THỰC HÀNH 2: BIG DATA

Họ và tên: Lê Thị Ngọc Ly_64HTTT1_2251162068

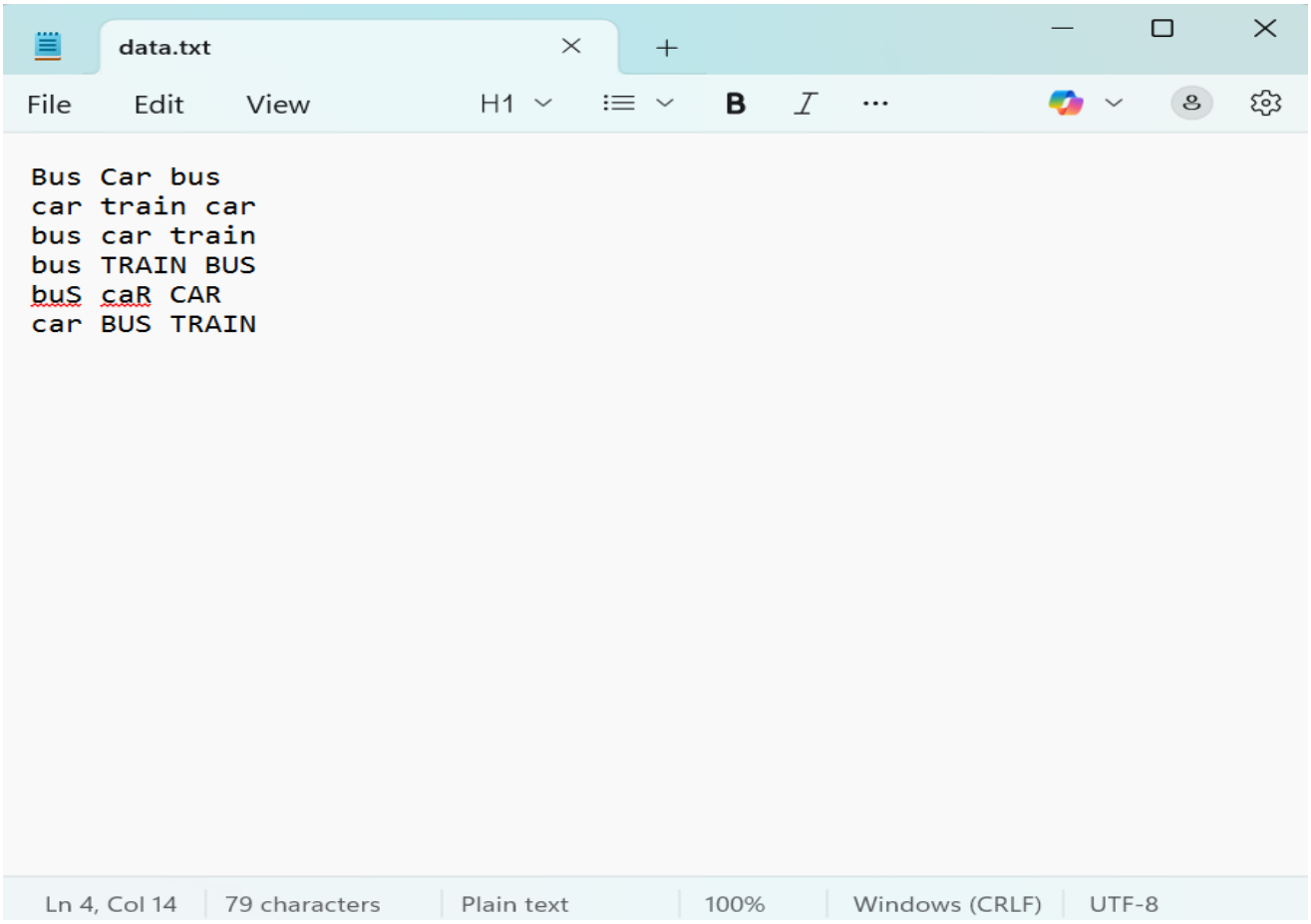
MapReduce - Lập trình chương trình WordCount

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]**

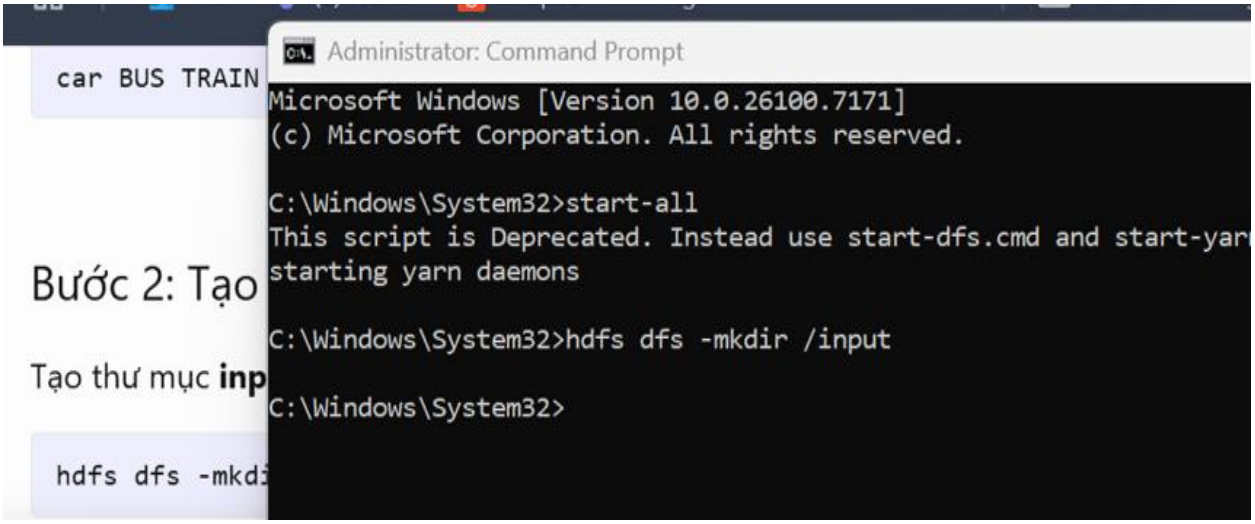
Bước 1: Tạo file data.txt

Nội dung của file data.txt là:

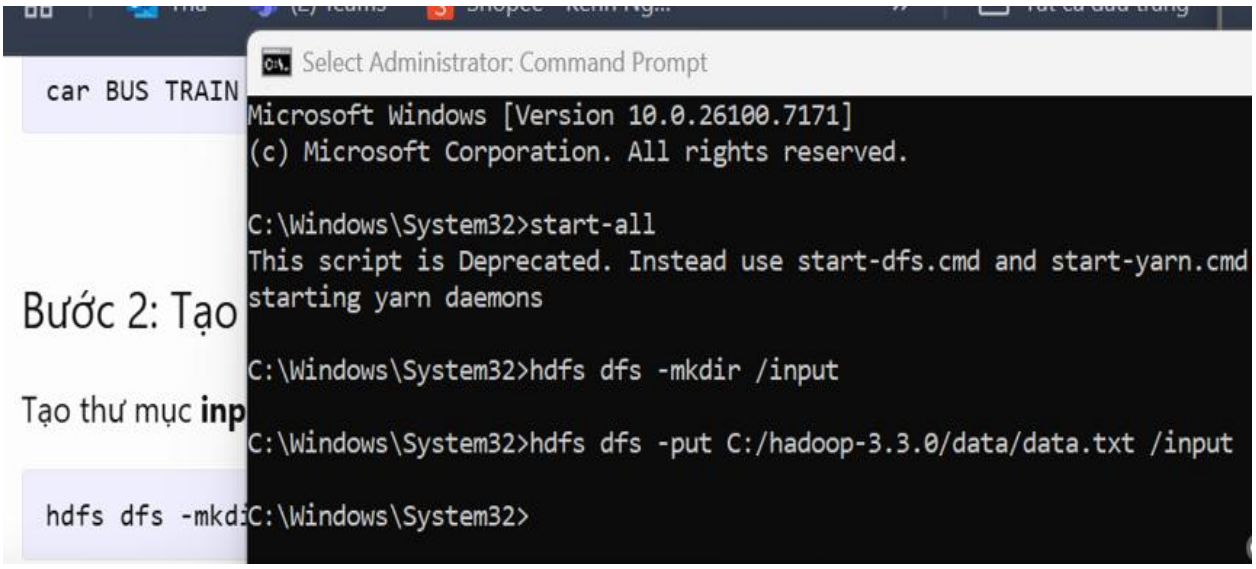


Bước 2: Tạo thư mục input tại hdfs và lưu file data.txt

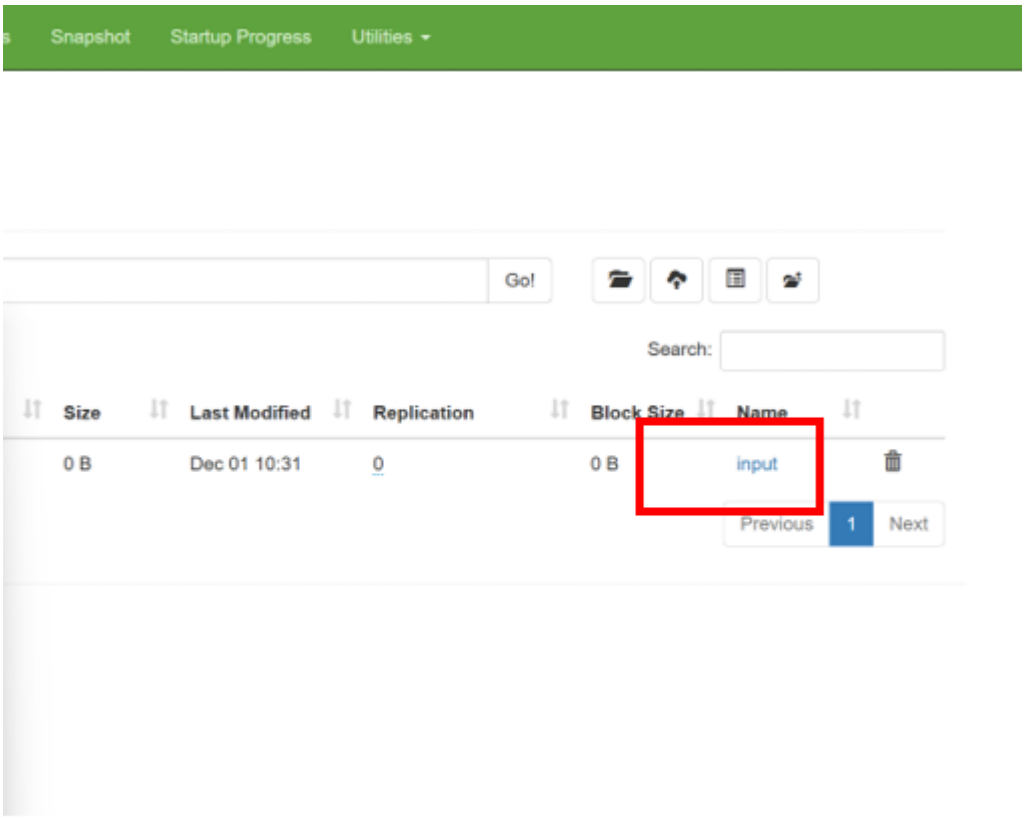
Tạo thư mục **input** trong hdfs với câu lệnh:

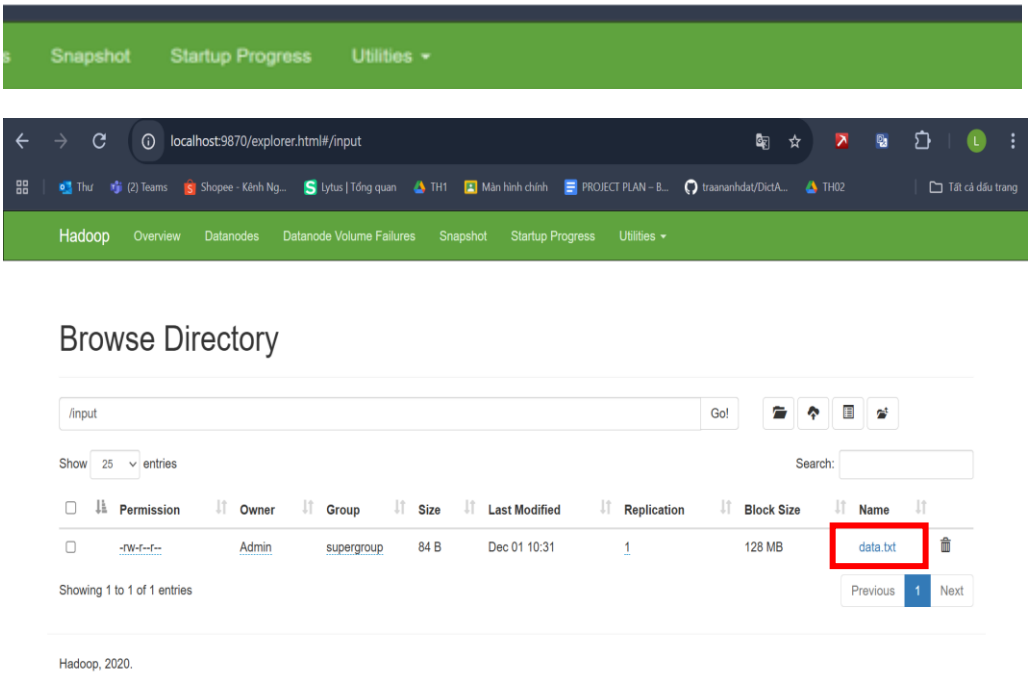


Đẩy file **data.txt** vào folder **input** vừa tạo:



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

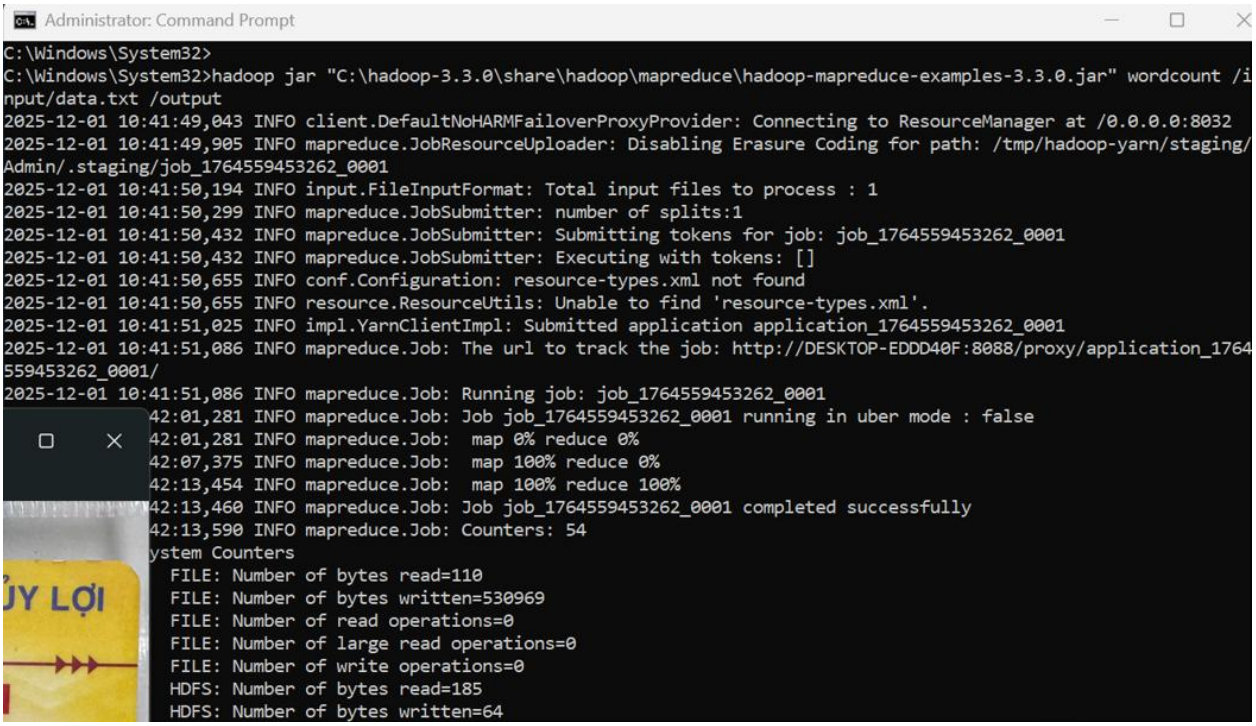




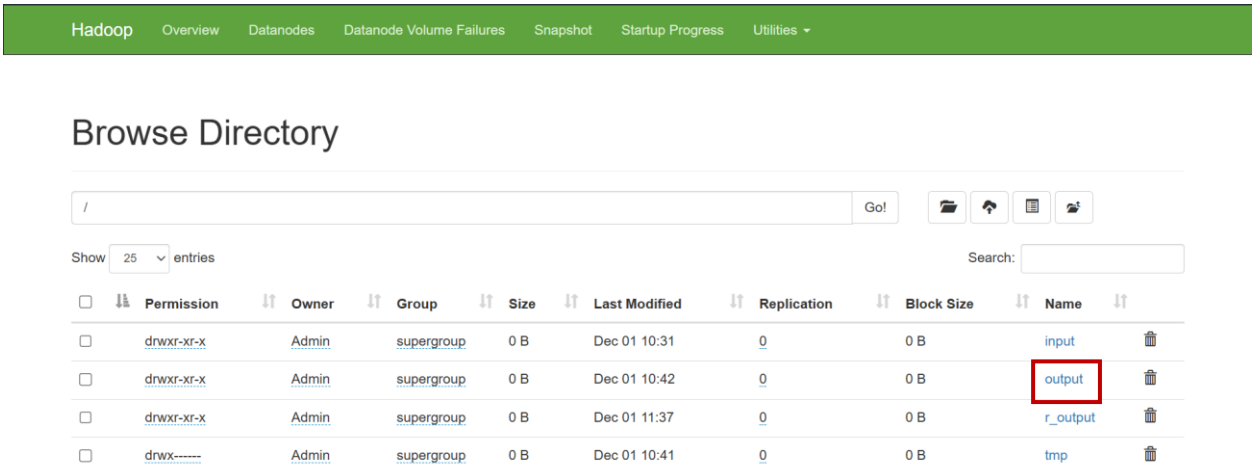
Bước 3: Chạy chương trình MapReduce và xem kết quả

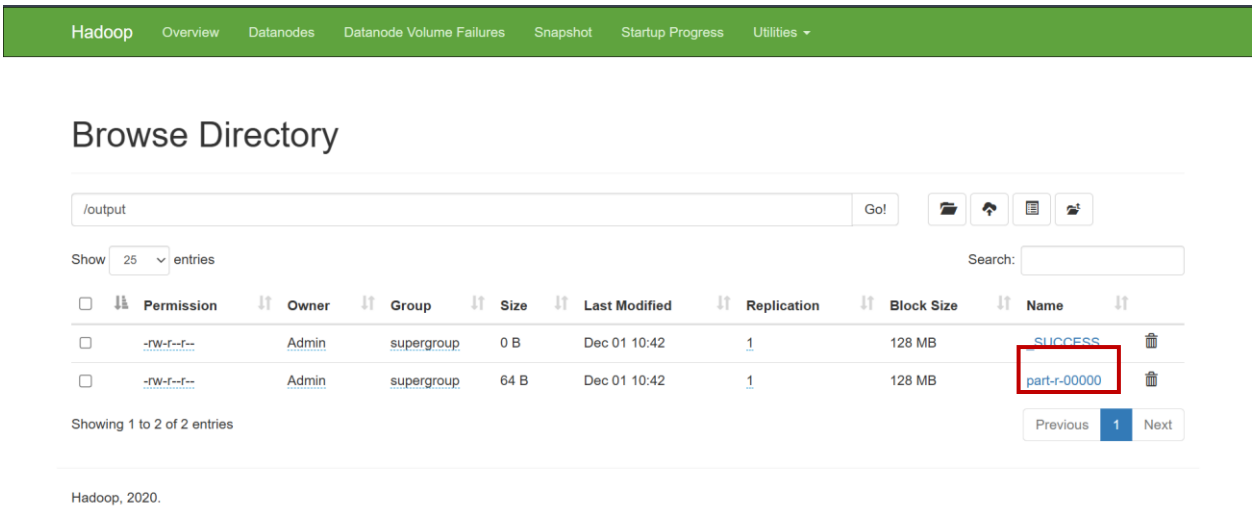
Chương trình mẫu MapReduce của Hadoop nằm tại **C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar**. Ta sẽ thử nghiệm đầu vào chương trình là file **data.txt** và kết quả sẽ được lưu tại folder **/output**, lệnh thực hiện”

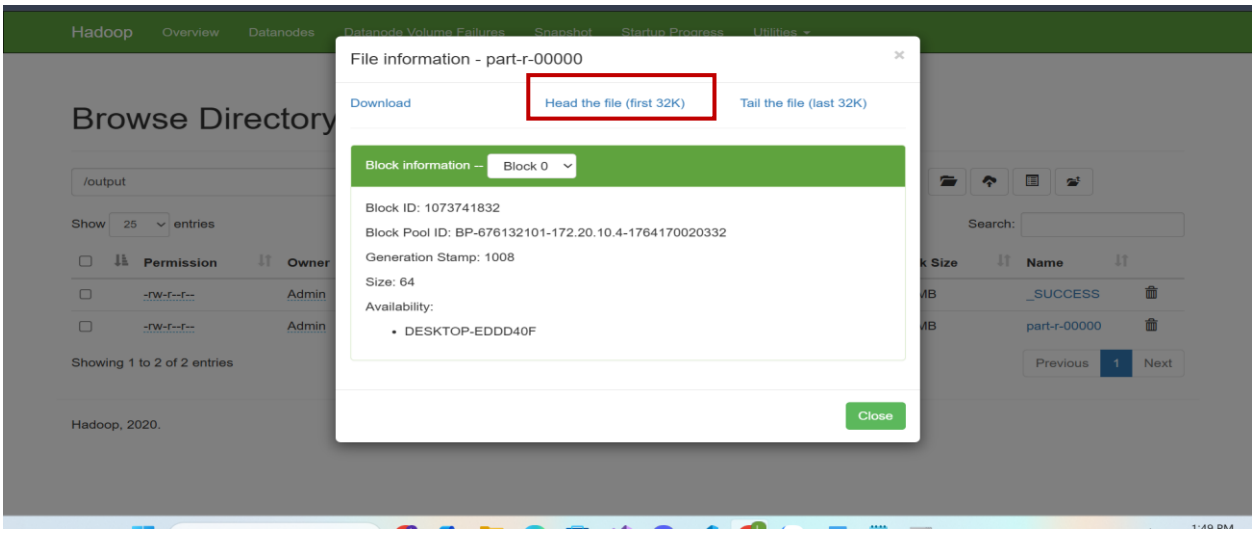
```
hadoop jar "C:\hadoop-3.3.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.0.jar" wordcount /input/data.txt /output
```

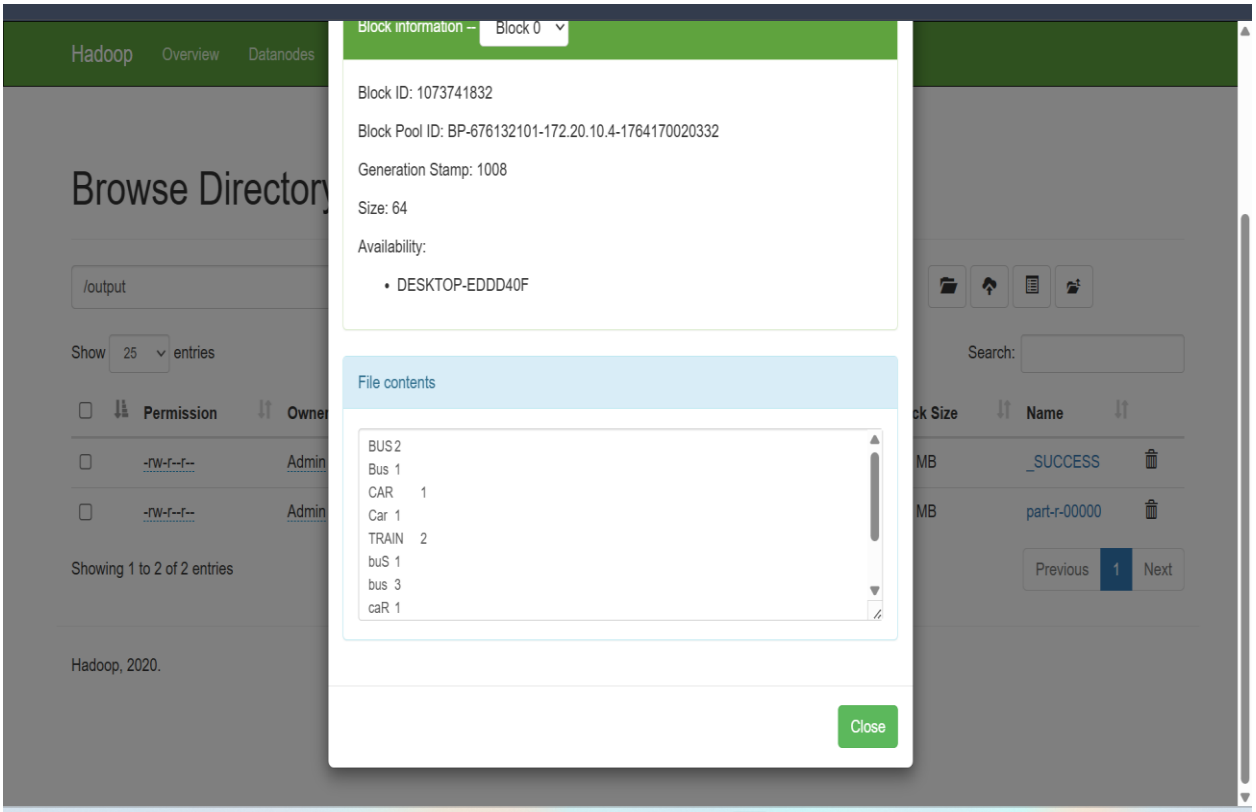


Xem kết quả thu được:



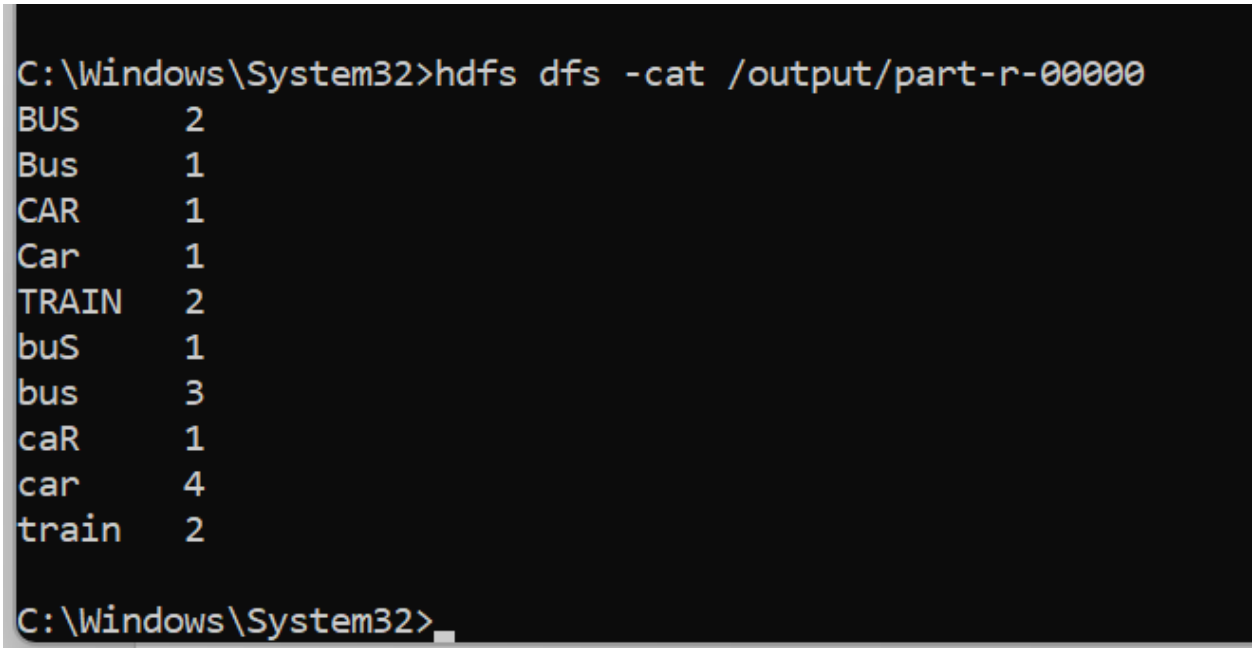



Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/output
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
Admin
_SUCCESS
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head 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


Hadoop
Overview
Datanodes
Browse Directory
/output
Show 25 entries
Search:
Permission
Name
-rw-r--r--
Admin
_SUCCESS
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
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
BUS 2
Bus 1
CAR 1
Car 1
TRAIN 2
buS 1
bus 3
caR 1
Close

Hoặc dùng lệnh cmd:

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



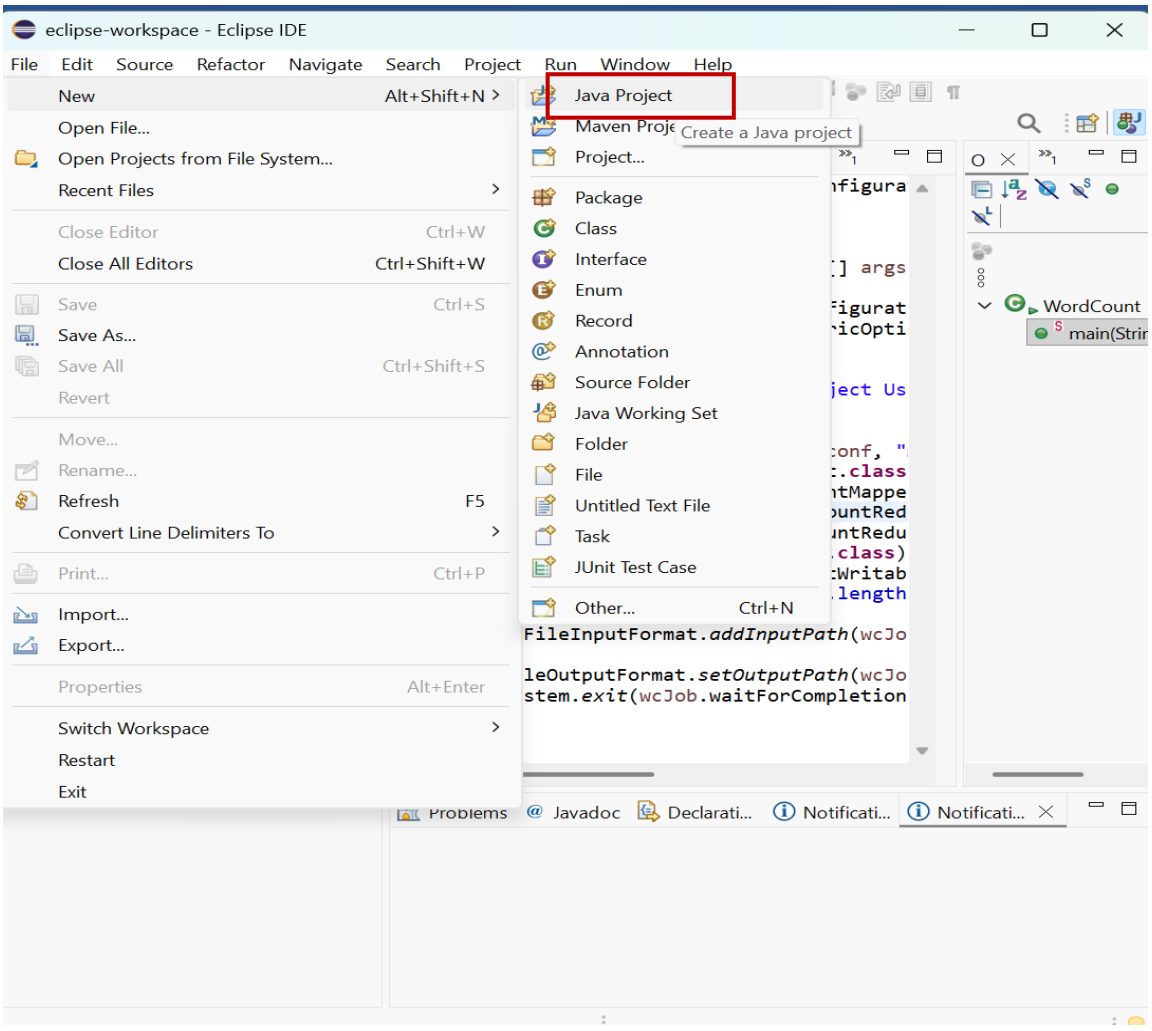
Như vậy ta đã chạy thành công chương trình mẫu MapReduce của Hadoop cung cấp.

II. Lập trình chương trình WordCount bằng Eclipse

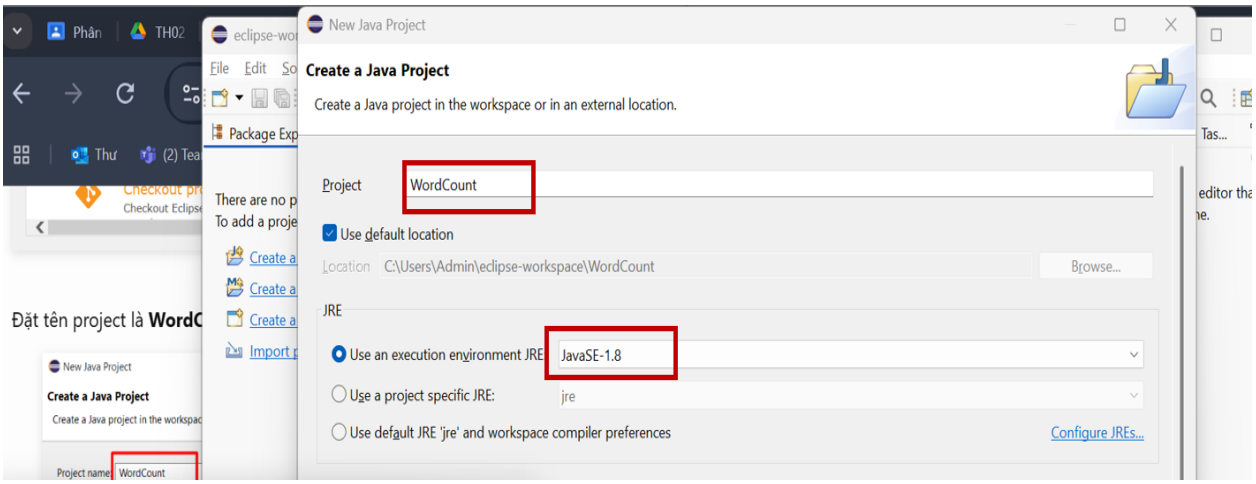
Bước 1: Tạo project

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

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

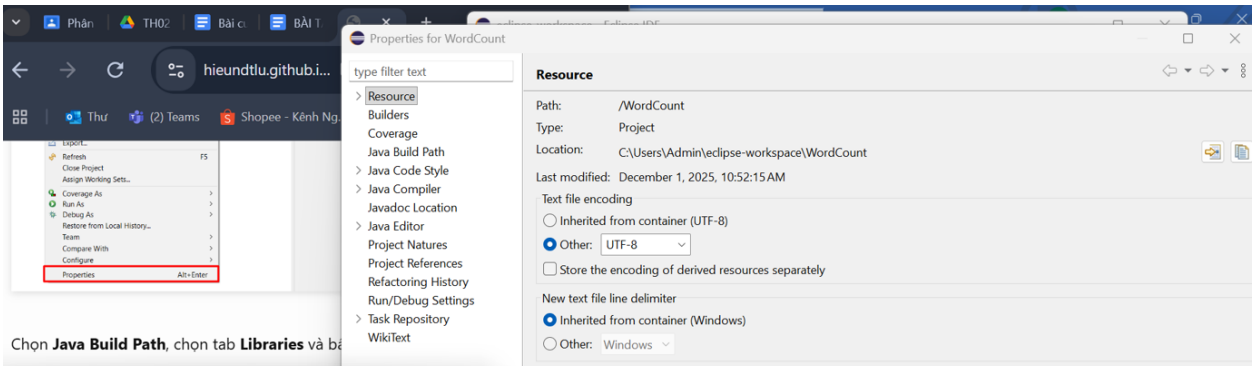


Đặt tên project là **WordCount** và chọn môi trường là **JavaSE-1.8**. Xong ấn **Finish**.



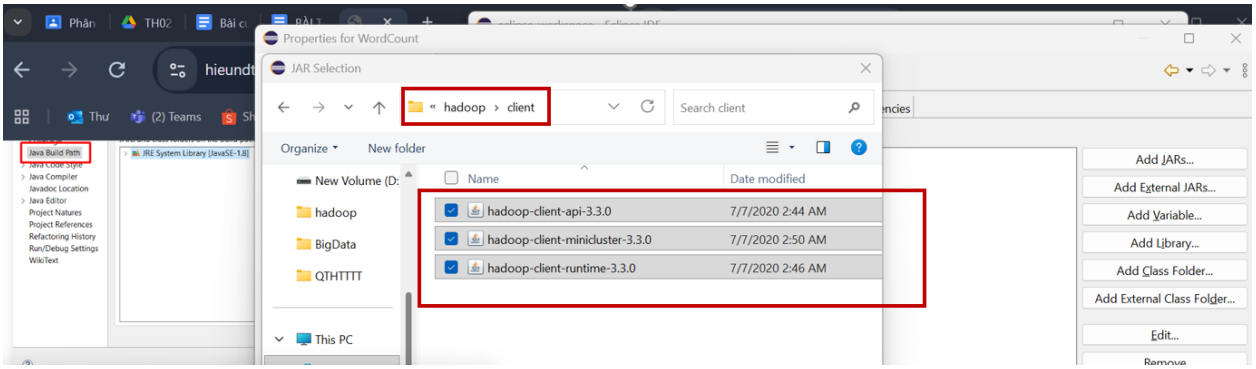
Bước 2: Thêm thư viện cần thiết để chạy MapReduce

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

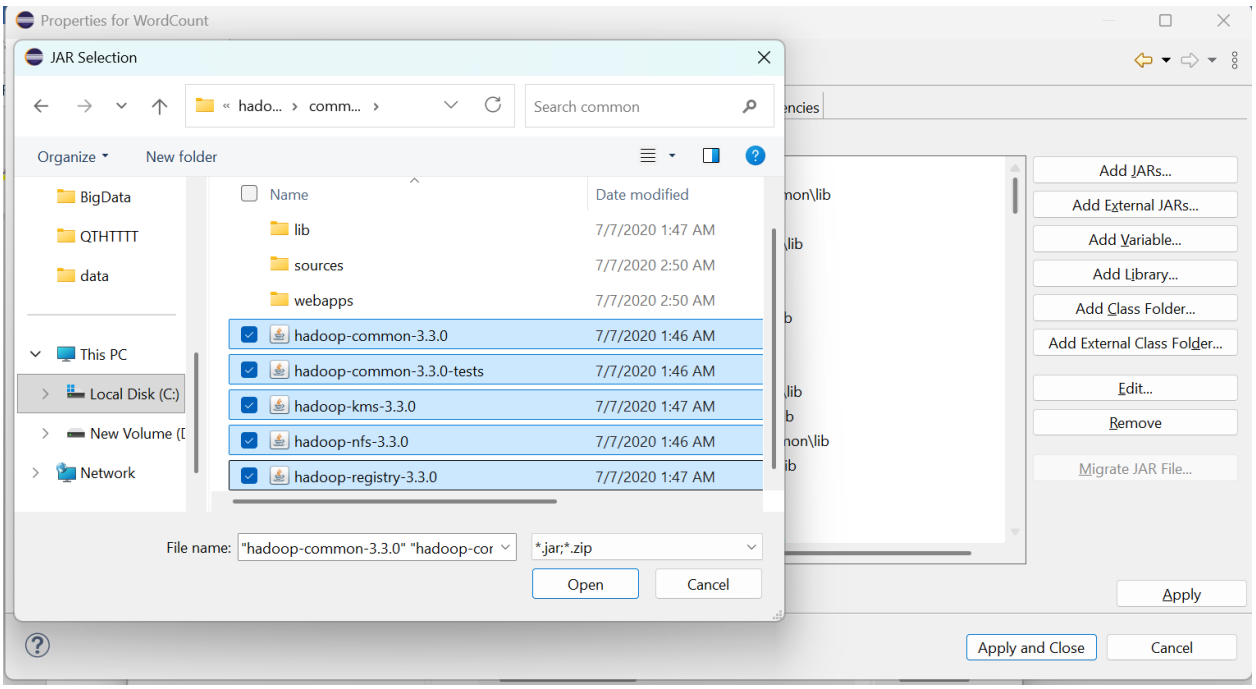


Chọn **Java Build Path**, chọn tab **Libraries** và bấm **Add External JARs**

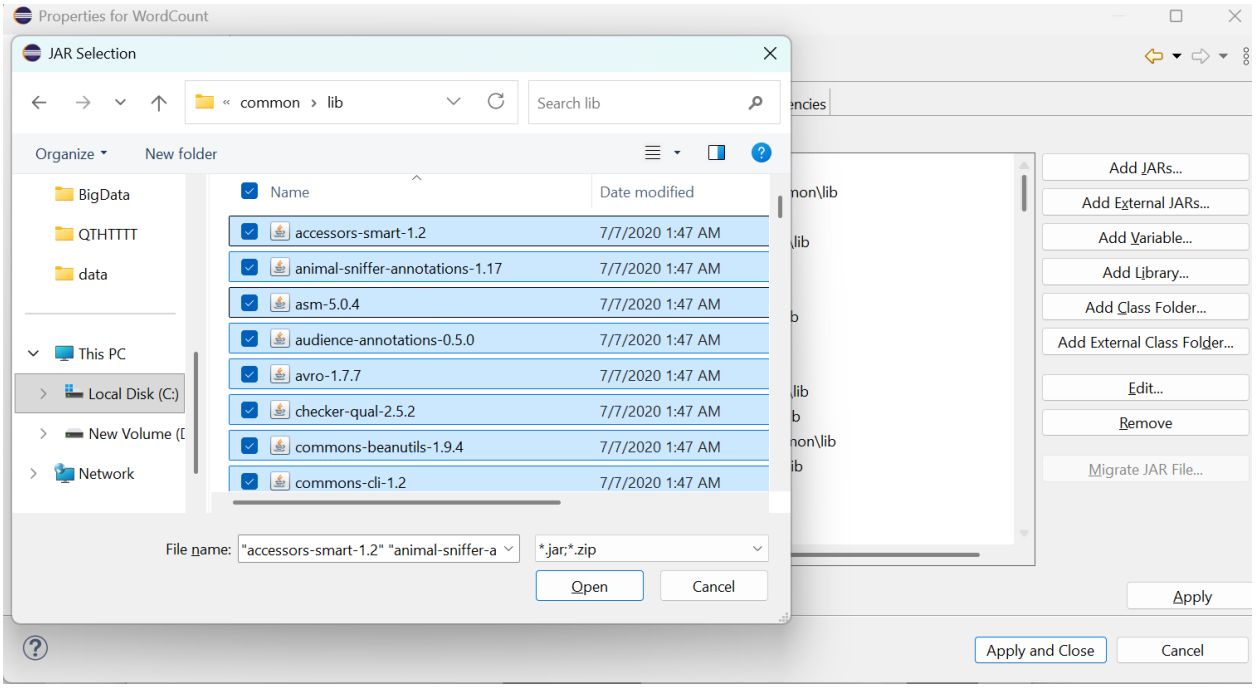
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\client** và ấn **Open**



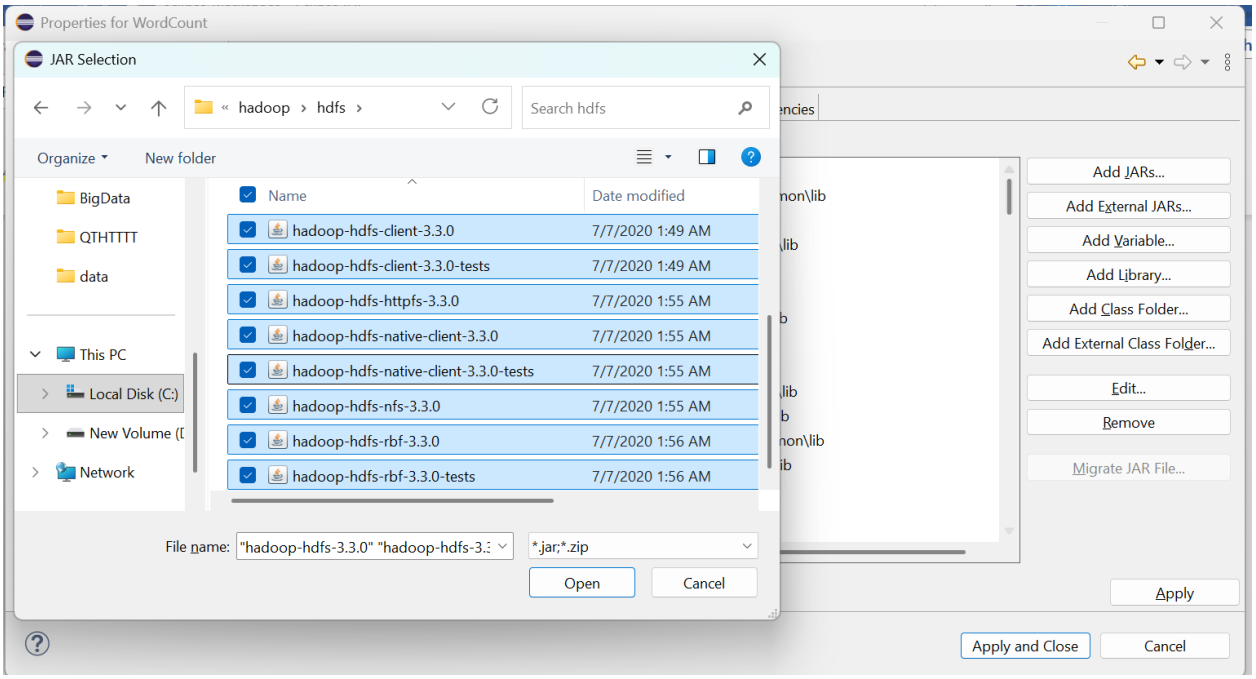
Tương tự chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common** và ấn **Open**



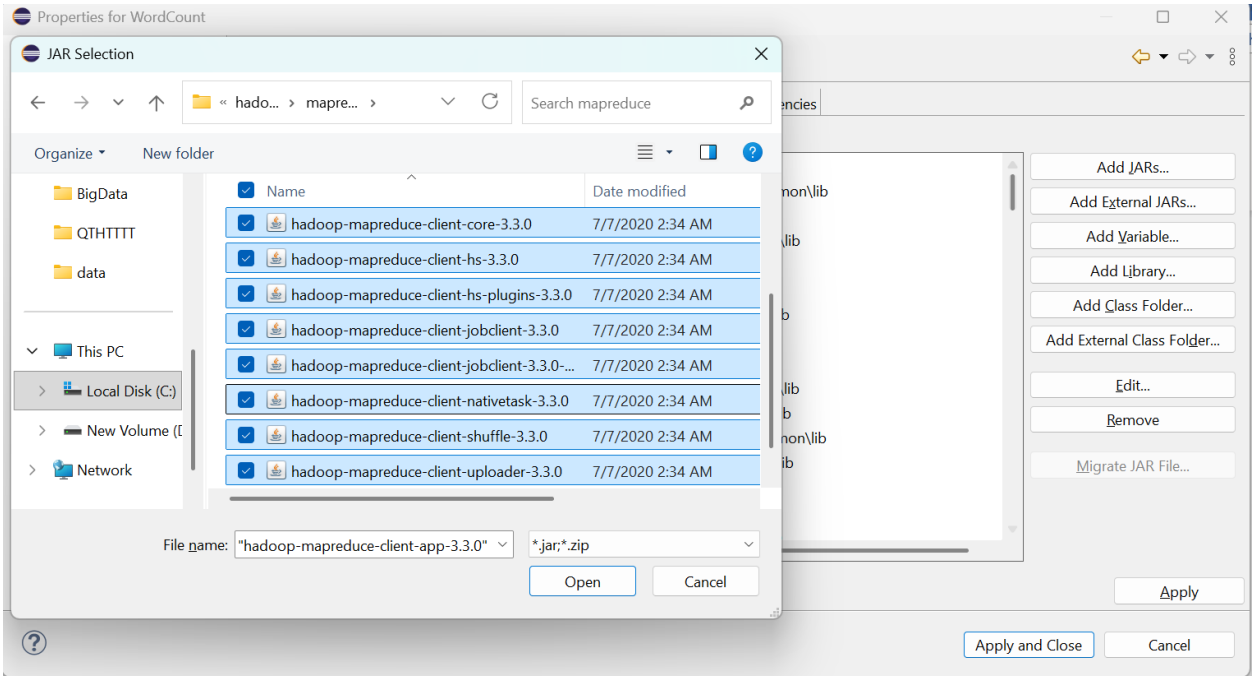
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\common\lib** và ấn **Open**



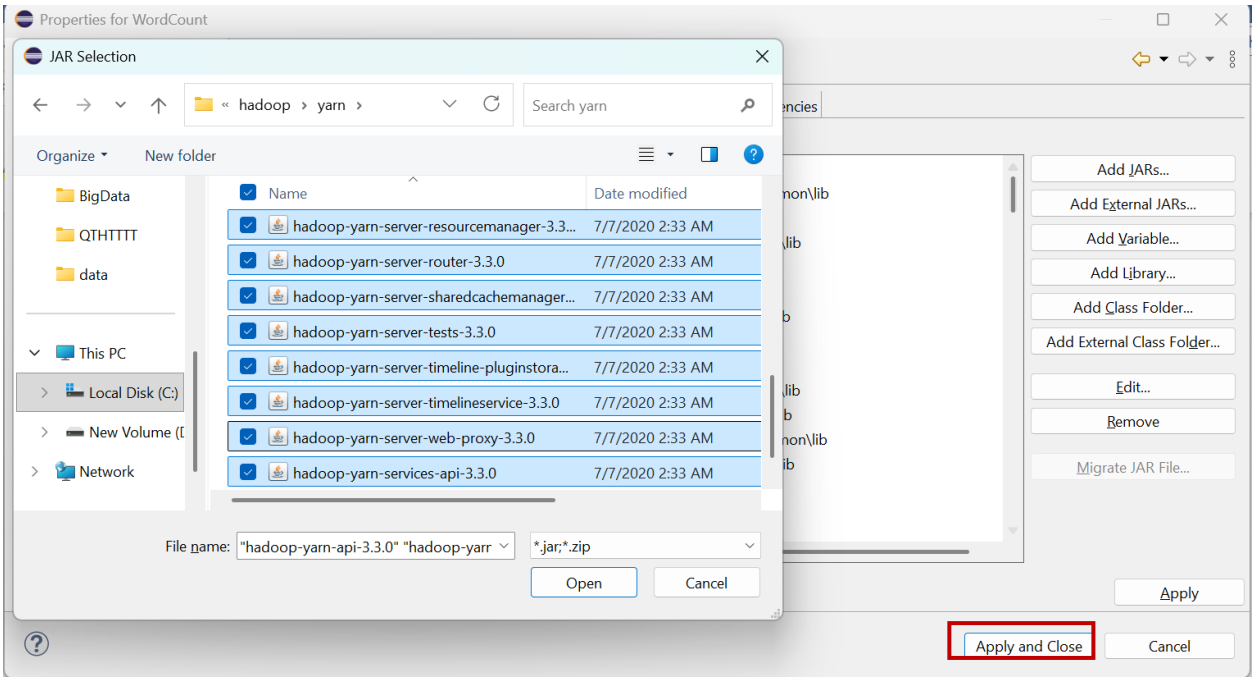
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\hdfs** và ấn **Open**



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



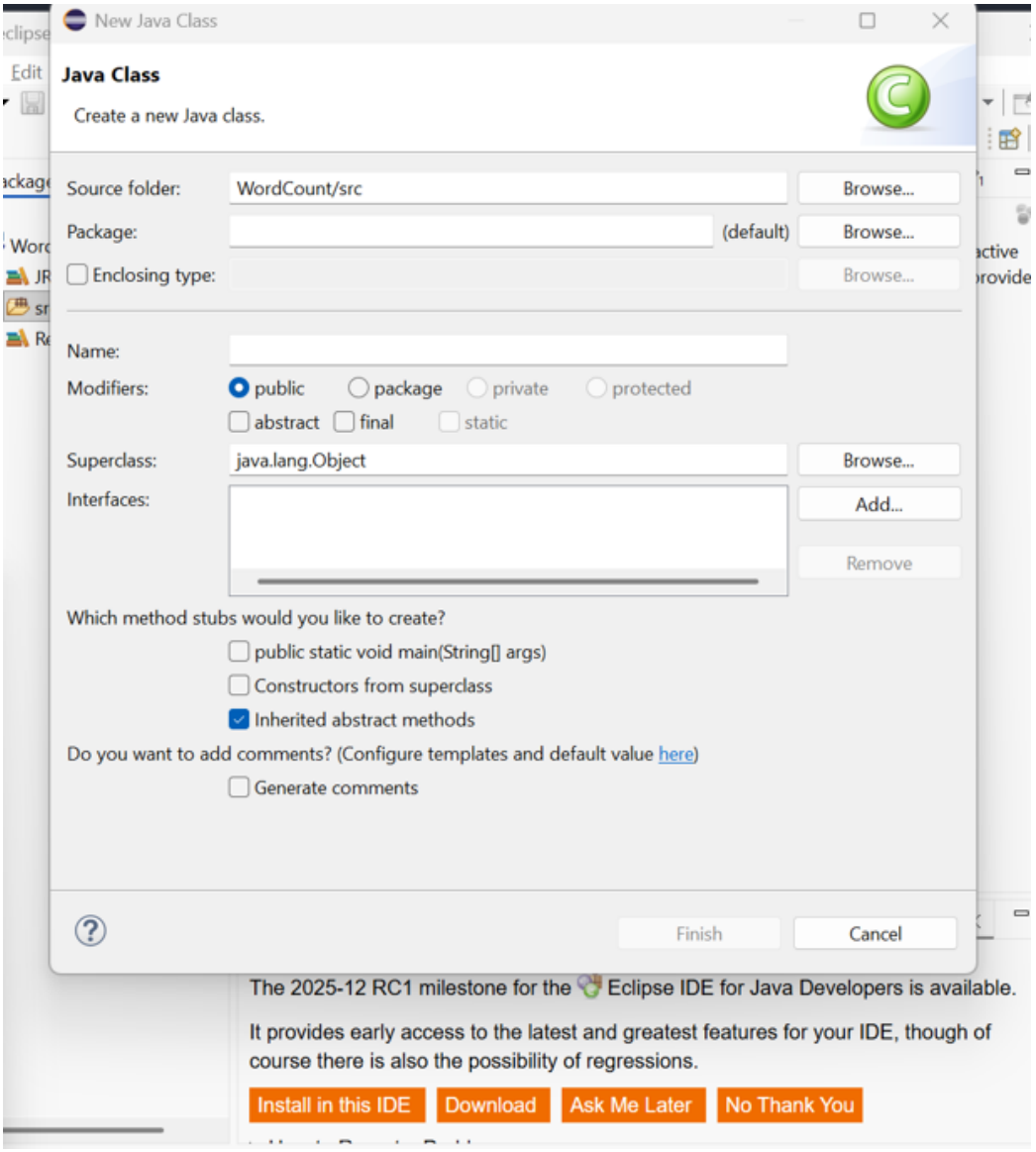
Chọn tất cả file trong thư mục **C:\hadoop-3.3.0\share\hadoop\yarn** và ấn **Open**



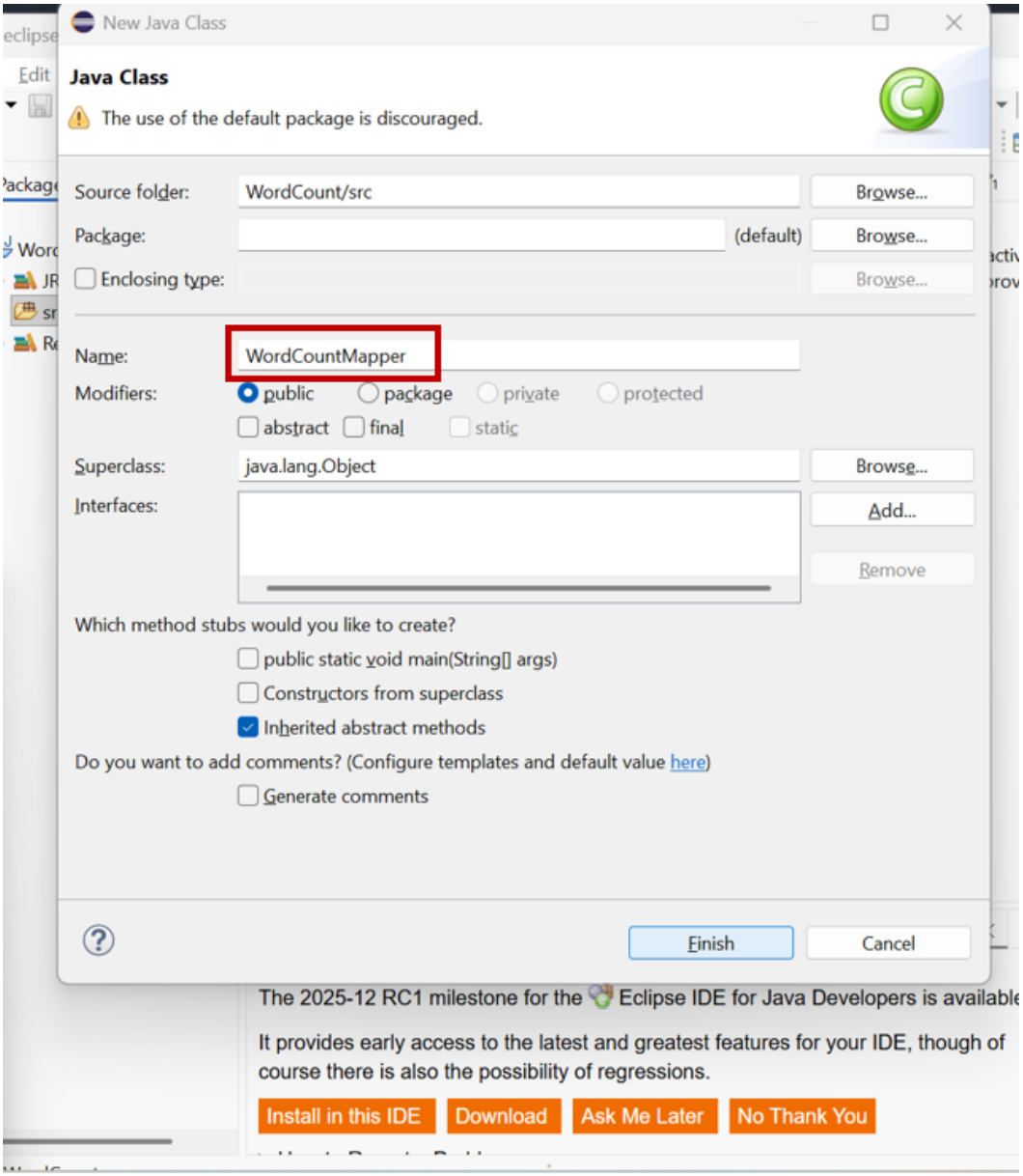
Ấn **Apply and Close**

Bước 3: Tạo class xử lý tác vụ MapReduce

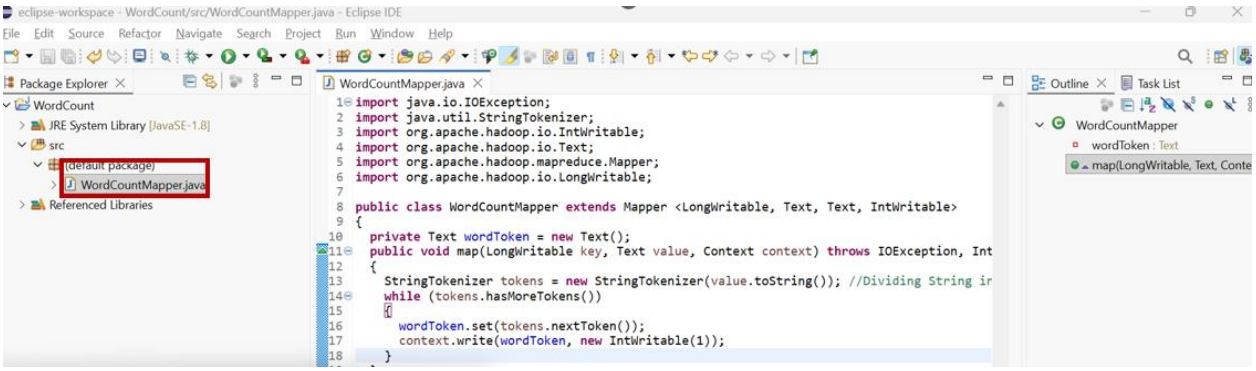
Double click vào project **WordCount**, chuột phải vào **src** và chọn **New > Class**



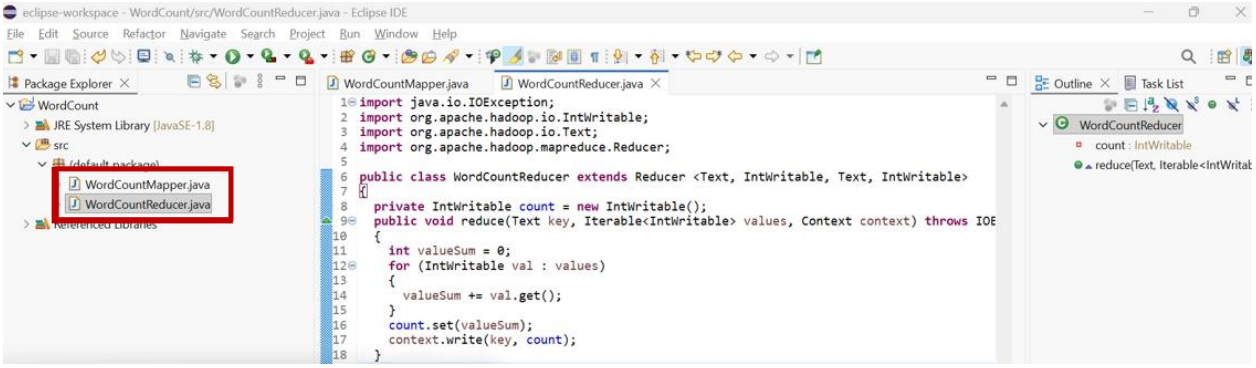
Tạo class để xử lý nhiệm vụ **Map**, đặt tên là **WordCountMapper**



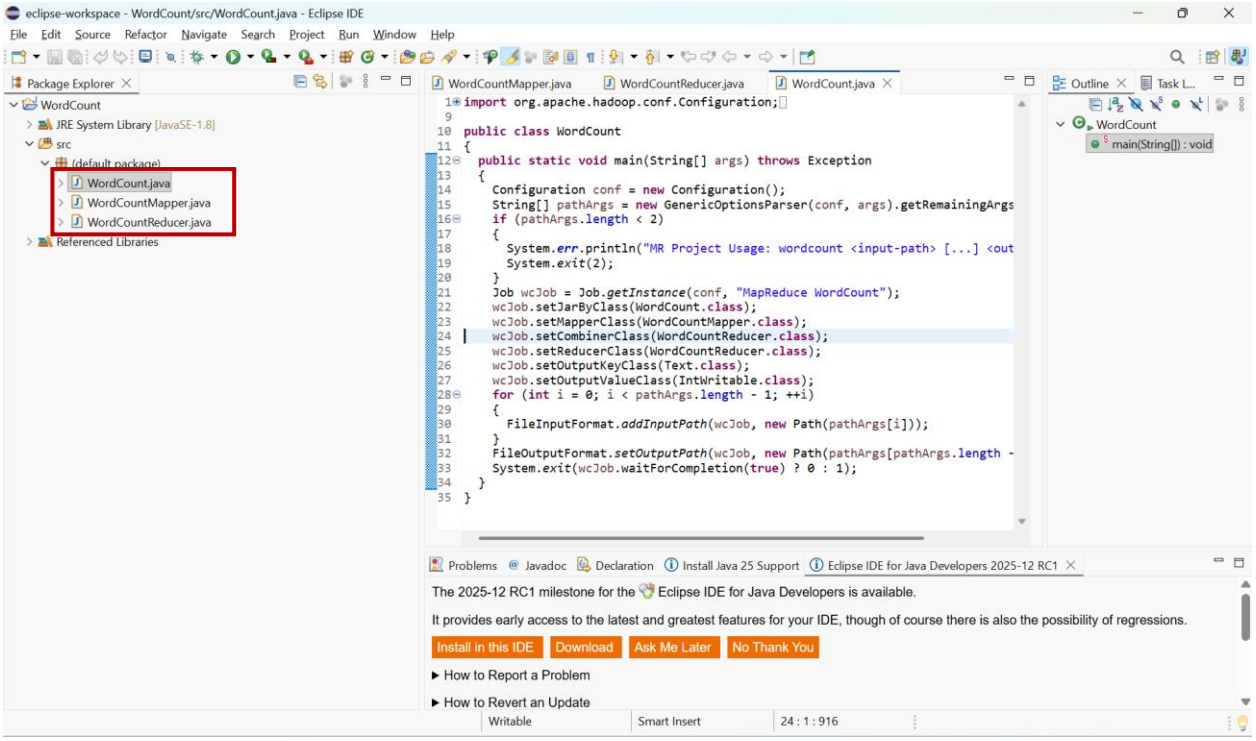
Nội dung bên trong file **WordCountMapper.java**:



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

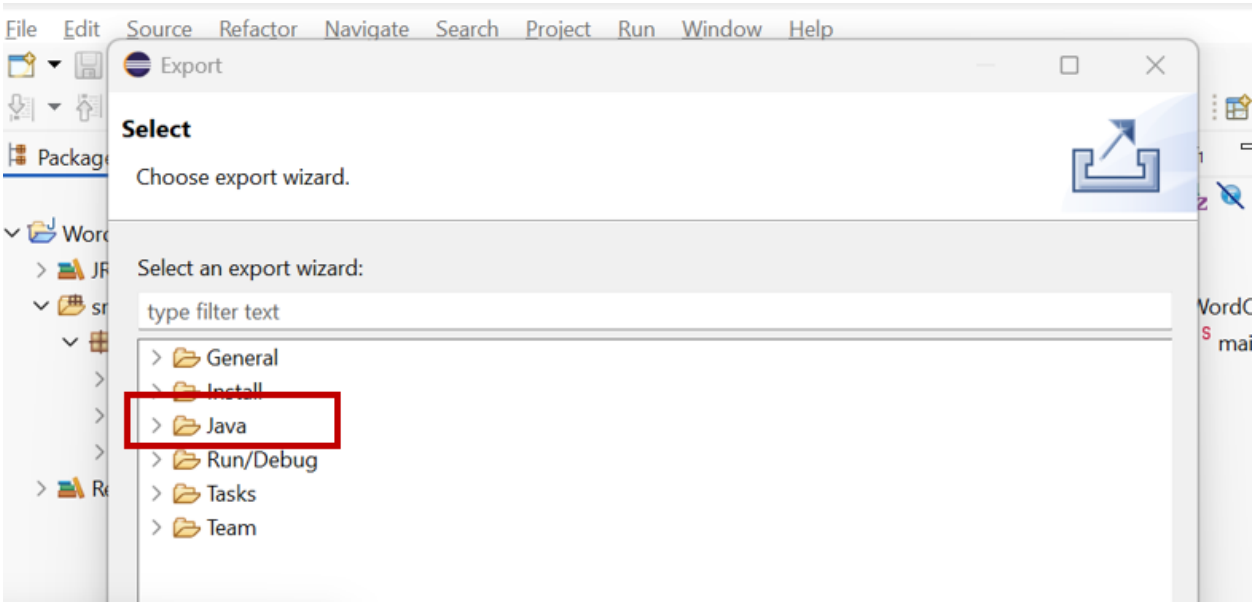


Và tạo class **WordCount** chứa hàm **main** để khởi chạy chương trình:

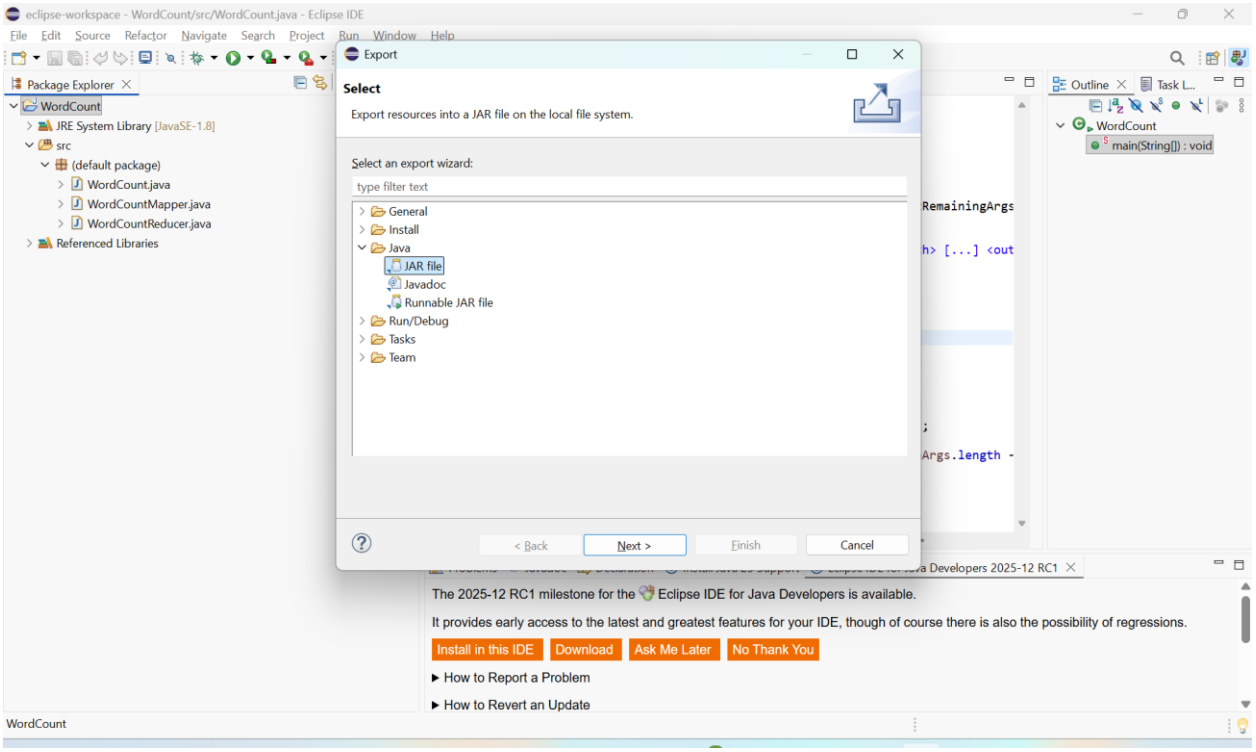


Bước 4: Tạo file JAR

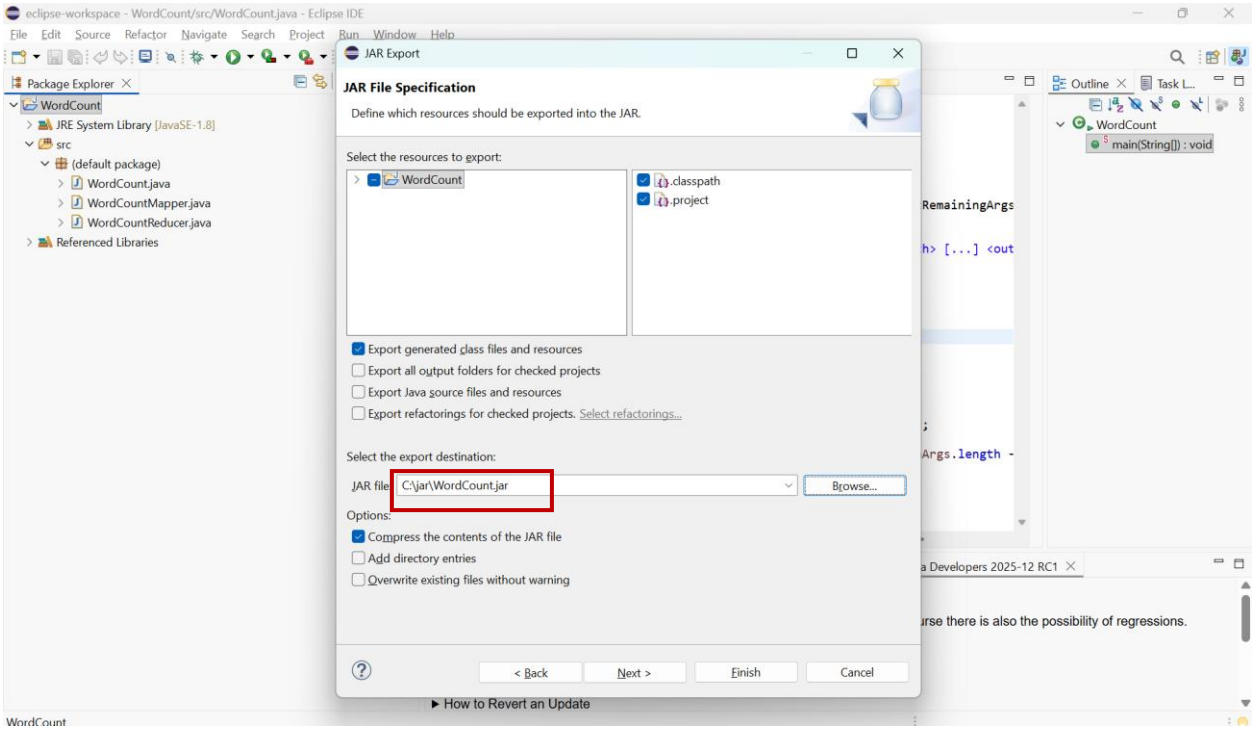
Chuột phải vào project **WordCount** chọn **Export**



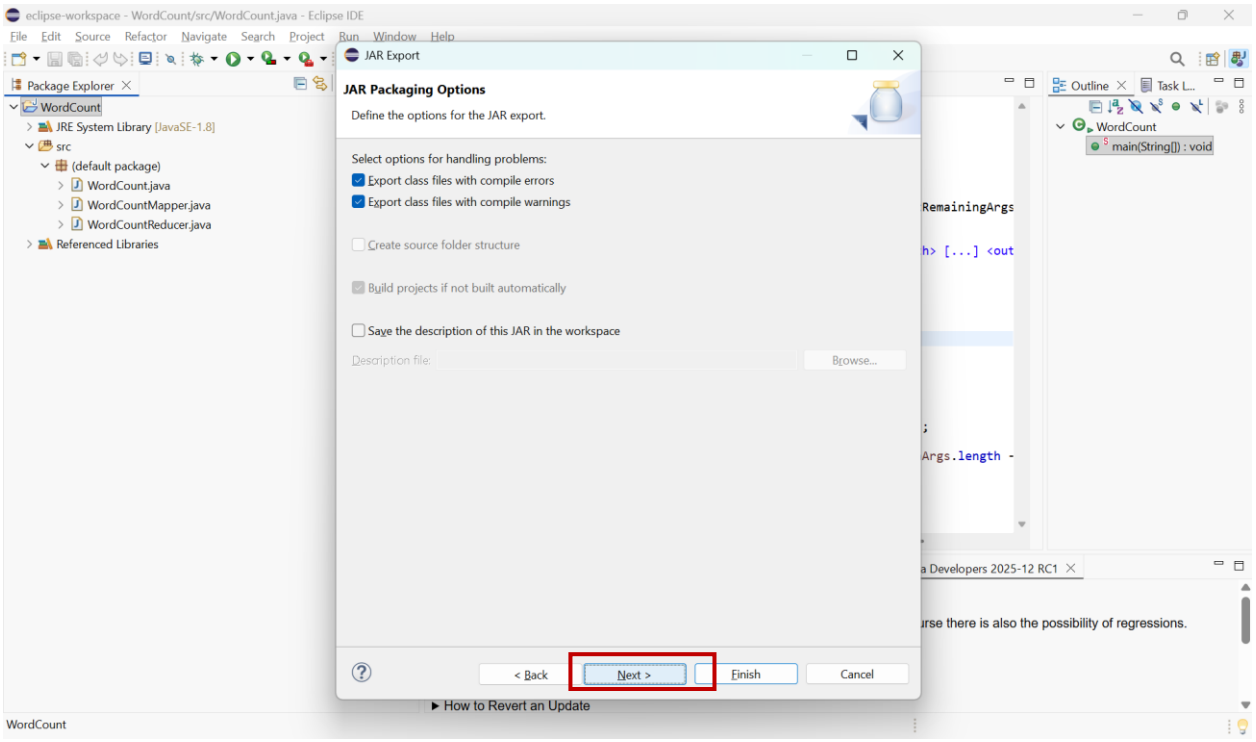
Chọn **Java > JAR File** rồi bấm **Next**



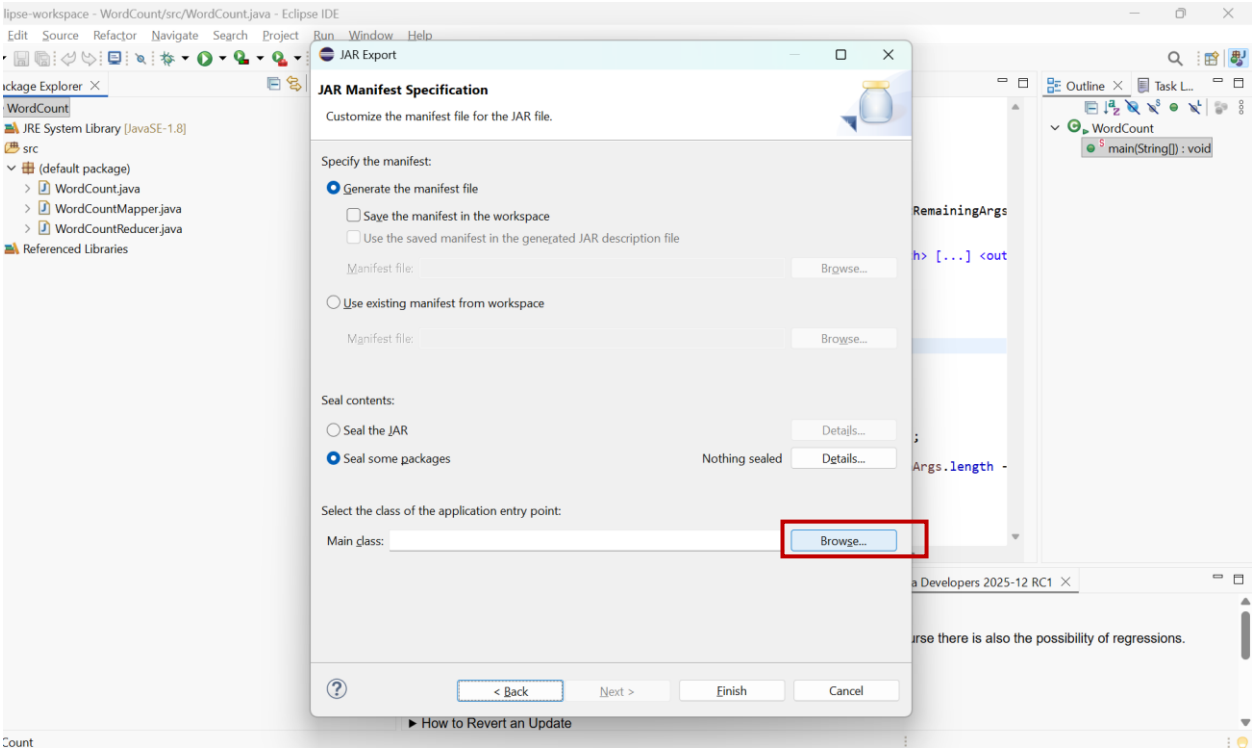
Chọn đường dẫn lưu file JAR và bấm **Next**



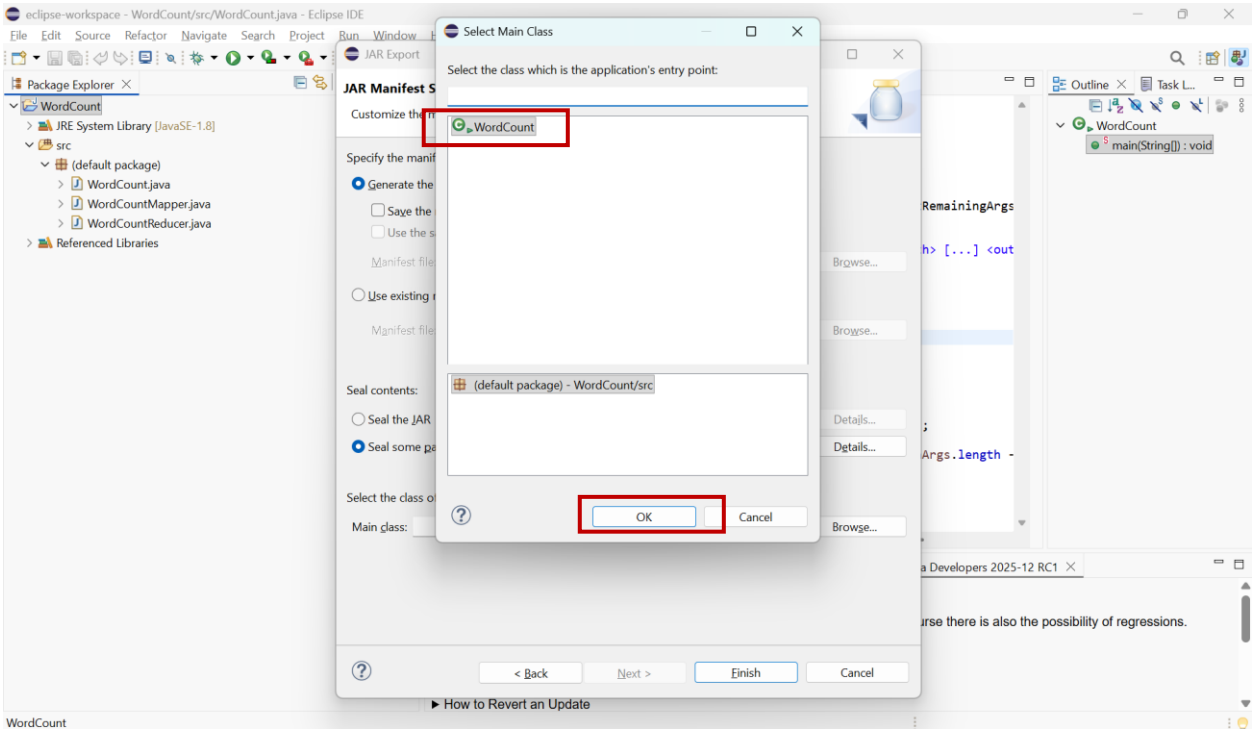
Bấm **Next**



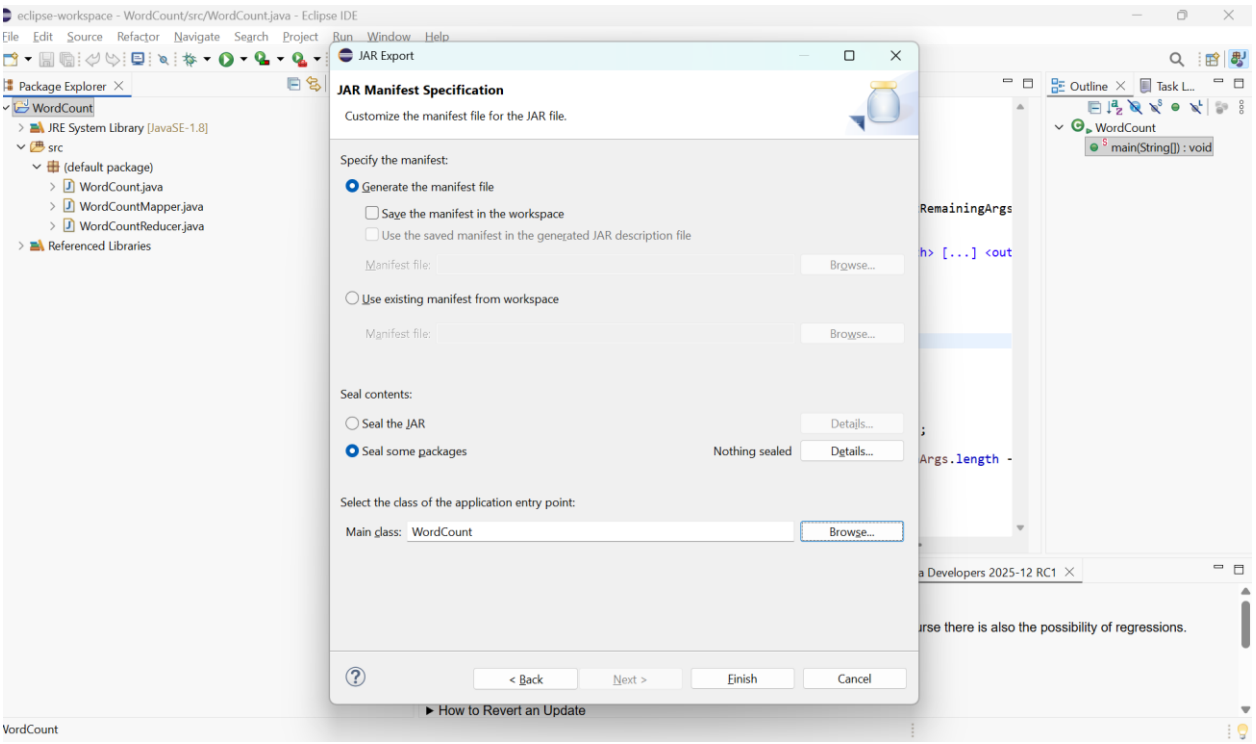
Bấm **Browser** để chọn file **main**



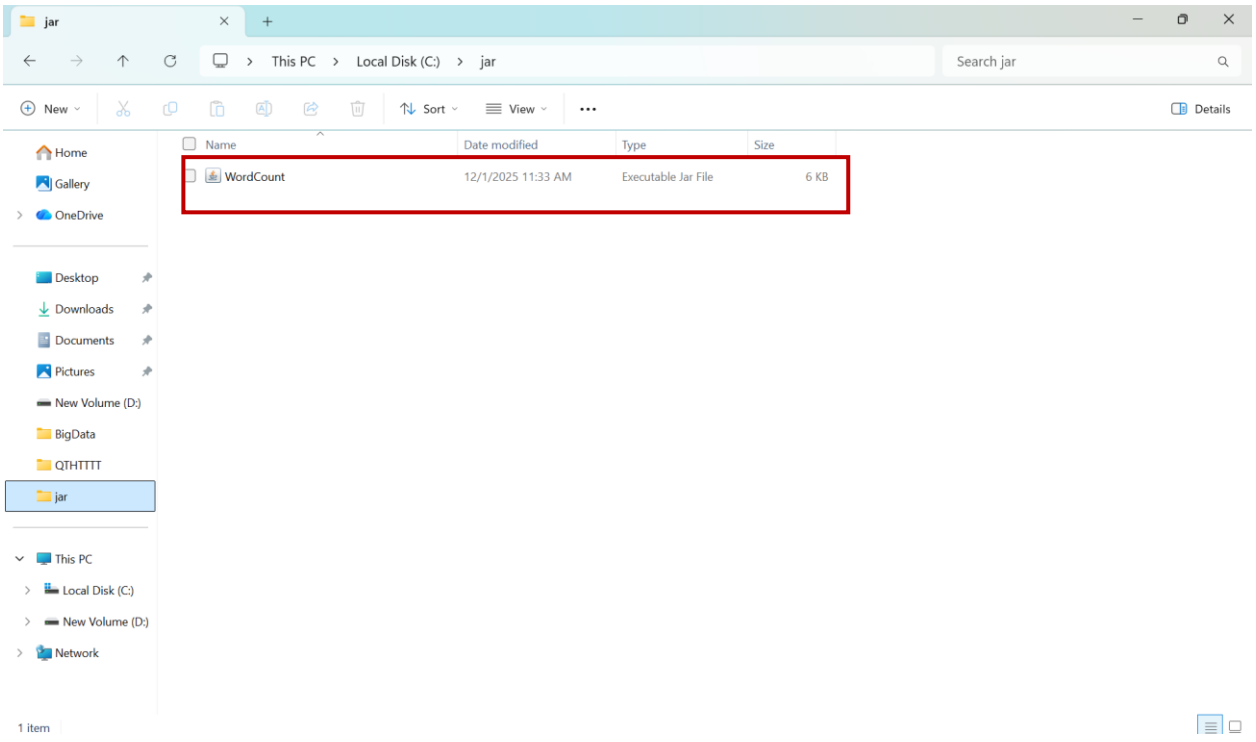
Chọn **WordCount** và bấm **OK**



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

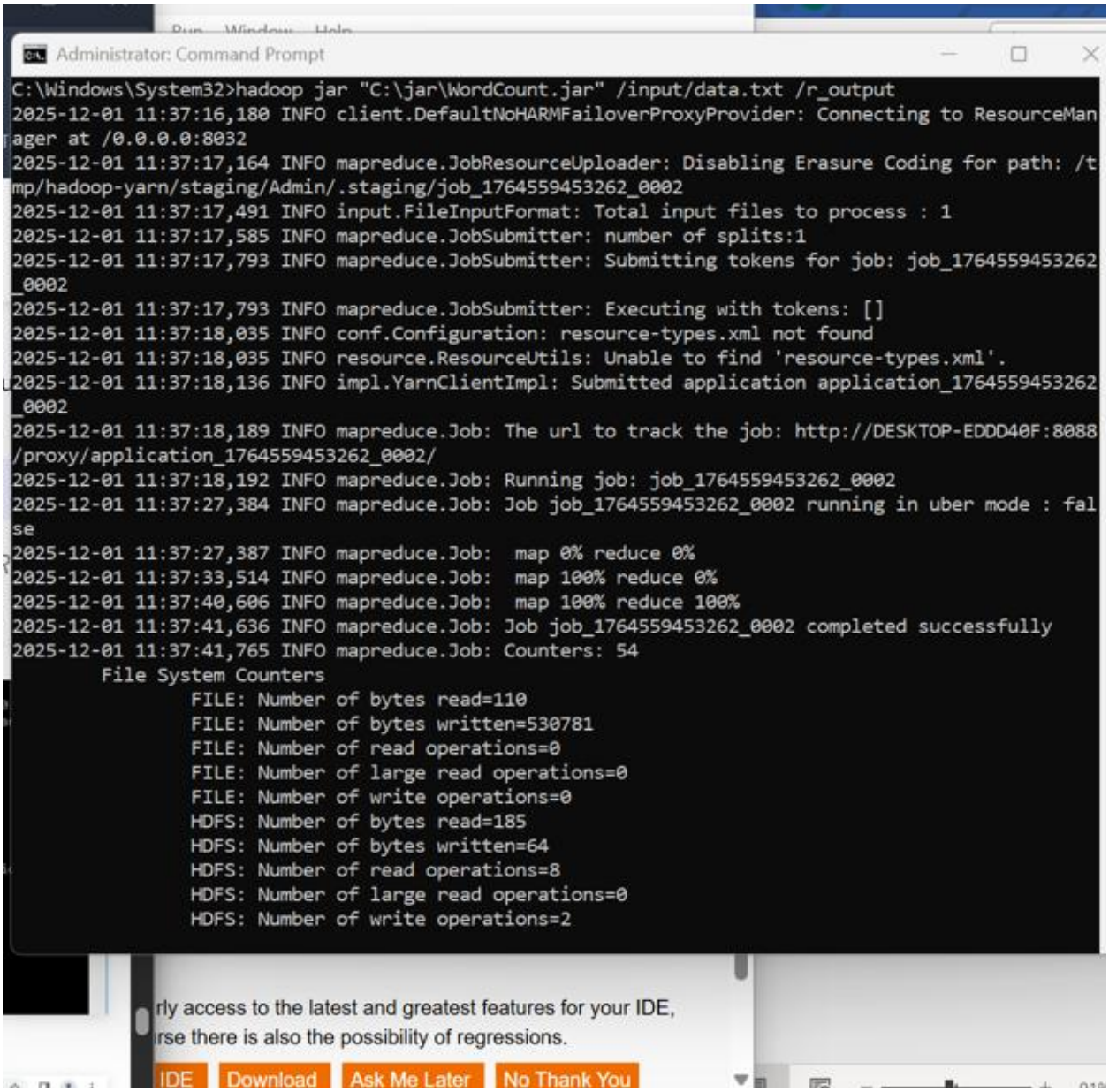


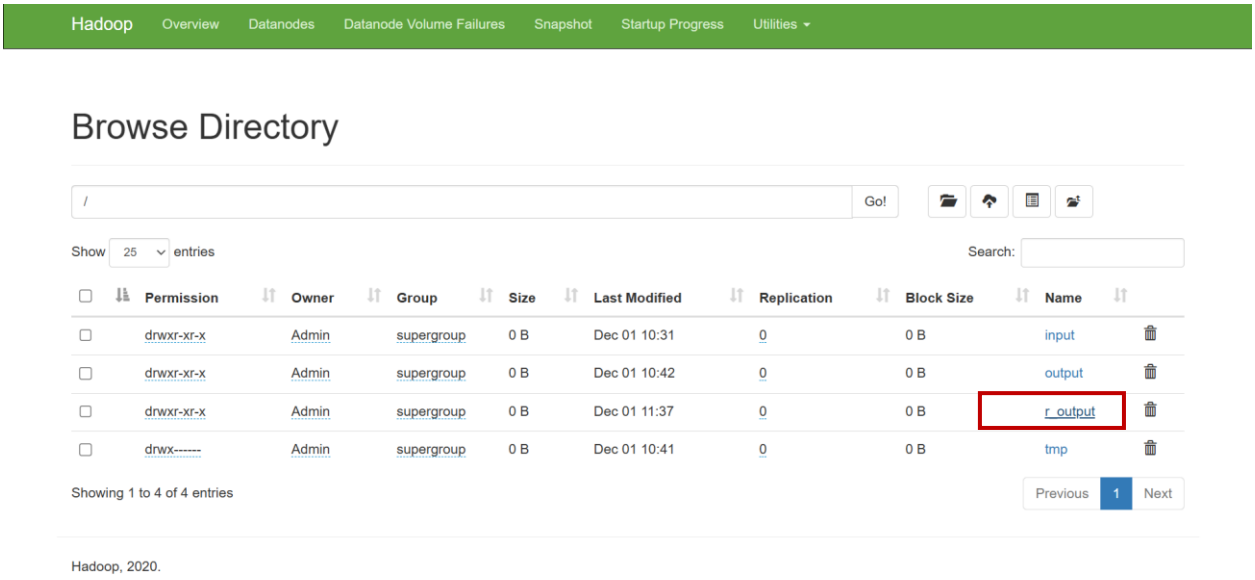
Vào thư mục chứa lưu file JAR vừa tạo và kiểm tra kết quả

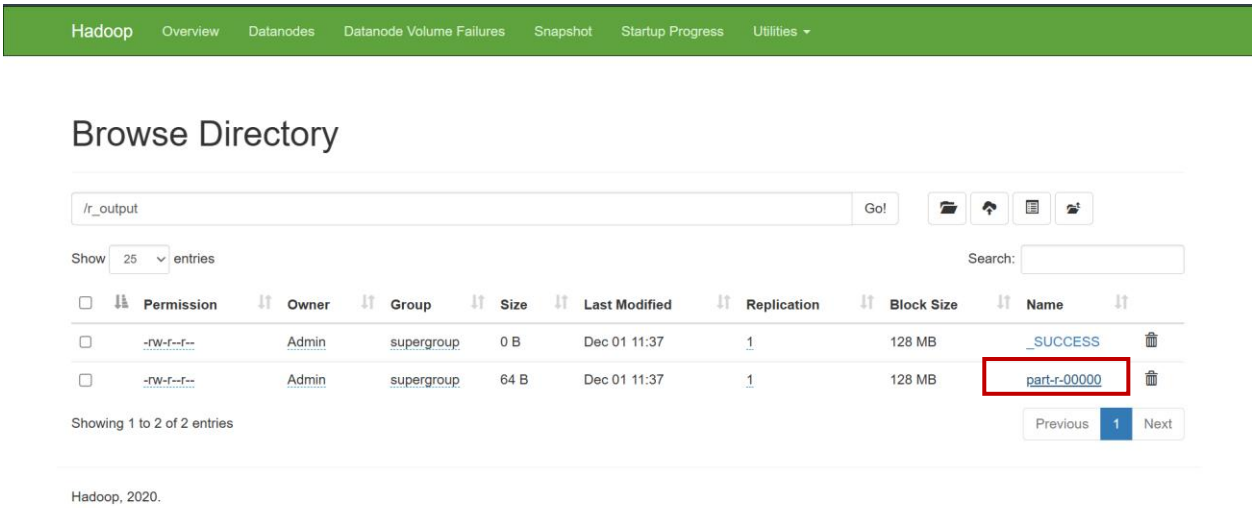


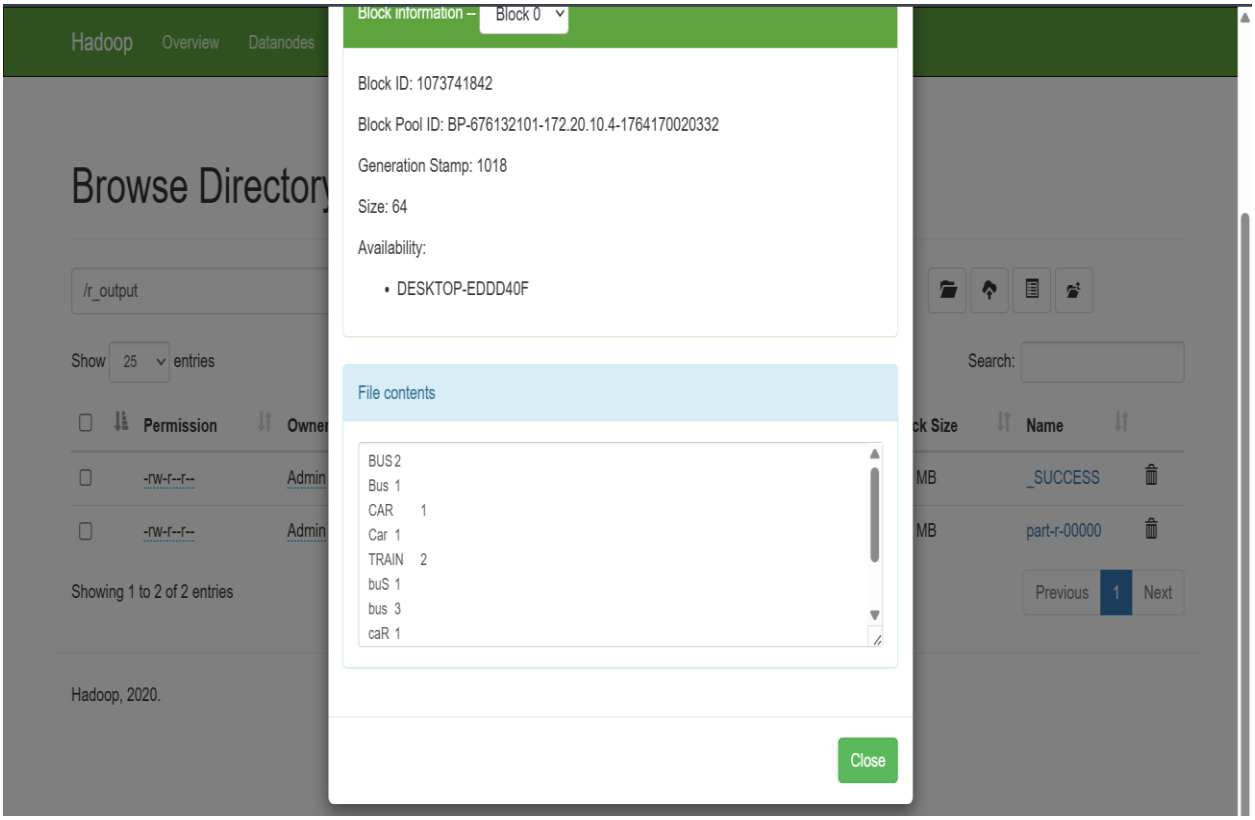
Thử nghiệm trên file dữ liệu **data.txt** đã tạo ở trên, và kết quả thu được lưu tại thư mục **r_output**. Chạy lệnh sau:

```
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
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!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Admin supergroup 0 B Dec 01 10:31 0 0 B input
drwxr-xr-x Admin supergroup 0 B Dec 01 10:42 0 0 B output
drwxr-xr-x Admin supergroup 0 B Dec 01 11:37 0 0 B r_output
drwx------ Admin supergroup 0 B Dec 01 10:41 0 0 B tmp
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!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r-- Admin supergroup 0 B Dec 01 11:37 1 128 MB _SUCCESS
-rw-r--r-- Admin supergroup 64 B Dec 01 11:37 1 128 MB part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.


Hadoop
Overview
Datanodes
Browse Directory
/r_output
Show 25 entries
Search:
Permission
Owner
Block Size
Name
-rw-r--r--
Admin
MB
_SUCCESS
-rw-r--r--
Admin
MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
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
BUS 2
Bus 1
CAR 1
Car 1
TRAIN 2
buS 1
bus 3
caR 1
Close