

# THỰC HÀNH 2: BIG DATA

Họ và tên: Lê Thị Ngọc Ly\_64HTTT1\_2251162068

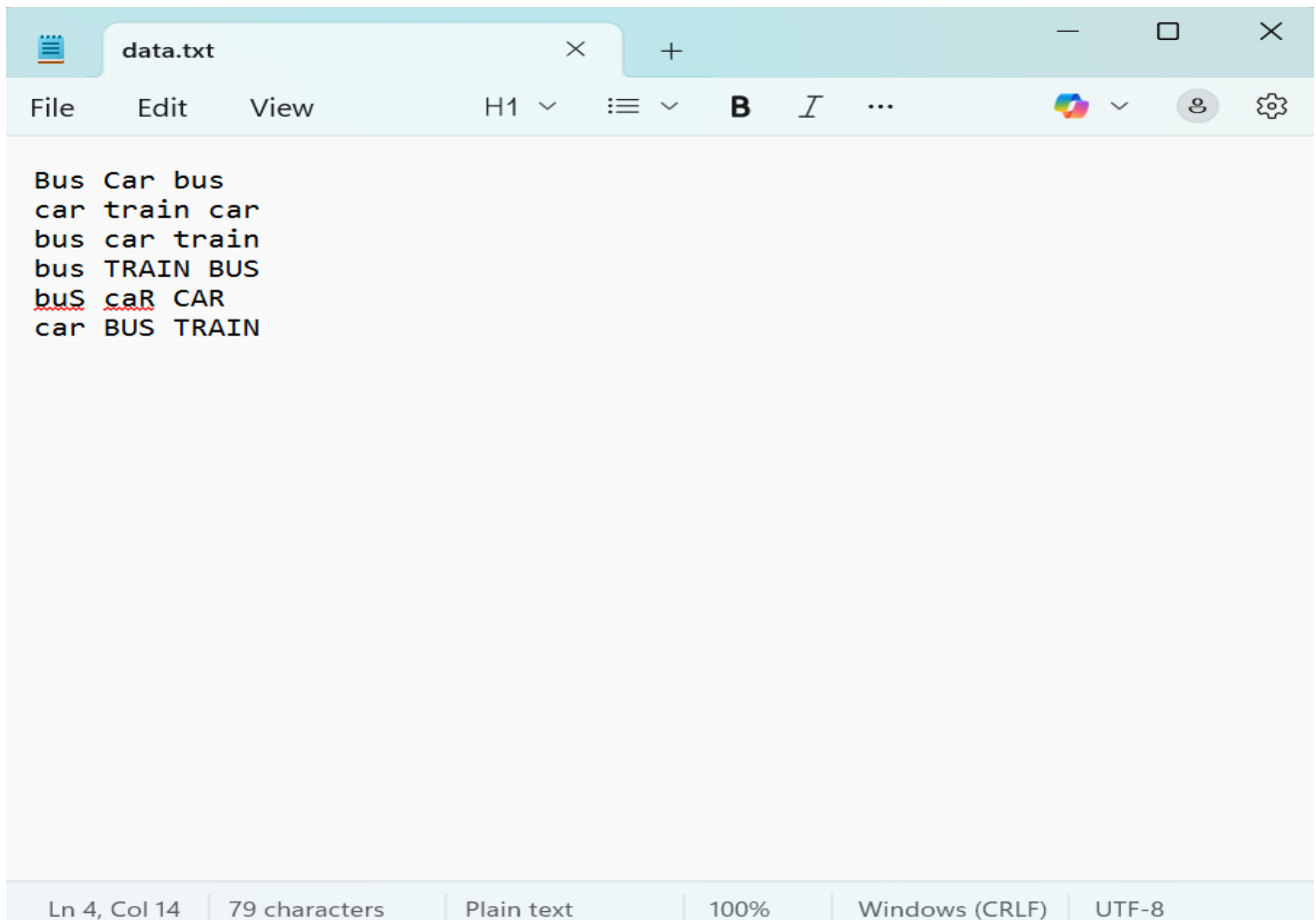
## MapReduce - Lập trình chương trình WordCount

### 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:

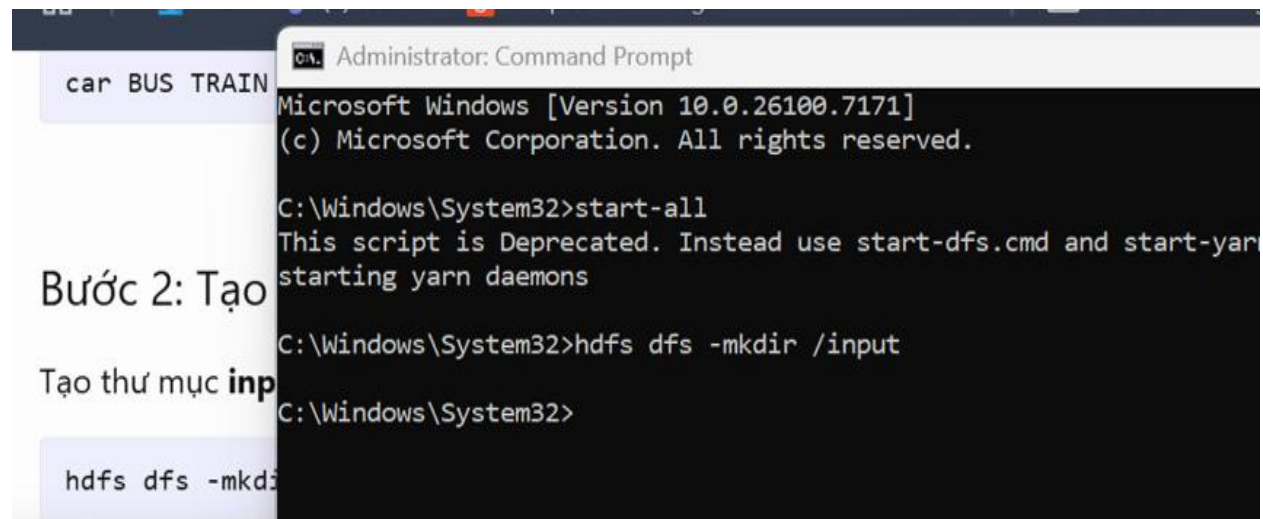
A screenshot of a text editor window titled 'data.txt'. The editor has a menu bar with 'File', 'Edit', and 'View'. Below the menu bar is a toolbar with various icons including a font face 'H1', a list icon, bold 'B', italic 'I', and others. The main text area contains the following content:

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

The words 'buS' and 'caR' in the fifth line are underlined in red. At the bottom of the window, a status bar shows: 'Ln 4, Col 14', '79 characters', 'Plain text', '100%', 'Windows (CRLF)', and 'UTF-8'.

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

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



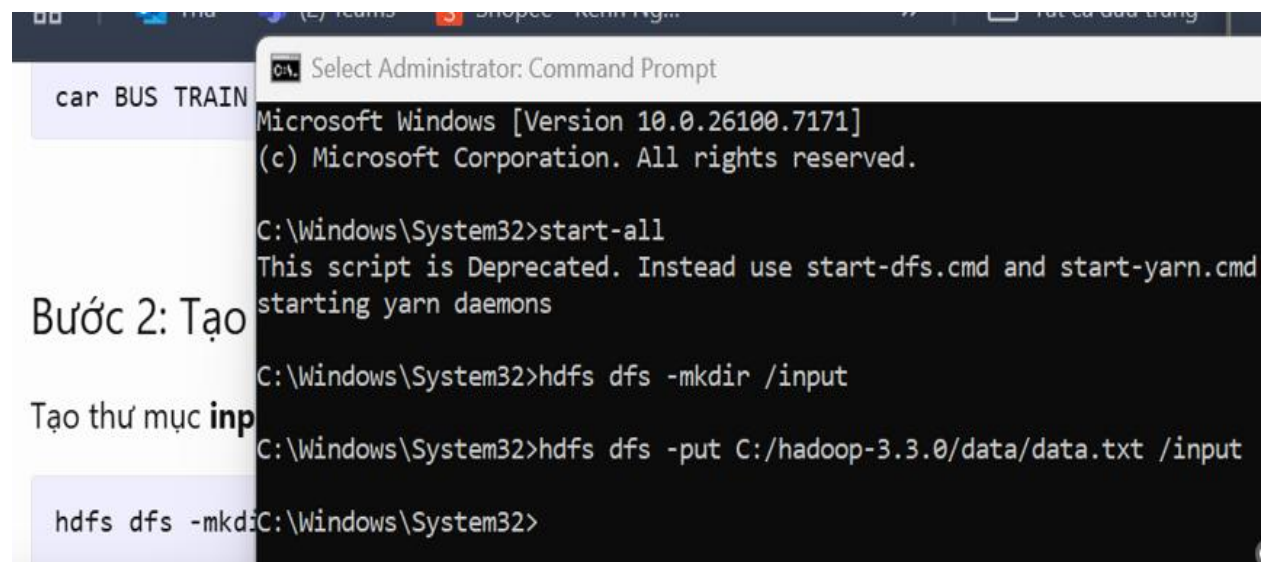
```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.26100.7171]
(c) Microsoft Corporation. All rights reserved.

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

C:\Windows\System32>hdfs dfs -mkdir /input

C:\Windows\System32>
```

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



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

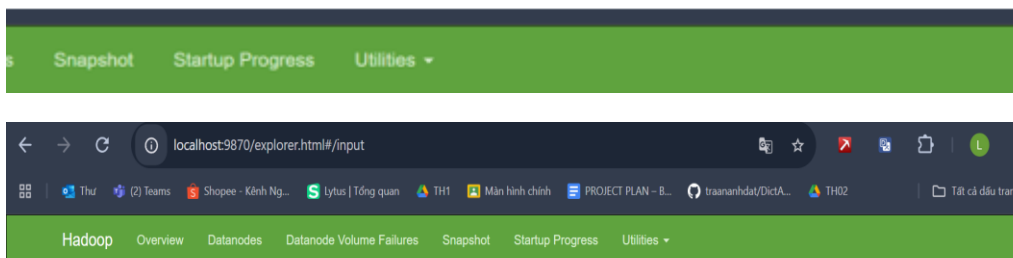
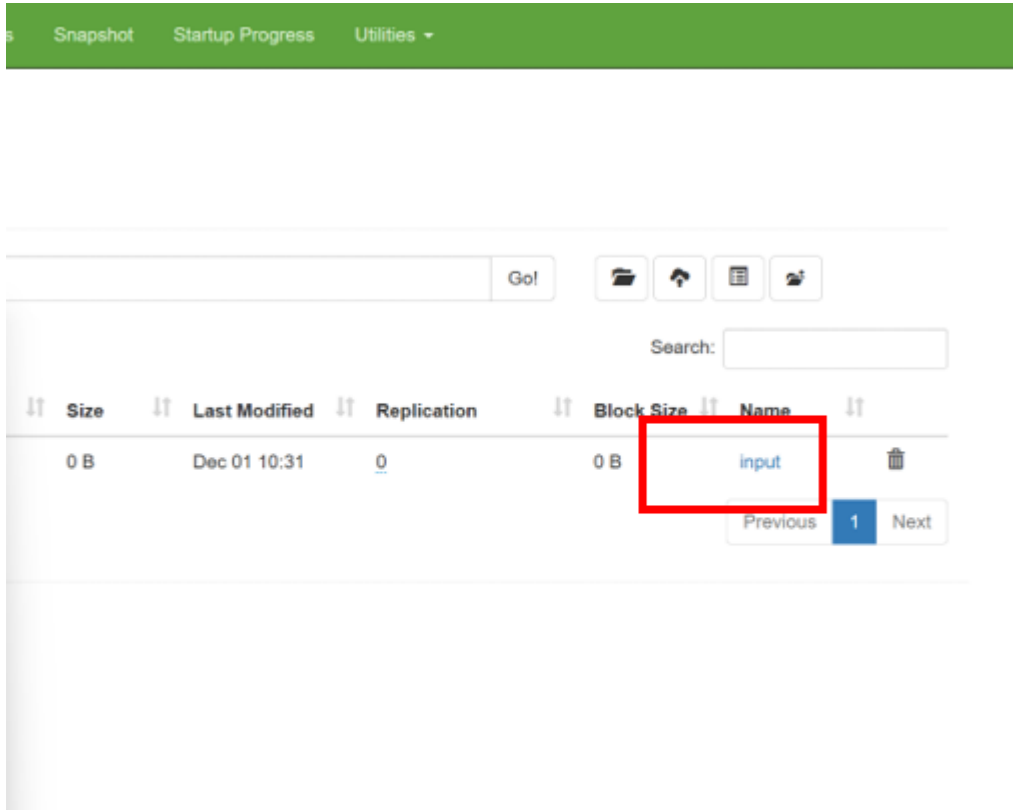
C:\Windows\System32>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\System32>hdfs dfs -mkdir /input

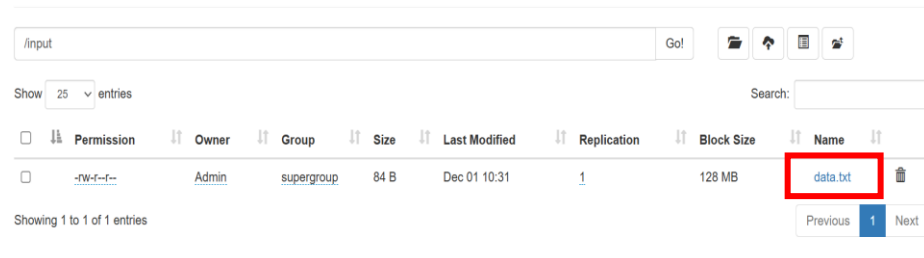
C:\Windows\System32>hdfs dfs -put C:/hadoop-3.3.0/data/data.txt /input

C:\Windows\System32>
```

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



## Browse Directory

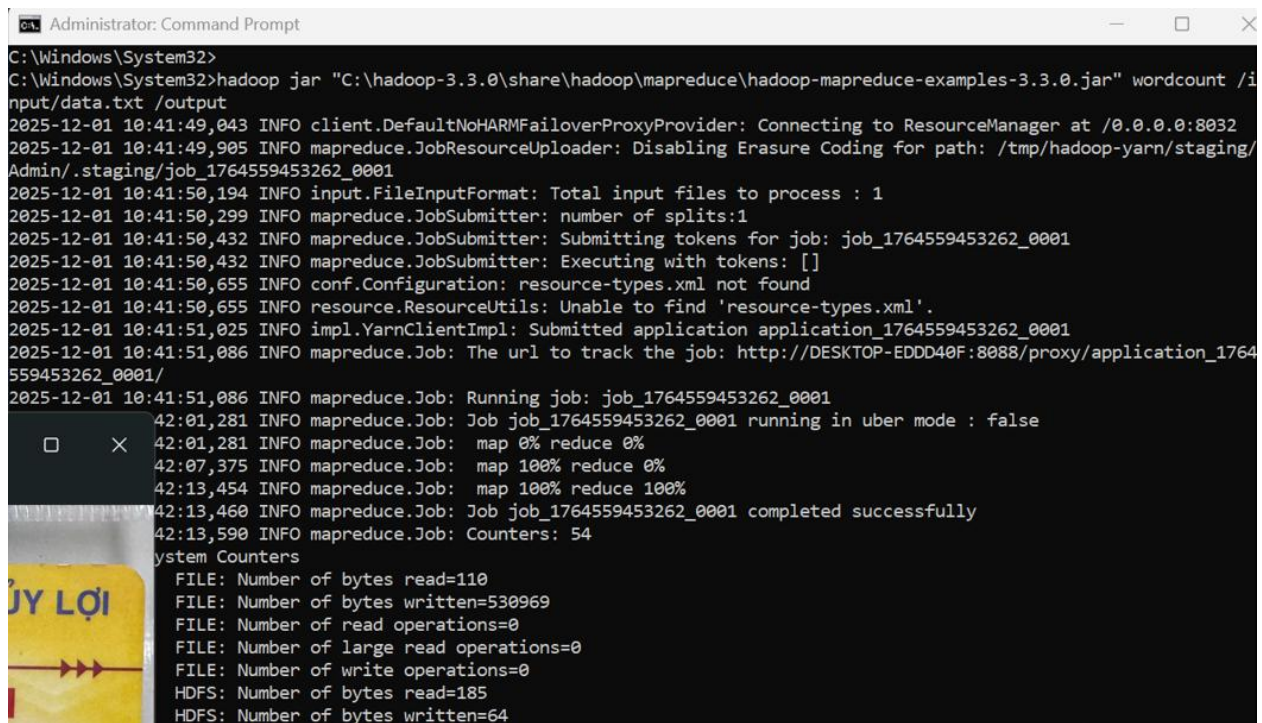


Hadoop, 2020.

### 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
```



```
Administrator: Command Prompt
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.DefaultNoHARMAFailoverProxyProvider: 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
2025-12-01 10:42:01,281 INFO mapreduce.Job: Job job_1764559453262_0001 running in uber mode : false
2025-12-01 10:42:01,281 INFO mapreduce.Job: map 0% reduce 0%
2025-12-01 10:42:07,375 INFO mapreduce.Job: map 100% reduce 0%
2025-12-01 10:42:13,454 INFO mapreduce.Job: map 100% reduce 100%
2025-12-01 10:42:13,460 INFO mapreduce.Job: Job job_1764559453262_0001 completed successfully
2025-12-01 10:42:13,590 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

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	0 B	Dec 01 10:31	<a href="#">0</a>	0 B	<a href="#">input</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	0 B	Dec 01 10:42	<a href="#">0</a>	0 B	<a href="#">output</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	0 B	Dec 01 11:37	<a href="#">0</a>	0 B	<a href="#">r_output</a>	
<input type="checkbox"/>	<a href="#">drwx-----</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	0 B	Dec 01 10:41	<a href="#">0</a>	0 B	<a href="#">tmp</a>	

[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"/>	<a href="#">-rw-r--r--</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	0 B	Dec 01 10:42	<a href="#">1</a>	128 MB	<a href="#">SUCCESS</a>	
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">Admin</a>	<a href="#">supergroup</a>	64 B	Dec 01 10:42	<a href="#">1</a>	128 MB	<a href="#">part-r-00000</a>	

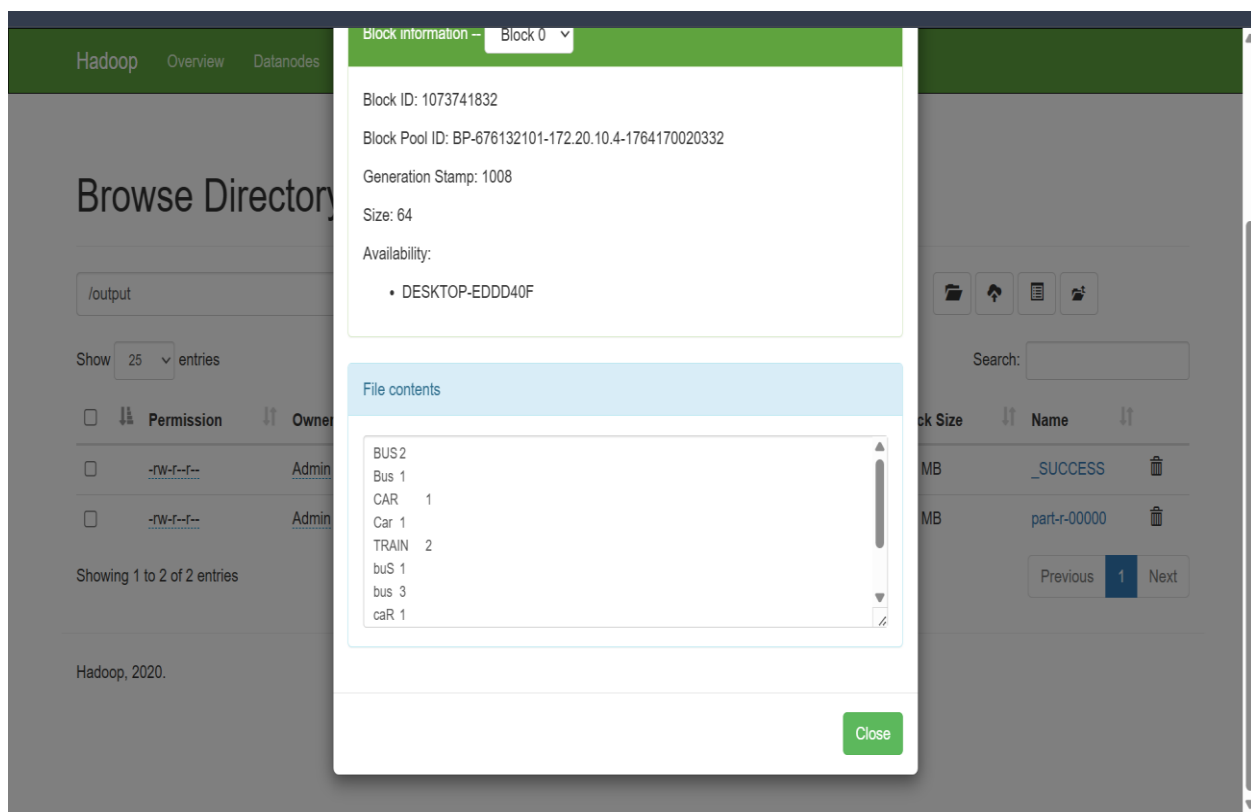
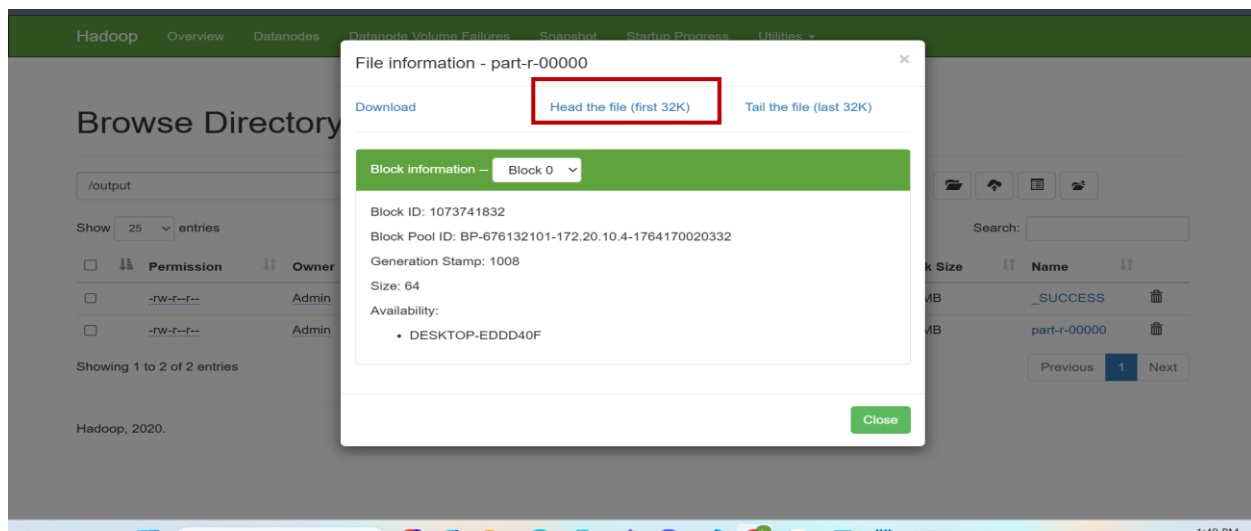
Showing 1 to 2 of 2 entries

Previous

1

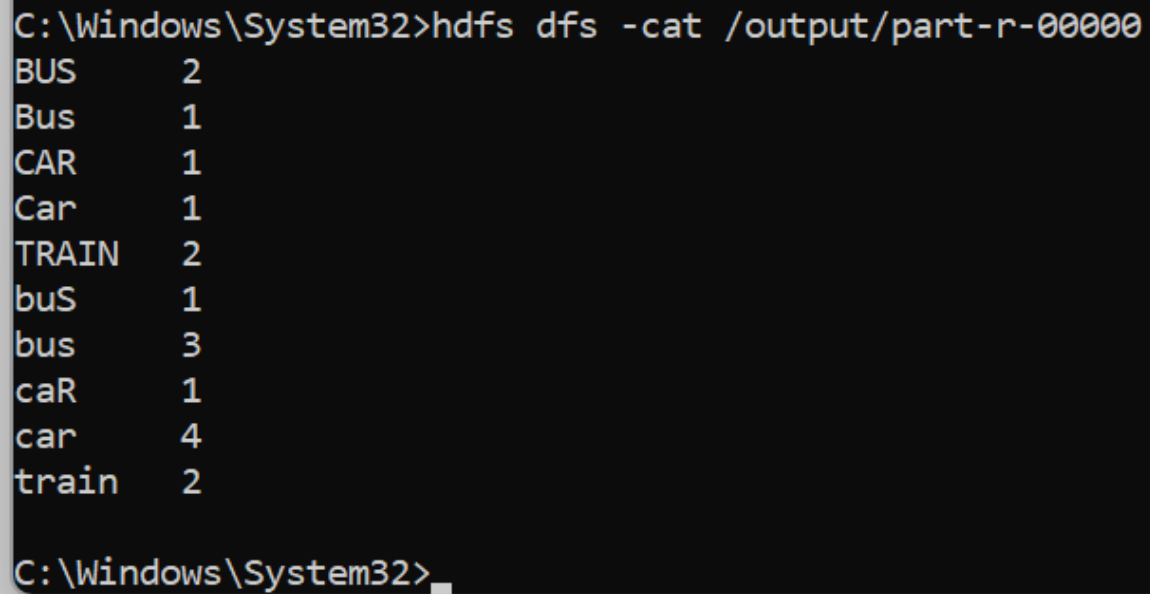
Next

Hadoop, 2020.



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>
```

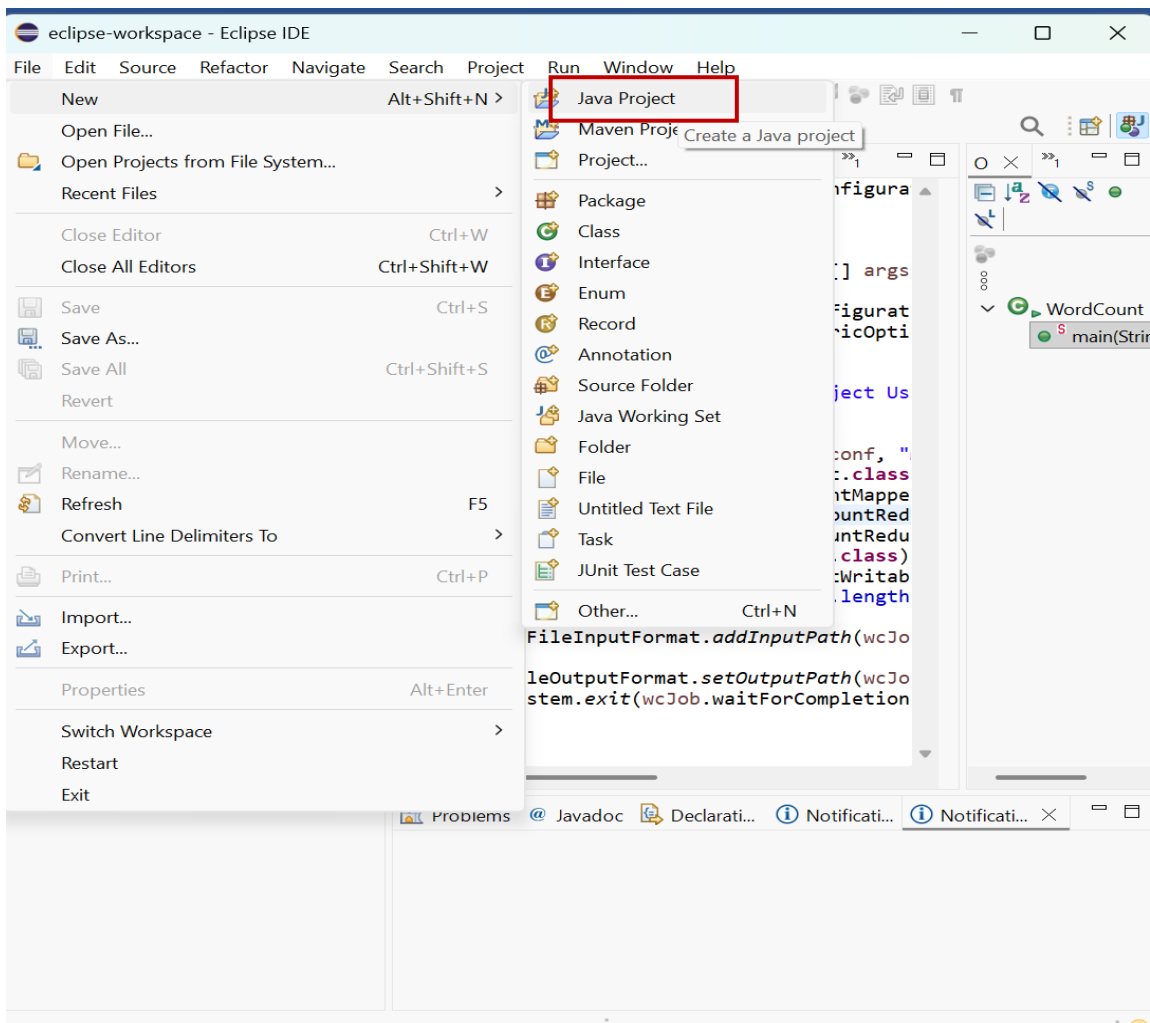
Như vậy ta đã chạy thành công chương trình mẫu MapReduce của Hadoop cung cấp.

## II. Lập trình chương trình WordCount bằng Eclipse

### Bước 1: Tạo project

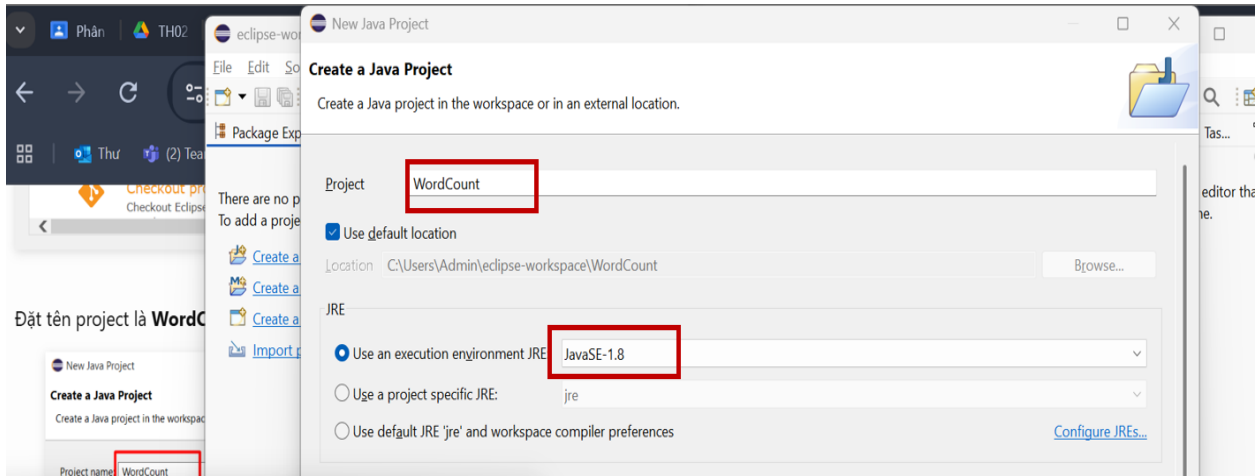
Mở chương trình Eclipse. Chọn **workspace** (nên để mặc định)

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



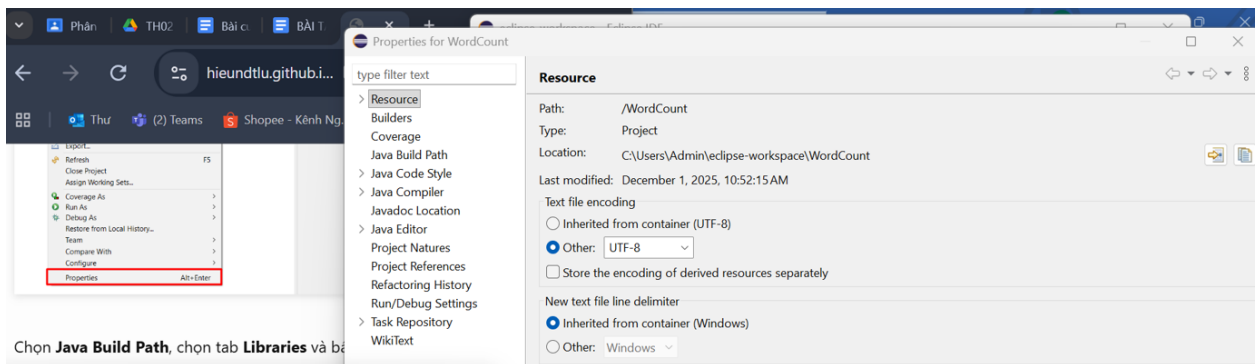


Đặt tên project là **WordCount** và chọn môi trường là **JavaSE-1.8**. Xong ấn **Finish**.



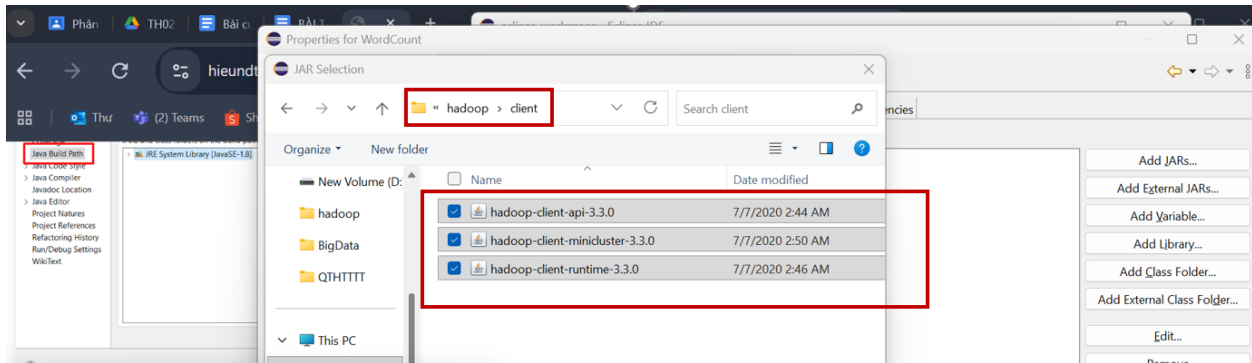
## Bước 2: Thêm thư viện cần thiết để chạy MapReduce

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

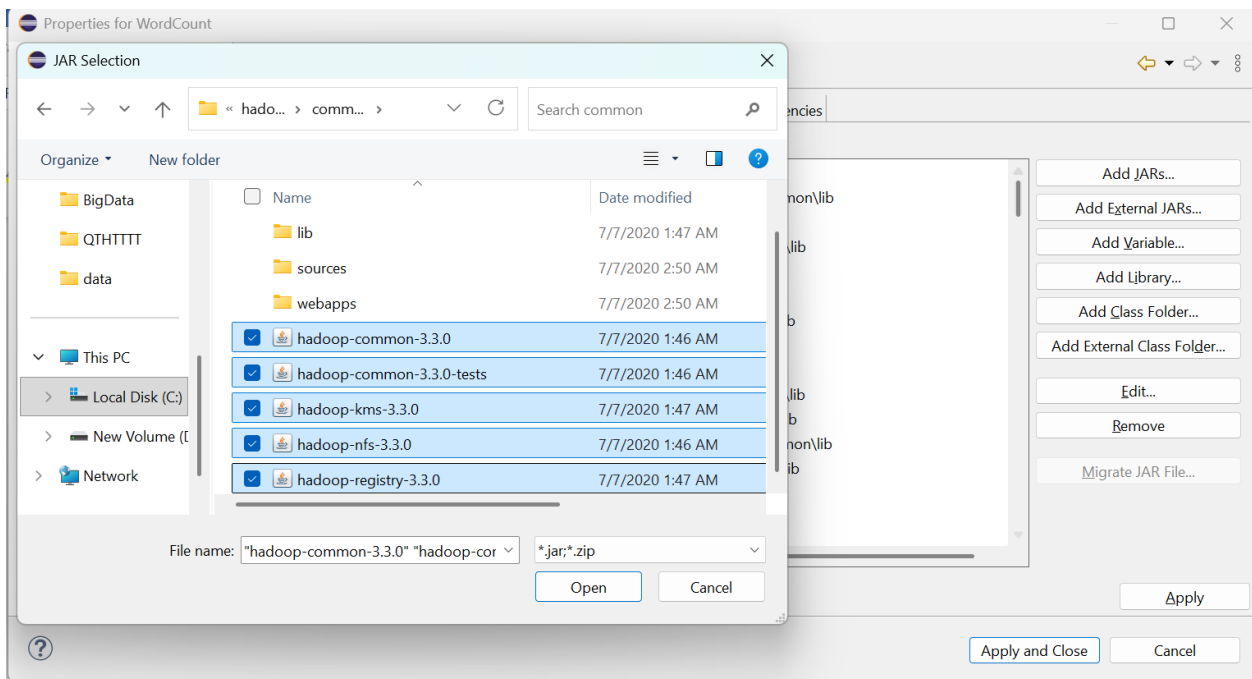


Chọn **Java Build Path**, chọn tab **Libraries** và bấm **Add External JARs**

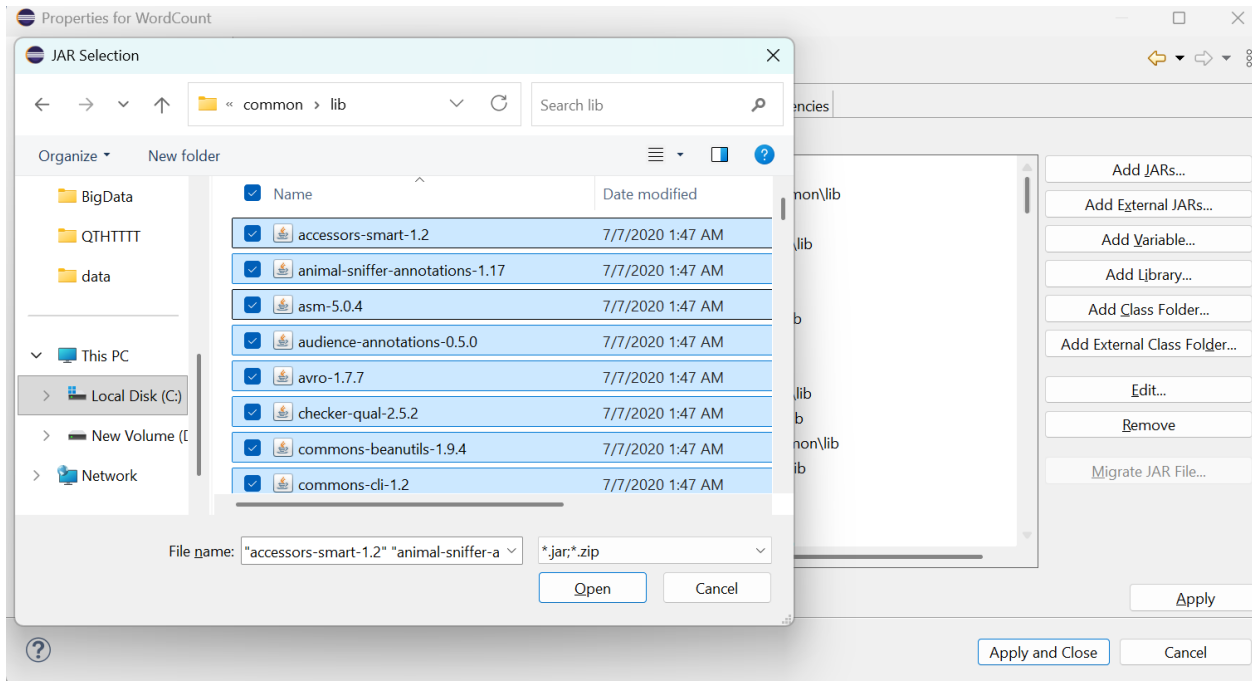
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\client** và ấn **Open**



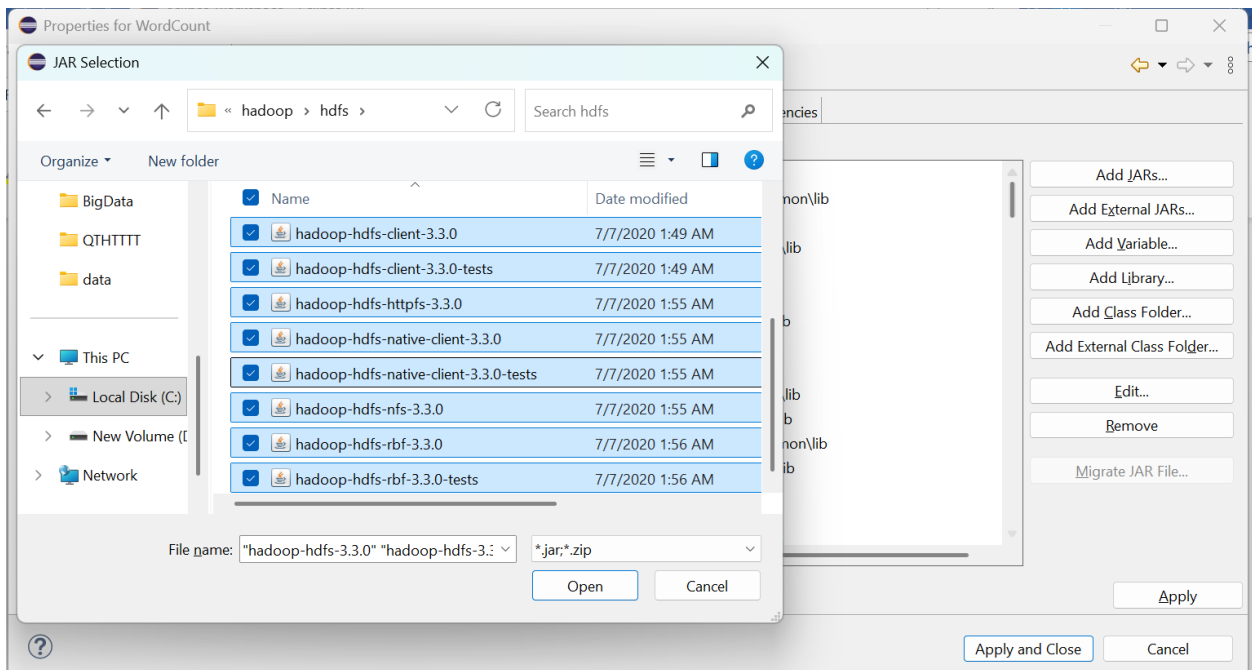
Tương tự chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common** và ấn **Open**



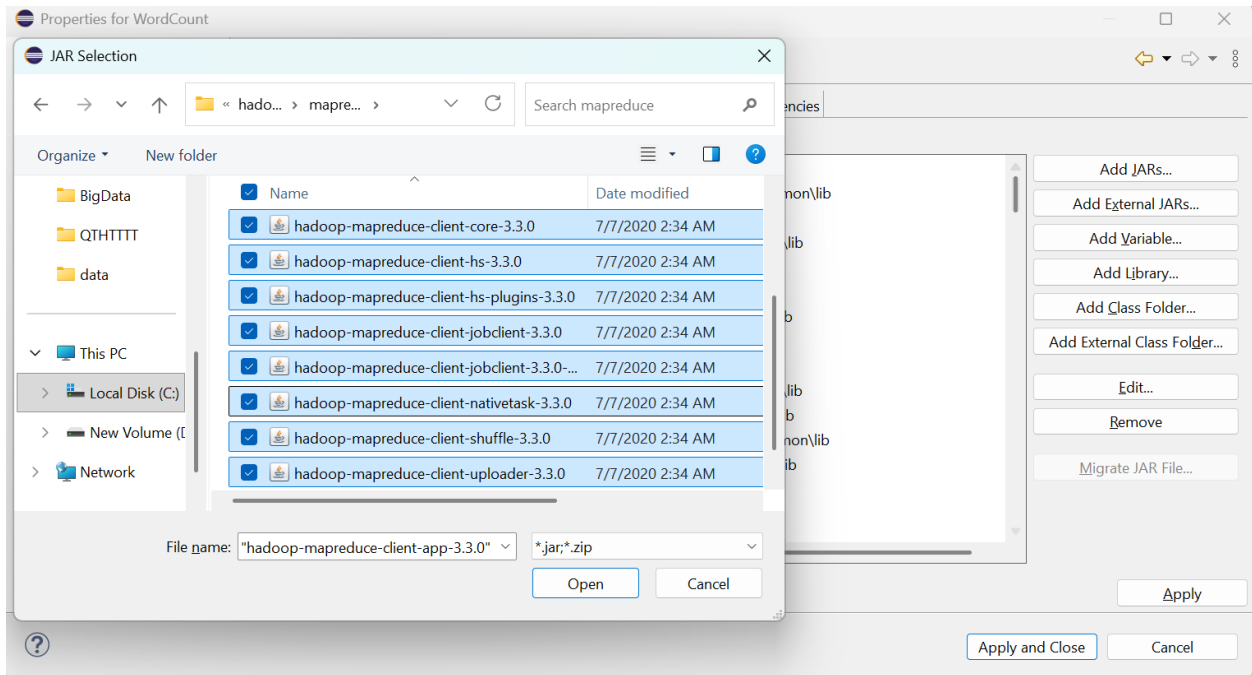
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common\lib** và ấn **Open**



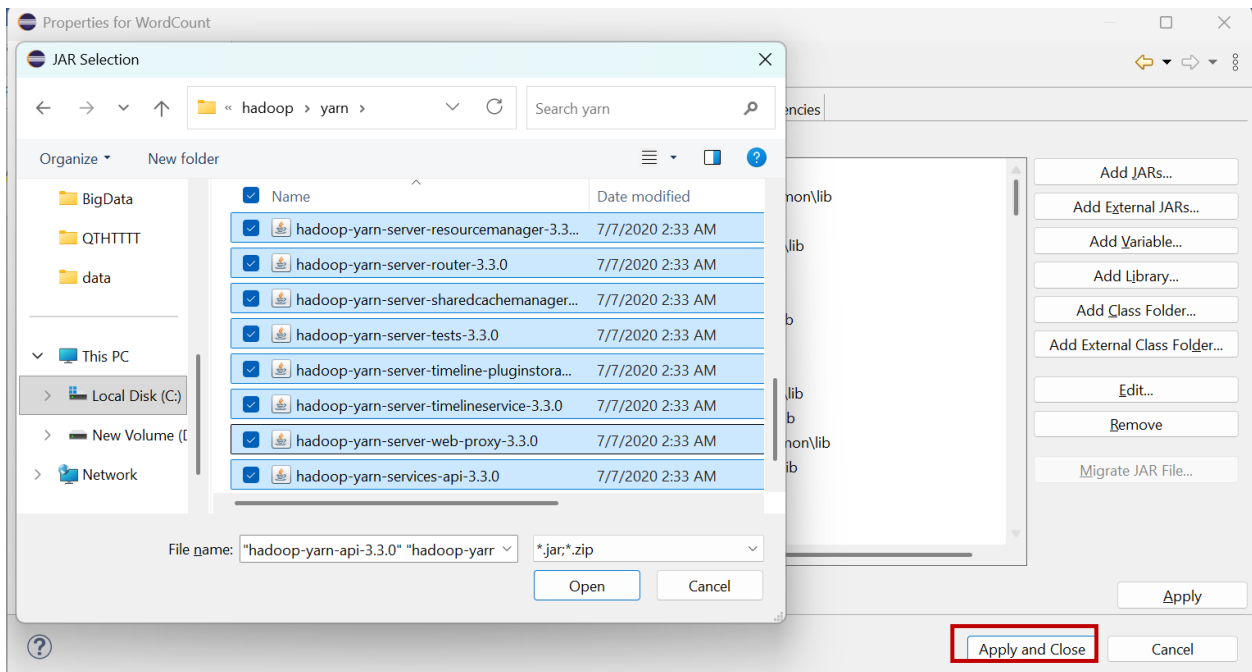
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\hdfs** và ấn **Open**



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



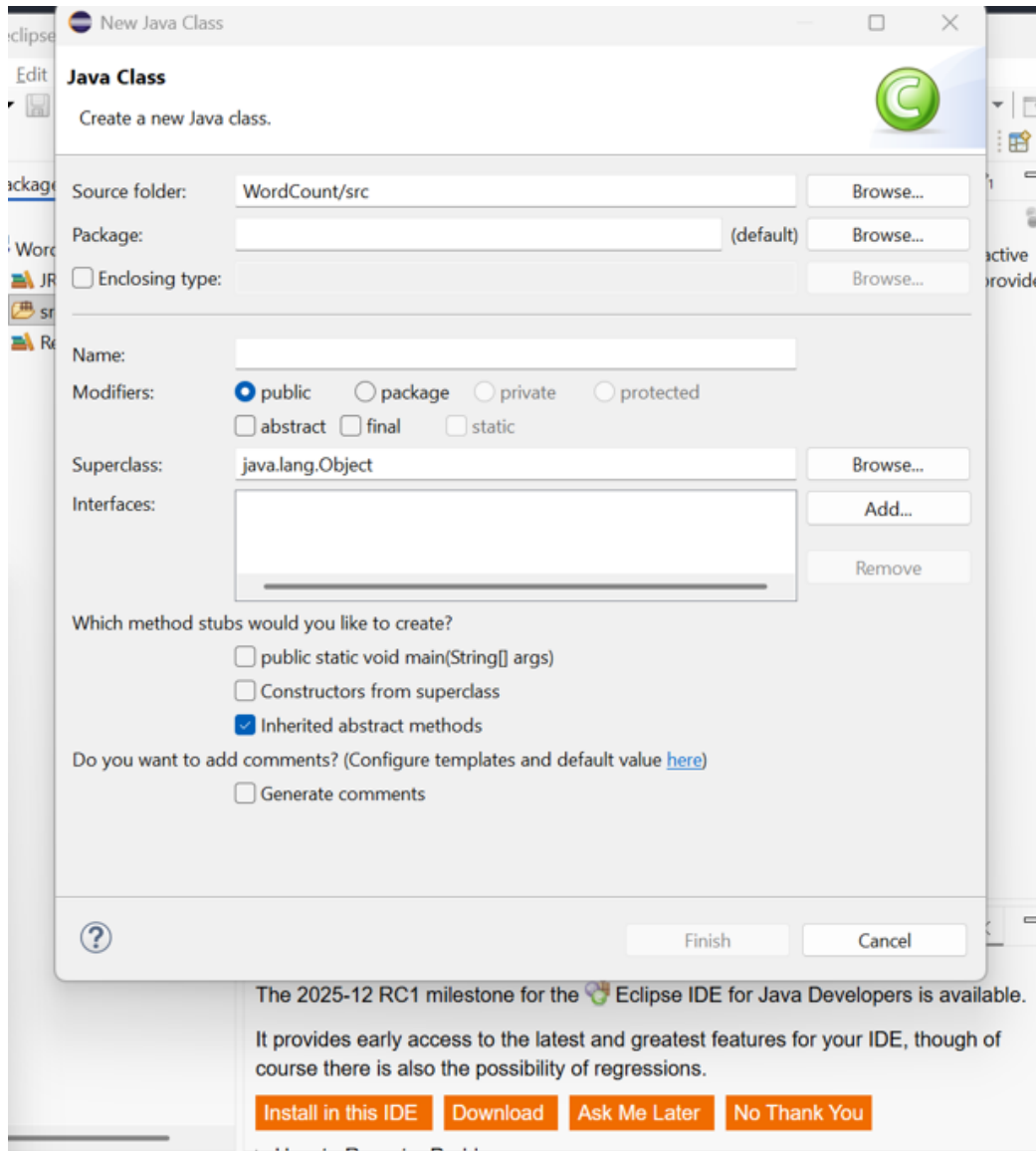
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\yarn** và ấn **Open**



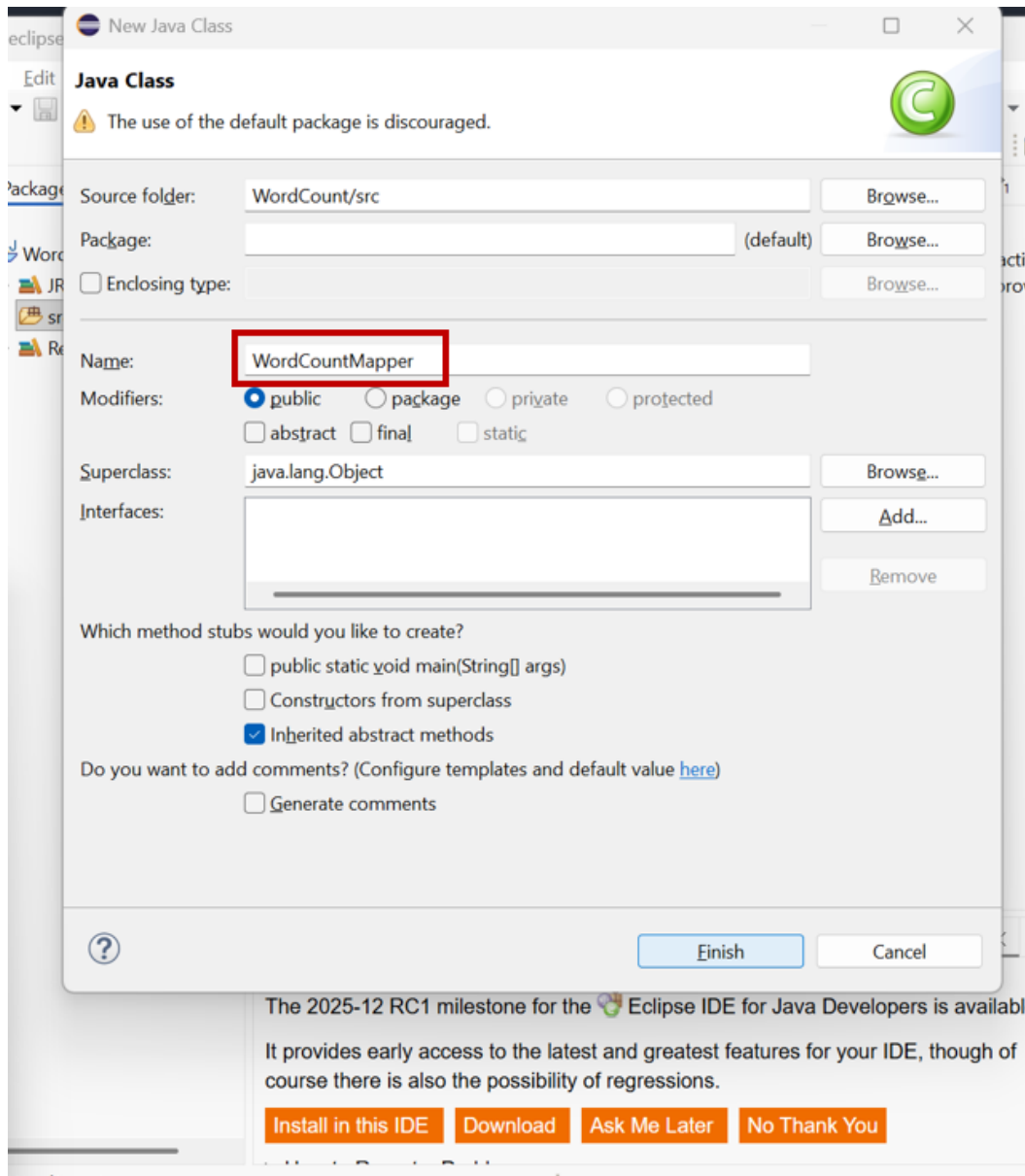
Ấn **Apply and Close**

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

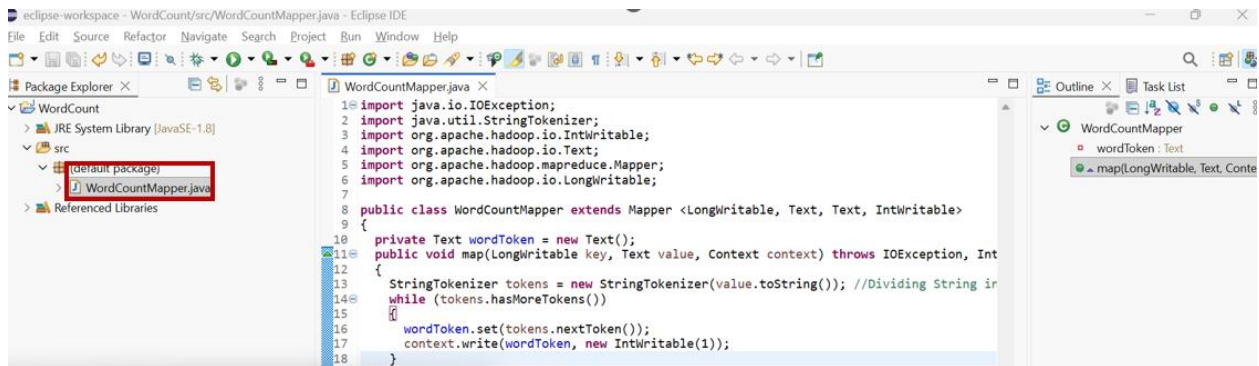
Double click vào project **WordCount**, chuột phải vào **src** và chọn **New > Class**



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



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

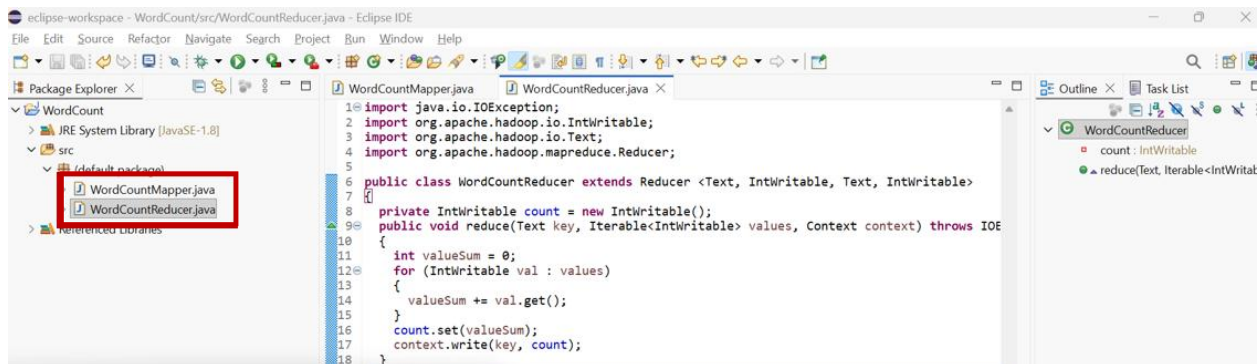


The screenshot shows the Eclipse IDE with the file `WordCountMapper.java` open. The Package Explorer on the left shows the project structure: `WordCount` > `src` > `WordCountMapper.java`. The main editor displays the following code:

```
1 import java.io.IOException;
2 import java.util.StringTokenizer;
3 import org.apache.hadoop.io.IntWritable;
4 import org.apache.hadoop.io.Text;
5 import org.apache.hadoop.mapreduce.Mapper;
6 import org.apache.hadoop.io.LongWritable;
7
8 public class WordCountMapper extends Mapper<LongWritable, Text, Text, IntWritable>
9 {
10     private Text wordToken = new Text();
11     public void map(LongWritable key, Text value, Context context) throws IOException, Int
12     {
13         StringTokenizer tokens = new StringTokenizer(value.toString()); //Dividing String in
14         while (tokens.hasMoreTokens())
15         {
16             wordToken.set(tokens.nextToken());
17             context.write(wordToken, new IntWritable(1));
18         }
19     }
20 }
```

The Outline view on the right shows the class `WordCountMapper` with a method `map(LongWritable, Text, Context)`.

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

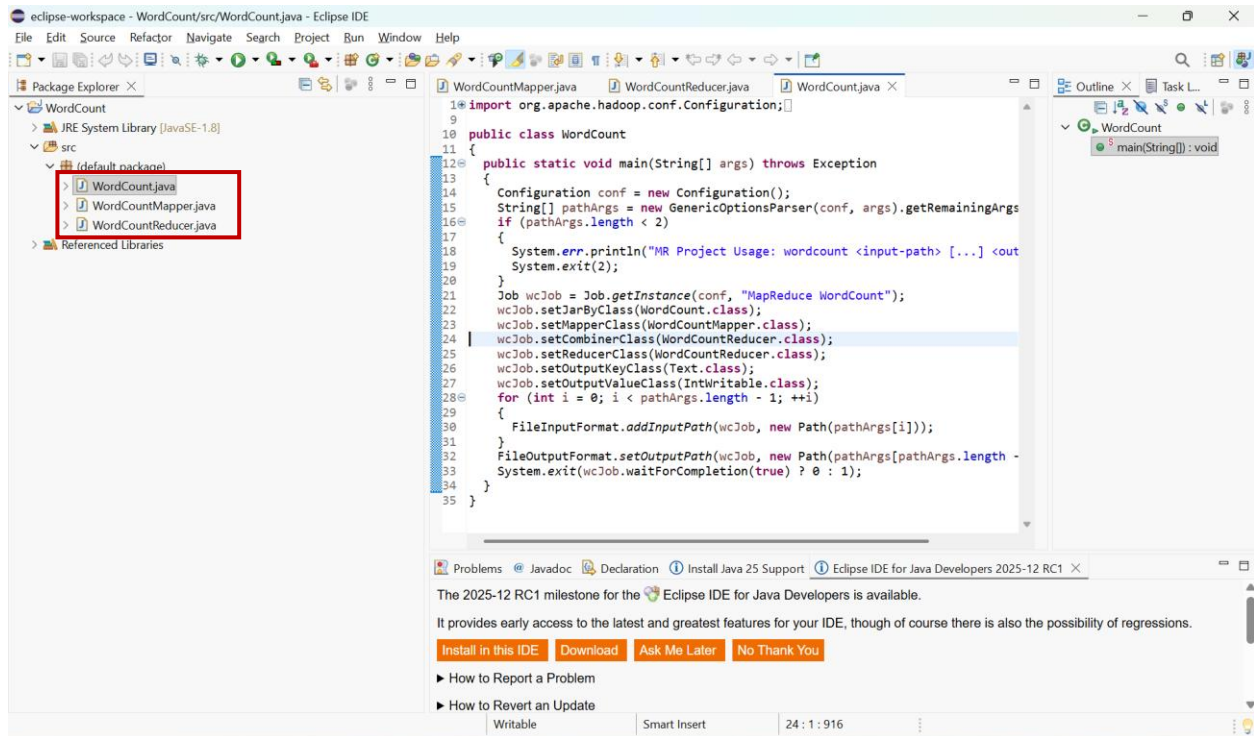


The screenshot shows the Eclipse IDE with the file `WordCountReducer.java` open. The Package Explorer on the left shows the project structure: `WordCount` > `src` > `WordCountMapper.java` and `WordCountReducer.java`. The main editor displays the following code:

```
1 import java.io.IOException;
2 import org.apache.hadoop.io.IntWritable;
3 import org.apache.hadoop.io.Text;
4 import org.apache.hadoop.mapreduce.Reducer;
5
6 public class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable>
7 {
8     private IntWritable count = new IntWritable();
9     public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOE
10     {
11         int valueSum = 0;
12         for (IntWritable val : values)
13         {
14             valueSum += val.get();
15         }
16         count.set(valueSum);
17         context.write(key, count);
18     }
19 }
```

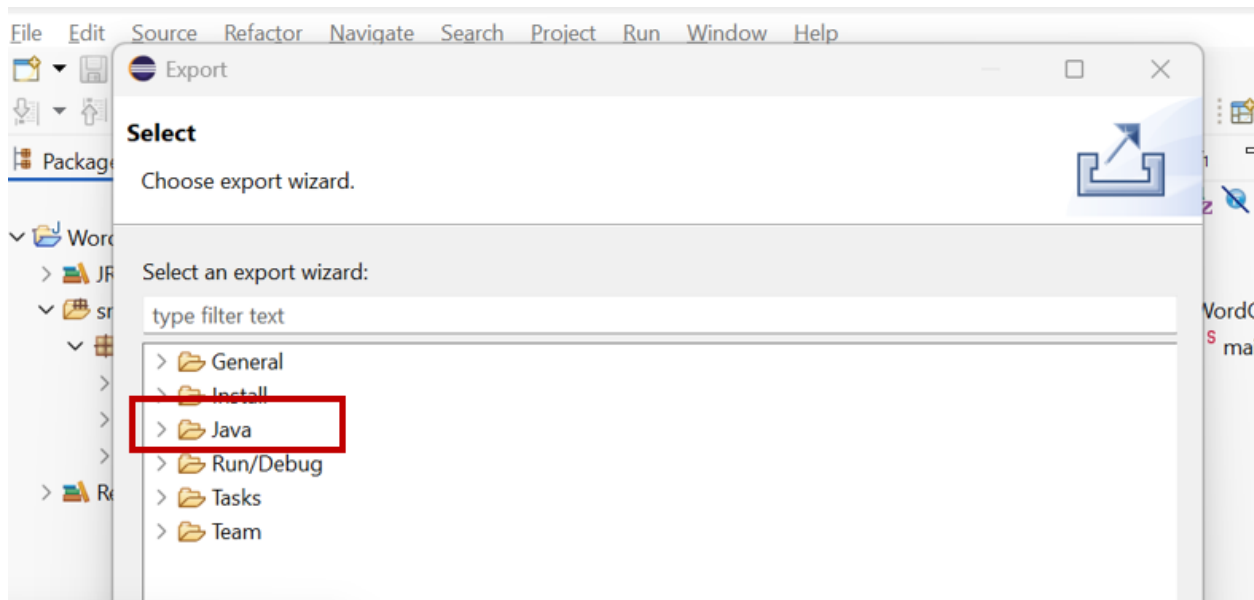
The Outline view on the right shows the class `WordCountReducer` with a method `reduce(Text, Iterable<IntWritable>)`.

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



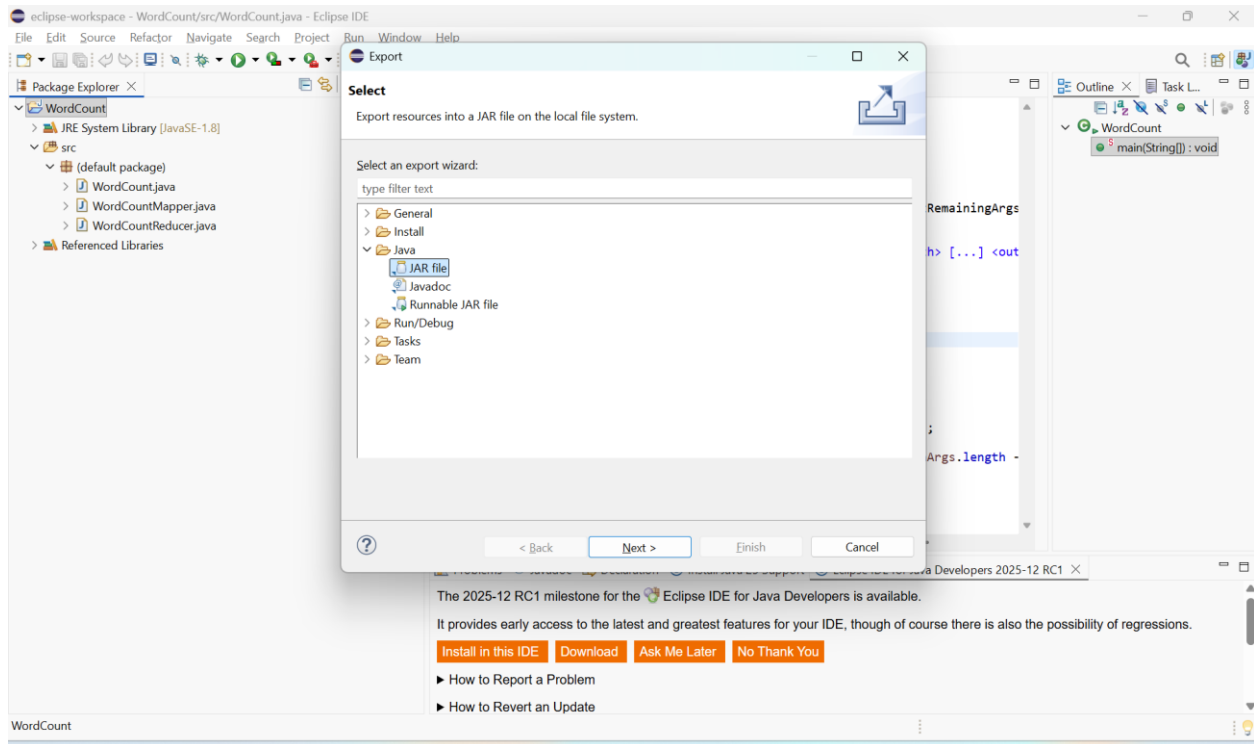
## Bước 4: Tạo file JAR

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

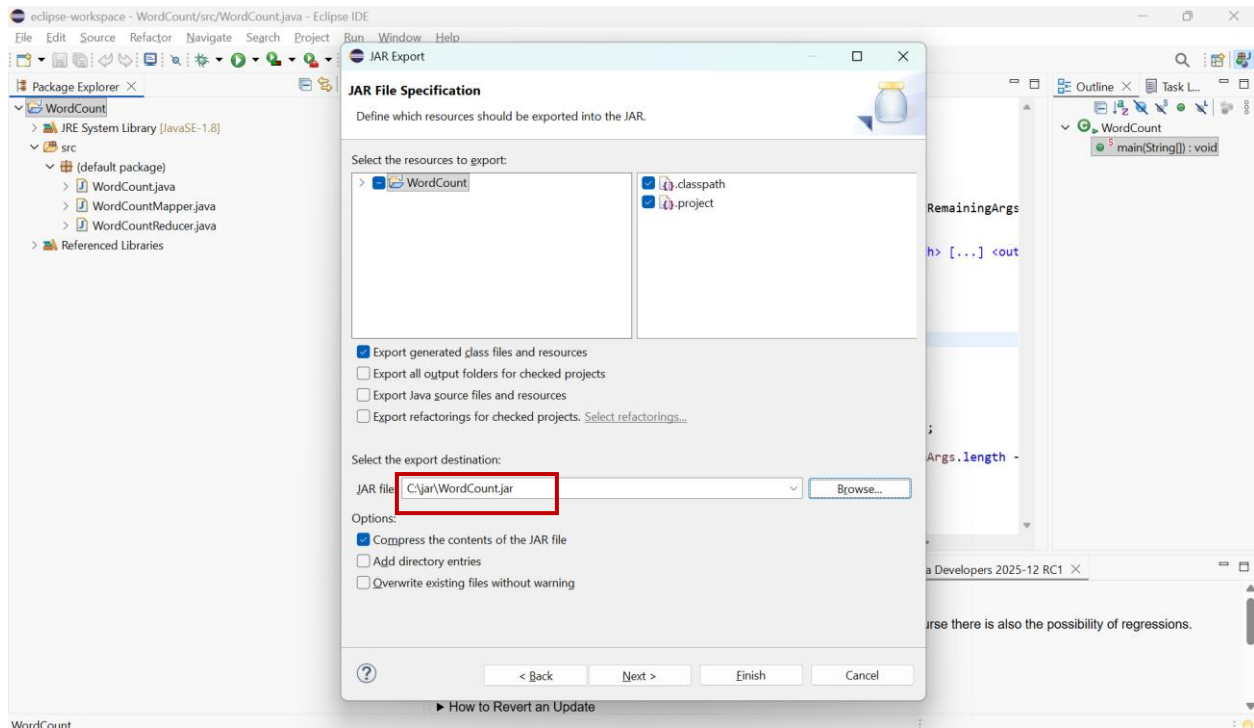




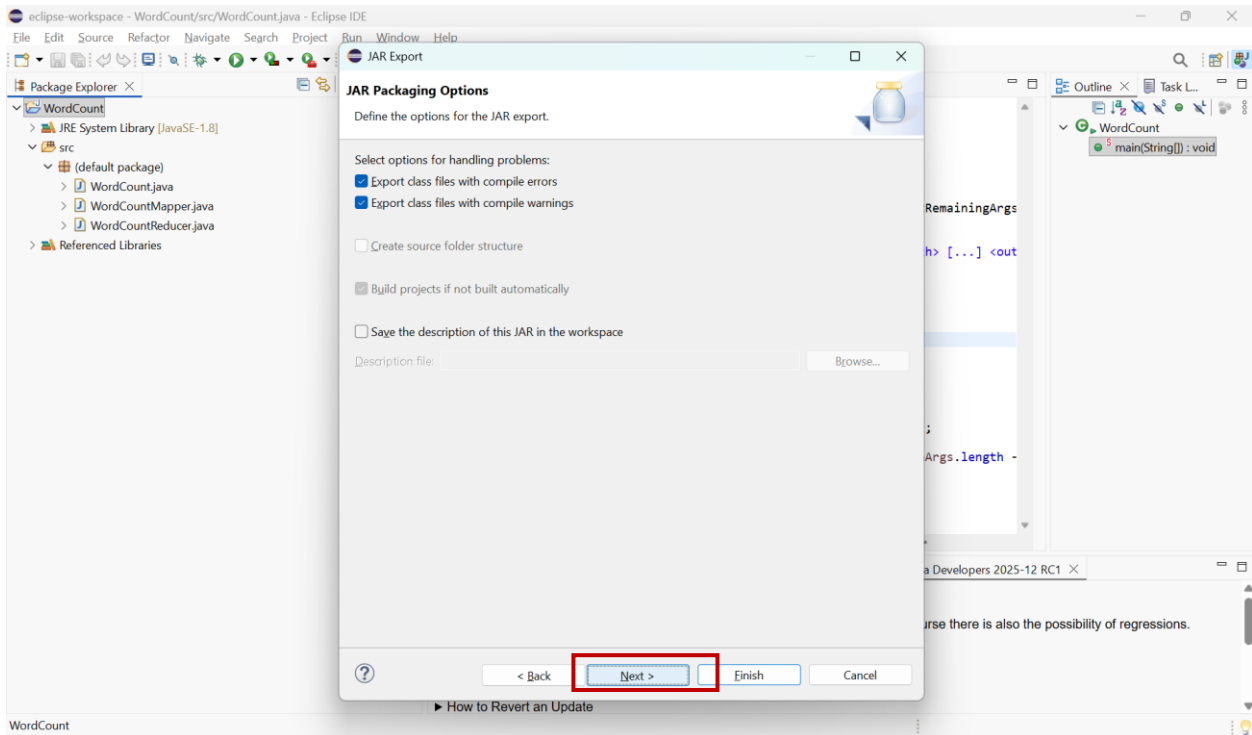
## Chọn **Java > JAR File** rồi bấm **Next**



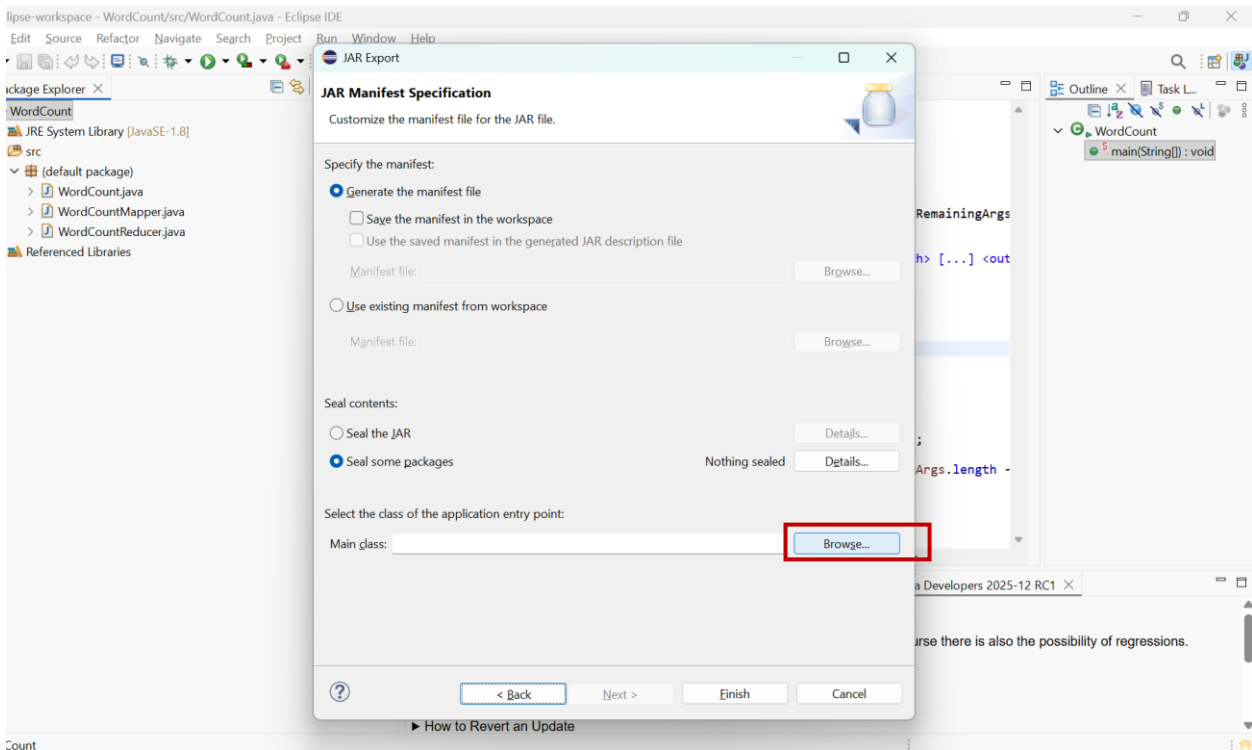
## Chọn đường dẫn lưu file JAR và bấm **Next**



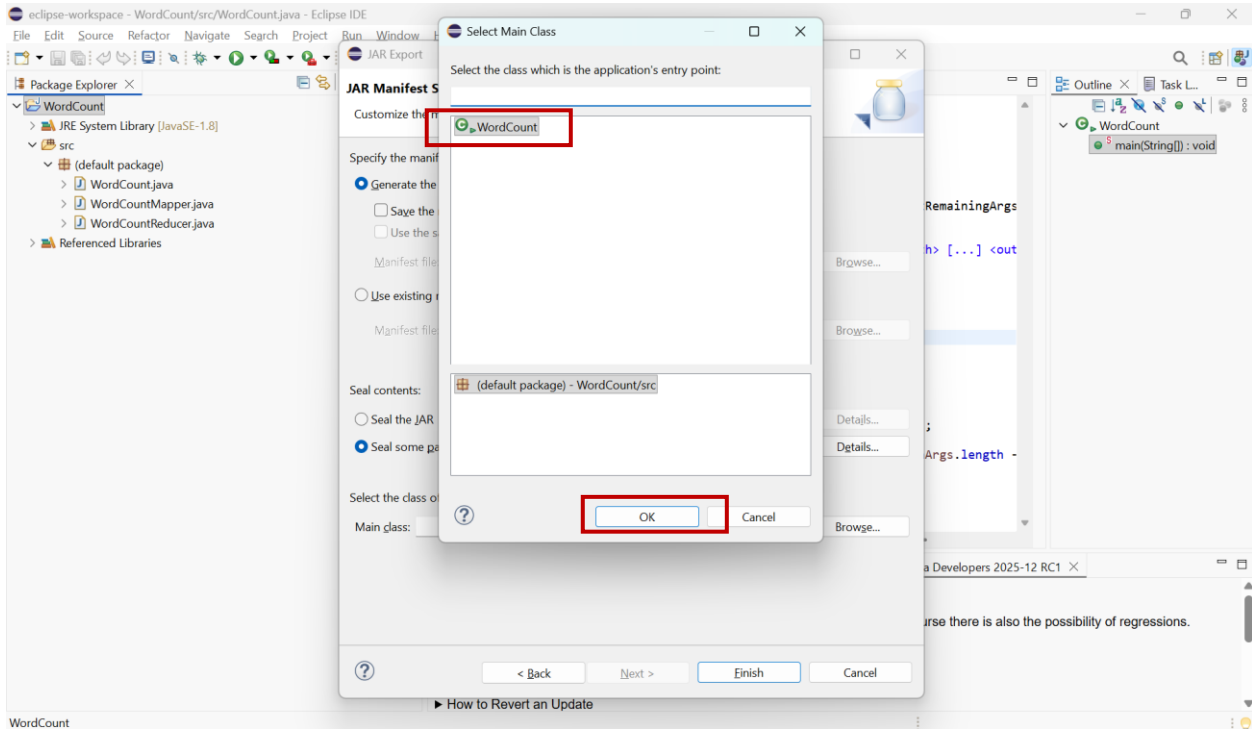
## Bấm Next



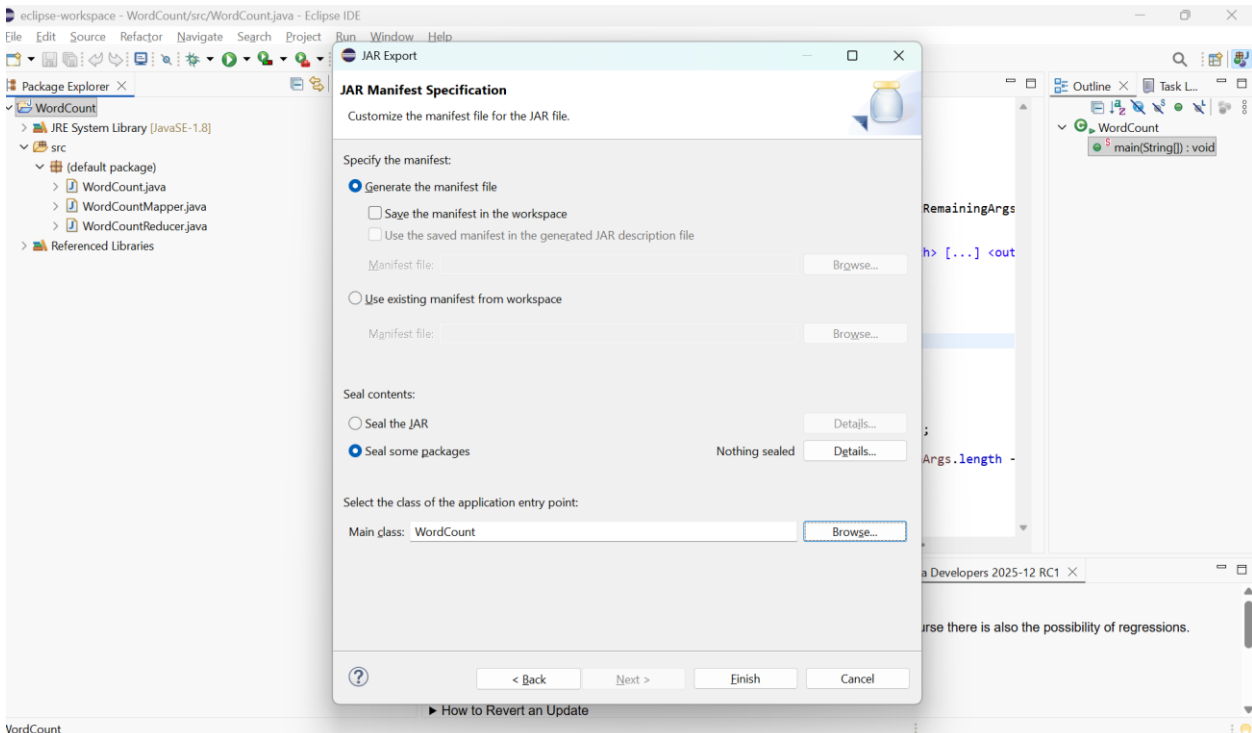
## Bấm Browser để chọn file main



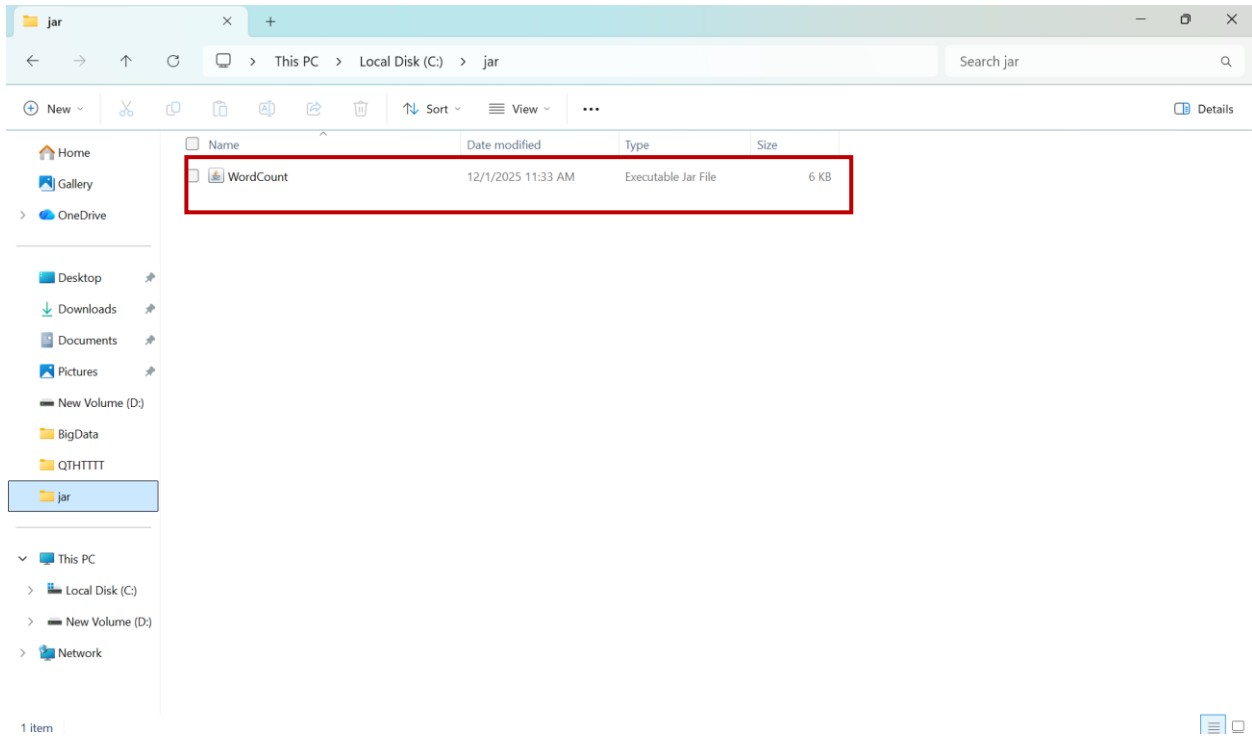
## Chọn **WordCount** và bấm **OK**



## 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
```

```
Administrator: Command Prompt
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 ResourceMan
ager at /0.0.0.0:8032
2025-12-01 11:37:17,164 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /t
mp/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 : fal
se
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
```

ly 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](#)

## Browse Directory

/ Go!    





Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:31	0	0 B	input	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 10:42	0	0 B	output	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	Admin	supergroup	0 B	Dec 01 11:37	0	0 B	r_output	<input type="checkbox"/>
<input type="checkbox"/>	drwx-----	Admin	supergroup	0 B	Dec 01 10:41	0	0 B	tmp	<input type="checkbox"/>

Showing 1 to 4 of 4 entries Previous 1 Next

Hadoop, 2020.

## 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"/>
<input type="checkbox"/>	-rwxr-xr-x	Admin	supergroup	0 B	Dec 01 11:37	1	128 MB	_SUCCESS	<input type="checkbox"/>
<input type="checkbox"/>	-rwxr-xr-x	Admin	supergroup	64 B	Dec 01 11:37	1	128 MB	part-r-00000	<input type="checkbox"/>

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2020.

Hadoop

Overview

Datanodes

Browse Directory

/r\_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: 1073741842

Block Pool ID: BP-676132101-172.20.10.4-1764170020332

Generation Stamp: 1018

Size: 64

Availability:

- DESKTOP-EDDD40F

File contents

BUS2

Bus 1

CAR 1

Car 1

TRAIN 2

bus 1

bus 3

caR 1

Close

Search:

Block Size

Name

MB

\_SUCCESS

MB

part-r-00000

Previous

1

Next