

THỰC HÀNH 2: BIG DATA

Họ và tên: Lê Thị Ngọc Ly_64HTTT1_2251162068

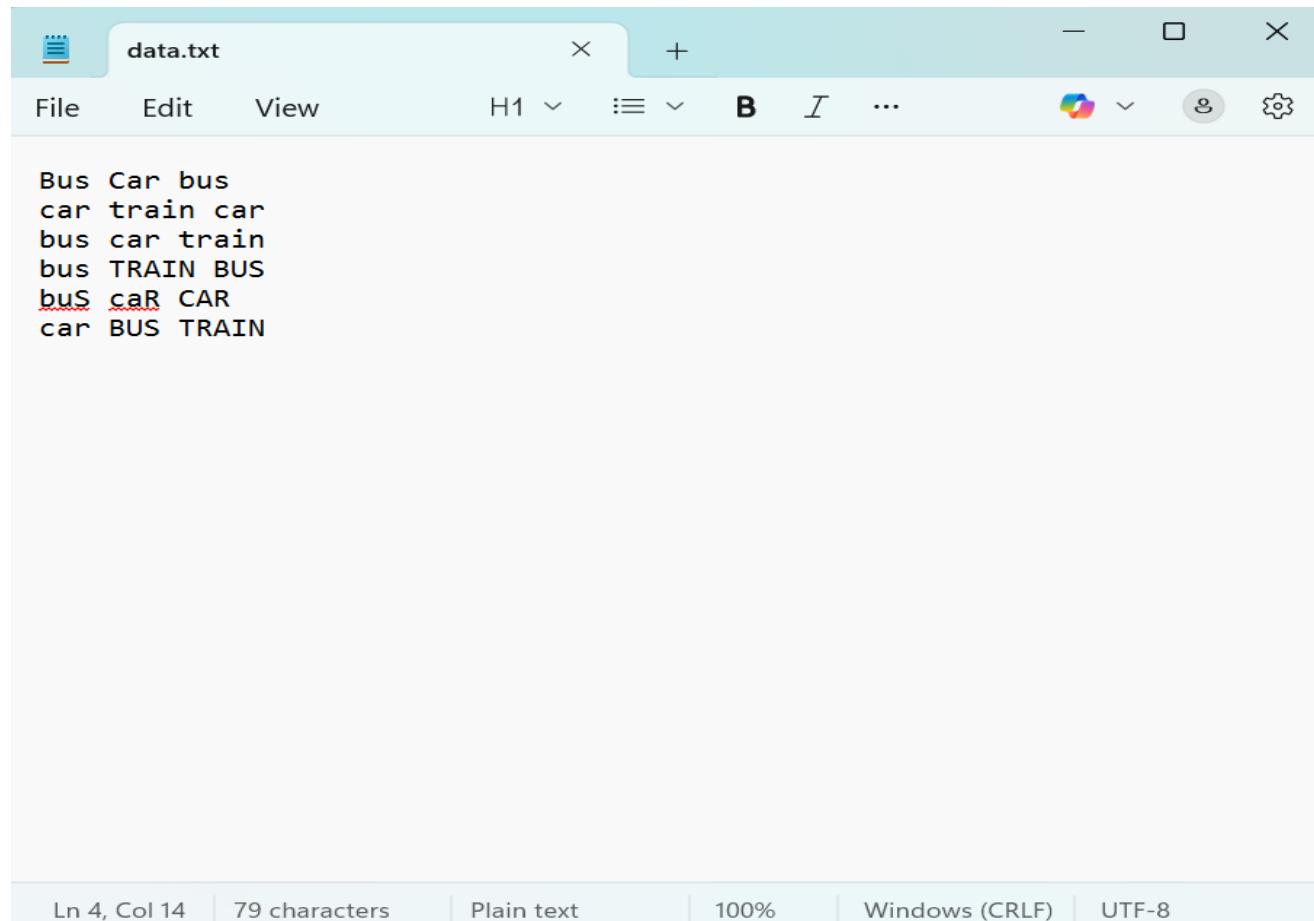
MapReduce - Lập trình chương trình WordCount

I. Thủ nghiêm 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:



The screenshot shows a Windows Notepad window with the title bar 'data.txt'. The window contains the following text:

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

The word 'bus' in the first line and 'buS' in the fifth line are highlighted in red, while 'CAR' in the fifth line is highlighted in blue. The status bar at the bottom of the window displays: Ln 4, Col 14 | 79 characters | Plain text | 100% | Windows (CRLF) | UTF-8

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

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

```
C:\> Administrator: Command Prompt
Microsoft Windows [Version 10.0.26100.7171]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

Bước 2: Tạo

Tạo thư mục inp
C:\Windows\System32>hdfs dfs -mkdir /input
C:\Windows\System32>

hdfs dfs -mkdC:\Windows\System32>
```

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

```
C:\> Select Administrator: Command Prompt
Microsoft Windows [Version 10.0.26100.7171]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

Bước 2: Tạo

Tạo thư mục inp
C:\Windows\System32>hdfs dfs -mkdir /input
C:\Windows\System32>hdfs dfs -put C:/hadoop-3.3.0/data/data.txt /input
C:\Windows\System32>hdfs dfs -mkdC:\Windows\System32>
```

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

The screenshot shows the HDFS NameNode web interface. At the top, there is a green header bar with tabs for 'Snapshot', 'Startup Progress', and 'Utilities'. Below the header is a search bar and a set of icons. The main area displays a table of files in the '/input' directory. The columns are labeled 'Size', 'Last Modified', 'Replication', 'Block Size', and 'Name'. A single row is shown for a file named 'input' which is 0 B in size, last modified on Dec 01 10:31, has a replication factor of 0, and a block size of 0 B. The 'Name' column is highlighted with a red box. At the bottom of the table are buttons for 'Previous', '1', and 'Next'.

This screenshot shows the browser's address bar with the URL 'localhost:9870/explorer.html#/input'. The page content is identical to the one above, displaying the '/input' directory listing with a single file named 'input'.

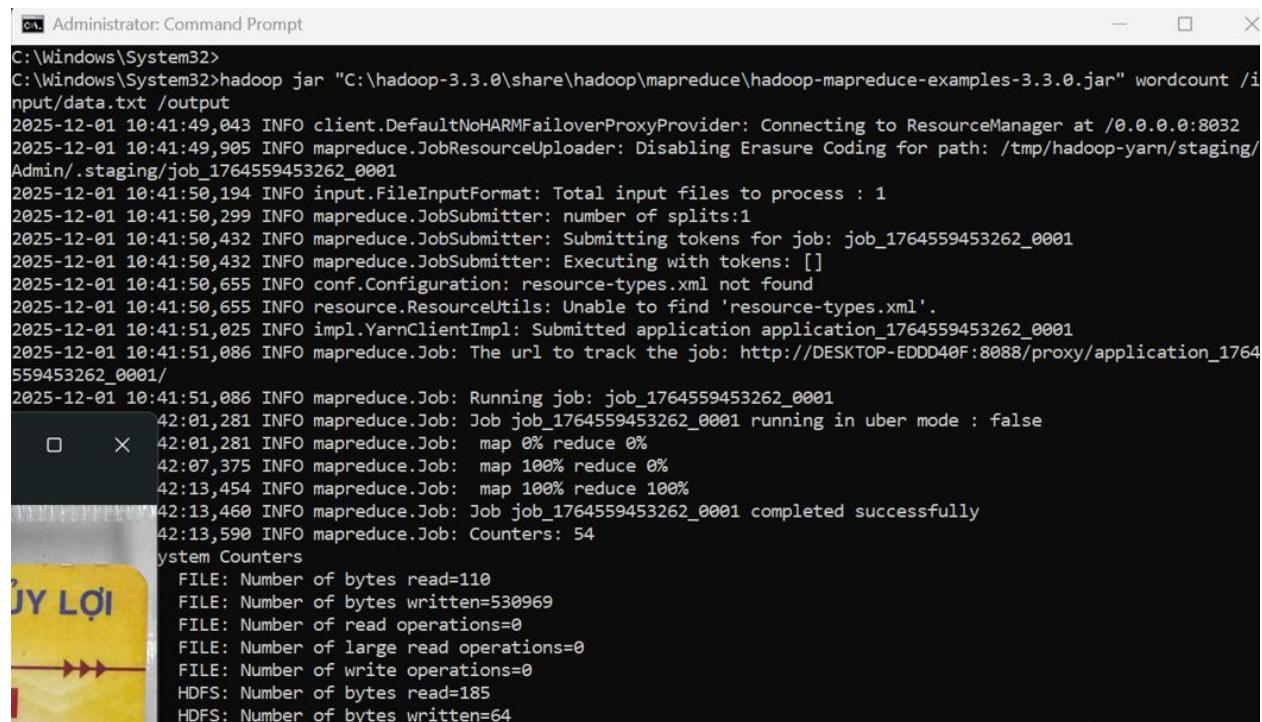
Browse Directory

This screenshot shows a detailed view of the 'data.txt' file within the '/input' directory. The table has columns for 'Permission', 'Owner', 'Group', 'Size', 'Last Modified', 'Replication', 'Block Size', and 'Name'. The 'Name' column is highlighted with a red box. The file 'data.txt' is listed with the following details: permission '-rw-r--r--', owner 'Admin', group 'supergroup', size '84 B', last modified 'Dec 01 10:31', replication '1', block size '128 MB', and name 'data.txt'. At the bottom, it says 'Showing 1 to 1 of 1 entries'.

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:\Windows\System32>
C:\Windows\System32>hadoop jar "C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar" wordcount /input/data.txt /output
2025-12-01 10:41:49,043 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-12-01 10:41:49,905 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_1764559453262_0001
2025-12-01 10:41:50,194 INFO input.FileInputFormat: Total input files to process : 1
2025-12-01 10:41:50,299 INFO mapreduce.JobSubmitter: number of splits:1
2025-12-01 10:41:50,432 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1764559453262_0001
2025-12-01 10:41:50,432 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-12-01 10:41:50,655 INFO conf.Configuration: resource-types.xml not found
2025-12-01 10:41:50,655 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-12-01 10:41:51,025 INFO impl.YarnClientImpl: Submitted application application_1764559453262_0001
2025-12-01 10:41:51,086 INFO mapreduce.Job: The url to track the job: http://DESKTOP-EDDD40F:8088/proxy/application_1764559453262_0001/
2025-12-01 10:41:51,086 INFO mapreduce.Job: Running job: job_1764559453262_0001
    42:01,281 INFO mapreduce.Job: Job job_1764559453262_0001 running in uber mode : false
    □ X 42:01,281 INFO mapreduce.Job: map 0% reduce 0%
    42:07,375 INFO mapreduce.Job: map 100% reduce 0%
    42:13,454 INFO mapreduce.Job: map 100% reduce 100%
    42:13,460 INFO mapreduce.Job: Job job_1764559453262_0001 completed successfully
    42:13,590 INFO mapreduce.Job: Counters: 54
        system Counters
            FILE: Number of bytes read=110
            FILE: Number of bytes written=530969
            FILE: Number of read operations=0
            FILE: Number of large read operations=0
            FILE: Number of write operations=0
            HDFS: Number of bytes read=185
            HDFS: Number of bytes written=64
```

Xem kết quả thu được:

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/									Go!	File	Up	Table	CSV	JSON
Show 25 entries									Search:					
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name						
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:31	0	0 B	input						
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:42	0	0 B	output						
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 11:37	0	0 B	r_output						
<input type="checkbox"/>	drwx-----	Admin	supergroup	0 B	Dec 01 10:41	0	0 B	tmp						

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/output									Go!	File	Up	Table	CSV	JSON
Show 25 entries									Search:					
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name						
<input type="checkbox"/>	-rw-r--r--	Admin	supergroup	0 B	Dec 01 10:42	1	128 MB	SUCCESS						
<input type="checkbox"/>	-rw-r--r--	Admin	supergroup	64 B	Dec 01 10:42	1	128 MB	part-r-00000						

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2020.

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/output

Show 25 entries

	Permission	Owner
□	-rw-r--r--	Admin
□	-rw-r--r--	Admin

Showing 1 to 2 of 2 entries

Hadoop, 2020.

File information - part-r-00000

Download Head the file (first 32K) Tail the file (last 32K)

Block information - Block 0

Block ID: 1073741832
Block Pool ID: BP-676132101-172.20.10.4-1764170020332
Generation Stamp: 1008
Size: 64
Availability:

- DESKTOP-EDDD40F

Close

Block Size Name

MB _SUCCESS

MB part-r-00000

Previous 1 Next

Hadoop Overview Datanodes

Browse Directory

/output

Show 25 entries

	Permission	Owner
□	-rw-r--r--	Admin
□	-rw-r--r--	Admin

Showing 1 to 2 of 2 entries

Hadoop, 2020.

Block information - Block 0

Block ID: 1073741832
Block Pool ID: BP-676132101-172.20.10.4-1764170020332
Generation Stamp: 1008
Size: 64
Availability:

- DESKTOP-EDDD40F

File contents

BUS2
Bus 1
CAR 1
Car 1
TRAIN 2
buS 1
bus 3
caR 1

Close

Block Size Name

MB _SUCCESS

MB part-r-00000

Previous 1 Next

Hoặc dùng lệnh cmd:

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

```
C:\Windows\System32>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
```

```
C:\Windows\System32>
```

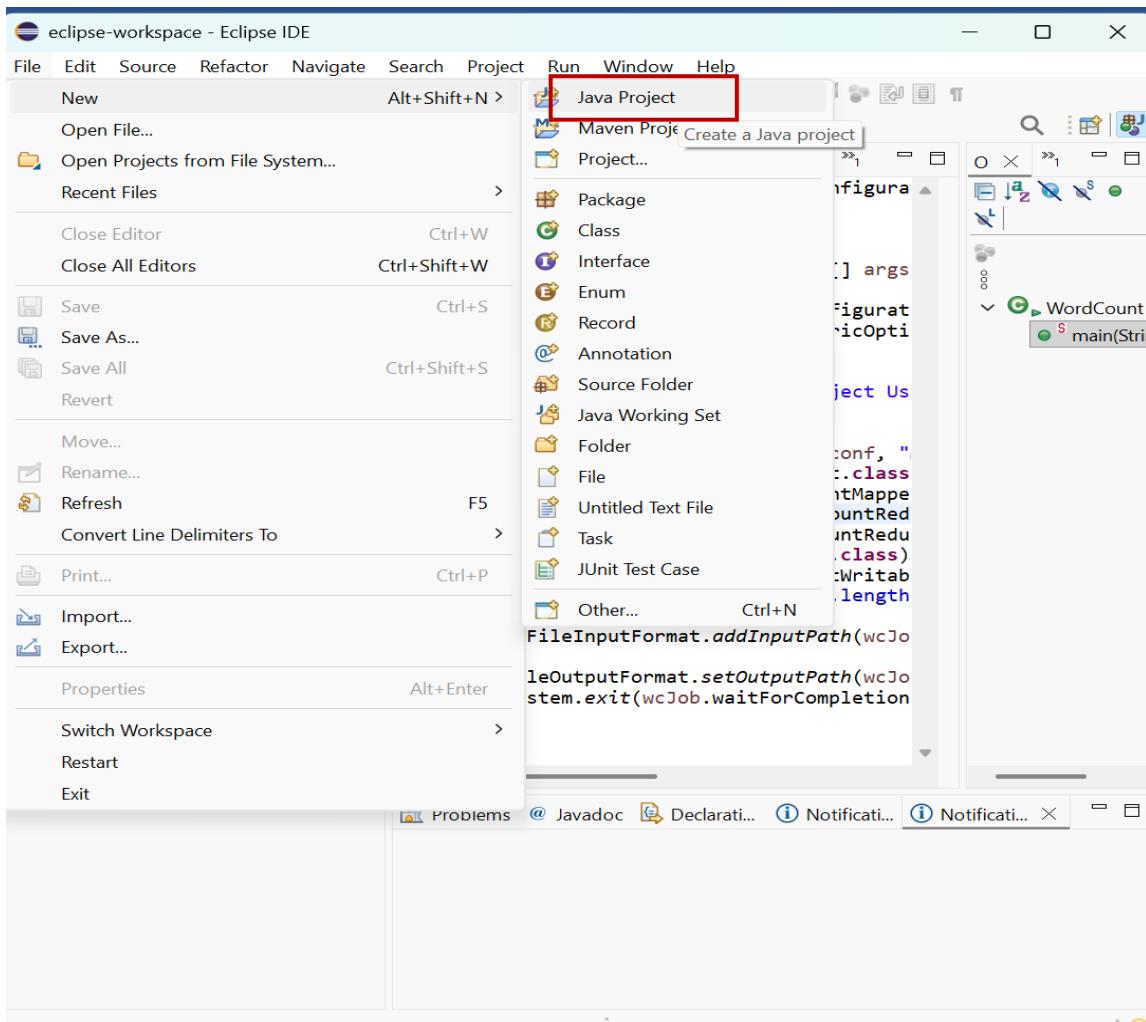
Như vậy ta đã chạy thành công chương trình mẫu MapReduce của Hadoop cung cấp.

II. Lập trình chương trình WordCount bằng Eclipse

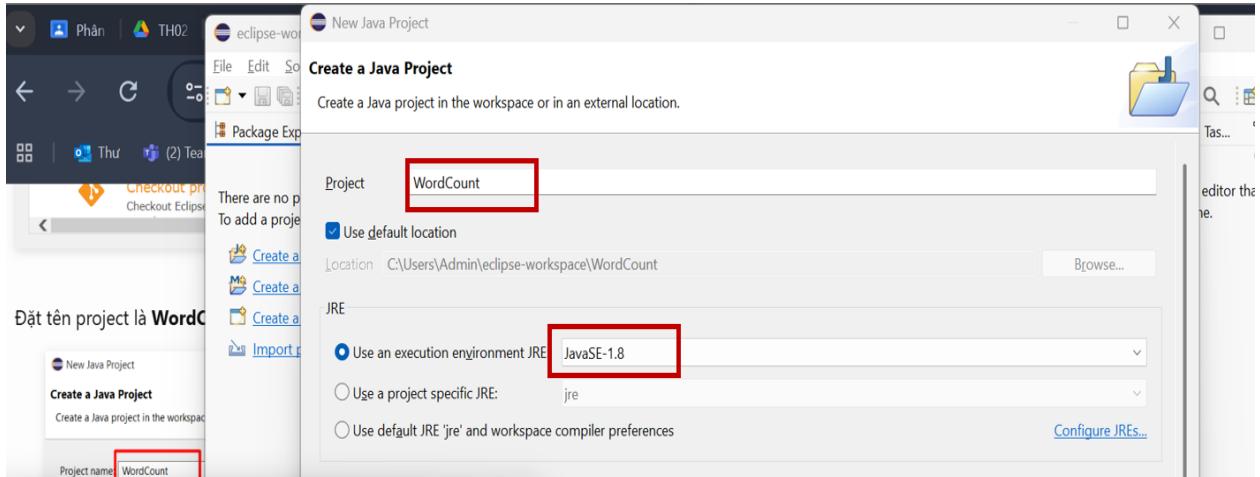
Bước 1: Tạo project

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

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

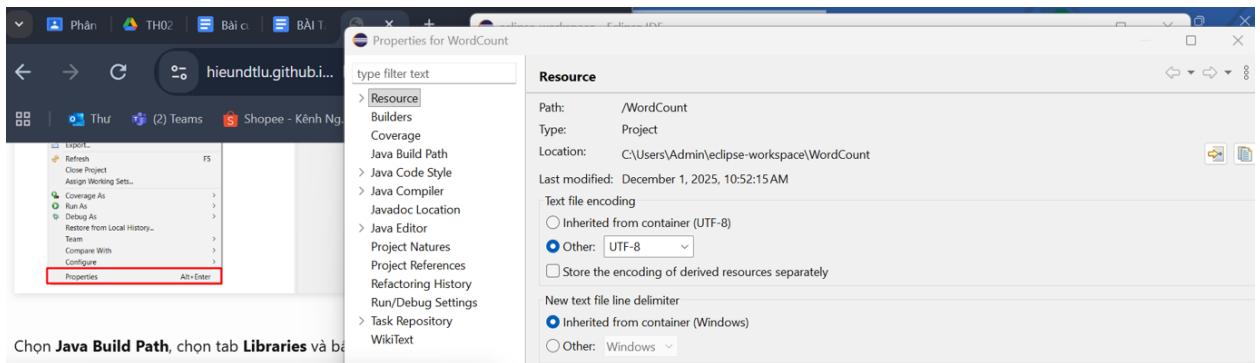


Đặt tên project là **WordCount** và chọn môi trường là **JavaSE-1.8**. Xong
ấn **Finish**.



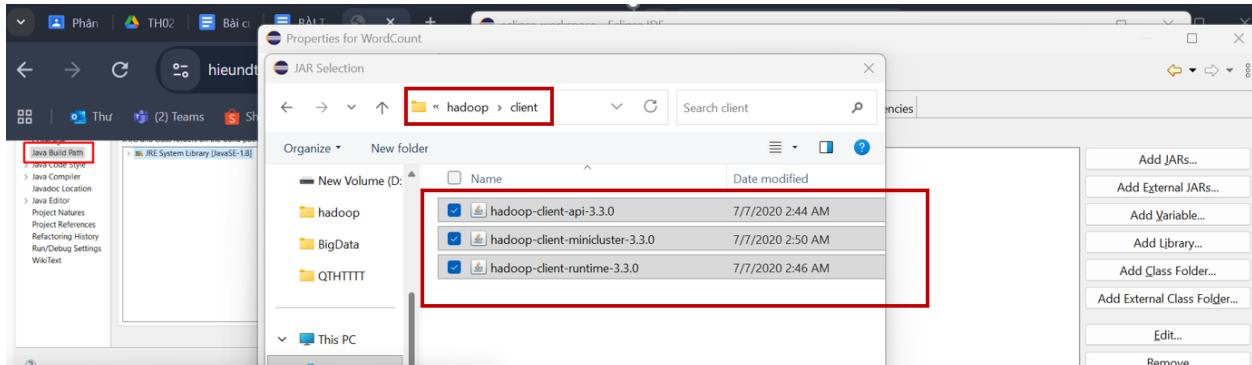
Bước 2: Thêm thư viện cần thiết để chạy MapReduce

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

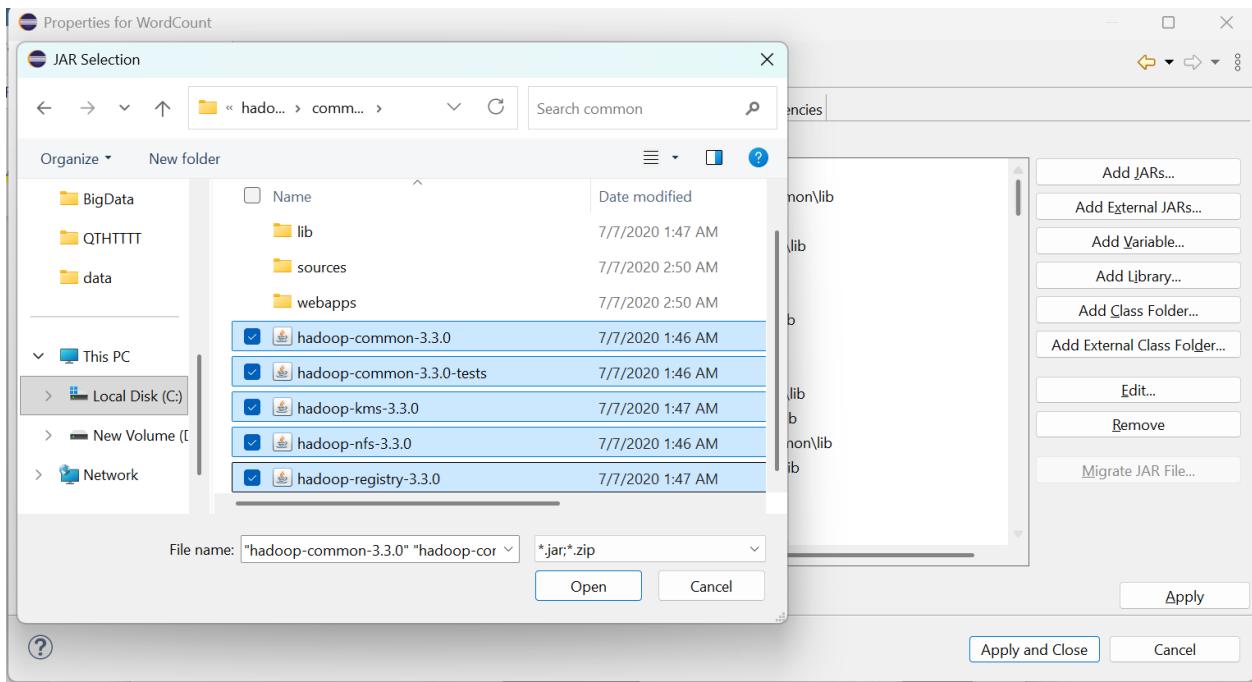


Chọn **Java Build Path**, chọn tab **Libraries** và bấm **Add External JARs**

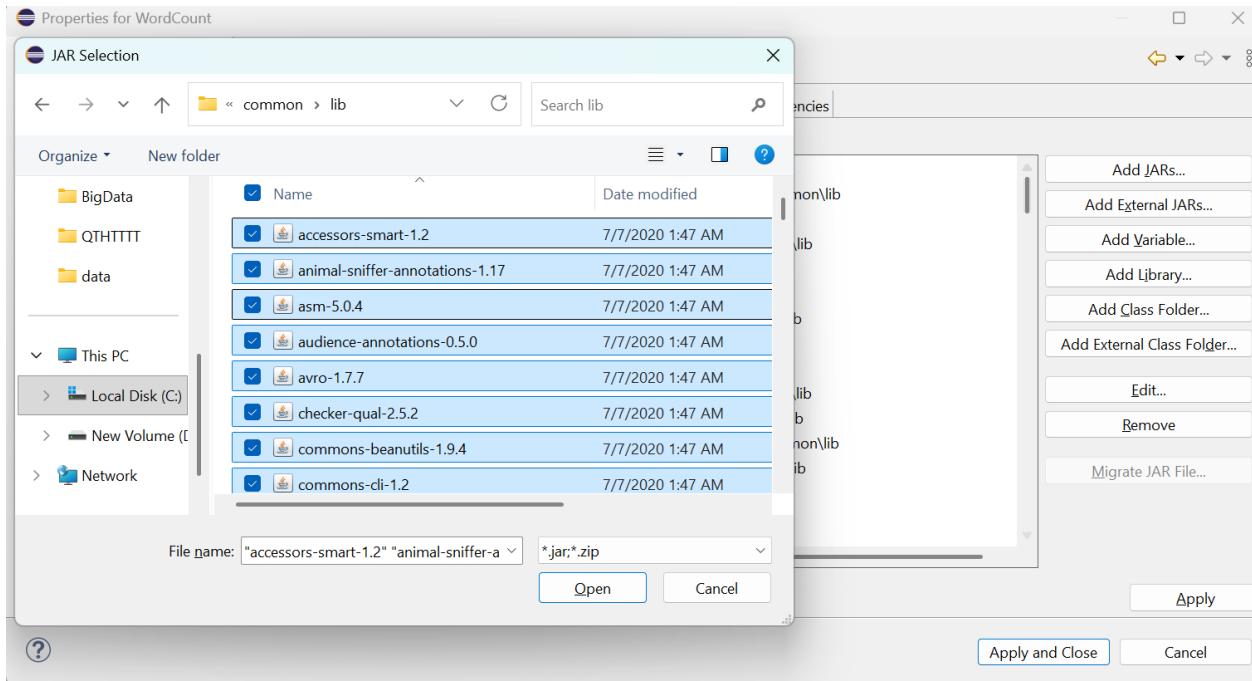
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\client** và ấn **Open**



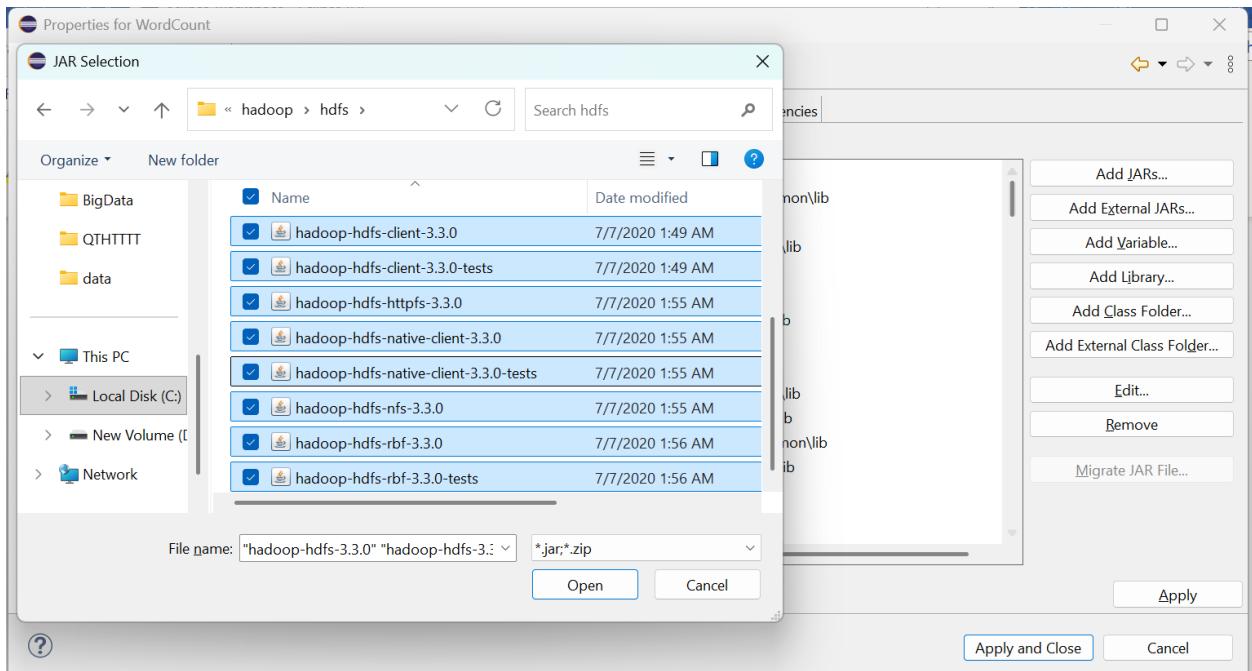
Tương tự chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common** và ấn **Open**



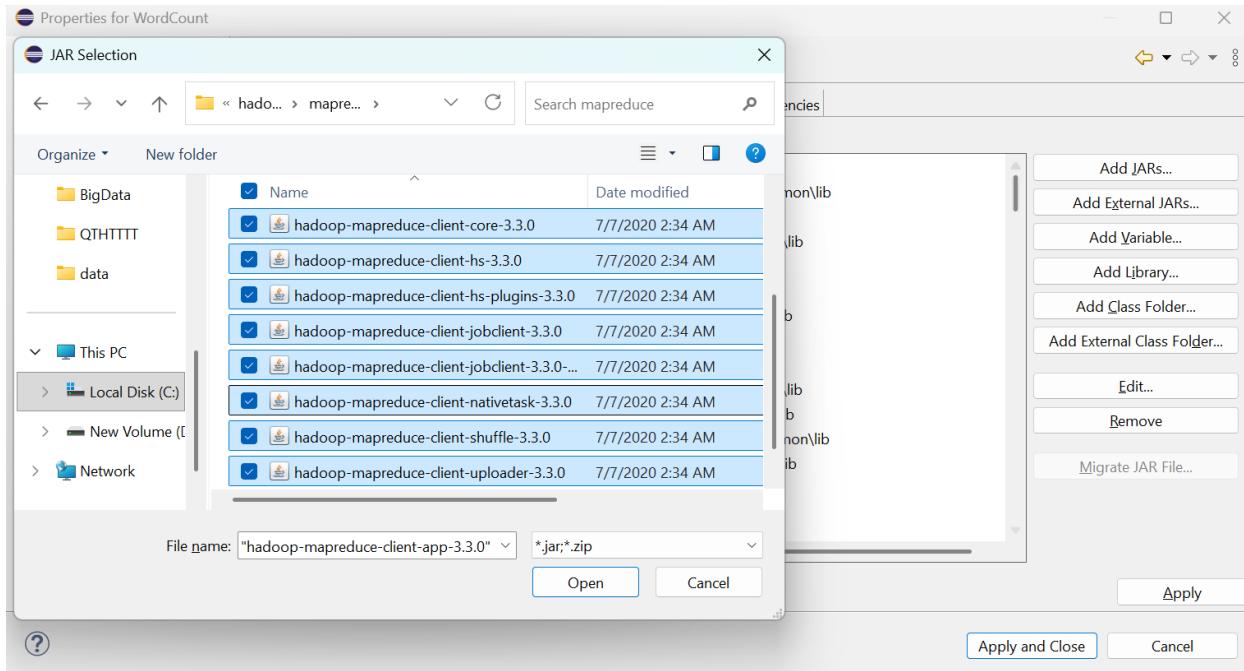
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common\lib** và
ấn **Open**



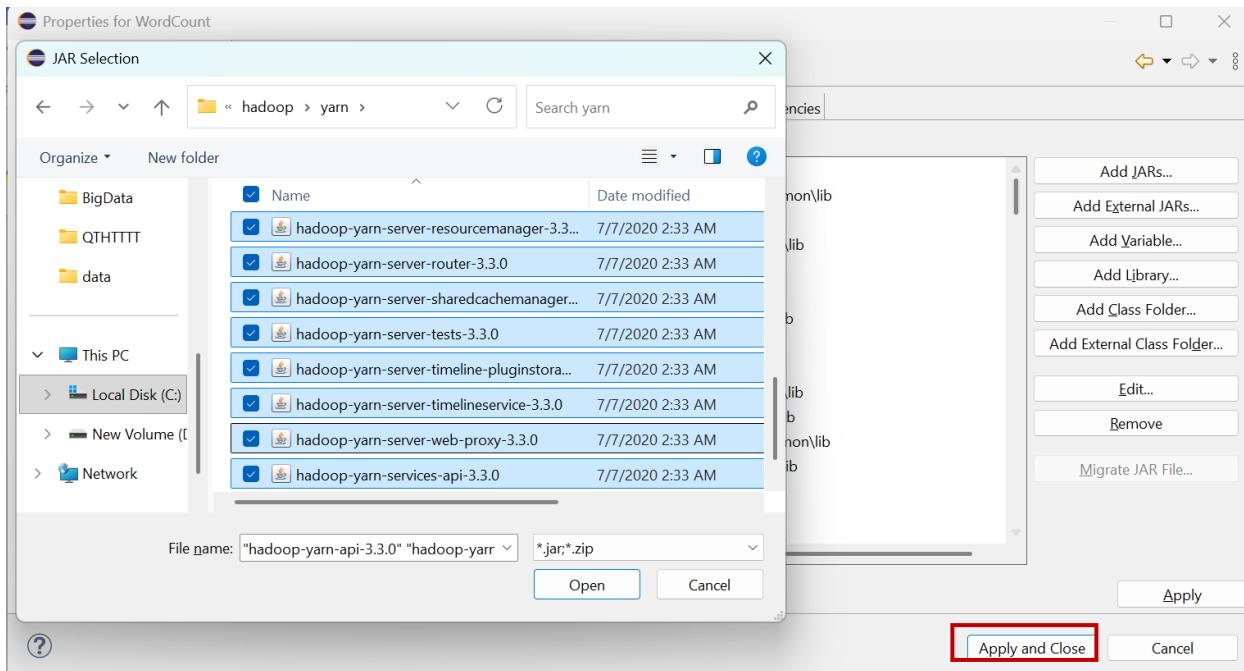
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\hdfs** và ấn **Open**



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



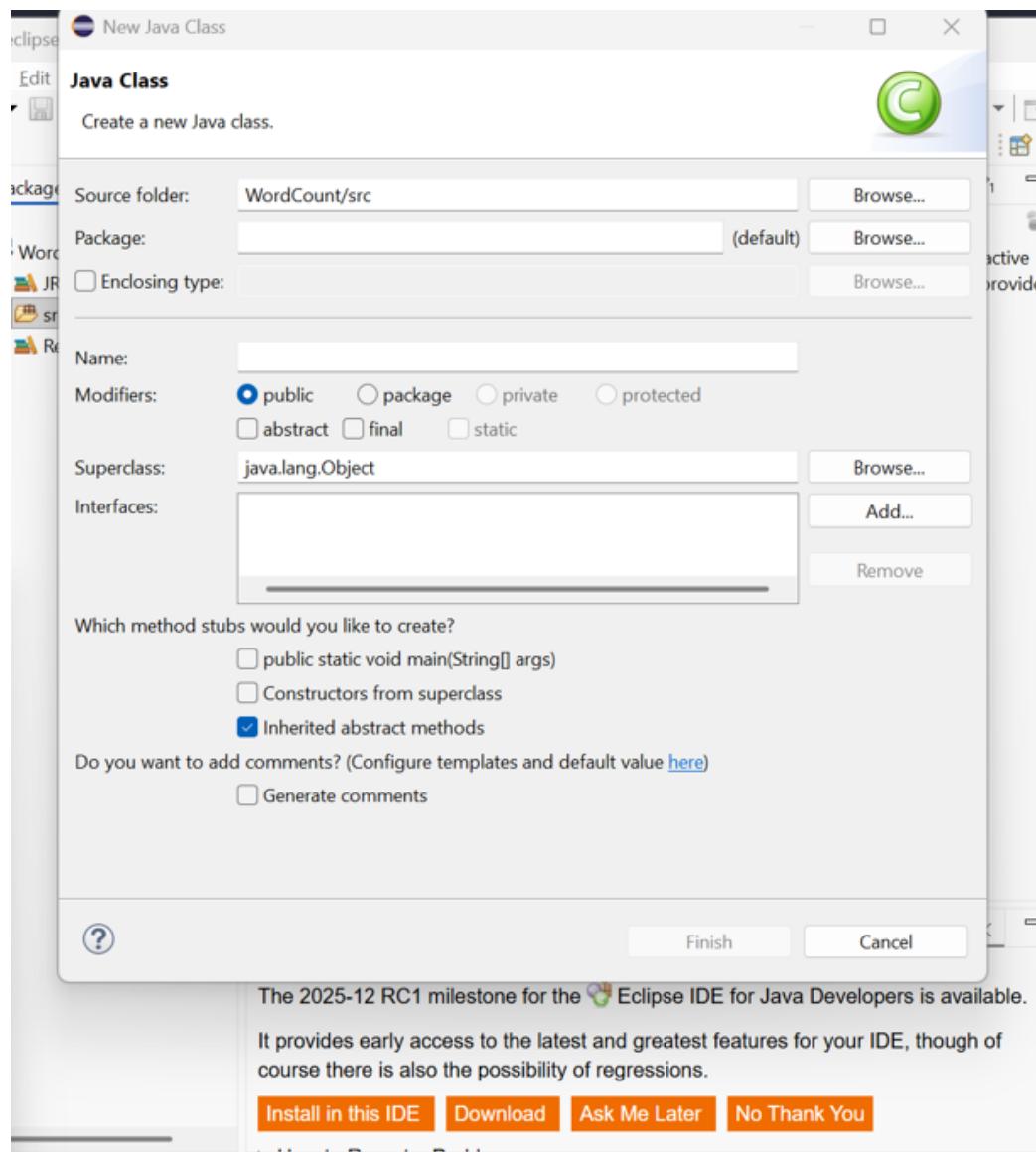
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\yarn** và **Ấn Open**



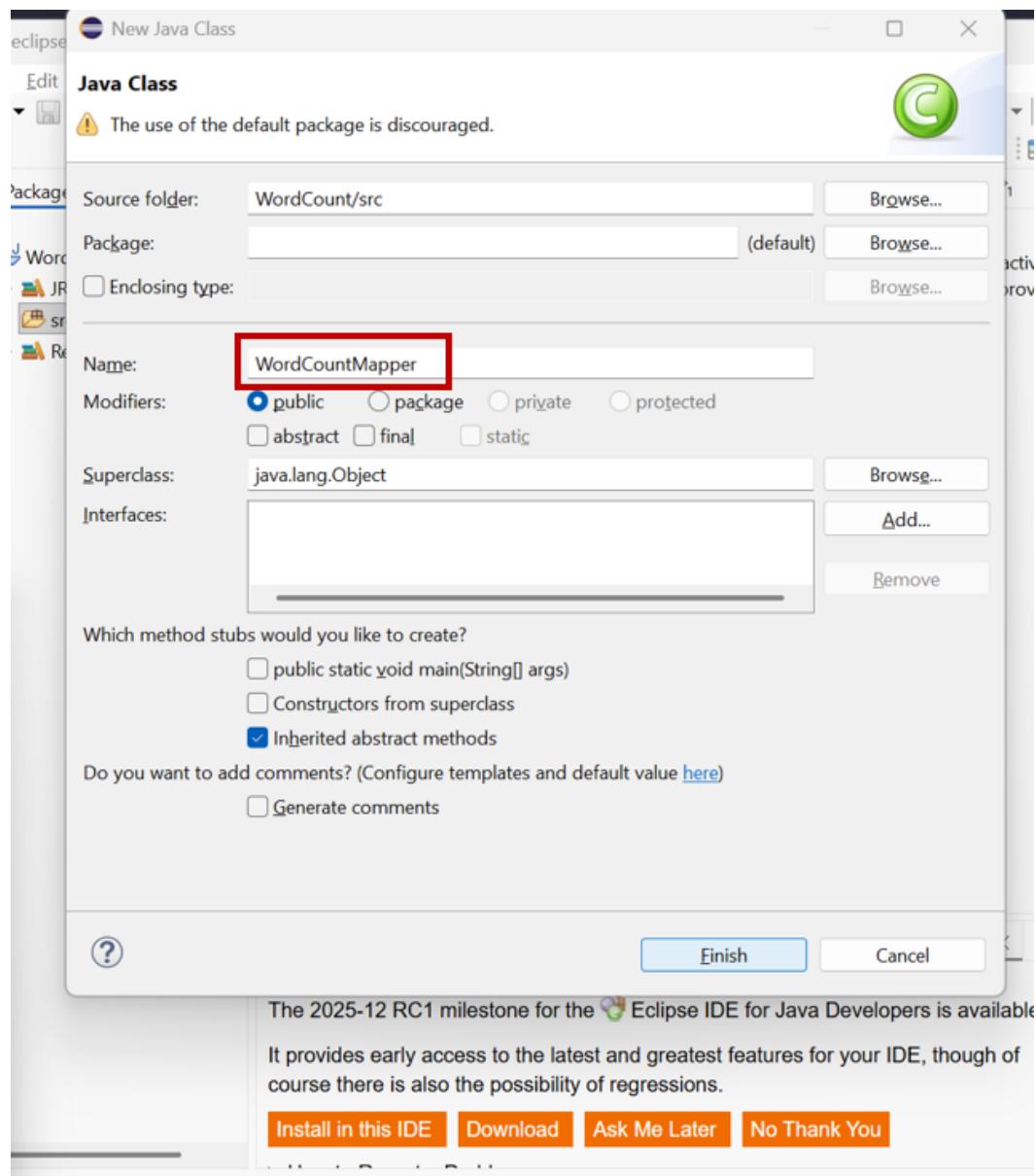
Ấn Apply and Close

Bước 3: Tạo class xử lý tác vụ MapReduce

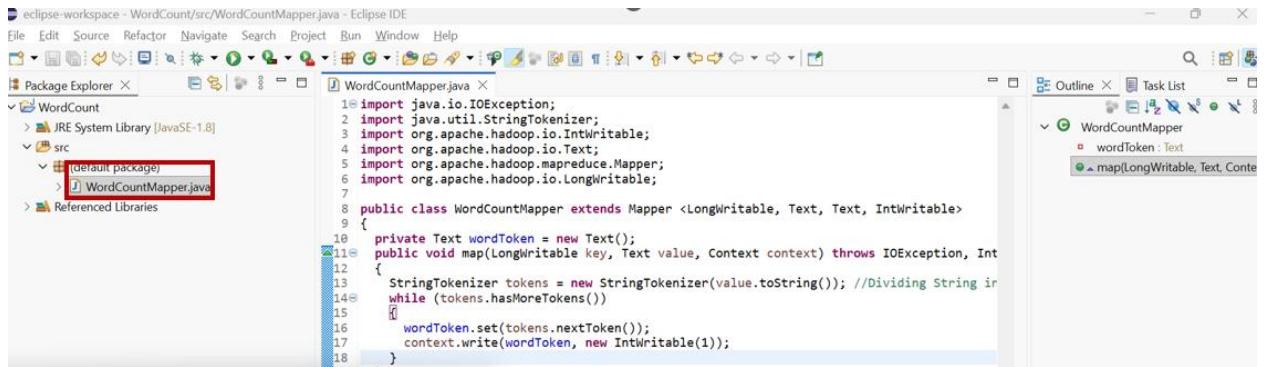
Double click vào project **WordCount**, chuột phải vào **src** và chọn **New > Class**



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

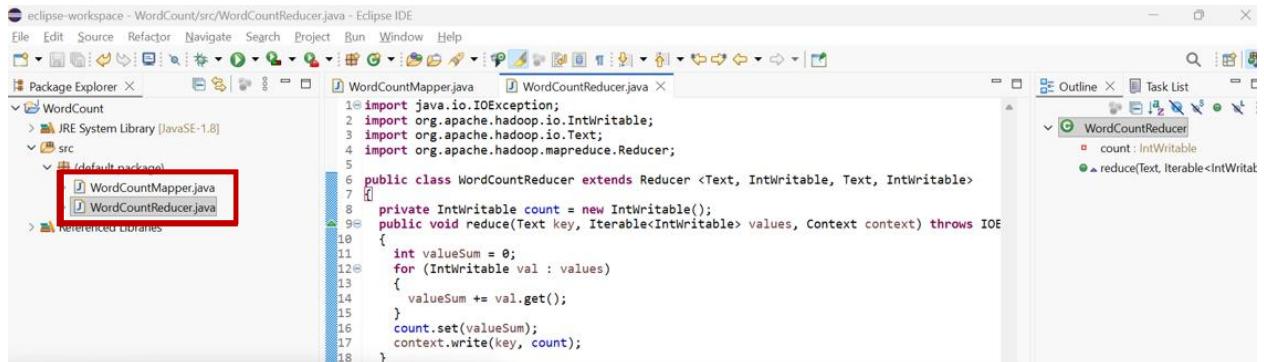


Nội dung bên trong file **WordCountMapper.java**:



```
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer X WordCountMapper.java X
WordCount JRE System Library [JavaSE-1.8]
src (default package) WordCountMapper.java
Referenced Libraries
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, IntWritableException
    {
        StringTokenizer tokens = new StringTokenizer(value.toString()); //Dividing String into words
        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**:



```
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer X WordCountMapper.java WordCountReducer.java X
WordCount JRE System Library [JavaSE-1.8]
src (default package) WordCountMapper.java WordCountReducer.java
Referenced Libraries
WordCountMapper.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 WordCountReducer extends Reducer <Text, IntWritable, Text, IntWritable>
{
    private IntWritable count = new IntWritable();
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, IntWritableException
    {
        int valueSum = 0;
        for (IntWritable val : values)
        {
            valueSum += val.get();
        }
        count.set(valueSum);
        context.write(key, count);
    }
}
```

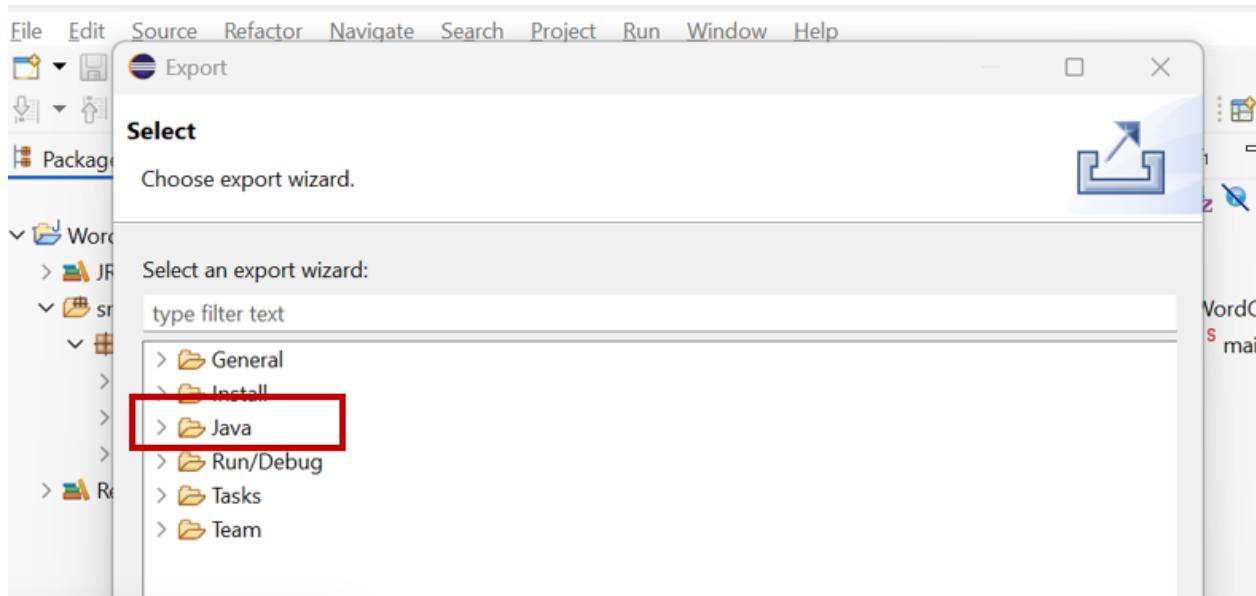
Và tạo class **WordCount** chứa hàm **main** để khởi chạy chương trình:

The screenshot shows the Eclipse IDE interface with the following details:

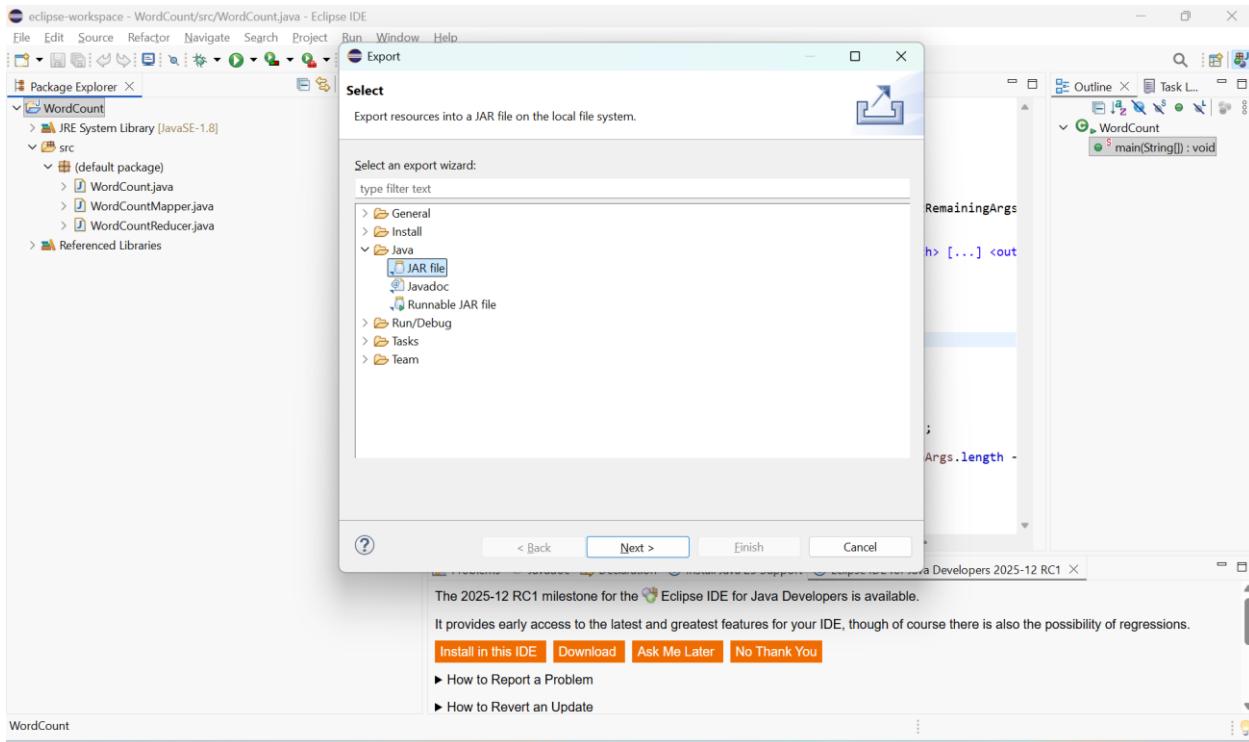
- Project Explorer:** Shows the `WordCount` project with three files: `WordCount.java`, `WordCountMapper.java`, and `WordCountReducer.java`. The `WordCount.java` file is currently selected and its code is displayed in the central editor.
- Code Editor:** Displays the Java code for `WordCount.java`. The code implements a `MapReduce` job for word counting. It includes imports for `org.apache.hadoop.conf.Configuration` and `org.apache.hadoop.mapreduce.Job`. The `main` method sets up the job with `WordCountMapper` as the mapper, `WordCountReducer` as the reducer, and `Text` as the output key class. It also adds input paths and a single output path.
- Outline View:** Shows the outline of the `WordCount` class, including the `main` method.
- Bottom Status Bar:** Shows the status bar with "Writable", "Smart Insert", and the current time "24:1:916".
- Bottom Footer:** Promotes the latest version of Eclipse IDE for Java Developers (2025-12 RC1) with links to "Install in this IDE", "Download", "Ask Me Later", and "No Thank You".

Bước 4: Tạo file JAR

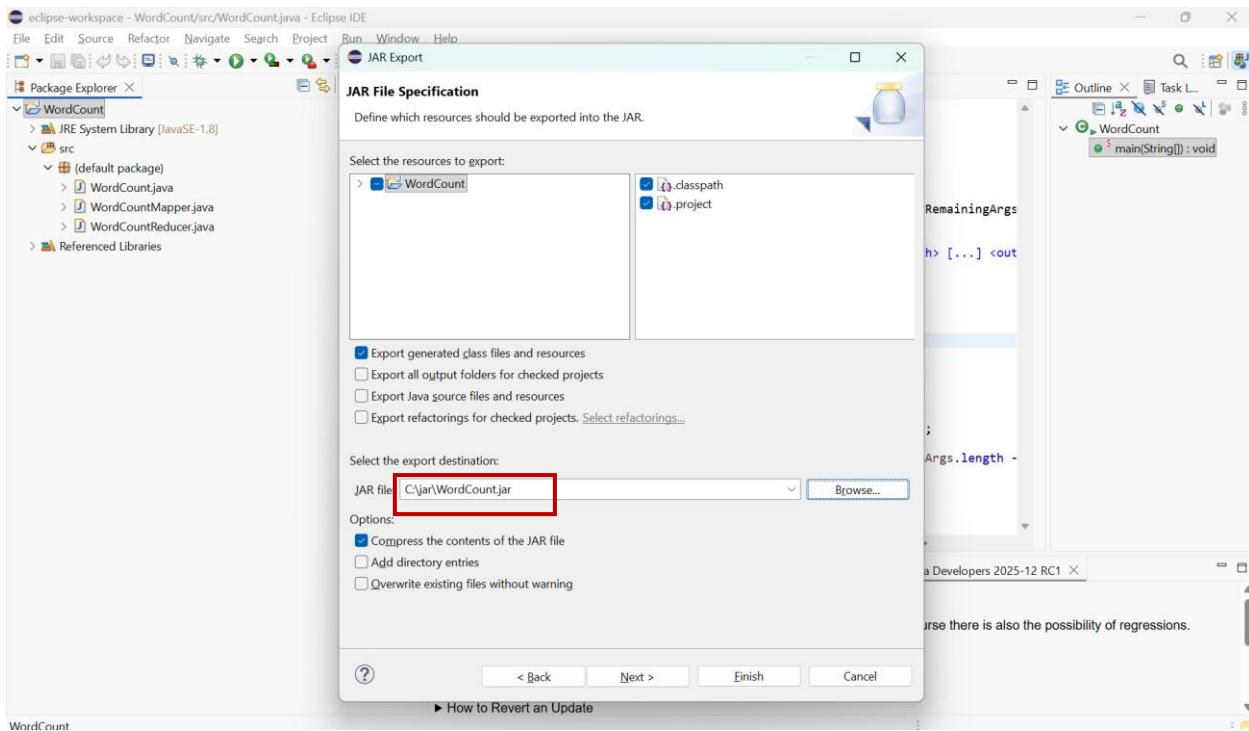
Chuột phải vào project **WordCount** chọn **Export**



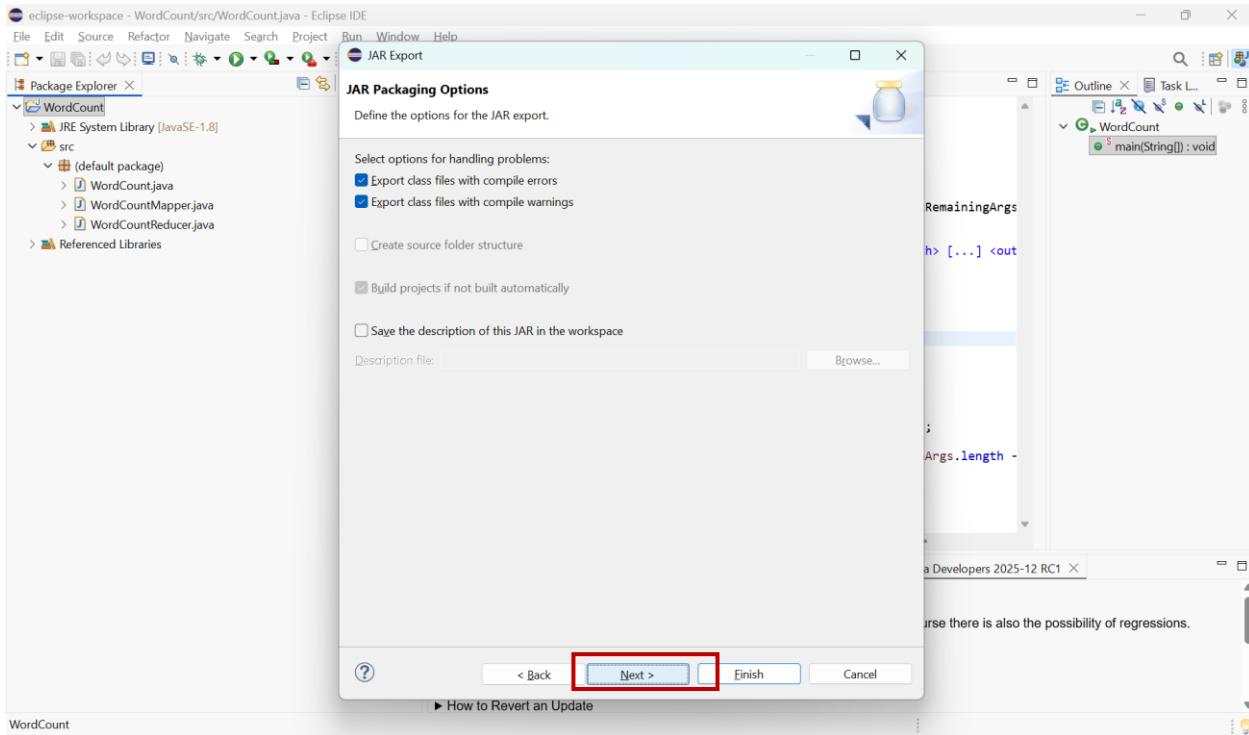
Chọn Java > JAR File rồi bấm Next



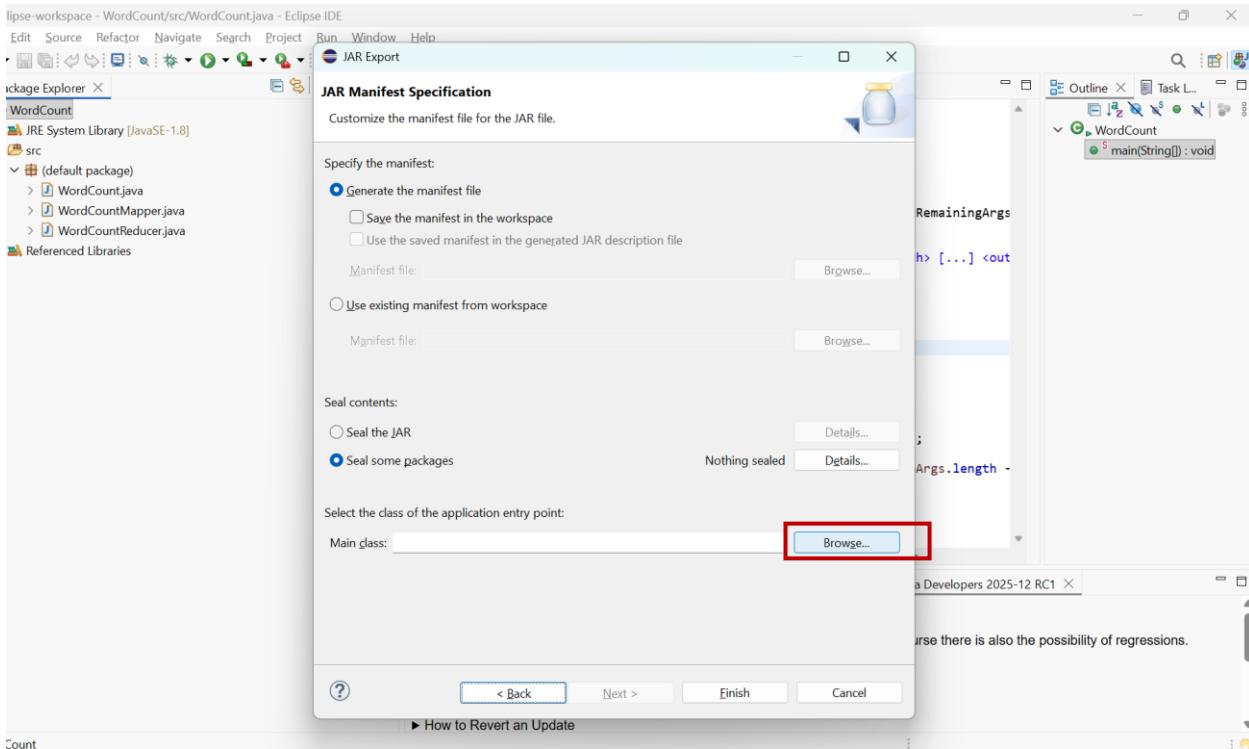
Chọn đường dẫn lưu file JAR và bấm Next



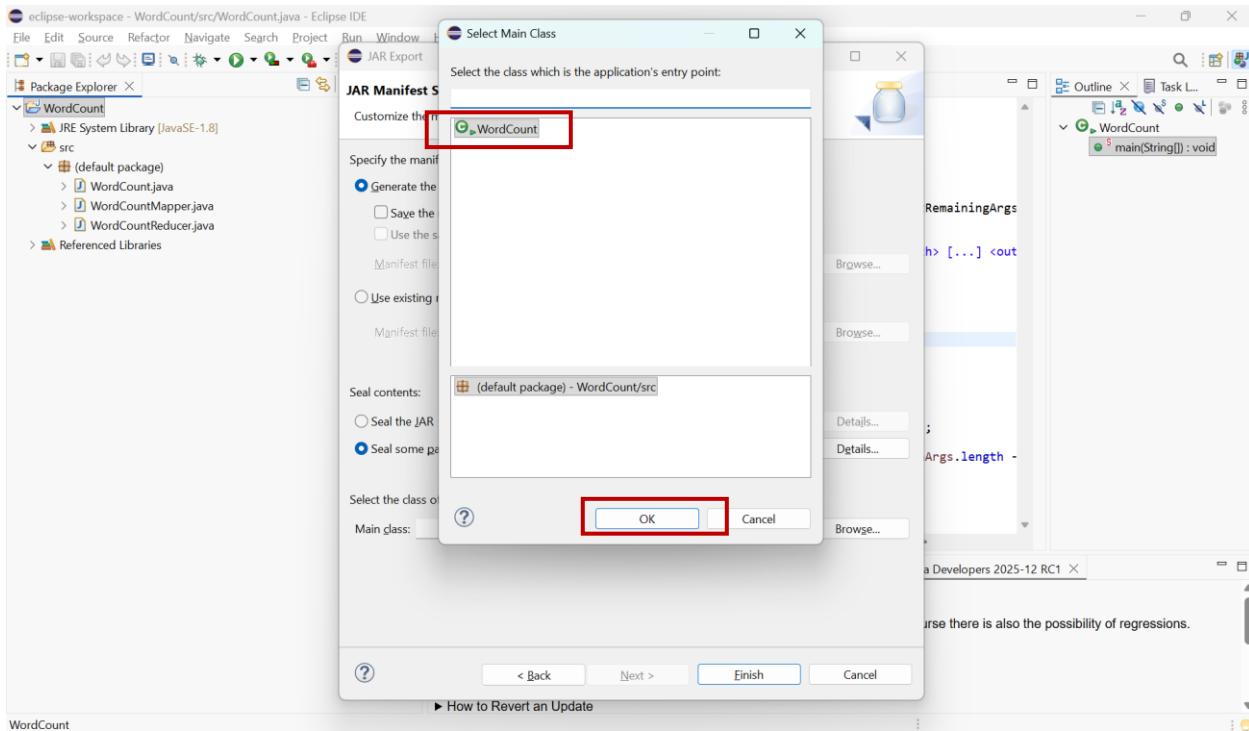
Bấm Next



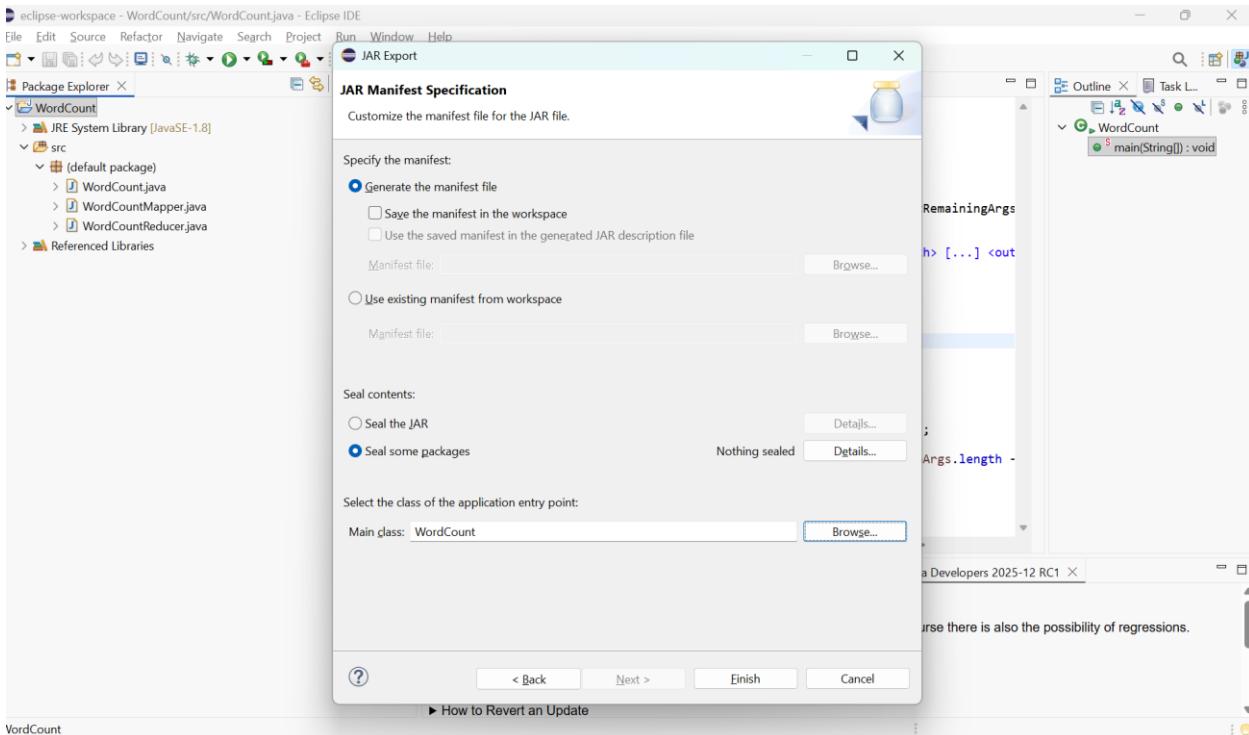
Bấm Browser để chọn file main



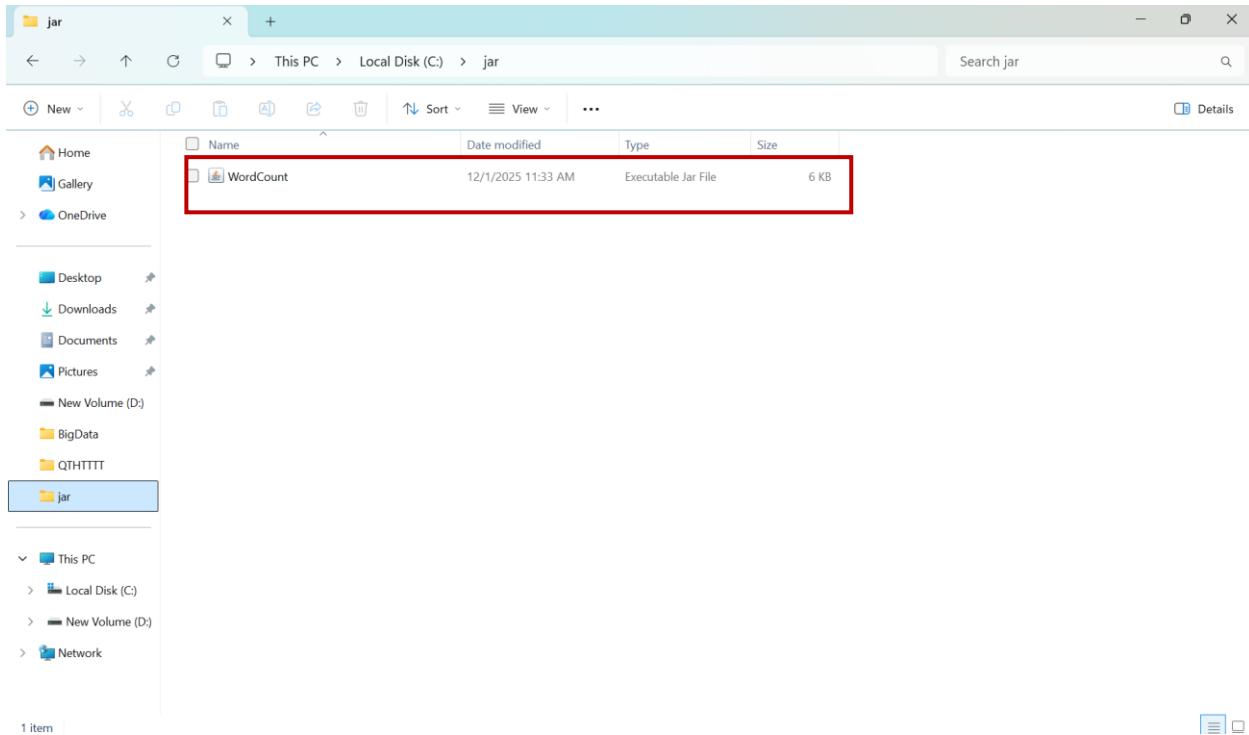
Chọn WordCount và bấm OK



Bấm Finish để thực hiện quá trình Export



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
```

```
C:\Windows\System32>hadoop jar "C:\jar\WordCount.jar" /input/data.txt /r_output
2025-12-01 11:37:16,180 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0:8032
2025-12-01 11:37:17,164 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_1764559453262_0002
2025-12-01 11:37:17,491 INFO input.FileInputFormat: Total input files to process : 1
2025-12-01 11:37:17,585 INFO mapreduce.JobSubmitter: number of splits:1
2025-12-01 11:37:17,793 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1764559453262_0002
2025-12-01 11:37:17,793 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-12-01 11:37:18,035 INFO conf.Configuration: resource-types.xml not found
2025-12-01 11:37:18,035 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-12-01 11:37:18,136 INFO impl.YarnClientImpl: Submitted application application_1764559453262_0002
2025-12-01 11:37:18,189 INFO mapreduce.Job: The url to track the job: http://DESKTOP-EDDD40F:8088/proxy/application_1764559453262_0002/
2025-12-01 11:37:18,192 INFO mapreduce.Job: Running job: job_1764559453262_0002
2025-12-01 11:37:27,384 INFO mapreduce.Job: Job job_1764559453262_0002 running in uber mode : false
2025-12-01 11:37:27,387 INFO mapreduce.Job: map 0% reduce 0%
2025-12-01 11:37:33,514 INFO mapreduce.Job: map 100% reduce 0%
2025-12-01 11:37:40,606 INFO mapreduce.Job: map 100% reduce 100%
2025-12-01 11:37:41,636 INFO mapreduce.Job: Job job_1764559453262_0002 completed successfully
2025-12-01 11:37:41,765 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=110
        FILE: Number of bytes written=530781
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=185
        HDFS: Number of bytes written=64
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
```

rly access to the latest and greatest features for your IDE,
rse there is also the possibility of regressions.

[IDE](#) [Download](#) [Ask Me Later](#) [No Thank You](#)

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/									Go!	File	Upload	Download	Preview
Show 25 entries									Search:				
checkbox	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	Actions				
checkbox	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:31	0	0 B	input	trash				
checkbox	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:42	0	0 B	output	trash				
checkbox	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 11:37	0	0 B	r_output	trash				
checkbox	drwx-----	Admin	supergroup	0 B	Dec 01 10:41	0	0 B	tmp	trash				

Showing 1 to 4 of 4 entries

Previous 1 Next

Hadoop, 2020.

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/r_output									Go!	File	Upload	Download	Preview
Show 25 entries									Search:				
checkbox	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	Actions				
checkbox	-rw-r--r--	Admin	supergroup	0 B	Dec 01 11:37	1	128 MB	_SUCCESS	trash				
checkbox	-rw-r--r--	Admin	supergroup	64 B	Dec 01 11:37	1	128 MB	part-r-00000	trash				

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2020.

Hadoop Overview Datanodes

Browse Directory

/r_output

Show 25 entries

Permission Owner

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

Showing 1 to 2 of 2 entries

Hadoop, 2020.

Block information - Block 0

Block ID: 1073741842
Block Pool ID: BP-676132101-172.20.10.4-1764170020332
Generation Stamp: 1018
Size: 64
Availability:
• DESKTOP-EDDD40F

File contents

Block Size	Name	
MB	_SUCCESS	
MB	part-r-00000	

Previous 1 Next

Close

The screenshot shows a Hadoop file browser interface. A modal window titled 'Block information - Block 0' is displayed in the center. The modal contains details about the block: Block ID (1073741842), Block Pool ID (BP-676132101-172.20.10.4-1764170020332), Generation Stamp (1018), Size (64), and Availability (DESKTOP-EDDD40F). Below this, a 'File contents' section shows the directory listing: BUS2, Bus 1, CAR 1, Car 1, TRAIN 2, buS 1, bus 3, caR 1. At the bottom of the modal is a 'Close' button. The background shows a 'Browse Directory' view with a list of files and a search bar.