Ngô Hồng Thông

22649011

REPORT LAB 5

5.1 Chuẩn bị môi trường

Để chuẩn bị một môi trường cho bài lab 5

Tạo thư mục cho lab:

hadoop@root:~$ cd bigdata-labs/

hadoop@root:~/bigdata-labs$ cd lab5/

hadoop@root:~/bigdata-labs/lab5$ cd src/

hadoop@root:~/bigdata-labs/lab5/src$

Ở những lab trước em đã setup:

Thư mục: $HADOOP_HOME/etc/hadoop/

core-site.xml

Cấu hình chung của Hadoop.

Thường chứa fs.defaultFS → ví dụ: hdfs://namenode:9000.

hdfs-site.xml

Cấu hình HDFS.

Các tham số quan trọng:

dfs.replication → số bản sao mặc định.

dfs.namenode.name.dir → nơi lưu metadata của NameNode.

dfs.datanode.data.dir → nơi lưu block của DataNode.

mapred-site.xml (thường copy từ mapred-site.xml.template)

Cấu hình MapReduce.

mapreduce.framework.name = yarn.

yarn-site.xml

Cấu hình YARN (ResourceManager, NodeManager).

Ví dụ: yarn.resourcemanager.hostname.

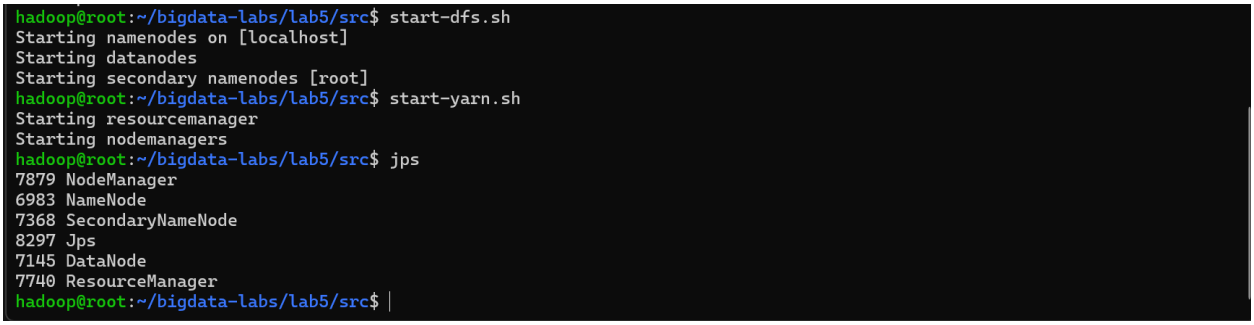
START lại HDFS và YARN

Đây là trường hợp em đã mở lại sau nên là HDFS và YARN đã đóng vì thế cần phải khởi động lại. Nếu mà chạy **JPS** mà ra đầy đủ các node thì không cần phải chạy lại

Lệnh :

start-dfs.sh: bật HDFS (lưu trữ dữ liệu)

start-yarn.sh: bật YARN (quản lý tài nguyên & job)



Sau khi đã bật lại HDFS và YARN thì ta bắt đầu copy lại dữ liệu để làm các bước tiếp theo

Copy dataset:

Việc lỗi safe mode là khi mới chạy lại Name node ở trong chế độ safe mode nên cần phải đợi một chút

```
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -mkdir -p /user/hadoop/lab5/input
mkdir: Cannot create directory /user/hadoop/lab5/input. Name node is in safe mode.
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfsadmin -safemode get
Safe mode is OFF
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -mkdir -p /user/hadoop/lab5/input
```

Việc dùng lệnh hdfs dfs để tạo thư mục trong DFS

Còn để copy ta dùng lệnh -put trong hdfs dfs thì để copy file dữ liệu vào hdfs

Kết quả:

```
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -ls /lab5/input
Found 1 items
-rw-r--r--   1 hadoop supergroup     144453 2025-09-24 19:23 /lab5/input/weblogs.txt
hadoop@root:~/bigdata-labs/lab5/src$
```

```
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:16 lab5
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -ls /lab5
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:23 /lab5/input
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:50 /lab5/input_students
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:23 /lab5/output
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:54 /lab5/output-students
hadoop@root:~/bigdata-labs/lab5/src$
```

5.2 Viết chương trình WordCount (Java)

Tạo file WordCount.java:

```
hadoop@root:~/bigdata-labs/lab5/src$ nano WordCount.java
hadoop@root:~/bigdata-labs/lab5/src$ cat WordCount.java
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();

        @Override
```

Luồng xử lý của hàm wordcount

1. Input: Một hoặc nhiều file văn bản trên HDFS.
    a. Ví dụ: weblogs.txt hoặc input.txt.
2. Mapper:
    a. Đọc từng dòng văn bản.
    b. Tách thành các từ.
    c. Phát ra cặp (từ, 1) cho mỗi từ.
3. Shuffle & Sort (Hadoop runtime):
    a. Tự động gom nhóm tất cả các cặp (từ, 1) có cùng key (từ).
    b. Đưa về cho Reducer.
4. Reducer:
    a. Nhận (từ, [1,1,1,...]).
    b. Cộng dồn lại → (từ, số_lần_xuất_hiện).
5. Output: File kết quả trên HDFS chứa danh sách từ và số lần xuất hiện.

Ví dụ cho hàm wordcount:

Input file:

Hadoop is fast

Hadoop is scalable

Mapper output:

(Hadoop,1) (is,1) (fast,1) (Hadoop,1) (is,1) (scalable,1)

Reducer output:

(Hadoop,2) (is,2) (fast,1) (scalable,1)

⇨ Kết quả đạt được sẽ đếm được tổng số lần xuất hiện chung của những từ ngữ có trong file input đầu vào

5.3 Compile và đóng gói

Sau khi đã có được file wordcount bằng java thì phải compile và đóng gói lại thì mới chạy được hàm wordcount trong hdfs

Tạo thư mục build:

mkdir -p ~/bigdata-labs/lab5/build

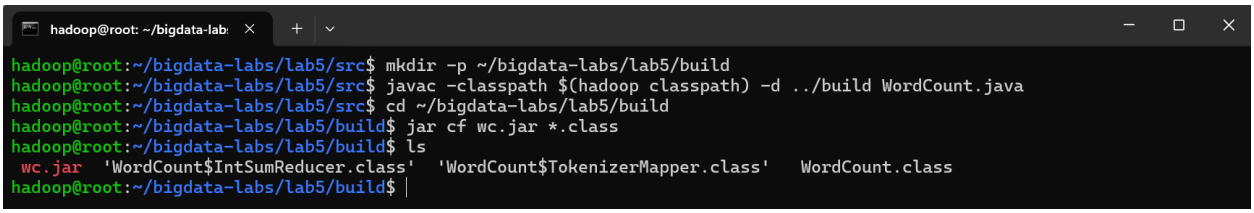
Compile:

cd ~/bigdata-labs/lab5/src

javac -classpath $(hadoop classpath) -d ../build WordCount.java Đóng gói JAR:

cd ~/bigdata-labs/lab5/build

jar cf wc.jar *.class



5.4 Chạy MapReduce Job

Sau khi đã có hàm wordcount thì nếu chưa put file dữ liệu vào input thì ta sẽ put dữ liệu là web.log vào input để làm data cho chạy mapreduce.

Bước tiếp theo nếu đã có dữ liệu thì chạy mapreduce trên file đó

Lệnh:

Hdfs dfs -put ~/bigdata-labs/lab5/src/web.logs.txt /lab5/input

Hadoop jar ~/bigdata-labs/lab5/build/wc.jar Wourdcount / lab5/input / lab5/output

Giải thích lệnh:

Chạy chương trình MapReduce Wordcount được đóng gói trong wc.jar, lấy dữ liệu đầu vào từ thư mục /lab5/input trên HDFS, xử lý và lưu kết quả vào /lab5/output trên HDFS.

Kết quả:

```
hadoop@root:~/bigdata-labs/lab5$ hdfs dfs -put ~/bigdata-labs/lab5/src/weblogs.txt /lab5/input/
hadoop@root:~/bigdata-labs/lab5$ hadoop jar ~/bigdata-labs/lab5/build/wc.jar WordCount /lab5/input /lab5/output
2025-09-24 19:23:35,544 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-24 19:23:36,205 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
 with ToolRunner to remedy this.
2025-09-24 19:23:36,215 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1758715548556_0002
2025-09-24 19:23:36,365 INFO input.FileInputFormat: Total input files to process : 1
2025-09-24 19:23:37,203 INFO mapreduce.JobSubmitter: number of splits:1
2025-09-24 19:23:37,672 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758715548556_0002
2025-09-24 19:23:37,672 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-24 19:23:37,771 INFO conf.Configuration: resource-types.xml not found
2025-09-24 19:23:37,771 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-24 19:23:37,837 INFO impl.YarnClientImpl: Submitted application application_1758715548556_0002
2025-09-24 19:23:37,874 INFO mapreduce.Job: The url to track the job: http://192.168.204.135:8088/proxy/application_1758715548556_0002/
2025-09-24 19:23:37,874 INFO mapreduce.Job: Running job: job_1758715548556_0002
2025-09-24 19:23:42,937 INFO mapreduce.Job: Job job_1758715548556_0002 running in uber mode : false
2025-09-24 19:23:42,938 INFO mapreduce.Job:  map 0% reduce 0%
2025-09-24 19:23:46,984 INFO mapreduce.Job:  map 100% reduce 0%
2025-09-24 19:23:50,001 INFO mapreduce.Job:  map 100% reduce 100%
2025-09-24 19:23:51,013 INFO mapreduce.Job: Job job_1758715548556_0002 completed successfully
2025-09-24 19:23:51,064 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=1275
                FILE: Number of bytes written=619965
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=144562
                HDFS: Number of bytes written=1121
                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
        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)=1671
                Total time spent by all reduces in occupied slots (ms)=1506
                Total time spent by all map tasks (ms)=1671
                Total time spent by all reduce tasks (ms)=1506
                Total vcore-milliseconds taken by all map tasks=1671
```

```
hadoop@root: ~/bigdata-lab
                Total vcore-milliseconds taken by all map tasks=1671
                Total vcore-milliseconds taken by all reduce tasks=1506
                Total megabyte-milliseconds taken by all map tasks=1711104
                Total megabyte-milliseconds taken by all reduce tasks=1542144
        Map-Reduce Framework
                Map input records=2000
                Map output records=18000
                Map output bytes=214453
                Map output materialized bytes=1275
                Input split bytes=109
                Combine input records=18000
                Combine output records=58
                Reduce input groups=58
                Reduce shuffle bytes=1275
                Reduce input records=58
                Reduce output records=58
                Spilled Records=116
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=186
                CPU time spent (ms)=1700
                Physical memory (bytes) snapshot=750153728
                Virtual memory (bytes) snapshot=5210574848
                Total committed heap usage (bytes)=722993152
                Peak Map Physical memory (bytes)=434483200
                Peak Map Virtual memory (bytes)=2600923136
                Peak Reduce Physical memory (bytes)=315670528
                Peak Reduce Virtual memory (bytes)=2609651712
        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=144453
        File Output Format Counters
                Bytes Written=1121
hadoop@root:~/bigdata-labs/lab5$
```

Kết quả output dữ liệu đầu ra:

```
hadoop@root:~/bigdata-labs/lab5$ hdfs dfs -ls /lab5
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:23 /lab5/input
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 19:23 /lab5/output
hadoop@root:~/bigdata-labs/lab5$
```

Kết quả đọc file dữ liệu đầu ra:

```
hadoop@root:~/bigdata-labs/lab5$ hdfs dfs -cat /lab5/output/part-r-00000
"GET    2000
+0000]  2000
-       4000
/       410
/cart   393
/checkout       398
/login  401
/products       398
192.168.1.1     101
192.168.1.10    110
192.168.1.11    104
192.168.1.12    121
192.168.1.13    82
192.168.1.14    104
192.168.1.15    91
192.168.1.16    98
192.168.1.17    105
192.168.1.18    106
192.168.1.19    101
192.168.1.2     84
192.168.1.20    111
192.168.1.3     103
192.168.1.4     115
192.168.1.5     94
192.168.1.6     93
192.168.1.7     94
192.168.1.8     84
192.168.1.9     99
200     2000
HTTP/1.1"       2000
[1/Jul/2024:10:23:45    61
[10/Jul/2024:10:23:45   75
[11/Jul/2024:10:23:45   65
[12/Jul/2024:10:23:45   71
[13/Jul/2024:10:23:45   61
[14/Jul/2024:10:23:45   72
[15/Jul/2024:10:23:45   79
[16/Jul/2024:10:23:45   69
[17/Jul/2024:10:23:45   66
[18/Jul/2024:10:23:45   87
[19/Jul/2024:10:23:45   71
```

Thống kê theo method

"GET": xuất hiện 2000 lần → nghĩa là trong log có 2000 request GET.

200: xuất hiện 2000 lần → số lượng request thành công (HTTP status code 200).

HTTP/1.1": xuất hiện 2000 lần → tất cả request dùng giao thức HTTP/1.1.

Thống kê theo engpoint

/ → trang chủ, được truy cập 410 lần.

/cart → trang giỏ hàng, 393 lần.

/checkout → trang thanh toán, 398 lần.

/login → trang đăng nhập, 401 lần.

Thông kế theo địa chỉ IP

192.168.1.1 -> 101 lần

192.168.1.10 -> 110 lần

192.168.1.11 -> 104 lần

192.168.1.9 -> 99 lần

Thống kế theo ngày

Ngày 10/07/2024 có 75 request.

Ngày 15/07/2024 có 79 request

Ngày 1/07/2024 có 61 request

Chạy wordcount trên tập dữ liệu student.txt

```
hadoop@root:~/bigdata-labs/lab5$ hadoop jar ~/bigdata-labs/lab5/build/wc.jar WordCount /lab5/input-students /lab5/output-students
2025-09-24 19:54:00,908 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-24 19:54:01,616 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
 with ToolRunner to remedy this.
2025-09-24 19:54:01,626 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1758715548556_0003
2025-09-24 19:54:01,769 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/hadoop/.staging/job_1758715548556_0003
Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://localhost:9000/lab5/input-students
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:342)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:281)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:445)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:311)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:328)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:201)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1677)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1674)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1953)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1674)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1695)
        at WordCount.main(WordCount.java:61)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:330)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:245)
Caused by: java.io.IOException: Input path does not exist: hdfs://localhost:9000/lab5/input-students
```

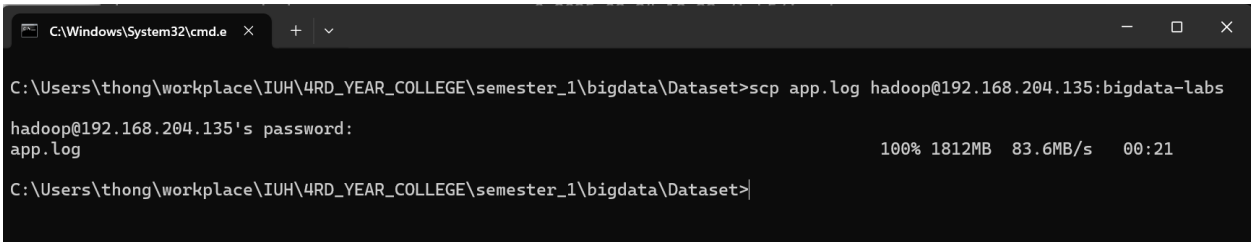
Kết quả đọc file dữ liệu đầu ra:

```
hadoop@root:~/bigdata-labs/lab5/src$ hdfs dfs -cat /lab5/output-students/part-r-00000
Alice   1
Bob     1
Charlie 1
David   1
Emma    1
Frank   1
Grace   1
Helen   1
Hoang   2
Ivan    1
Julia   1
Le      2
Nguyen  2
Pham    2
Tran    2
hadoop@root:~/bigdata-labs/lab5/src$
```

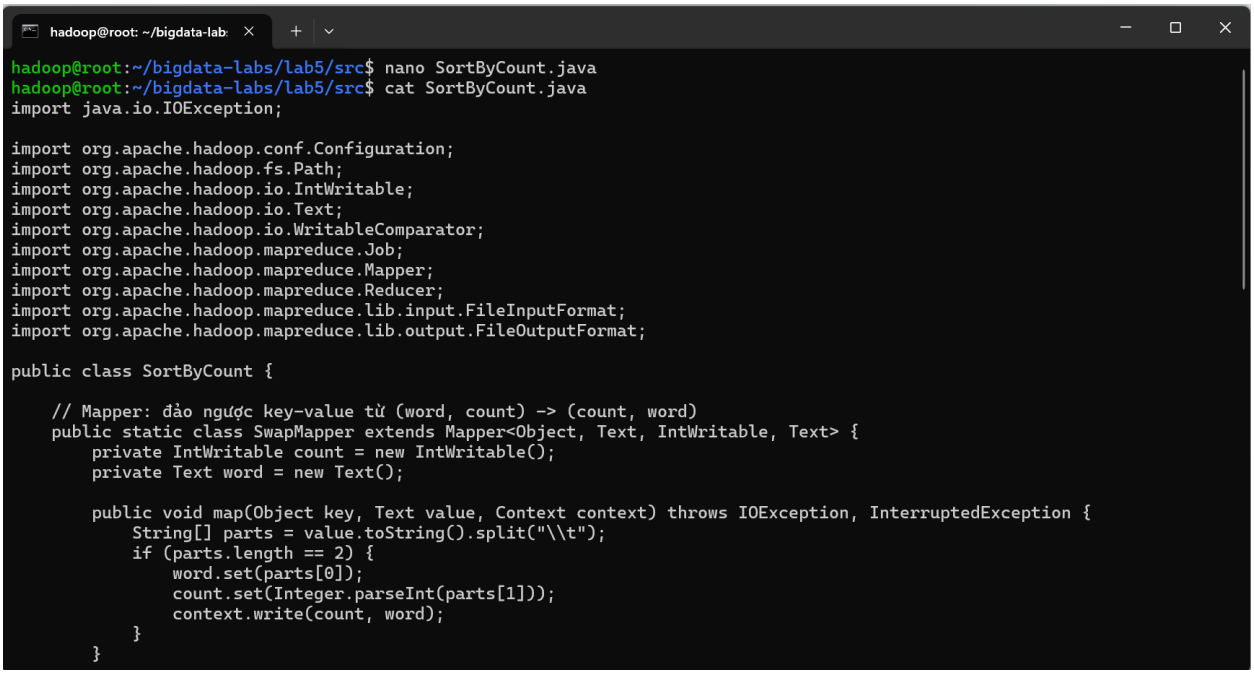
⇨ Dữ liệu của tập student ít hơn và ít phức tạp hơn

5.7 MapReduce nâng cao – SortByCount

Để chuẩn bị cho bài nâng cao em đã cho vào bộ dữ liệu ~2G để kiểm tra hiệu năng trên bộ dữ liệu lớn. Đây là bộ data app.log của một Backend đùng để lưu trữ lại tất cả các request mà dưới client gọi đến



Trước tiên hãy build trước hàm SortByCount



Buildl và compile hàm SortByCount để có thể chạy trên Hadoop

javac -classpath $(hadoop classpath) -d ../build SortByCount.java

Sau khi chạy xong sẽ có hàm trong build

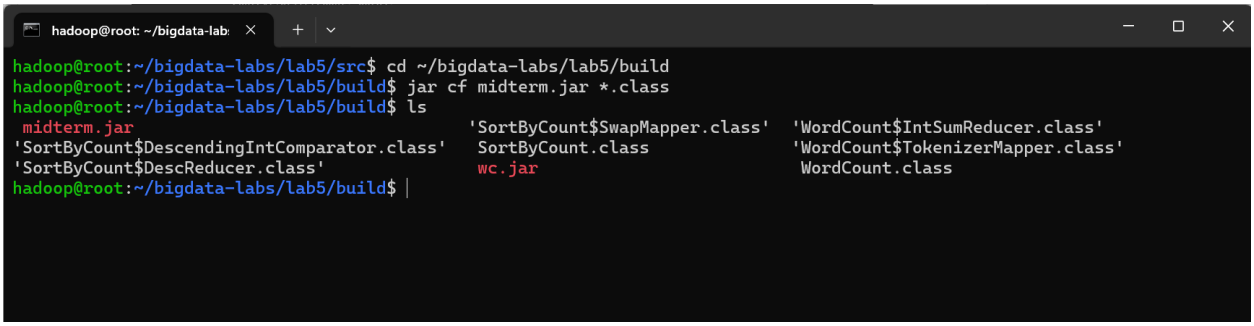
cd ~/bigdata-labs/lab5/build

Đóng gói thành jar

jar cf midterm.jar *.class

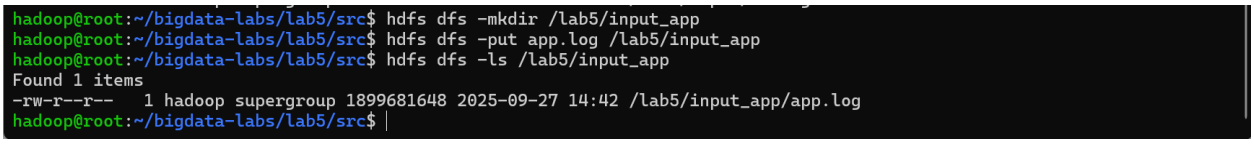
Kết quả

Có cả 2 hàm WordCount và SortByCount.

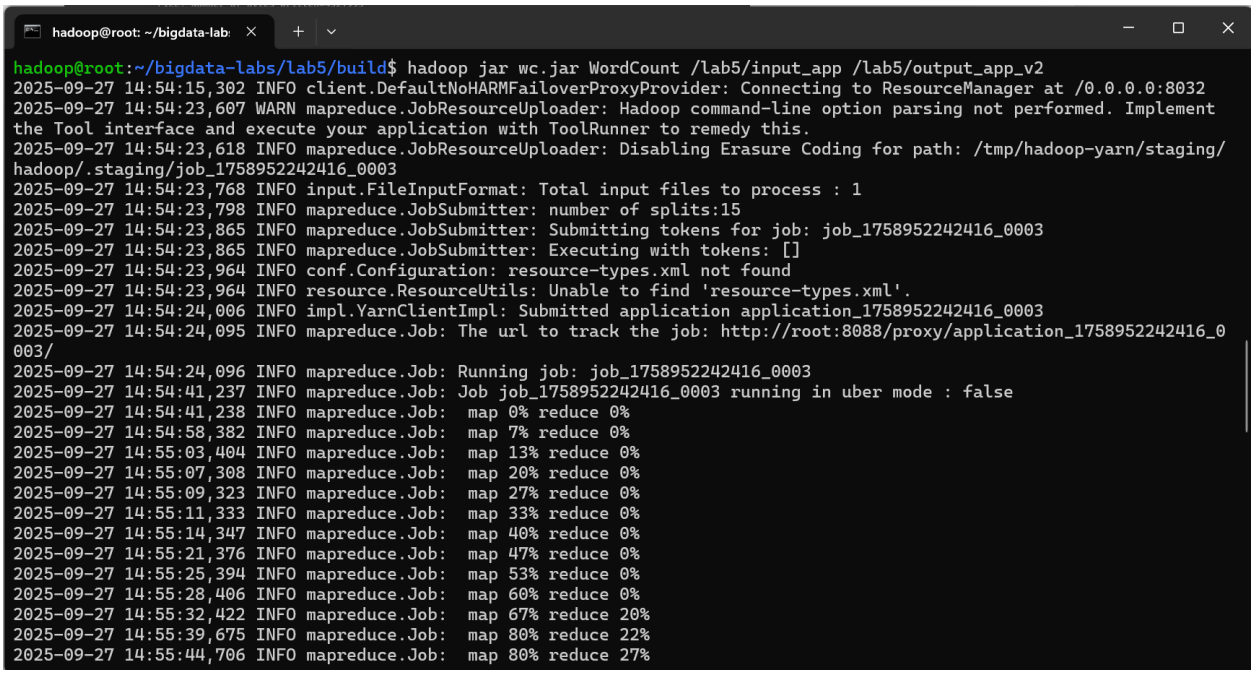


Chạy Worcount và SortByCount trên bộ data bên ngoài với khoảng 10 triệu dòng dữ liệu

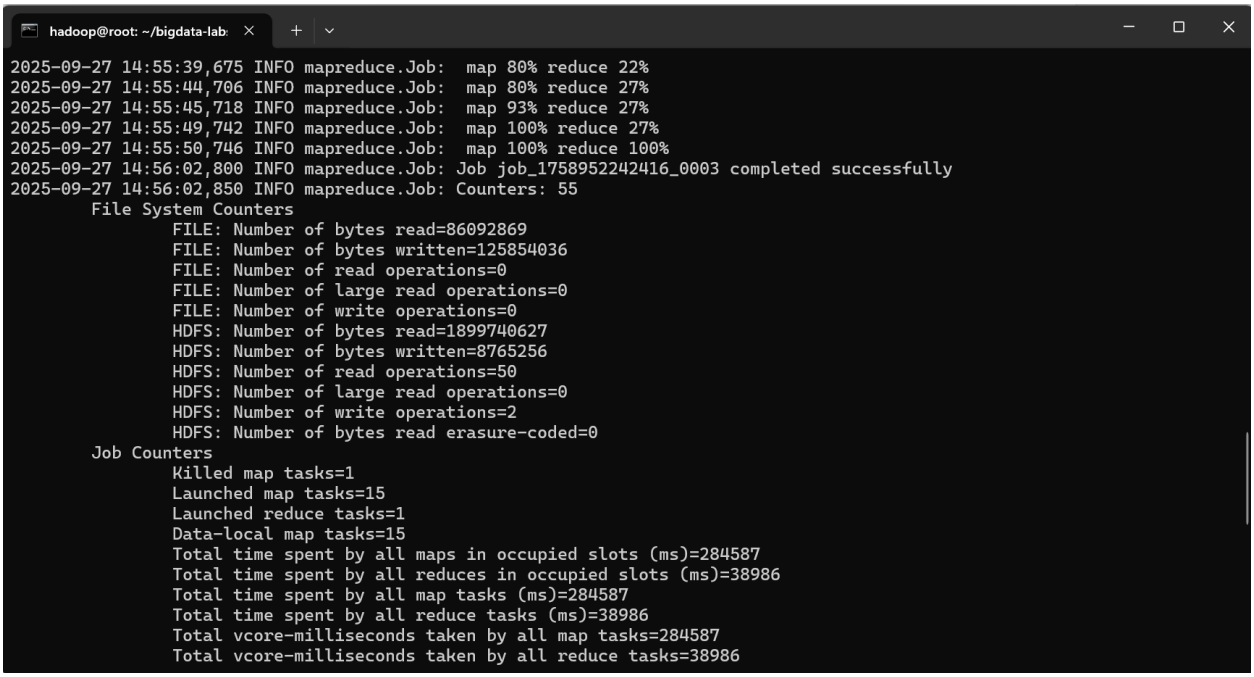
Tạo thư mục và put file dữ liệu vào thư mục để chuẩn bị chạy HDFS



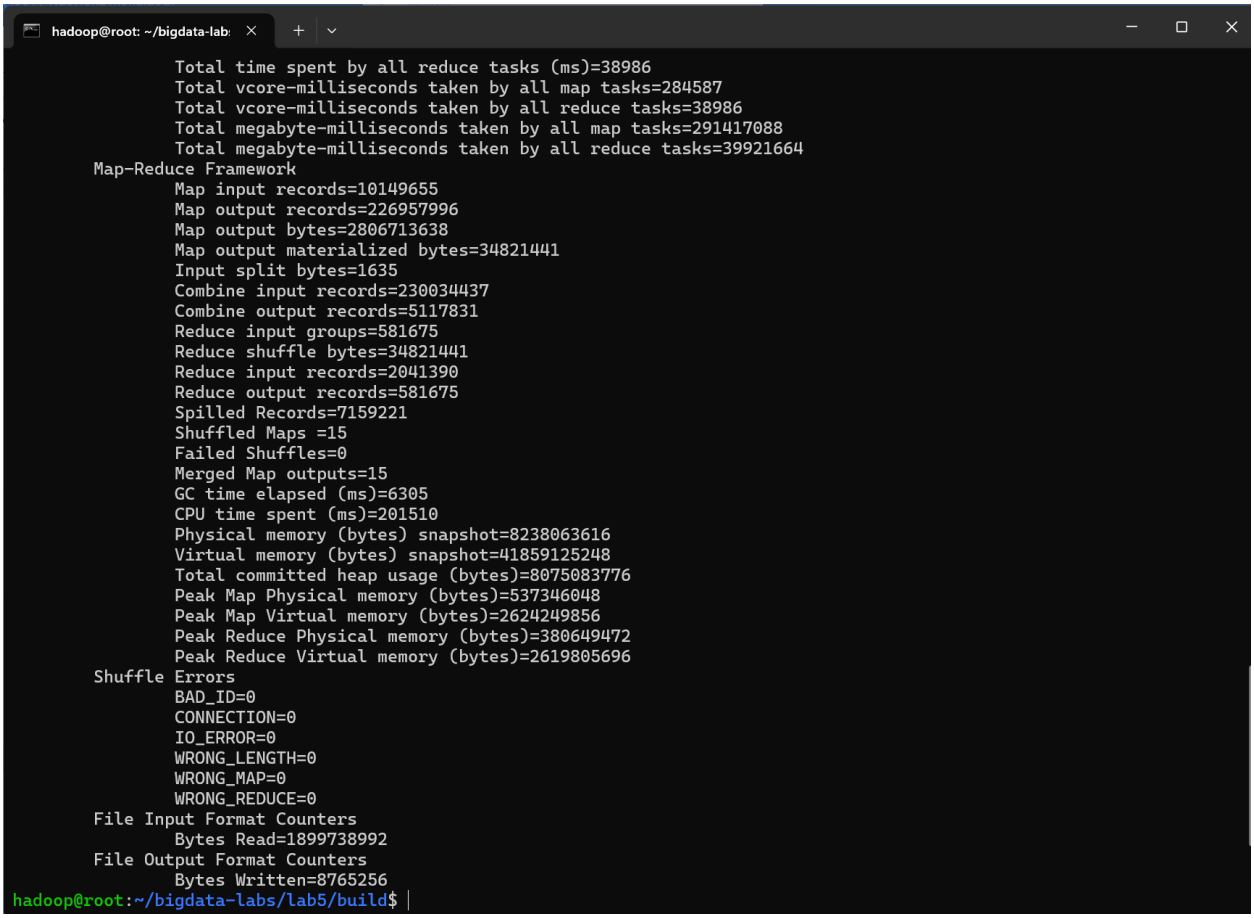
Đầu tiên chạy trước wordcount trên file app.log (10tr dòng)



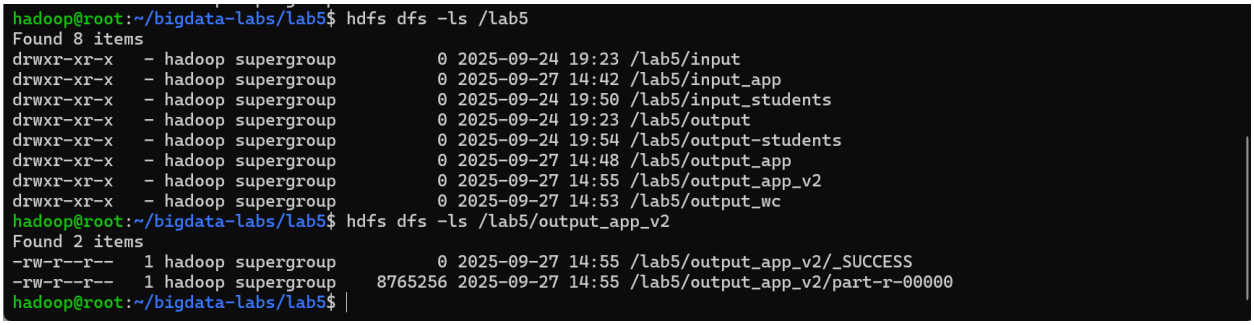
```
hadoop@root: ~/bigdata-lab
hadoop@root:~/bigdata-labs/lab5/build$ hadoop jar wc.jar WordCount /lab5/input_app /lab5/output_app_v2
2025-09-27 14:54:15,302 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-27 14:54:23,607 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-09-27 14:54:23,618 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hadoop/.staging/job_1758952242416_0003
2025-09-27 14:54:23,768 INFO input.FileInputFormat: Total input files to process : 1
2025-09-27 14:54:23,798 INFO mapreduce.JobSubmitter: number of splits:15
2025-09-27 14:54:23,865 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758952242416_0003
2025-09-27 14:54:23,865 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-27 14:54:23,964 INFO conf.Configuration: resource-types.xml not found
2025-09-27 14:54:23,964 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-27 14:54:24,006 INFO impl.YarnClientImpl: Submitted application application_1758952242416_0003
2025-09-27 14:54:24,095 INFO mapreduce.Job: The url to track the job: http://root:8088/proxy/application_1758952242416_0
003/
2025-09-27 14:54:24,096 INFO mapreduce.Job: Running job: job_1758952242416_0003
2025-09-27 14:54:41,237 INFO mapreduce.Job: Job job_1758952242416_0003 running in uber mode : false
2025-09-27 14:54:41,238 INFO mapreduce.Job:  map 0% reduce 0%
2025-09-27 14:54:58,382 INFO mapreduce.Job:  map 7% reduce 0%
2025-09-27 14:55:03,404 INFO mapreduce.Job:  map 13% reduce 0%
2025-09-27 14:55:07,308 INFO mapreduce.Job:  map 20% reduce 0%
2025-09-27 14:55:09,323 INFO mapreduce.Job:  map 27% reduce 0%
2025-09-27 14:55:11,333 INFO mapreduce.Job:  map 33% reduce 0%
2025-09-27 14:55:14,347 INFO mapreduce.Job:  map 40% reduce 0%
2025-09-27 14:55:21,376 INFO mapreduce.Job:  map 47% reduce 0%
2025-09-27 14:55:25,394 INFO mapreduce.Job:  map 53% reduce 0%
2025-09-27 14:55:28,406 INFO mapreduce.Job:  map 60% reduce 0%
2025-09-27 14:55:32,422 INFO mapreduce.Job:  map 67% reduce 20%
2025-09-27 14:55:39,675 INFO mapreduce.Job:  map 80% reduce 22%
2025-09-27 14:55:44,706 INFO mapreduce.Job:  map 80% reduce 27%
```

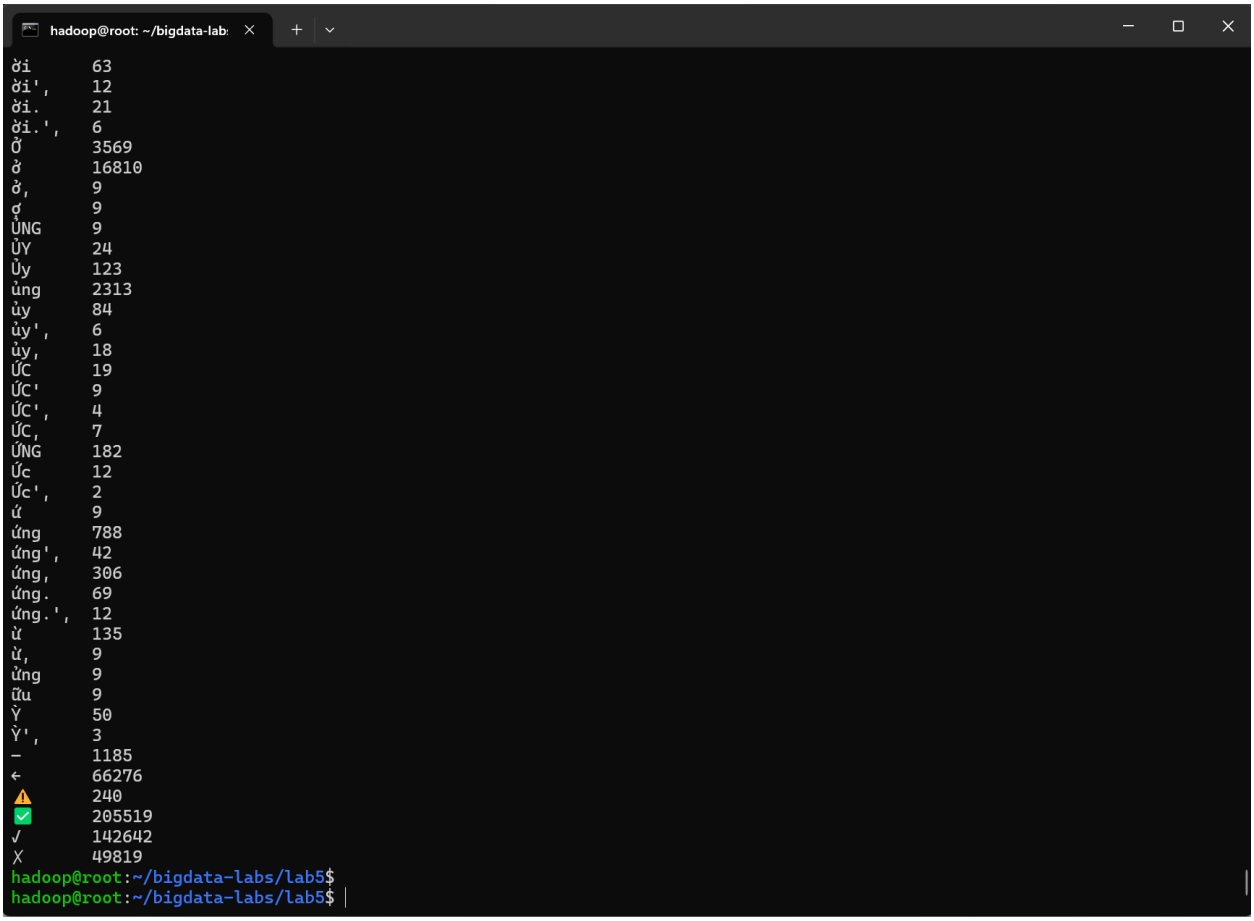


```
hadoop@root: ~/bigdata-lab
2025-09-27 14:55:39,675 INFO mapreduce.Job:  map 80% reduce 22%
2025-09-27 14:55:44,706 INFO mapreduce.Job:  map 80% reduce 27%
2025-09-27 14:55:45,718 INFO mapreduce.Job:  map 93% reduce 27%
2025-09-27 14:55:49,742 INFO mapreduce.Job:  map 100% reduce 27%
2025-09-27 14:55:50,746 INFO mapreduce.Job:  map 100% reduce 100%
2025-09-27 14:56:02,800 INFO mapreduce.Job: Job job_1758952242416_0003 completed successfully
2025-09-27 14:56:02,850 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=86092869
                FILE: Number of bytes written=125854036
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1899740627
                HDFS: Number of bytes written=8765256
                HDFS: Number of read operations=50
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=15
                Launched reduce tasks=1
                Data-local map tasks=15
                Total time spent by all maps in occupied slots (ms)=284587
                Total time spent by all reduces in occupied slots (ms)=38986
                Total time spent by all map tasks (ms)=284587
                Total time spent by all reduce tasks (ms)=38986
                Total vcore-milliseconds taken by all map tasks=284587
                Total vcore-milliseconds taken by all reduce tasks=38986
```



Đọc kết quả từ file kết quả

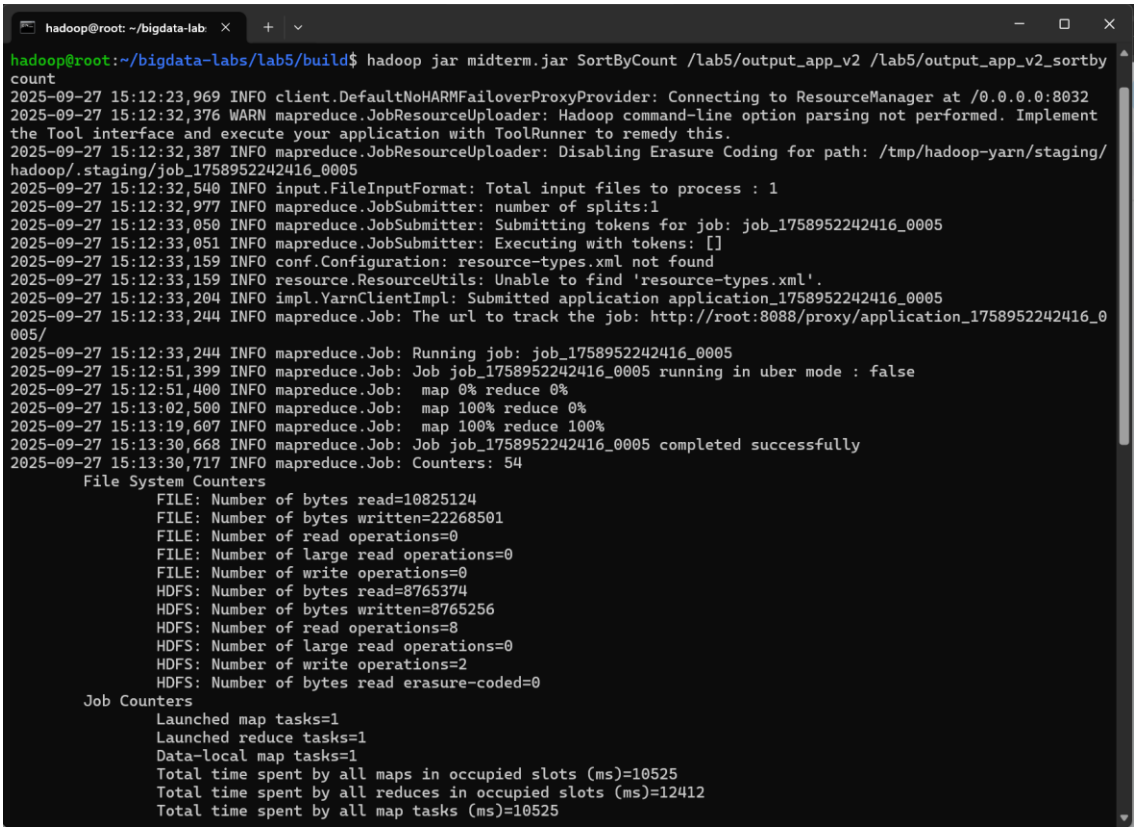


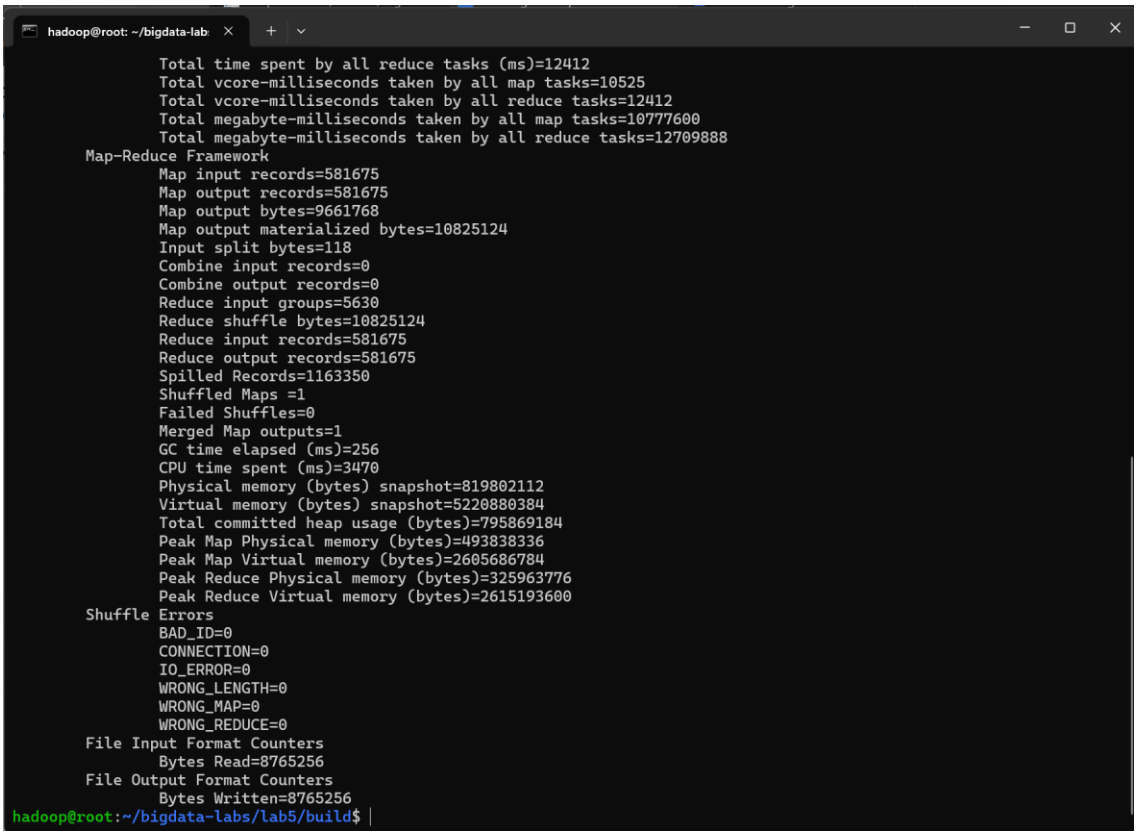


⇨ **Vì là data log nên là sẽ lấy được từng kí tự cũng như là từng chữ theo bài toán wordcount**

Tiếp theo chạy SortByCount trên bộ dữ liệu đã được chạy SortByCount output đầu ra của hàm CountWord

hadoop jar midterm.jar SortByCount /lab5/output_app_v2 /lab5/output_app_v2_sortbycount


hadoop@root: ~/bigdata-lab
hadoop@root:~/bigdata-labs/lab5/build$ hadoop jar midterm.jar SortByCount /lab5/output_app_v2 /lab5/output_app_v2_sortby
count
2025-09-27 15:12:23,969 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-27 15:12:32,376 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-09-27 15:12:32,387 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hadoop/.staging/job_1758952242416_0005
2025-09-27 15:12:32,540 INFO input.FileInputFormat: Total input files to process : 1
2025-09-27 15:12:32,977 INFO mapreduce.JobSubmitter: number of splits:1
2025-09-27 15:12:33,050 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758952242416_0005
2025-09-27 15:12:33,051 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-27 15:12:33,159 INFO conf.Configuration: resource-types.xml not found
2025-09-27 15:12:33,159 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-27 15:12:33,204 INFO impl.YarnClientImpl: Submitted application application_1758952242416_0005
2025-09-27 15:12:33,244 INFO mapreduce.Job: The url to track the job: http://root:8088/proxy/application_1758952242416_0
005/
2025-09-27 15:12:33,244 INFO mapreduce.Job: Running job: job_1758952242416_0005
2025-09-27 15:12:51,399 INFO mapreduce.Job: Job job_1758952242416_0005 running in uber mode : false
2025-09-27 15:12:51,400 INFO mapreduce.Job: map 0% reduce 0%
2025-09-27 15:13:02,500 INFO mapreduce.Job: map 100% reduce 0%
2025-09-27 15:13:19,607 INFO mapreduce.Job: map 100% reduce 100%
2025-09-27 15:13:30,668 INFO mapreduce.Job: Job job_1758952242416_0005 completed successfully
2025-09-27 15:13:30,717 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=10825124
FILE: Number of bytes written=22268501
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=8765374
HDFS: Number of bytes written=8765256
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
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)=10525
Total time spent by all reduces in occupied slots (ms)=12412
Total time spent by all map tasks (ms)=10525


hadoop@root: ~/bigdata-lab
Total time spent by all reduce tasks (ms)=12412
Total vcore-milliseconds taken by all map tasks=10525
Total vcore-milliseconds taken by all reduce tasks=12412
Total megabyte-milliseconds taken by all map tasks=10777600
Total megabyte-milliseconds taken by all reduce tasks=12709888
Map-Reduce Framework
Map input records=581675
Map output records=581675
Map output bytes=9661768
Map output materialized bytes=10825124
Input split bytes=118
Combine input records=0
Combine output records=0
Reduce input groups=5630
Reduce shuffle bytes=10825124
Reduce input records=581675
Reduce output records=581675
Spilled Records=1163350
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=256
CPU time spent (ms)=3470
Physical memory (bytes) snapshot=819802112
Virtual memory (bytes) snapshot=5220880384
Total committed heap usage (bytes)=795869184
Peak Map Physical memory (bytes)=493838336
Peak Map Virtual memory (bytes)=2605686784
Peak Reduce Physical memory (bytes)=325963776
Peak Reduce Virtual memory (bytes)=2615193600
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=8765256
File Output Format Counters
Bytes Written=8765256
hadoop@root:~/bigdata-labs/lab5/build$

Đọc kết quả sau khi đã SortByCount

```
hadoop@root:~/bigdata-labs/lab5/build$ hdfs dfs -cat /lab5/output_app_v2_sortbycount/part-r-00000 | head
1       cus_id=#2525102961,
1       cus_id=#2525102936,
1       cus_id=#2525102921,
1       cus_id=#2525102816,
1       cus_id=#2525102726,
1       cus_id=#2525102440,
1       cus_id=#2525101713,
1       cus_id=#2525101386,
1       cus_id=#2525101323,
1       cus_id=#2525101276,
cat: Unable to write to output stream.
hadoop@root:~/bigdata-labs/lab5/build$
```


```
hadoop@root:~/bigdata-labs/lab5/build$ hdfs dfs -cat /lab5/output_app_v2_sortbycount/part-r-00000 | head
1       cus_id=#2525102961,
1       cus_id=#2525102936,
1       cus_id=#2525102921,
1       cus_id=#2525102816,
1       cus_id=#2525102726,
1       cus_id=#2525102440,
1       cus_id=#2525101713,
1       cus_id=#2525101386,
1       cus_id=#2525101323,
1       cus_id=#2525101276,
cat: Unable to write to output stream.
hadoop@root:~/bigdata-labs/lab5/build$
```


```
hadoop@root:~/bigdata-labs/lab5/build$ hdfs dfs -cat /lab5/output_app_v2_sortbycount/part-r-00000 \
  | egrep "INFO|DEBUG|ERROR"
97739   ERROR
2525406 DEBUG:
9910437 INFO
hadoop@root:~/bigdata-labs/lab5/build$
```

Mục đích của việc em áp dụng WordCount và SortByCount để có thể đếm được số lượng log được ghi ra, là INFO, DEBUG, ERROR

Tóm lại

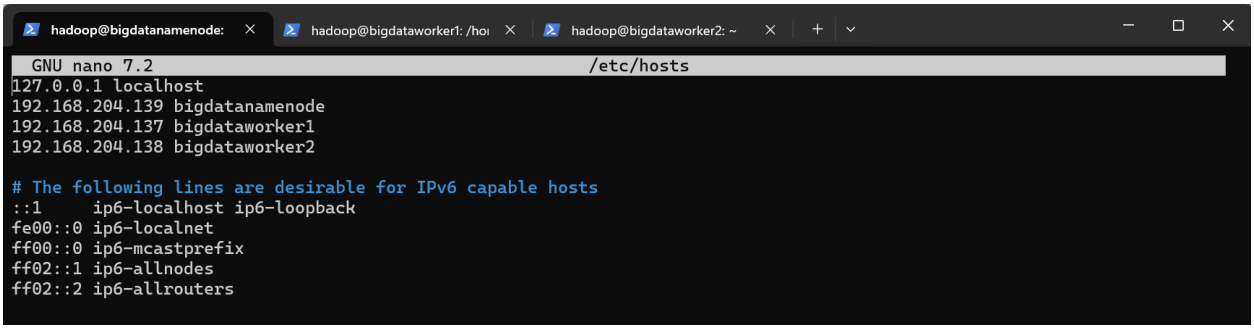
1. Thời gian thực chạy job = dựa vào dấu mốc Running job và completed successfully (≈ 57s).
2. Thời gian CPU tiêu tốn = CPU time spent (≈ 34.7s).
3. Thời gian của từng phase:
4. Map slots: 10.5s
5. Reduce slots: 12.4s

NÂNG CAO CHẠY TRÊN NHIỀU NODE MÀ NHIỀU MÁY ẢO

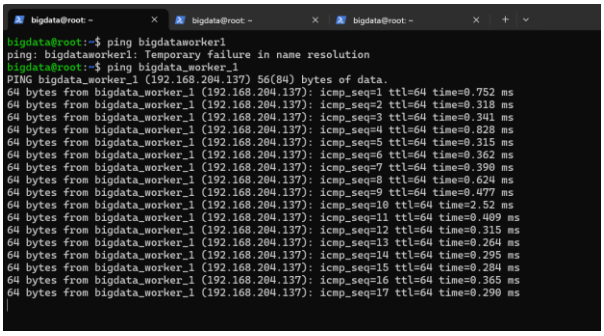
Chuẩn bị 3 máy cùng phiên bản java và hadoop

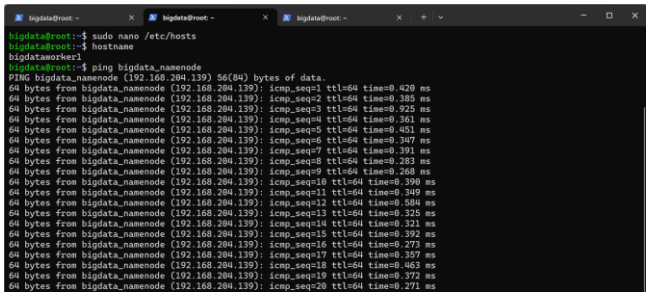
1 Máy NodeName (24G)

2 Máy worker (8g)



ở trên ba máy phải setup để có thể nhận biết được từng máy

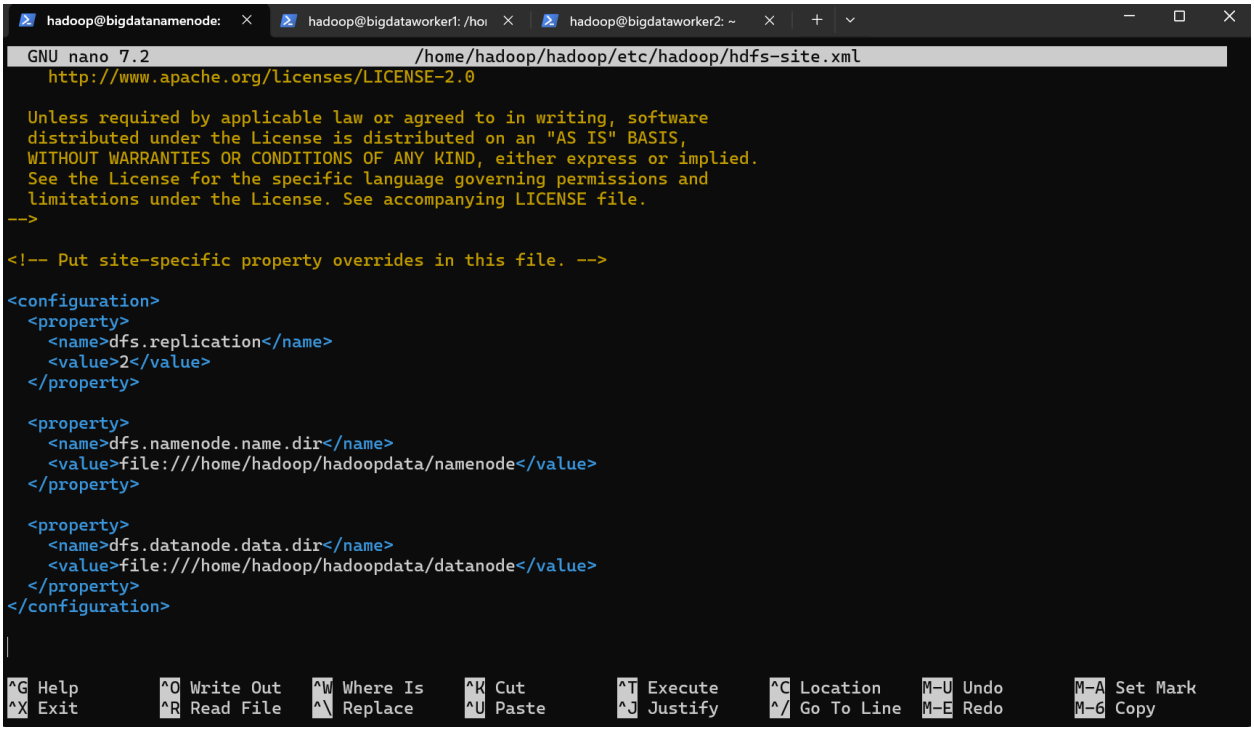




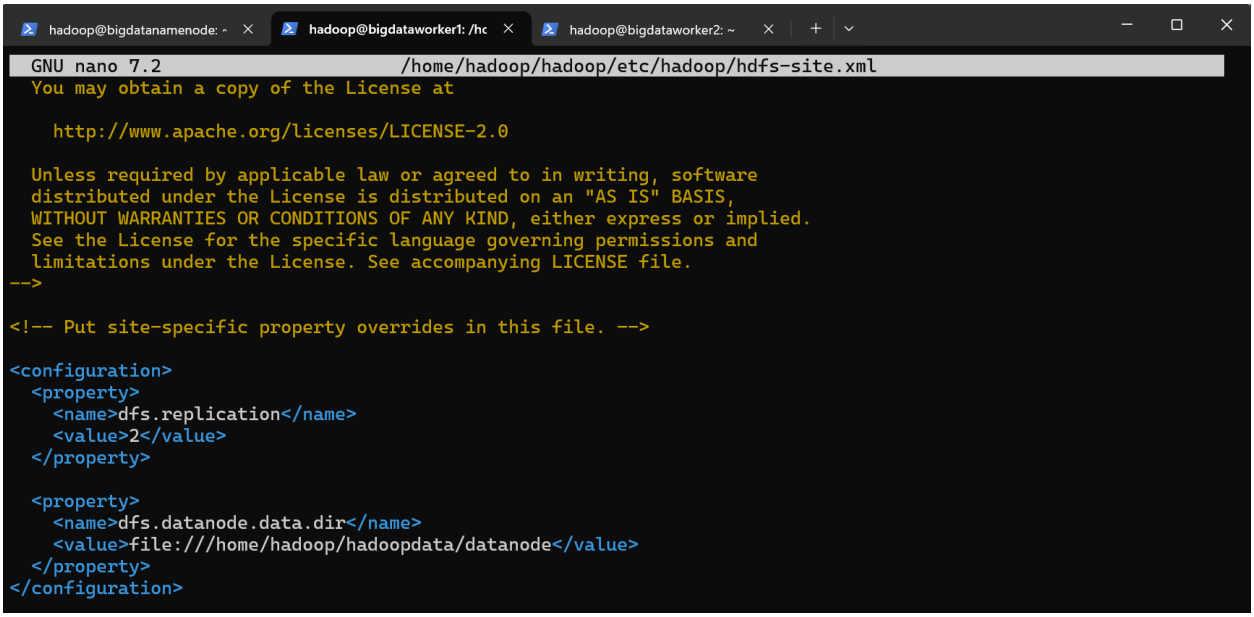
Chỉ khi ping được giữa các máy là bạn đã có thể connect giữa 3 máy

Cài đạt lại các file hdfs-site và core-site cho worker để không chạy namenode

Hdfs-site cho namenode



Hdfs-site cho worker



Core-site trên namenode cho cả trên woker

```
hadoop@bigdatanamenode:   hadoop@bigdataworker1: /hoi   hadoop@bigdataworker2: ~
  GNU nano 7.2                          /home/hadoop/hadoop/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://bigdatanamenode:9000</value>
  </property>
</configuration>
```

Chạy lại hdfs và yarn trên namenode

stop-dfs.sh/ start-dfs.sh

stop-yarn.sh/ start-yarn.sh

Sau đó chạy để quản dung lượng sử lý, số lượng Datanode, tình trạng block

hdfs dfsadmin -report

```
hadoop@bigdatanamenode:/home/bigdata$ hdfs dfsadmin -report
Configured Capacity: 105036840960 (97.82 GB)
Present Capacity: 65715965952 (61.20 GB)
DFS Remaining: 65715916800 (61.20 GB)
DFS Used: 49152 (48 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (2):

Name: 192.168.204.137:9866 (bigdataworker1)
Hostname: bigdataworker1
Decommission Status : Normal
Configured Capacity: 52518420480 (48.91 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 16961744896 (15.80 GB)
DFS Remaining: 32855678976 (30.60 GB)
DFS Used%: 0.00%
DFS Remaining%: 62.56%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sat Sep 27 17:13:10 ICT 2025
Last Block Report: Sat Sep 27 17:12:13 ICT 2025
Num of Blocks: 0


Name: 192.168.204.138:9866 (bigdataworker2)
Hostname: bigdataworker2
Decommission Status : Normal
Configured Capacity: 52518420480 (48.91 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 16957186048 (15.79 GB)
DFS Remaining: 32860237824 (30.60 GB)
```

yarn node -list

```
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ yarn node -list
2025-09-27 17:32:16,445 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at bigdatanamenode
/192.168.204.139:8032
Total Nodes:2
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
bigdataworker1:45549            RUNNING bigdataworker1:8042                               0
bigdataworker2:42355            RUNNING bigdataworker2:8042                               0
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$
```

⇨ **Chỉ khi chạy và thấy rằng datanode(2) thì ok**

Chạy lại WordCount trên file 2g mới fully distributed (3 máy)

Chạy lại file 2G xem tốc độ và cách chia plit

```
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ hdfs dfs -put app.log /lab5/input_full/
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ hadoop jar ~/bigdata-labs/lab5/build/wc.jar WordCount /lab5/input_full /
lab5/output_full
2025-09-27 17:22:12,419 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at bigdatanamenode
/192.168.204.139:8032
2025-09-27 17:22:12,692 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-09-27 17:22:12,704 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hadoop/.staging/job_1758967940226_0004
2025-09-27 17:22:12,871 INFO input.FileInputFormat: Total input files to process : 1
2025-09-27 17:22:12,914 INFO mapreduce.JobSubmitter: number of splits:15
2025-09-27 17:22:12,980 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758967940226_0004
2025-09-27 17:22:12,980 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-27 17:22:13,081 INFO conf.Configuration: resource-types.xml not found
2025-09-27 17:22:13,081 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-27 17:22:13,164 INFO impl.YarnClientImpl: Submitted application application_1758967940226_0004
2025-09-27 17:22:13,183 INFO mapreduce.Job: The url to track the job: http://bigdatanamenode:8088/proxy/application_1758
967940226_0004/
2025-09-27 17:22:13,184 INFO mapreduce.Job: Running job: job_1758967940226_0004
2025-09-27 17:22:18,261 INFO mapreduce.Job: Job job_1758967940226_0004 running in uber mode : false
2025-09-27 17:22:18,261 INFO mapreduce.Job:  map 0% reduce 0%
2025-09-27 17:22:37,614 INFO mapreduce.Job:  map 41% reduce 0%
2025-09-27 17:22:42,679 INFO mapreduce.Job:  map 44% reduce 0%
2025-09-27 17:22:43,703 INFO mapreduce.Job:  map 71% reduce 0%
2025-09-27 17:22:44,711 INFO mapreduce.Job:  map 81% reduce 0%
2025-09-27 17:22:45,719 INFO mapreduce.Job:  map 93% reduce 0%
2025-09-27 17:22:47,737 INFO mapreduce.Job:  map 100% reduce 0%
2025-09-27 17:22:49,747 INFO mapreduce.Job:  map 100% reduce 100%
2025-09-27 17:22:49,756 INFO mapreduce.Job: Job job_1758967940226_0004 completed successfully
2025-09-27 17:22:49,817 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=86092869
                FILE: Number of bytes written=125854660
```

```
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ ls
app.log  midterm.jar  SortByCount.java  weblogs.txt  WordCount.java
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ hdfs dfs -put app.log /lab5/input_full/
hadoop@bigdatanamenode:~/bigdata-labs/lab5/src$ hadoop jar ~/bigdata-labs/lab5/build/wc.jar WordCount /lab5/input_full /
lab5/output_full
2025-09-27 17:22:12,419 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at bigdatanamenode
/192.168.204.139:8032
2025-09-27 17:22:12,692 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-09-27 17:22:12,704 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hadoop/.staging/job_1758967940226_0004
2025-09-27 17:22:12,871 INFO input.FileInputFormat: Total input files to process : 1
2025-09-27 17:22:12,914 INFO mapreduce.JobSubmitter: number of splits:15
2025-09-27 17:22:12,980 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758967940226_0004
2025-09-27 17:22:12,980 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-27 17:22:13,081 INFO conf.Configuration: resource-types.xml not found
2025-09-27 17:22:13,081 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-27 17:22:13,164 INFO impl.YarnClientImpl: Submitted application application_1758967940226_0004
2025-09-27 17:22:13,183 INFO mapreduce.Job: The url to track the job: http://bigdatanamenode:8088/proxy/application_1758
967940226_0004/
2025-09-27 17:22:13,184 INFO mapreduce.Job: Running job: job_1758967940226_0004
2025-09-27 17:22:18,261 INFO mapreduce.Job: Job job_1758967940226_0004 running in uber mode : false
2025-09-27 17:22:18,261 INFO mapreduce.Job:  map 0% reduce 0%
2025-09-27 17:22:37,614 INFO mapreduce.Job:  map 41% reduce 0%
2025-09-27 17:22:42,679 INFO mapreduce.Job:  map 44% reduce 0%
2025-09-27 17:22:43,703 INFO mapreduce.Job:  map 71% reduce 0%
2025-09-27 17:22:44,711 INFO mapreduce.Job:  map 81% reduce 0%
2025-09-27 17:22:45,719 INFO mapreduce.Job:  map 93% reduce 0%
2025-09-27 17:22:47,737 INFO mapreduce.Job:  map 100% reduce 0%
2025-09-27 17:22:49,747 INFO mapreduce.Job:  map 100% reduce 100%
2025-09-27 17:22:49,756 INFO mapreduce.Job: Job job_1758967940226_0004 completed successfully
```