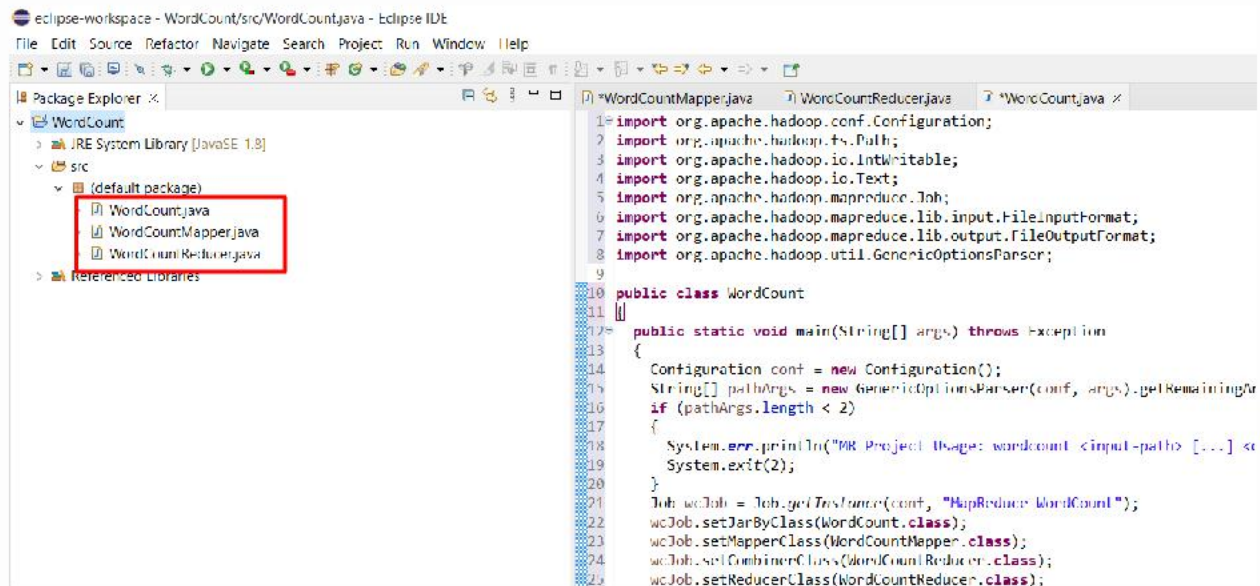
```

```

import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

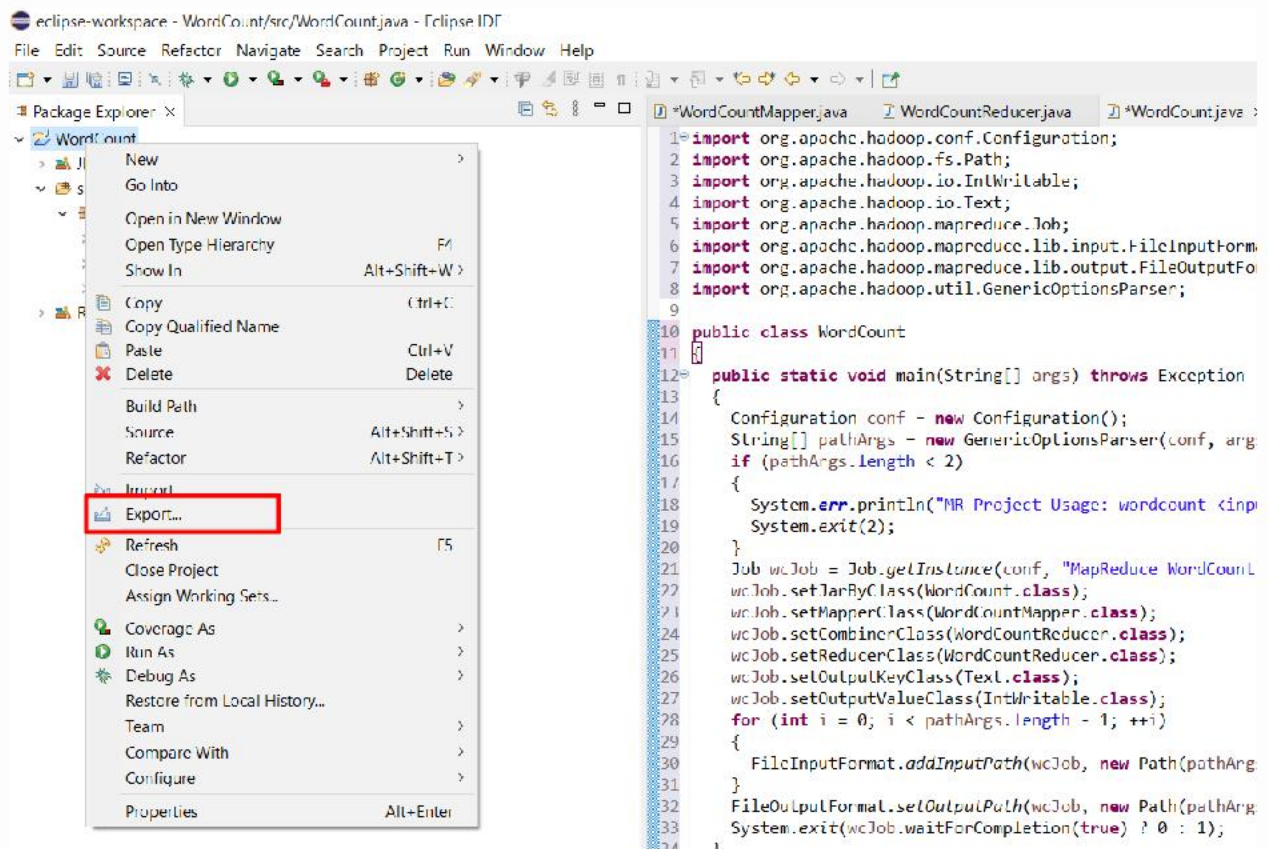
public class WordCount
{
    public static void main(String[] args) throws Exception
    {
        Configuration conf = new Configuration();
        String[] pathArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
        if (pathArgs.length < 2)
        {
            System.err.println("MR Project Usage: wordcount <input-path> [...] <output-path>");
            System.exit(2);
        }
        Job wcJob = Job.getInstance(conf, "MapReduce WordCount");
        wcJob.setJarByClass(WordCount.class);
        wcJob.setMapperClass(WordCountMapper.class);
        wcJob.setCombinerClass(WordCountReducer.class);
        wcJob.setReducerClass(WordCountReducer.class);
        wcJob.setOutputKeyClass(Text.class);
        wcJob.setOutputValueClass(IntWritable.class);
        for (int i = 0; i < pathArgs.length - 1; ++i)
        {
            FileInputFormat.addInputPath(wcJob, new Path(pathArgs[i]));
        }
        FileOutputFormat.setOutputPath(wcJob, new Path(pathArgs[pathArgs.length - 1]));
        System.exit(wcJob.waitForCompletion(true) ? 0 : 1);
    }
}

```

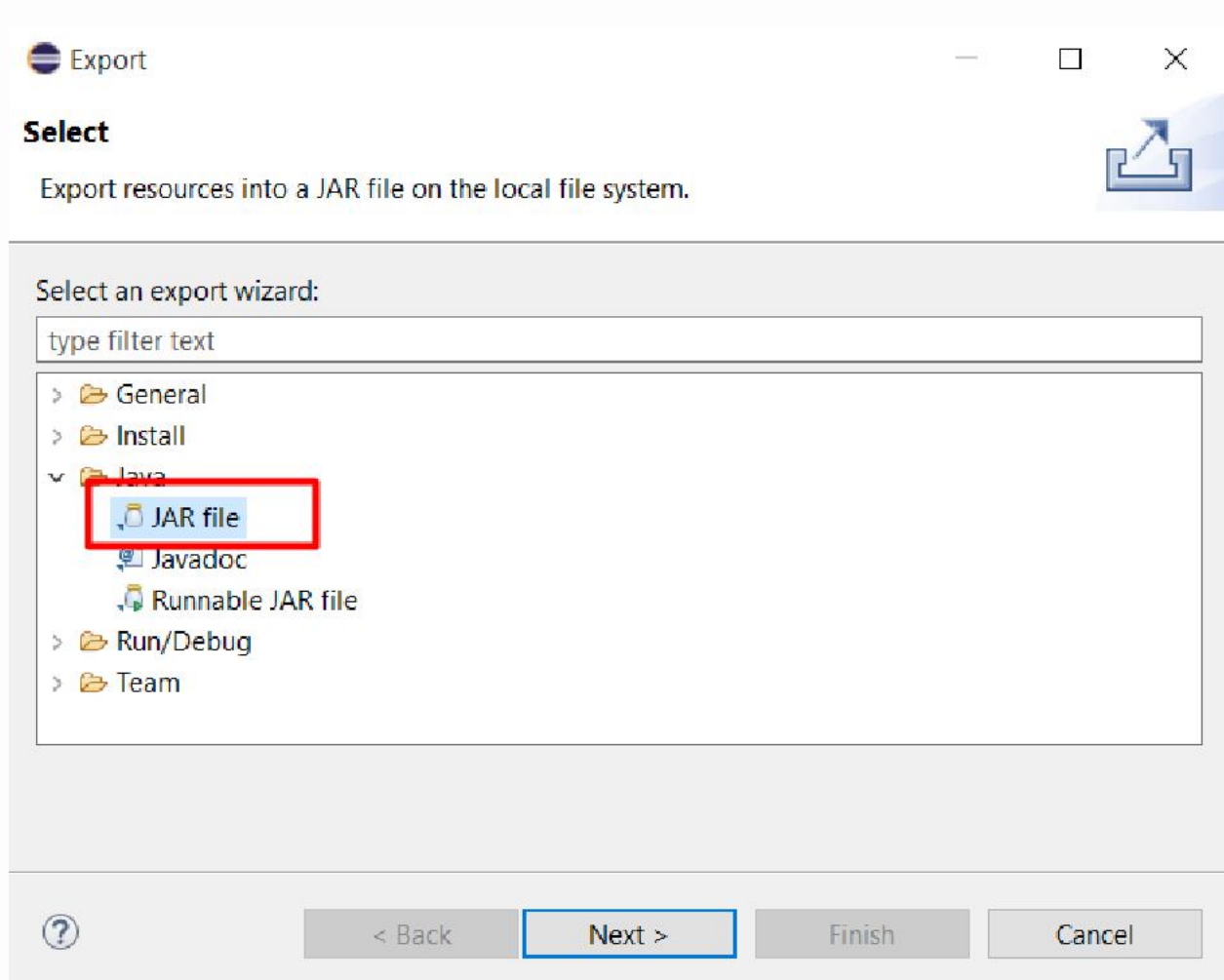


## Bước 4: Tạo file JAR

Chuột phải vào project **WordCount** chọn **Export**



Chọn **Java** > **JAR File** rồi bấm **Next**



Chọn đường dẫn lưu file JAR và bấm **Next**






## JAR File Specification

Define which resources should be exported into the JAR.



Select the resources to export:

<p>&gt;  WordCount</p>	<p><input checked="" type="checkbox"/>  .classpath</p> <p><input checked="" type="checkbox"/>  .project</p>
---	---

- ☒ Export generated class files and resources
- ☐ Export all output folders for checked projects
- ☐ Export Java source files and resources
- ☐ Export refactorings for checked projects. [Select refactorings...](#)

Select the export destination:

JAR file:   

Options:

- ☒ Compress the contents of the JAR file
- ☐ Add directory entries
- ☐ Overwrite existing files without warning



< Back

Next >

Finish

Cancel

## JAR Packaging Options

Define the options for the JAR export.



Select options for handling problems:

- ☒ Export class files with compile errors
- ☒ Export class files with compile warnings

☐ Create source folder structure

☒ Build projects if not built automatically

☐ Save the description of this JAR in the workspace

Description file:

Browse...



< Back

Next >

Finish

Cancel

Bấm **Browser** để chọn file **main**



## JAR Manifest Specification

Customize the manifest file for the JAR file.



Specify the manifest:

☒ Generate the manifest file

☐ Save the manifest in the workspace

☐ Use the saved manifest in the generated JAR description file

Manifest file:

Browse...

☐ Use existing manifest from workspace

Manifest file:

Browse...

Seal contents:

☐ Seal the JAR

Details...

☒ Seal some packages

Nothing sealed

Details...

Select the class of the application entry point:

Main class:

Browse...



< Back

Next >

Finish

Cancel

Chọn **WordCount** và bấm **OK**



Select Main Class



Select the class which is the application's entry point:



WordCount



(default package) - WordCount/src



OK


Cancel

Bấm **Finish** để thực hiện quá trình **Export**

JAR Export

### JAR Manifest Specification

Customize the manifest file for the JAR file.



Specify the manifest:

☒ **Generate the manifest file**

☐ Save the manifest in the workspace

☐ Use the saved manifest in the generated JAR description file

Manifest file:

☐ **Use existing manifest from workspace**

Manifest file:


Seal contents:

☐ Seal the JAR

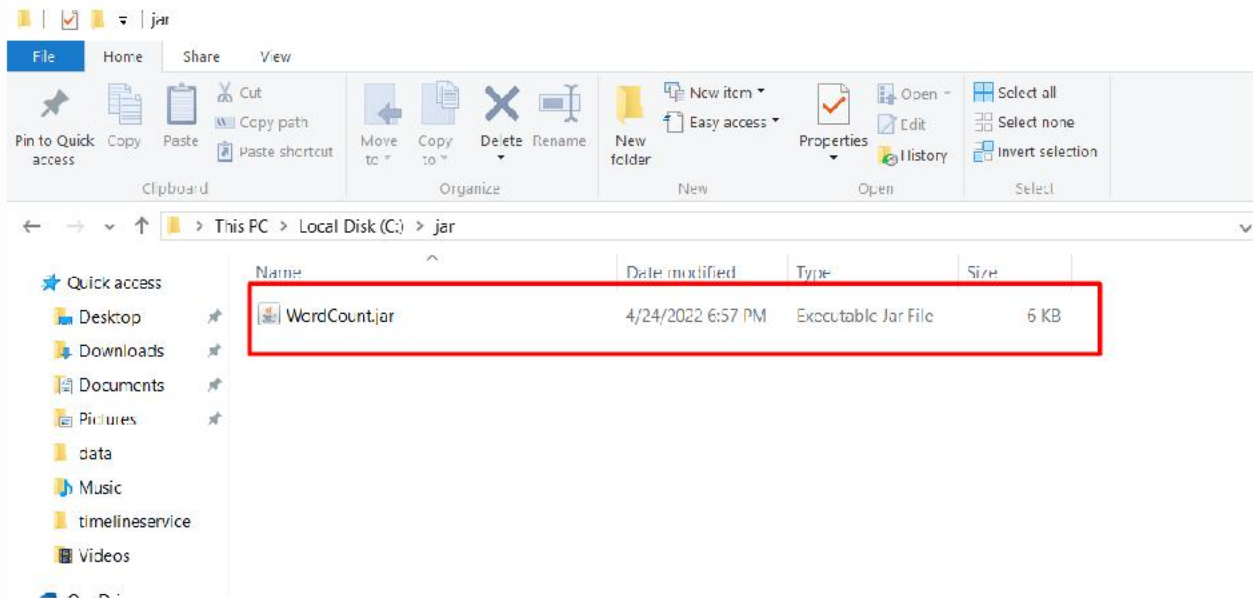
☒ Seal some packages Nothing sealed

Select the class of the application entry point:

Main class:



Vào thư mục chứa lưu file JAR vừa tạo và kiểm tra kết quả



Thử nghiệm trên file dữ liệu **data.txt** đã tạo ở trên, và kết quả thu được lưu tại thư mục **r\_output**. Chạy lệnh sau:

```
hadoop jar "C:\jar\WordCount.jar" /input/data.txt /r_output
```

*Lưu ý: Thay "C:\jar\WordCount.jar" bằng đường dẫn chứa file JAR ở trên máy*





```
Administrator: Command Prompt
C:\hadoop-3.3.0\shin>hadoop jar "C:\jar\WordCount.jar" /input/data.txt /r_output
2022-04-24 21:41:54,228 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-04-24 21:41:54,999 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
dachi/.staging/job_1650811011093_0002
2022-04-24 21:41:55,498 INFO input.FileInputFormat: Total input files to process : 1
2022-04-24 21:41:55,663 INFO mapreduce.JobSubmitter: number of splits:1
2022-04-24 21:41:56,273 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1650811011093_0002
2022-04-24 21:41:56,276 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-24 21:41:56,690 INFO conf.Configuration: resource-types.xml not found
2022-04-24 21:41:56,691 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-24 21:41:57,566 INFO impl.YarnClientImpl: Submitted application application_1650811011093_0002
2022-04-24 21:41:57,611 INFO mapreduce.Job: The url to track the job: http://DFSSTOP-QBKQI72:8088/proxy/application_1650
811011093_0002/
2022-04-24 21:41:57,613 INFO mapreduce.Job: Running job: job_1650811011093_0002
2022-04-24 21:42:12,190 INFO mapreduce.Job: Job job_1650811011093_0002 running in uber mode : false
2022-04-24 21:42:12,192 INFO mapreduce.Job: map 0% reduce 0%
2022-04-24 21:42:19,541 INFO mapreduce.Job: map 100% reduce 0%
2022-04-24 21:42:28,704 INFO mapreduce.Job: map 100% reduce 100%
2022-04-24 21:42:28,712 INFO mapreduce.Job: Job job_1650811011093_0002 completed successfully
2022-04-24 21:42:28,795 INFO mapreduce.Job: Counters: 54
```

← → ↻ localhost:9870/explorer.html/#/






Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

## Browse Directory

/

Go!    

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 17:58	0	0 B	input	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 18:01	0	0 B	output	
<input type="checkbox"/>	drwxr-xr-x	dachi	supergroup	0 B	Apr 24 21:42	0	0 B	r_output	
<input type="checkbox"/>	drwx----	dachi	supergroup	0 B	Apr 24 18:00	0	0 B	tmp	

Showing 1 to 4 of 4 entries





Previous 1 Next

← → ↻ localhost:9870/explorer.html/#/r\_output




Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

## Browse Directory

/r\_output

Go!    

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	dachi	supergroup	0 B	Apr 24 21:42	1	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-r--r--	dachi	supergroup	64 B	Apr 24 21:42	1	128 MB	part-r-00000	

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2020.

←

→

↻

localhost:9870/explorer.html#/r\_output

🔖

☆

🗖

👤

⋮

Hadoop

Overview

Datanodes

# Browse Directory

/r\_output

Show 25 entries

<input type="checkbox"/>	Permission	Owner
<input type="checkbox"/>	-rw-r--r--	dachl
<input type="checkbox"/>	-rw-r--r--	dachl

Showing 1 to 2 of 2 entries

Hadoop, 2020.

Block Information — Block 0

Block ID: 1079741844

Block Pool ID: BP-469985238-192.168.23.128-1650351230140

Generation Stamp: 1020

Size: 64

Availability:

- DESKTOP-GBKQL72.localdomain

File contents

BUS 2

Rink 1

CAR 1

Car 1

TRAIN 2

bus 1

bus 3

car 1

🗖

📄

📁

🗑

Search:

Block Size	Name
MB	SUCCESS
MB	part-r-00000

Previous

1

Next