

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN**



NHẬP MÔN DỮ LIỆU LỚN
LAB 02: BÀI TOÁN ĐẾM TỪ VỚI MAPREDUCE

Giảng viên lý thuyết
TS. Nguyễn Ngọc Thảo

Giảng viên hướng dẫn thực hành
ThS. Lê Ngọc Thành

Sinh viên thực hiện
Nguyễn Thị Thu Hằng – 18120027
Nguyễn Tấn Thìn – 18120085
Phạm Nguyên Minh Thy – 18120090

Tháng 11 năm 2021

Mục lục

1.	Thông tin nhóm và phân công	3
1.1.	Thông tin thành viên.....	3
1.2.	Phân công và đánh giá công việc	3
2.	Nội dung báo cáo.....	3
2.1.	Chương trình MapReduce mức 1.....	3
2.1.1.	Thiết kế quá trình thực thi chạy phân tán trên nhiều node	3
2.1.2.	Phiên bản 1.0	4
2.1.3.	Phiên bản 2.0	15
2.1.4.	Phiên bản 3.0	22
2.2.	Chương trình MapReduce mức 2.....	30
2.2.1.	Trường hợp đếm phân biệt hoa thường.....	32
2.2.2.	Trường hợp đếm không phân biệt hoa thường.....	37
2.2.3.	Tìm từ xuất hiện nhiều nhất trong tài liệu (không phân biệt hoa thường).....	40
3.	Tài liệu tham khảo	46

1. Thông tin nhóm và phân công

1.1. Thông tin thành viên

STT	MSSV	Họ và tên	Email
1	18120027	Nguyễn Thị Thu Hằng	18120027@student.hcmus.edu.vn
2	18120085	Nguyễn Tấn Thìn	18120085@student.hcmus.edu.vn
3	18120090	Phạm Nguyên Minh Thy	18120090@student.hcmus.edu.vn

1.2. Phân công và đánh giá công việc

Thành viên	Phân công	Đánh giá mức độ hoàn thành
Nguyễn Thị Thu Hằng	<ul style="list-style-type: none">• Tìm hiểu, viết và chạy chương trình mapreduce mức 1 phiên bản 1.0 và 2.0• Tìm testcase cho mapreduce mức 1• Viết báo cáo phần chạy chương trình mapreduce mức 1• Chỉnh sửa video	100%
Nguyễn Tấn Thìn	<ul style="list-style-type: none">• Tìm hiểu, viết và chạy chương trình mapreduce mức 2 sử dụng MRJob• Viết báo cáo phần chạy chương trình mapreduce mức 2• Quay video chạy chương trình mapreduce mức 2	100%
Phạm Nguyên Minh Thy	<ul style="list-style-type: none">• Tìm hiểu, viết và chạy chương trình mapreduce mức 1 phiên bản 3.0• Tìm hiểu cách chạy phân tán trên nhiều node cho mapreduce mức 1• Viết báo cáo phần chạy chương trình mapreduce mức 1• Quay video chạy chương trình mapreduce mức 1	100%

2. Nội dung báo cáo

2.1. Chương trình MapReduce mức 1

2.1.1. Thiết kế quá trình thực thi chạy phân tán trên nhiều node

Thiết lập 2 Datanode chạy trên cùng 1 Namenode `hdfs://hadoop-namenode:9820/`



Hình 1: Minh chứng có 2 Datanode đang chạy trên cùng 1 Namenode

Hiện thị 2 users trong hệ thống HDFS với user **pnmth** là máy chứa Namenode và Datanode 1, user **thy** là máy chứa Datanode 2.

Browse Directory

/user | Go! | [Icons]

Show: 25 entries | Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	pnmth	supergroup	0 B	Nov 16 12:25	0	0 B	pnmth
drwxr-xr-x	thy	supergroup	0 B	Nov 16 12:42	0	0 B	thy

Showing 1 to 2 of 2 entries | Previous 1 Next

Hình 2: Minh chứng 2 users truy cập vào hệ thống HDFS của Namenode

2.1.2. Phiên bản 1.0

a. Mã nguồn và giải thích

Kế thừa mã nguồn

Cấu trúc khai báo và xử lý đối với Map, Reduce task.

Cách tách từ dựa trên Pattern regular expression.

Thay đổi – Bổ sung mã nguồn

Cho phép nhận nhiều đường dẫn input trên command line theo cấu trúc `<input_1> <input_2> ... <input_n> <output>`.

Giải thích mã nguồn

```

// Kế thừa từ class Configured và interface Tool để định nghĩa cho Hadoop
cách chạy chương trình.
public class WordCount extends Configured implements Tool {

    public static void main(String[] args) throws Exception {
        // Dùng ToolRunner tạo 1 instance mới của WordCount để chạy MapReduce.
        int res = ToolRunner.run(new WordCount(), args);
        System.exit(res);
    }

    public int run(String[] args) throws Exception {
        // Tạo 1 instance Job mới.
        // Dùng getConf() để lấy configuration object của class WordCount và đặt
        tên cho Job là wordcount.
        Job job = Job.getInstance(getConf(), "wordcount");
        // Thiết lập Jar file cho class WordCount.
        job.setJarByClass(this.getClass());

        // Thêm đường dẫn tới input là các argument trừ argument cuối cùng trong
        command.
        for (int i = 0; i < args.length - 1; i += 1) {
            FileInputFormat.addInputPath(job, new Path(args[i]));
        }

        // Thêm đường dẫn tới output là argument cuối cùng trong command.
        FileOutputFormat.setOutputPath(job, new Path(args[args.length - 1]));

        // Thiết lập Map class, Reduce class tương ứng với custom Map và Reduce
        class bên trong WordCount.
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);

        // Thiết lập output key ở dạng chuỗi nên là class Text
        // và output value ở dạng số nguyên nên là class IntWritable.
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        // Chờ cho tới khi hoàn thành task, 0 là không có lỗi xảy ra, 1 là có
        lỗi.
        return job.waitForCompletion(true) ? 0 : 1;
    }

    public static class Map extends Mapper<LongWritable, Text, Text,
    IntWritable> {
        // Tạo biến one kiểu IntWritable và có giá trị là 1
        private final static IntWritable one = new IntWritable(1);

```

```

    // Tạo WORD_BOUNDARY là regular expression để tách từ trong câu bằng
    // khoảng trắng và kí tự đặc biệt.
    // Với \s* là nhiều kí tự khoảng trắng, \b là bắt buộc trùng khớp với
    // thành phần trước nó.
    private static final Pattern WORD_BOUNDARY =
    Pattern.compile("\\s*\\b\\s*");

    // Đầu vào của Map là từng dòng của input, và dùng Context để ghi output
    // cho Map task.
    public void map(LongWritable offset, Text inputLine, Context output)
        throws IOException, InterruptedException {

        // Ép kiểu cho inputLine từ Text sang String
        String line = inputLine.toString();

        // Tách line thành nhiều từ với Pattern WORD_BOUNDARY.
        for (String word : WORD_BOUNDARY.split(line)) {
            // Nếu word rỗng thì bỏ qua
            if (word.isEmpty())
                continue;

            // Tạo 1 Text outputWord từ word
            // rồi ghi outputWord vào trong output với key là outputWord và
            // value là 1.
            Text outputWord = new Text(word);
            output.write(outputWord, one);
        }
    }

    // Đầu vào của Reduce là (từ, mảng chứa các số lần xuất hiện),
    // và dùng Context để ghi output cho Reduce task.
    public static class Reduce extends Reducer<Text, IntWritable, Text,
    IntWritable> {
        public void reduce(Text word, Iterable<IntWritable> counts, Context
        output)
            throws IOException, InterruptedException {
            // Tính tổng số lần xuất hiện của word
            int sum = 0;
            for (IntWritable count : counts)
                sum += count.get();

            // Ghi vào output key là word, value là số lần xuất hiện của word
            output.write(word, new IntWritable(sum));
        }
    }
}

```

b. Quá trình thực thi

Testcase 0

Tại máy chứa Namenode, copy file test0-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-0.txt input
```

Tại máy chứa Datanode 2, copy file test0-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-1.txt input
```

Tại máy chứa Namenode, biên dịch mã nguồn của WordCount phiên bản 1.0 ra file jar

```
javac -classpath jar_files\* -d archive
```

```
Hadoop_tutorial\WordCount1\WordCount.java
```

```
jar -cvf WordCount.jar -C archive\ .
```

Tại máy chứa Namenode, chạy MapReduce

```
bin\yarn jar WordCount.jar WordCount /user/pnmth/input  
/user/thy/input output
```

Tại máy chứa Namenode, xem kết quả trong thư mục chứa output

```
bin\hdfs dfs -cat output/*
```

Nội dung file test0-0.txt

```
Install Hadoop3.
```

```
Run Hadoop Wordcount. Mapreduce Example.
```

```
The hadoop mapreduce!
```

Nội dung file test0-1.txt


```
My name is Thy.
```


```
Thy is a name, and I have successfully installed multiple node  
on Hadoop.
```


Browse Directory


/user/pnmth/input

Go!



















Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div>Permission</div>	<div><div></div>Owner</div>	<div><div></div>Group</div>	<div><div></div>Size</div>	<div><div></div>Last Modified</div>	<div><div></div>Replication</div>	<div><div></div>Block Size</div>	<div><div></div>Name</div>	<div><div></div></div>
<input type="checkbox"/>	<div><div><div>-rw-r--r--</div></div></div>	<div><div><div>pnmth</div></div></div>	<div><div><div>supergroup</div></div></div>	<div><div><div>81 B</div></div></div>	<div><div><div>Nov 16 13:53</div></div></div>	<div><div><div>3</div></div></div>	<div><div><div>128 MB</div></div></div>	<div><div><div>test0-0.txt</div></div></div>	<div><div><div></div></div></div>

Hình 3: Minh chứng đã upload file test0-0.txt trên máy user pnmth

Browse Directory

/user/thy/input

Go!

Show

25

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		thy		supergroup		90 B		Nov 16 13:55		3		128 MB		test0-1.txt	

Hình 4: Minh chứng đã upload file test0-1.txt trên máy user thy

```
e:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 13:07:12,093 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 13:07:12,837 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
pnmth/.staging/job_1637039448001_0001
2021-11-16 13:07:14,278 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 13:07:14,479 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 13:07:14,891 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637039448001_0001
2021-11-16 13:07:14,891 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 13:07:15,101 INFO conf.Configuration: resource-types.xml not found
2021-11-16 13:07:15,102 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 13:07:15,706 INFO impl.YarnClientImpl: Submitted application application_1637039448001_0001
2021-11-16 13:07:15,744 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SH4IL:8088/proxy/application_1637
039448001_0001/
2021-11-16 13:07:15,745 INFO mapreduce.Job: Running job: job_1637039448001_0001
2021-11-16 13:07:34,738 INFO mapreduce.Job: Job job_1637039448001_0001 running in uber mode : false
2021-11-16 13:07:34,741 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 13:07:42,873 INFO mapreduce.Job: map 50% reduce 0%
2021-11-16 13:07:43,877 INFO mapreduce.Job: map 100% reduce 0%
2021-11-16 13:07:49,940 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 5: Quá trình chạy task trên testcase 0 của phiên bản 1.0 trong khoảng 15s


```

c:\hadoop-3.3.1>bin\hdfs dfs -cat output/*
!      1
.      1
.      5
Example 1
Hadoop 2
Hadoop3 1
I      1
Install 1
Mapreduce      1
My      1
Run      1
The      1
Thy      2
Wordcount      1
a      1
and      1
hadoop 1
have 1
installed      1
is      2
mapreduce      1
multiple      1
name      2
node 1
on      1
successfully 1

```

Hình 6: Kết quả testcase 0 của phiên bản 1.0

Testcase 1

Tại máy chứa Namenode, copy file test1-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test1/test1-0.txt input
```

Tại máy chứa Datanode 2, copy file test1-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test1/test1-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

Kích thước input

File test1-0.txt ~ 58MB

File test1-1.txt ~ 38MB

Do testcase khá lớn nên khó có thể thấy toàn bộ input và output.

Browse Directory

/user/pnmth/input

Go!

Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div><div></div><div></div></div> Permission	<div><div></div><div></div><div></div></div> Owner	<div><div></div><div></div><div></div></div> Group	<div><div></div><div></div><div></div></div> Size	<div><div></div><div></div><div></div></div> Last Modified	<div><div></div><div></div><div></div></div> Replication	<div><div></div><div></div><div></div></div> Block Size	<div><div></div><div></div><div></div></div> Name	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	-rw-r--r--	pnmth	supergroup	57.33 MB	Nov 16 13:36	3	128 MB	test1-0.txt	<div><div></div></div>

Hình 7: Minh chứng đã upload file test1-0.txt trên máy user pnmth

Browse Directory

/user/thy/input

Go!

Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div><div></div></div> Permission	<div><div></div><div></div></div> Owner	<div><div></div><div></div></div> Group	<div><div></div><div></div></div> Size	<div><div></div><div></div></div> Last Modified	<div><div></div><div></div></div> Replication	<div><div></div><div></div></div> Block Size	<div><div></div><div></div></div> Name	<div><div></div><div></div></div>
<input type="checkbox"/>	-rw-r--r--	thy	supergroup	37.68 MB	Nov 16 13:39	3	128 MB	test1-1.txt	<div><div></div></div>

Hình 8: Minh chứng đã upload file test1-1.txt trên máy user thy

```
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 13:42:30,371 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 13:42:30,921 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
pnmth/.staging/job_1637039448001_0003
2021-11-16 13:42:31,483 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 13:42:32,443 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 13:42:34,117 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637039448001_0003
2021-11-16 13:42:34,117 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 13:42:34,281 INFO conf.Configuration: resource-types.xml not found
2021-11-16 13:42:34,282 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 13:42:34,343 INFO impl.YarnClientImpl: Submitted application application_1637039448001_0003
2021-11-16 13:42:34,379 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_1637
039448001_0003/
2021-11-16 13:42:34,379 INFO mapreduce.Job: Running job: job_1637039448001_0003
2021-11-16 13:42:49,956 INFO mapreduce.Job: Job job_1637039448001_0003 running in uber mode : false
2021-11-16 13:42:49,959 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 13:43:08,249 INFO mapreduce.Job: map 33% reduce 0%
2021-11-16 13:43:09,288 INFO mapreduce.Job: map 83% reduce 0%
2021-11-16 13:43:13,335 INFO mapreduce.Job: map 100% reduce 0%
2021-11-16 13:43:23,453 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 9: Quá trình chạy task trên testcase 1 của phiên bản 1.0 trong khoảng 34s

```

zilch 4
zillion 1
zillow 6
zinc 2
zion 24
zip 217
zipcode 13
zipcodes 2
zipped 3
zipper 6
zipxxx 3
zircon 1
zlimen 3
zocaloan 1
zocaloans 2
zombi 1
zombie 30
zomie 1
zone 176
zoned 4
zoning 11
zoo 4
zoom 10
zoomed 1
zoomsup 1
zsaleh 1
zuntafi 4
zwicker 36

```

Hình 10: Một phần kết quả testcase 1 của phiên bản 1.0 do testcase khá dài nên không thể hiện rõ được sự khác biệt giữa 3 phiên bản

Testcase 2

Tại máy chứa Namenode, copy file test2-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

Kích thước input

File test1-0.txt ~ 180MB

File test1-1.txt ~ 167MB

Do testcase khá lớn nên khó có thể thấy toàn bộ input và output.

Browse Directory

/user/pnmth/input

Go!

Show

25

entries

Search:

Permission

Owner

Group

Size

Last Modified

Replication

Block Size

Name

-rw-r--r--

pnmth

supergroup

176.28 MB

Nov 16 15:42

3

128 MB

test2-0.txt

Hình 11: Minh chứng đã upload file test2-0.txt trên máy user pnmth

Browse Directory

/user/thy/input

Go!

Show

25

entries

Search:

Permission

Owner

Group

Size

Last Modified

Replication

Block Size

Name

-rw-r--r--

thy

supergroup

163.14 MB

Nov 16 15:52

3

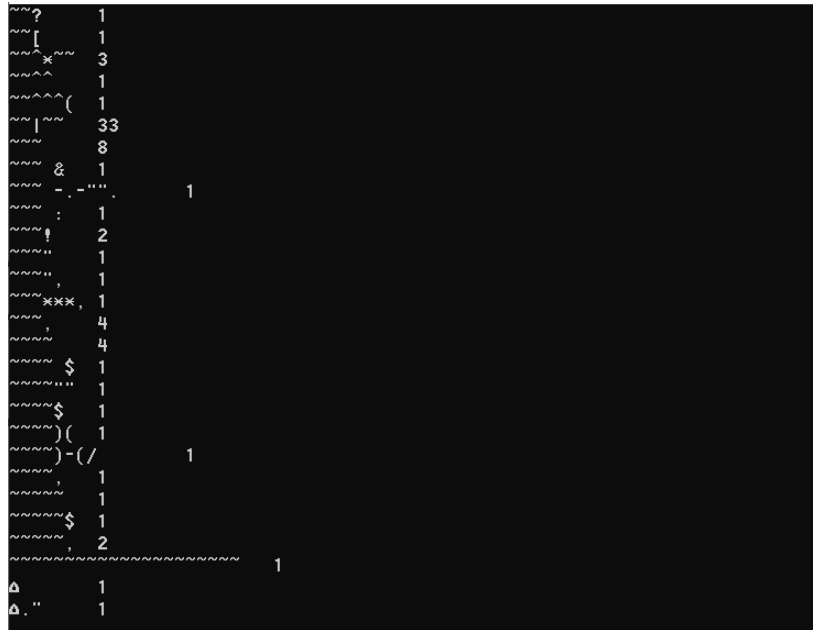
128 MB

test2-1.txt

Hình 12: Minh chứng đã upload file test2-1.txt trên máy user thy

```
e:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 15:53:38,840 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 15:53:39,523 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pnmth/.staging/j
2021-11-16 15:53:42,774 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 15:53:51,403 INFO mapreduce.JobSubmitter: number of splits:4
2021-11-16 15:54:02,366 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637052105009_0001
2021-11-16 15:54:02,366 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 15:54:02,543 INFO conf.Configuration: resource-types.xml not found
2021-11-16 15:54:02,543 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 15:54:03,075 INFO impl.YarnClientImpl: Submitted application application_1637052105009_0001
2021-11-16 15:54:03,116 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_1637052105009_0001/
2021-11-16 15:54:03,117 INFO mapreduce.Job: Running job: job_1637052105009_0001
2021-11-16 15:54:26,220 INFO mapreduce.Job: Job job_1637052105009_0001 running in uber mode : false
2021-11-16 15:54:26,222 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 15:54:46,612 INFO mapreduce.Job: map 6% reduce 0%
2021-11-16 15:54:47,624 INFO mapreduce.Job: map 10% reduce 0%
2021-11-16 15:54:50,652 INFO mapreduce.Job: map 19% reduce 0%
2021-11-16 15:54:51,669 INFO mapreduce.Job: map 31% reduce 0%
2021-11-16 15:54:52,685 INFO mapreduce.Job: map 34% reduce 0%
2021-11-16 15:54:53,714 INFO mapreduce.Job: map 36% reduce 0%
2021-11-16 15:54:56,751 INFO mapreduce.Job: map 40% reduce 0%
2021-11-16 15:54:57,764 INFO mapreduce.Job: map 44% reduce 0%
2021-11-16 15:54:58,775 INFO mapreduce.Job: map 47% reduce 0%
2021-11-16 15:54:59,790 INFO mapreduce.Job: map 49% reduce 0%
2021-11-16 15:55:02,835 INFO mapreduce.Job: map 52% reduce 0%
2021-11-16 15:55:03,880 INFO mapreduce.Job: map 60% reduce 0%
2021-11-16 15:55:08,738 INFO mapreduce.Job: map 69% reduce 0%
2021-11-16 15:55:10,780 INFO mapreduce.Job: map 73% reduce 0%
2021-11-16 15:55:12,846 INFO mapreduce.Job: map 75% reduce 0%
2021-11-16 15:55:16,906 INFO mapreduce.Job: map 76% reduce 0%
2021-11-16 15:55:18,946 INFO mapreduce.Job: map 78% reduce 0%
2021-11-16 15:55:22,996 INFO mapreduce.Job: map 79% reduce 0%
2021-11-16 15:55:25,020 INFO mapreduce.Job: map 81% reduce 0%
2021-11-16 15:55:28,063 INFO mapreduce.Job: map 81% reduce 17%
2021-11-16 15:55:29,078 INFO mapreduce.Job: map 85% reduce 17%
2021-11-16 15:55:31,101 INFO mapreduce.Job: map 87% reduce 17%
2021-11-16 15:55:34,147 INFO mapreduce.Job: map 91% reduce 17%
2021-11-16 15:55:37,192 INFO mapreduce.Job: map 92% reduce 17%
2021-11-16 15:55:40,233 INFO mapreduce.Job: map 92% reduce 25%
2021-11-16 15:55:43,277 INFO mapreduce.Job: map 94% reduce 25%
2021-11-16 15:55:49,352 INFO mapreduce.Job: map 97% reduce 25%
2021-11-16 15:55:55,399 INFO mapreduce.Job: map 100% reduce 25%
2021-11-16 15:55:58,420 INFO mapreduce.Job: map 100% reduce 39%
2021-11-16 15:56:04,491 INFO mapreduce.Job: map 100% reduce 72%
2021-11-16 15:56:10,558 INFO mapreduce.Job: map 100% reduce 81%
2021-11-16 15:56:16,627 INFO mapreduce.Job: map 100% reduce 91%
2021-11-16 15:56:22,706 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 13: Quá trình chạy task trên testcase 2 của phiên bản 1.0 trong khoảng 116s, số input splits là 4 và reduce task được bắt đầu thực hiện khi một phần map task đã hoàn thành



Hình 14: Một phần kết quả testcase 2 của phiên bản 1.0 thấy được WordCount 1.0 vẫn còn đếm các kí tự không phải từ

Testcase 3

Tại máy chứa Namenode, copy file test3-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

Kích thước input

File test1-0.txt ~ 293MB

File test1-1.txt ~ 210MB

Do testcase khá lớn nên khó có thể thấy toàn bộ input và output.

Browse Directory

/user/pnmth/input

Go!

Show

25

entries

Search:

<div><input type="checkbox"/></div>	<div><div><div></div></div>Permission</div>	<div><div><div></div></div>Owner</div>	<div><div><div></div></div>Group</div>	<div><div><div></div></div>Size</div>	<div><div><div></div></div>Last Modified</div>	<div><div><div></div></div>Replication</div>	<div><div><div></div></div>Block Size</div>	<div><div><div></div></div>Name</div>	<div><div><div></div></div></div>
<div><input type="checkbox"/></div>	<div>-rw-f--f--</div>	<div>pnmth</div>	<div>supergroup</div>	<div>286.97 MB</div>	<div>Nov 16 16:21</div>	<div>3</div>	<div>128 MB</div>	<div>test3-0.txt</div>	<div><div><div></div></div></div>

Hình 15: Minh chứng đã upload file test3-0.txt trên máy user pnmth

Browse Directory

/user/thy/input

Go!

Show

25

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		thy		supergroup		205.77 MB		Nov 16 16:28		3		128 MB		test3-1.txt	

Hình 16: Minh chứng đã upload file test3-1.txt trên máy user thy

```
c:\hadoop-3.3.1\bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 16:30:04,029 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 16:30:04,791 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pnmth/.staging/job_1637052105009_0002
2021-11-16 16:30:05,183 INFO Input.FileInputFormat: Total input files to process : 2
2021-11-16 16:30:05,369 INFO mapreduce.JobSubmitter: number of splits:5
2021-11-16 16:30:06,010 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637052105009_0002
2021-11-16 16:30:06,011 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 16:30:06,184 INFO conf.Configuration: resource-types.xml not found
2021-11-16 16:30:06,185 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 16:30:06,308 INFO impl.YarnClientImpl: Submitted application application_1637052105009_0002
2021-11-16 16:30:06,371 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SH4IL:8088/proxy/application_1637052105009_0002/
2021-11-16 16:30:06,372 INFO mapreduce.Job: Running job: job_1637052105009_0002
2021-11-16 16:30:37,821 INFO mapreduce.Job: Job job_1637052105009_0002 running in uber mode : false
2021-11-16 16:30:37,822 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 16:30:59,529 INFO mapreduce.Job: map 2% reduce 0%
2021-11-16 16:31:02,690 INFO mapreduce.Job: map 4% reduce 0%
2021-11-16 16:31:05,831 INFO mapreduce.Job: map 5% reduce 0%
2021-11-16 16:31:08,037 INFO mapreduce.Job: map 7% reduce 0%
2021-11-16 16:31:13,255 INFO mapreduce.Job: map 8% reduce 0%
2021-11-16 16:31:14,343 INFO mapreduce.Job: map 18% reduce 0%
2021-11-16 16:31:16,424 INFO mapreduce.Job: map 19% reduce 0%
2021-11-16 16:31:19,561 INFO mapreduce.Job: map 21% reduce 0%
2021-11-16 16:31:20,573 INFO mapreduce.Job: map 26% reduce 0%
2021-11-16 16:31:22,640 INFO mapreduce.Job: map 28% reduce 0%
2021-11-16 16:31:26,752 INFO mapreduce.Job: map 33% reduce 0%
2021-11-16 16:31:32,903 INFO mapreduce.Job: map 34% reduce 0%
2021-11-16 16:31:33,973 INFO mapreduce.Job: map 39% reduce 0%
2021-11-16 16:31:36,063 INFO mapreduce.Job: map 41% reduce 0%
2021-11-16 16:31:43,105 INFO mapreduce.Job: map 49% reduce 0%
2021-11-16 16:31:46,207 INFO mapreduce.Job: map 50% reduce 0%
2021-11-16 16:31:47,224 INFO mapreduce.Job: map 52% reduce 0%
2021-11-16 16:31:48,257 INFO mapreduce.Job: map 53% reduce 0%
2021-11-16 16:31:52,513 INFO mapreduce.Job: map 55% reduce 0%
2021-11-16 16:31:53,659 INFO mapreduce.Job: map 56% reduce 0%
2021-11-16 16:31:58,174 INFO mapreduce.Job: map 57% reduce 0%
2021-11-16 16:32:00,316 INFO mapreduce.Job: map 60% reduce 0%
2021-11-16 16:32:06,600 INFO mapreduce.Job: map 62% reduce 0%
2021-11-16 16:32:08,752 INFO mapreduce.Job: map 64% reduce 0%
2021-11-16 16:32:13,926 INFO mapreduce.Job: map 66% reduce 7%
2021-11-16 16:32:15,013 INFO mapreduce.Job: map 68% reduce 7%
2021-11-16 16:32:21,226 INFO mapreduce.Job: map 70% reduce 7%
2021-11-16 16:32:22,289 INFO mapreduce.Job: map 71% reduce 7%
2021-11-16 16:32:27,501 INFO mapreduce.Job: map 74% reduce 7%
2021-11-16 16:32:28,505 INFO mapreduce.Job: map 76% reduce 7%
2021-11-16 16:32:28,525 INFO mapreduce.Job: map 76% reduce 7%
2021-11-16 16:32:34,738 INFO mapreduce.Job: map 78% reduce 7%
2021-11-16 16:32:35,788 INFO mapreduce.Job: map 79% reduce 7%
2021-11-16 16:32:39,896 INFO mapreduce.Job: map 80% reduce 13%
2021-11-16 16:32:40,921 INFO mapreduce.Job: map 81% reduce 13%
2021-11-16 16:32:47,101 INFO mapreduce.Job: map 82% reduce 13%
2021-11-16 16:32:49,152 INFO mapreduce.Job: map 83% reduce 13%
2021-11-16 16:32:53,308 INFO mapreduce.Job: map 85% reduce 13%
2021-11-16 16:32:55,365 INFO mapreduce.Job: map 87% reduce 13%
2021-11-16 16:32:59,514 INFO mapreduce.Job: map 88% reduce 13%
2021-11-16 16:33:01,596 INFO mapreduce.Job: map 89% reduce 13%
2021-11-16 16:33:06,760 INFO mapreduce.Job: map 92% reduce 13%
2021-11-16 16:33:07,793 INFO mapreduce.Job: map 93% reduce 13%
2021-11-16 16:33:13,948 INFO mapreduce.Job: map 95% reduce 13%
2021-11-16 16:33:15,007 INFO mapreduce.Job: map 97% reduce 13%
2021-11-16 16:33:20,170 INFO mapreduce.Job: map 98% reduce 13%
2021-11-16 16:33:25,358 INFO mapreduce.Job: map 100% reduce 27%
2021-11-16 16:33:31,446 INFO mapreduce.Job: map 100% reduce 52%
2021-11-16 16:33:37,534 INFO mapreduce.Job: map 100% reduce 69%
2021-11-16 16:33:43,594 INFO mapreduce.Job: map 100% reduce 73%
2021-11-16 16:33:49,661 INFO mapreduce.Job: map 100% reduce 74%
2021-11-16 16:33:55,715 INFO mapreduce.Job: map 100% reduce 75%
2021-11-16 16:34:01,770 INFO mapreduce.Job: map 100% reduce 77%
2021-11-16 16:34:07,851 INFO mapreduce.Job: map 100% reduce 82%
2021-11-16 16:34:13,909 INFO mapreduce.Job: map 100% reduce 88%
2021-11-16 16:34:19,978 INFO mapreduce.Job: map 100% reduce 93%
2021-11-16 16:34:26,051 INFO mapreduce.Job: map 100% reduce 98%
2021-11-16 16:34:29,093 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 17: Quá trình chạy task trên testcase 3 của phiên bản 1.0 trong khoảng 233s, số input splits là 5 và reduce task được bắt đầu thực hiện khi một phần map task đã hoàn thành


```

for (int i = 0; i < args.length - 1; i += 1) {
    FileInputFormat.addInputPath(job, new Path(args[i]));
}

FileOutputFormat.setOutputPath(job, new Path(args[args.length - 1]));

job.setMapperClass(Map.class);
job.setCombinerClass(Reduce.class);
job.setReducerClass(Reduce.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
return job.waitForCompletion(true) ? 0 : 1;
}

public static class Map extends Mapper<LongWritable, Text, Text,
IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    // Biến caseSensitive = true: phân biệt hoa thường
    // Biến caseSensitive = false: không phân biệt hoa thường
    private boolean caseSensitive = false;
    // Tạo regular expression chỉ lấy từ có kí tự [a..z], [A..Z]
    private static final Pattern PATTERN = Pattern.compile("[a-zA-Z]+");

    protected void setup(Mapper.Context context)
        throws IOException,
        InterruptedException {

        // caseSensitive được gán bằng giá trị biến hệ thống -
        Dwordcount.case.sensitive ở trên command
        // và mặc định nếu biến -Dwordcount.case.sensitive không có giá trị
        thì gán bằng false.
        Configuration config = context.getConfiguration();
        this.caseSensitive = config.getBoolean("wordcount.case.sensitive",
false);
    }

    public void map(LongWritable offset, Text inputLine, Context output)
        throws IOException, InterruptedException {
        String line = inputLine.toString();

        // Nếu caseSensitive = false thì biến tất cả kí tự của line thành chu
        thuong
        if (!caseSensitive) {
            line = line.toLowerCase();
        }

        // Dùng biến PATTERN để lấy ra các từ match với regular expression

```



```

    Matcher m = PATTERN.matcher(line);
    // Sử dụng Macher.find() để truy xuất từng từ trong list các matcher
    while (m.find()) {
        String word = m.group();

        Text outputWord = new Text(word);
        output.write(outputWord, one);
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text,
IntWritable> {
    public void reduce(Text word, Iterable<IntWritable> counts, Context
output)
        throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable count : counts)
            sum += count.get();

        output.write(word, new IntWritable(sum));
    }
}
}

```

b. Quá trình thực thi

Testcase 0

Tại máy chứa Namenode, copy file test0-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-0.txt input
```

Tại máy chứa Datanode 2, copy file test0-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-1.txt input
```

Tại máy chứa Namenode, biên dịch mã nguồn của WordCount phiên bản 2.0 ra file jar

```
javac -classpath jar_files\* -d archive
Hadoop_tutorial\WordCount2\WordCount.java
jar -cvf WordCount.jar -C archive\ .
```

Tại máy chứa Namenode, chạy MapReduce

```
bin\yarn jar WordCount.jar WordCount /user/pnmth/input
/user/thy/input output
```

Tại máy chứa Namenode, xem kết quả trong thư mục chứa output

```
bin\hdfs dfs -cat output/*
```

```
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 21:28:51,426 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceMan
2021-11-16 21:28:51,903 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /t
pnmth/.staging/job_1637072212719_0003
2021-11-16 21:28:52,910 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 21:28:53,568 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 21:28:55,362 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637072212719_
2021-11-16 21:28:55,363 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 21:28:55,530 INFO conf.Configuration: resource-types.xml not found
2021-11-16 21:28:55,531 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 21:28:55,618 INFO impl.YarnClientImpl: Submitted application application_1637072212719_
2021-11-16 21:28:55,669 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/
072212719_0003/
2021-11-16 21:28:55,670 INFO mapreduce.Job: Running job: job_1637072212719_0003
2021-11-16 21:29:11,196 INFO mapreduce.Job: Job job_1637072212719_0003 running in uber mode : false
2021-11-16 21:29:11,197 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 21:29:17,299 INFO mapreduce.Job: map 100% reduce 0%
2021-11-16 21:29:23,370 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 19: Quá trình chạy task trên testcase 0 của phiên bản 2.0 trong khoảng 12s, nhanh hơn phiên bản 1.0 vài giây

```
c:\hadoop-3.3.1>bin\hdfs dfs -cat output/*
a 1
and 1
example 1
hadoop 4
have 1
i 1
install 1
installed 1
is 2
mapreduce 2
multiple 1
my 1
name 2
node 1
on 1
run 1
successfully 1
the 1
thy 2
wordcount 1
```

Hình 20: Kết quả testcase 0 của phiên bản 2.0 không còn xuất hiện kí tự đặc biệt nữa

Testcase 1

Tại máy chứa Namenode, copy file test1-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

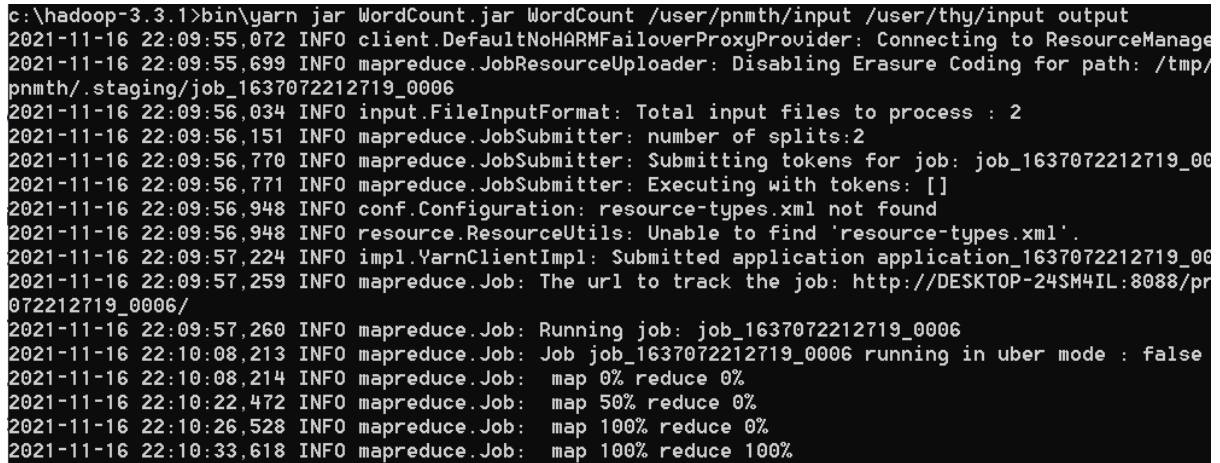
```
bin\hdfs dfs -copyFromLocal c:/test1/test1-0.txt input
```

Tại máy chứa Datanode 2, copy file test1-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test1/test1-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.



```
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 22:09:55,072 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager
2021-11-16 22:09:55,699 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/pnmth/.staging/job_1637072212719_0006
2021-11-16 22:09:56,034 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 22:09:56,151 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 22:09:56,770 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637072212719_0006
2021-11-16 22:09:56,771 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 22:09:56,948 INFO conf.Configuration: resource-types.xml not found
2021-11-16 22:09:56,948 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 22:09:57,224 INFO impl.YarnClientImpl: Submitted application application_1637072212719_0006
2021-11-16 22:09:57,259 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/wordcount/job_1637072212719_0006/
2021-11-16 22:09:57,260 INFO mapreduce.Job: Running job: job_1637072212719_0006
2021-11-16 22:10:08,213 INFO mapreduce.Job: Job job_1637072212719_0006 running in uber mode : false
2021-11-16 22:10:08,214 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-16 22:10:22,472 INFO mapreduce.Job:  map 50% reduce 0%
2021-11-16 22:10:26,528 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-16 22:10:33,618 INFO mapreduce.Job:  map 100% reduce 100%
```

Hình 21: Quá trình chạy task trên testcase 2 của phiên bản 2.0 trong khoảng 79s, nhanh hơn phiên bản 1.0

Do kích thước testcase 1 khá lớn nên không quan sát rõ được sự khác biệt output của phiên bản 2.0 so với phiên bản 1.0

Testcase 2

Tại máy chứa Namenode, copy file test2-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

```
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-17 09:41:11,546 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceMa
2021-11-17 09:41:13,238 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /
2021-11-17 09:41:14,112 INFO input.FileInputFormat: Total input files to process : 2
2021-11-17 09:41:14,415 INFO mapreduce.JobSubmitter: number of splits:4
2021-11-17 09:41:15,040 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_163711493387
2021-11-17 09:41:15,041 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-17 09:41:15,538 INFO conf.Configuration: resource-types.xml not found
2021-11-17 09:41:15,539 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-17 09:41:16,468 INFO impl.YarnClientImpl: Submitted application application_163711493387
2021-11-17 09:41:16,531 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:808
2021-11-17 09:41:16,532 INFO mapreduce.Job: Running job: job_1637114933876_0001
2021-11-17 09:41:52,988 INFO mapreduce.Job: Job job_1637114933876_0001 running in uber mode : fa
2021-11-17 09:41:52,993 INFO mapreduce.Job: map 0% reduce 0%
2021-11-17 09:42:19,677 INFO mapreduce.Job: map 9% reduce 0%
2021-11-17 09:42:20,696 INFO mapreduce.Job: map 20% reduce 0%
2021-11-17 09:42:21,708 INFO mapreduce.Job: map 32% reduce 0%
2021-11-17 09:42:25,761 INFO mapreduce.Job: map 36% reduce 0%
2021-11-17 09:42:26,781 INFO mapreduce.Job: map 41% reduce 0%
2021-11-17 09:42:31,665 INFO mapreduce.Job: map 68% reduce 0%
2021-11-17 09:42:37,778 INFO mapreduce.Job: map 73% reduce 0%
2021-11-17 09:42:46,683 INFO mapreduce.Job: map 78% reduce 0%
2021-11-17 09:42:49,721 INFO mapreduce.Job: map 81% reduce 0%
2021-11-17 09:42:52,753 INFO mapreduce.Job: map 81% reduce 17%
2021-11-17 09:42:55,861 INFO mapreduce.Job: map 89% reduce 17%
2021-11-17 09:42:56,887 INFO mapreduce.Job: map 92% reduce 17%
2021-11-17 09:42:58,913 INFO mapreduce.Job: map 92% reduce 25%
2021-11-17 09:43:01,941 INFO mapreduce.Job: map 100% reduce 25%
2021-11-17 09:43:04,999 INFO mapreduce.Job: map 100% reduce 98%
2021-11-17 09:43:11,085 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 22: Quá trình chạy task trên testcase 2 của phiên bản 2.0 trong khoảng 27s, nhanh hơn phiên bản 1.0 vài giây

```
zzzonked 2
zzzphar 2
zzzquil 3
zzztyu 2
zzzxxc 2
zzzz 12
zzzzp 1
zzzzz 8
zzzzziilil111 2
zzzzzz 7
zzzzzzz 3
zzzzzzzz 8
zzzzzzzzzz 3
zzzzzzzzzzeee 1
zzzzzzzzzzzz 2
zzzzzzzzzzzzap 1
zzzzzzzzzzzzzz 4
zzzzzzzzzzzzzzz 3
zzzzzzzzzzzzzzzz 2
zzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzz 10
zzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzz 3
zzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzzzzzzzzzz 1
zzzzzzzzzzzzzzzzzzzzzzzzzzzz 1
```

Hình 23: Kết quả testcase 2 của phiên bản 2.0 không còn xuất hiện kí tự đặc biệt nữa

Testcase 3

Tại máy chứa Namenode, copy file test3-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

```
c:\hadoop-3.3.1\bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output
2021-11-16 17:20:02.913 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 17:20:03.999 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pnmth/.staging/job_1637052105009
2021-11-16 17:20:04.627 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 17:20:05.000 INFO mapreduce.JobSubmitter: number of splits:5
2021-11-16 17:20:08.064 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637052105009_0003
2021-11-16 17:20:08.064 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 17:20:08.369 INFO conf.Configuration: resource-types.xml not found
2021-11-16 17:20:08.370 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 17:20:08.513 INFO impl.YarnClientImpl: Submitted application application_1637052105009_0003
2021-11-16 17:20:08.590 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_1637052105009_0003/
2021-11-16 17:20:08.593 INFO mapreduce.Job: Running job: job_1637052105009_0003
2021-11-16 17:20:42.954 INFO mapreduce.Job: Job job_1637052105009_0003 running in uber mode : false
2021-11-16 17:20:42.956 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 17:21:15.314 INFO mapreduce.Job: map 4% reduce 0%
2021-11-16 17:21:23.921 INFO mapreduce.Job: map 5% reduce 0%
2021-11-16 17:21:25.082 INFO mapreduce.Job: map 17% reduce 0%
2021-11-16 17:21:26.191 INFO mapreduce.Job: map 21% reduce 0%
2021-11-16 17:21:30.412 INFO mapreduce.Job: map 25% reduce 0%
2021-11-16 17:21:39.814 INFO mapreduce.Job: map 35% reduce 0%
2021-11-16 17:21:43.039 INFO mapreduce.Job: map 42% reduce 0%
2021-11-16 17:21:48.108 INFO mapreduce.Job: map 46% reduce 0%
2021-11-16 17:21:51.192 INFO mapreduce.Job: map 47% reduce 0%
2021-11-16 17:21:52.235 INFO mapreduce.Job: map 49% reduce 0%
2021-11-16 17:21:53.269 INFO mapreduce.Job: map 53% reduce 0%
2021-11-16 17:21:57.342 INFO mapreduce.Job: map 54% reduce 0%
2021-11-16 17:21:58.368 INFO mapreduce.Job: map 55% reduce 0%
2021-11-16 17:22:03.105 INFO mapreduce.Job: map 56% reduce 0%
2021-11-16 17:22:05.196 INFO mapreduce.Job: map 59% reduce 0%
2021-11-16 17:22:07.336 INFO mapreduce.Job: map 61% reduce 0%
2021-11-16 17:22:10.535 INFO mapreduce.Job: map 67% reduce 0%
2021-11-16 17:22:11.634 INFO mapreduce.Job: map 69% reduce 0%
2021-11-16 17:22:12.773 INFO mapreduce.Job: map 70% reduce 0%
2021-11-16 17:22:18.190 INFO mapreduce.Job: map 73% reduce 0%
2021-11-16 17:22:20.304 INFO mapreduce.Job: map 74% reduce 0%
2021-11-16 17:22:25.541 INFO mapreduce.Job: map 75% reduce 0%
2021-11-16 17:22:27.716 INFO mapreduce.Job: map 77% reduce 13%
2021-11-16 17:22:31.980 INFO mapreduce.Job: map 78% reduce 13%
2021-11-16 17:22:33.006 INFO mapreduce.Job: map 80% reduce 13%
2021-11-16 17:22:39.288 INFO mapreduce.Job: map 87% reduce 13%
2021-11-16 17:22:41.428 INFO mapreduce.Job: map 87% reduce 20%
2021-11-16 17:22:44.552 INFO mapreduce.Job: map 100% reduce 20%
2021-11-16 17:22:47.597 INFO mapreduce.Job: map 100% reduce 81%
2021-11-16 17:22:53.652 INFO mapreduce.Job: map 100% reduce 89%
2021-11-16 17:22:59.708 INFO mapreduce.Job: map 100% reduce 91%
2021-11-16 17:23:05.782 INFO mapreduce.Job: map 100% reduce 94%
2021-11-16 17:23:11.848 INFO mapreduce.Job: map 100% reduce 96%
2021-11-16 17:23:17.908 INFO mapreduce.Job: map 100% reduce 98%
2021-11-16 17:23:23.956 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 24: Quá trình chạy task trên testcase 3 của phiên bản 2.0 trong khoảng 182s, nhanh hơn so với phiên bản 1.0 do không xử lý các ký tự không phải là từ và không phân biệt hoa thường


```

public int run(String[] args) throws Exception {
    Configuration conf = new Configuration();

    // isSkip = true: có phần skip stop_words.
    boolean isSkip = false;

    // Biến dùng để lưu lại đường dẫn của file stop_words.txt.
    String stopWordPath = "";
    // Biến để giữ lại index của skip trong args.
    int indexSkip = 0;
    for (int i = 0; i < args.length; i += 1) {
        // Lấy argument nằm sau -skip là đường dẫn của stop_words.txt.
        if ("-skip".equals(args[i])) {
            isSkip = true;
            indexSkip = i;
            i += 1;
            logger.info("Path of stop_words.txt: " + args[i]);
            stopWordPath = args[i];
        }
    }
    Job job = new Job(conf, "wordcount");

    // Gán biến hệ thống "wordcount.skip.patterns" = isSkip.
    job.getConfiguration().setBoolean("wordcount.skip.patterns", isSkip);

    // Nếu có đường dẫn stop_word thì gán biến hệ thống
    "wordcount.cache.file" = stopWordPath.
    if (stopWordPath.length() > 0)
        job.getConfiguration().setStrings("wordcount.cache.file",
stopWordPath);
    job.setJarByClass(this.getClass());

    // Nếu indexSkip = 0 có nghĩa là người dùng không nhập đường dẫn
    stop_words.txt,
    // nên gán indexSkip = args.length để xử lý đường dẫn input output
    if (indexSkip == 0)
        indexSkip = args.length;
    for (int i = 0; i < indexSkip - 1; i += 1) {
        FileInputFormat.addInputPath(job, new Path(args[i]));
    }
    FileOutputFormat.setOutputPath(job, new Path(args[indexSkip - 1]));

    job.setMapperClass(Map.class);
    job.setCombinerClass(Reduce.class);
    job.setReducerClass(Reduce.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

```

```

        return job.waitForCompletion(true) ? 0 : 1;
    }

    public static class Map extends Mapper<LongWritable, Text, Text,
IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private boolean caseSensitive = false;

        // Biến chứa các từ trong stop_words.txt: ở dạng hashset (hash table)
        private Set<String> skippedPatterns = new HashSet<String>();
        private static final Pattern PATTERN = Pattern.compile("[a-zA-Z]+");

        protected void setup(Mapper.Context context)
            throws IOException,
            InterruptedException {
            Configuration config = context.getConfiguration();
            // Gán biến caseSensitive bằng biến hệ thống wordcount.case.sensitive,
            // mặc định là false
            this.caseSensitive = config.getBoolean("wordcount.case.sensitive",
            false);

            // Lấy biến hệ thống 'wordcount.skip.patterns', mặc định là false
            // Nếu có skip patterns thì lấy giá trị đường dẫn stop_words.txt từ
            "wordcount.cache.file"
            if (config.getBoolean("wordcount.skip.patterns", false)) {
                String[] path = config.getStrings("wordcount.cache.file", "");
                parseSkipFile(path[0]);
            }
        }

        // Hàm có đầu vào là đường dẫn stop_words.txt
        // Chức năng: Parse các từ trong file stop_words.txt bỏ vào
        skippedPatterns,
        // sử dụng FileSystem để lấy file từ HDFS và đọc bằng BufferedReader
        private void parseSkipFile(String path) {
            try {
                Path pt = new Path(path);
                FileSystem fs = FileSystem.get(new Configuration());
                BufferedReader br = new BufferedReader(new
InputStreamReader(fs.open(pt)));
                String line = br.readLine();
                while (line != null){
                    skippedPatterns.add(line);
                    line = br.readLine();
                }
            } catch (IOException e) {
                System.err.println("Caught exception while parsing stop_words.txt
file: '" + StringUtils.stringifyException(e));
            }
        }
    }

```



```

    }
}

public void map(LongWritable offset, Text inputLine, Context output)
    throws IOException, InterruptedException {
    String line = inputLine.toString();
    if (!caseSensitive) {
        line = line.toLowerCase();
    }

    Matcher m = PATTERN.matcher(line);
    while (m.find()) {
        String word = m.group();
        // Nếu word nằm trong skippedPatterns thì bỏ qua
        if (skippedPatterns.contains(word))
            continue;

        Text outputWord = new Text(word);
        output.write(outputWord, one);
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text,
IntWritable> {
    public void reduce(Text word, Iterable<IntWritable> counts, Context
output)
        throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable count : counts) {
            sum += count.get();
        }
        output.write(word, new IntWritable(sum));
    }
}
}

```

b. Quá trình thực thi

Testcase 0

Tại máy chứa Namenode, copy file test0-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-0.txt input
```

Tại máy chứa Datanode 2, copy file test0-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test0/test0-1.txt input
```

Tại máy chứa Namenode, biên dịch mã nguồn của WordCount phiên bản 3.0 ra file jar

```
javac -classpath jar_files\* -d archive
Hadoop_tutorial\WordCount2\WordCount.java
jar -cvf WordCount.jar -C archive\ .
```

Tại máy chứa Namenode, chạy MapReduce

```
bin\yarn jar WordCount.jar WordCount /user/pnmth/input
/user/thy/input output -skip cache/stop_words.txt
```

Tại máy chứa Namenode, xem kết quả trong thư mục chứa output

```
bin\hdfs dfs -cat output/*
```

Nội dung file stop_words.txt

a
an
and
but
is
or
the
to
.
,

```
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output -skip cache/stop_words.txt
2021-11-16 21:34:34,019 INFO WordCount: Path of stop_words.txt: cache/stop_words.txt
2021-11-16 21:34:34,862 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 21:34:35,326 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2021-11-16 21:34:35,349 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
pnmth/.staging/job_1637072212719_0004
2021-11-16 21:34:35,684 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 21:34:35,838 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 21:34:38,553 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637072212719_0004
2021-11-16 21:34:38,553 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 21:34:38,730 INFO conf.Configuration: resource-types.xml not found
2021-11-16 21:34:38,730 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 21:34:38,793 INFO impl.YarnClientImpl: Submitted application application_1637072212719_0004
2021-11-16 21:34:38,839 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_1637
072212719_0004/
2021-11-16 21:34:38,840 INFO mapreduce.Job: Running job: job_1637072212719_0004
2021-11-16 21:34:54,448 INFO mapreduce.Job: Job job_1637072212719_0004 running in uber mode : false
2021-11-16 21:34:54,449 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 21:35:00,562 INFO mapreduce.Job: map 100% reduce 0%
2021-11-16 21:35:06,636 INFO mapreduce.Job: map 100% reduce 100%
```

Hình 26: Quá trình chạy task trên testcase 0 của phiên bản 3.0 trong khoảng 12s, nhanh hơn phiên bản 1.0 vài giây

```
c:\hadoop-3.3.1>bin\hdfs dfs -cat output/*
example 1
hadoop 4
have 1
i 1
install 1
installed 1
mapreduce 2
multiple 1
my 1
name 2
node 1
on 1
run 1
successfully 1
thy 2
wordcount 1
```

Hình 27: Kết quả testcase 0 của phiên bản 3.0 đã không còn đếm các từ trong stop_words.txt

Testcase 1

Tại máy chứa Namenode, copy file test1-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test1/test1-0.txt input
```

Tại máy chứa Datanode 2, copy file test1-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test1/test1-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

```

c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input o
2021-11-16 22:21:27,160 INFO WordCount: Path of stop_words.txt: cache/stop_words.txt
2021-11-16 22:21:27,995 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to Res
2021-11-16 22:21:28,541 WARN mapreduce.JobResourceUploader: Hadoop command-line option pa
the Tool interface and execute your application with ToolRunner to remedy this.
2021-11-16 22:21:28,563 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for
pnmth/.staging/job_1637072212719_0009
2021-11-16 22:21:28,860 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 22:21:29,039 INFO mapreduce.JobSubmitter: number of splits:2
2021-11-16 22:21:29,968 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16370
2021-11-16 22:21:29,969 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 22:21:30,139 INFO conf.Configuration: resource-types.xml not found
2021-11-16 22:21:30,140 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'
2021-11-16 22:21:30,201 INFO impl.YarnClientImpl: Submitted application application_16370
2021-11-16 22:21:30,237 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24S
072212719_0009/
2021-11-16 22:21:30,237 INFO mapreduce.Job: Running job: job_1637072212719_0009
2021-11-16 22:21:42,205 INFO mapreduce.Job: Job job_1637072212719_0009 running in uber m
2021-11-16 22:21:42,206 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 22:21:56,440 INFO mapreduce.Job: map 50% reduce 0%
2021-11-16 22:22:00,503 INFO mapreduce.Job: map 100% reduce 0%
2021-11-16 22:22:05,570 INFO mapreduce.Job: map 100% reduce 100%

```

Hình 28: Quá trình chạy task trên testcase 1 của phiên bản 3.0 trong khoảng 23s

Do kích thước testcase 1 khá lớn nên không quan sát rõ được sự khác biệt output của phiên bản 3.0 so với phiên bản 1.0

Testcase 2

Tại máy chứa Namenode, copy file test2-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test2/test2-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

```

c:\hadoop-3.3.1>
c:\hadoop-3.3.1>bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output -skip cache/stop_words.txt
2021-11-17 09:48:58,619 INFO WordCount: Path of stop_words.txt: cache/stop_words.txt
2021-11-17 09:48:59,507 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-17 09:49:00,071 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement this
medy this.
2021-11-17 09:49:00,093 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/p
2021-11-17 09:49:00,435 INFO input.FileInputFormat: Total input files to process : 2
2021-11-17 09:49:00,607 INFO mapreduce.JobSubmitter: number of splits:4
2021-11-17 09:49:02,062 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637114933876_0002
2021-11-17 09:49:02,062 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-17 09:49:02,224 INFO conf.Configuration: resource-types.xml not found
2021-11-17 09:49:02,224 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-17 09:49:02,305 INFO impl.YarnClientImpl: Submitted application application_1637114933876_0002
2021-11-17 09:49:02,350 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_16371
2021-11-17 09:49:02,351 INFO mapreduce.Job: Running job: job_1637114933876_0002
2021-11-17 09:49:27,220 INFO mapreduce.Job: Job job_1637114933876_0002 running in uber mode : false
2021-11-17 09:49:27,221 INFO mapreduce.Job: map 0% reduce 0%
2021-11-17 09:49:46,752 INFO mapreduce.Job: map 6% reduce 0%
2021-11-17 09:49:47,766 INFO mapreduce.Job: map 13% reduce 0%
2021-11-17 09:49:50,861 INFO mapreduce.Job: map 30% reduce 0%
2021-11-17 09:49:52,900 INFO mapreduce.Job: map 33% reduce 0%
2021-11-17 09:49:57,898 INFO mapreduce.Job: map 68% reduce 0%
2021-11-17 09:49:58,908 INFO mapreduce.Job: map 71% reduce 0%
2021-11-17 09:49:59,927 INFO mapreduce.Job: map 76% reduce 0%
2021-11-17 09:50:05,040 INFO mapreduce.Job: map 79% reduce 0%
2021-11-17 09:50:06,051 INFO mapreduce.Job: map 82% reduce 0%
2021-11-17 09:50:09,094 INFO mapreduce.Job: map 91% reduce 0%
2021-11-17 09:50:11,114 INFO mapreduce.Job: map 100% reduce 0%
2021-11-17 09:50:15,159 INFO mapreduce.Job: map 100% reduce 100%

```

Hình 29: Quá trình chạy task trên testcase 2 của phiên bản 3.0 trong khoảng 48s, nhanh hơn phiên bản 2.0

Do kích thước testcase 2 khá lớn nên không quan sát rõ được sự khác biệt output của phiên bản 3.0 so với phiên bản 2.0

Testcase 3

Tại máy chứa Namenode, copy file test3-0.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-0.txt input
```

Tại máy chứa Datanode 2, copy file test2-1.txt vào thư mục input

```
bin\hdfs dfs -mkdir -p input
```

```
bin\hdfs dfs -copyFromLocal c:/test3/test3-1.txt input
```

Thực hiện các câu lệnh còn lại giống Testcase 0.

```

c:\hadoop-3.3.1\bin\yarn jar WordCount.jar WordCount /user/pnmth/input /user/thy/input output -skip cache/stop_words.txt
2021-11-16 17:36:08,743 INFO WordCount: Path of stop_words.txt: cache/stop_words.txt
2021-11-16 17:36:10,285 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 17:36:11,177 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface
  medy this.
2021-11-16 17:36:11,206 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pnmth/.staging/job_1637052105009_0005
2021-11-16 17:36:11,643 INFO input.FileInputFormat: Total input files to process : 2
2021-11-16 17:36:11,805 INFO mapreduce.JobSubmitter: number of splits:5
2021-11-16 17:36:12,252 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637052105009_0005
2021-11-16 17:36:12,253 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 17:36:12,541 INFO conf.Configuration: resource-types.xml not found
2021-11-16 17:36:12,542 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 17:36:12,660 INFO impl.YarnClientImpl: Submitted application application_1637052105009_0005
2021-11-16 17:36:12,724 INFO mapreduce.Job: The url to track the job: http://DESKTOP-24SM4IL:8088/proxy/application_1637052105009_0005/
2021-11-16 17:36:12,725 INFO mapreduce.Job: Running job: job_1637052105009_0005
2021-11-16 17:36:47,373 INFO mapreduce.Job: Job job_1637052105009_0005 running in uber mode : false
2021-11-16 17:36:47,374 INFO mapreduce.Job: map 0% reduce 0%
2021-11-16 17:37:13,788 INFO mapreduce.Job: map 3% reduce 0%
2021-11-16 17:37:16,959 INFO mapreduce.Job: map 5% reduce 0%
2021-11-16 17:37:19,085 INFO mapreduce.Job: map 8% reduce 0%
2021-11-16 17:37:20,150 INFO mapreduce.Job: map 26% reduce 0%
2021-11-16 17:37:23,382 INFO mapreduce.Job: map 28% reduce 0%
2021-11-16 17:37:27,677 INFO mapreduce.Job: map 32% reduce 0%
2021-11-16 17:37:31,877 INFO mapreduce.Job: map 39% reduce 0%
2021-11-16 17:37:36,542 INFO mapreduce.Job: map 49% reduce 0%
2021-11-16 17:37:39,661 INFO mapreduce.Job: map 50% reduce 0%
2021-11-16 17:37:40,703 INFO mapreduce.Job: map 53% reduce 0%
2021-11-16 17:37:43,829 INFO mapreduce.Job: map 55% reduce 0%
2021-11-16 17:37:45,863 INFO mapreduce.Job: map 57% reduce 0%
2021-11-16 17:37:46,933 INFO mapreduce.Job: map 60% reduce 0%
2021-11-16 17:37:53,674 INFO mapreduce.Job: map 71% reduce 0%
2021-11-16 17:37:57,781 INFO mapreduce.Job: map 72% reduce 0%
2021-11-16 17:38:00,876 INFO mapreduce.Job: map 73% reduce 0%
2021-11-16 17:38:06,059 INFO mapreduce.Job: map 75% reduce 0%
2021-11-16 17:38:08,194 INFO mapreduce.Job: map 77% reduce 0%
2021-11-16 17:38:09,249 INFO mapreduce.Job: map 79% reduce 0%
2021-11-16 17:38:11,354 INFO mapreduce.Job: map 79% reduce 13%
2021-11-16 17:38:12,397 INFO mapreduce.Job: map 80% reduce 13%
2021-11-16 17:38:19,949 INFO mapreduce.Job: map 84% reduce 13%
2021-11-16 17:38:23,125 INFO mapreduce.Job: map 90% reduce 13%
2021-11-16 17:38:24,170 INFO mapreduce.Job: map 100% reduce 27%
2021-11-16 17:38:30,244 INFO mapreduce.Job: map 100% reduce 92%
2021-11-16 17:38:36,301 INFO mapreduce.Job: map 100% reduce 100%

```

Hình 30: Quá trình chạy task trên testcase 3 của phiên bản 3.0 trong khoảng 109s, nhanh hơn so với phiên bản 2.0 do bỏ qua các từ trong stop_words.txt

Do testcase 3 quá lớn nên không thể thấy rõ được toàn bộ output nên kết quả testcase 3 có vẻ giống với output của phiên bản 2.0.

2.2. Chương trình MapReduce mức 2

Phần này trình bày cách thực thi và mã nguồn của 3 chương trình wordcount sử dụng thư viện MRJob trên hệ điều hành Ubuntu.

Phần chuẩn bị chung để chạy 3 chương trình:

- Thiết lập biến môi trường HADOOP_STREAMING

```
HADOOP_HOME=~/.hadoop-3.3.1
```

```
HADOOP_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-3.3.1.jar
```

- Bổ sung vào file mapred-site.xml 3 thuộc tính:

```

<property>
  <name>yarn.app.mapreduce.am.env</name>
  <value>HADOOP_MAPRED_HOME=$HADOOP_MAPRED_HOME</value>
</property>
<property>
  <name>mapreduce.map.env</name>

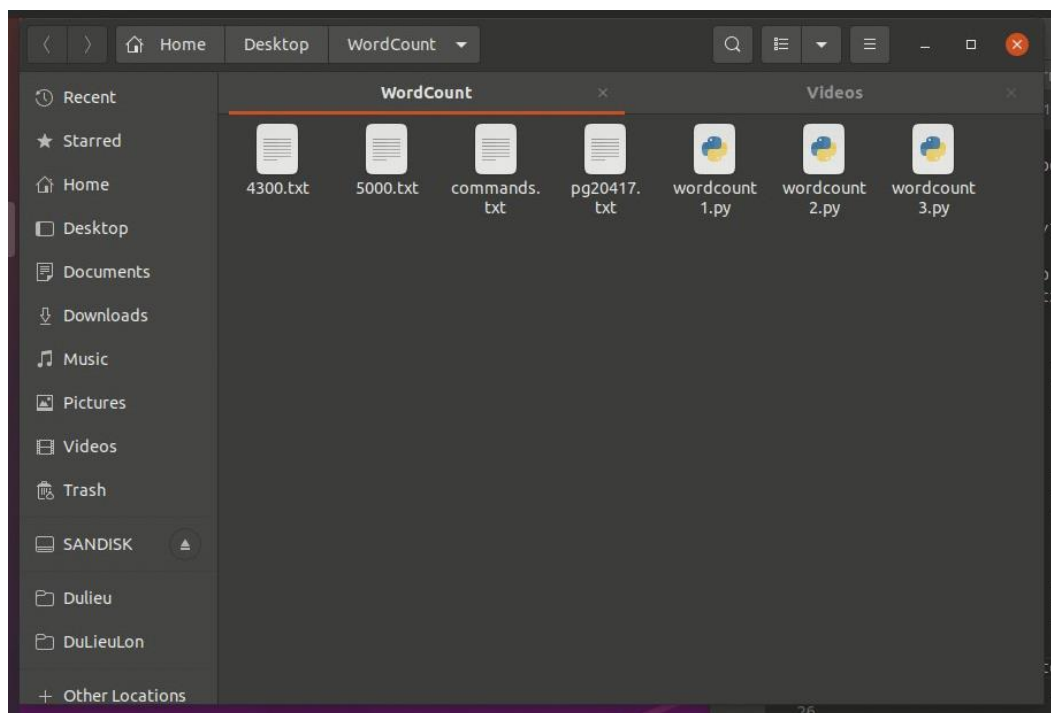
```

```

    <value>HADOOP_MAPRED_HOME=$HADOOP_MAPRED_HOME</value>
</property>
<property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_MAPRED_HOME</value>
</property>

```

- Tạo folder WordCount ở local tại vị trí ~/Desktop/, đây là nơi chứa các file mã nguồn python, các file text input và gọi lệnh thực thi chương trình.



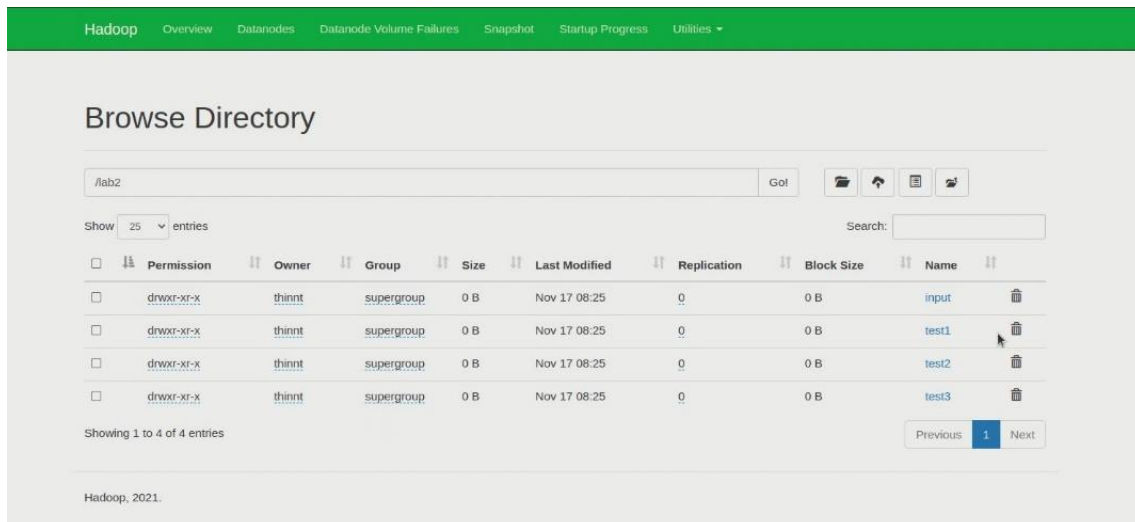
Hình 31. Thư mục WordCount chứa file python và text.

- Tạo thư mục /lab2/input, /lab2/test1, /lab2/test2, /lab2/test3 trên HDFS là nơi chứa input và chứa output thực thi chương trình:

```

hadoop fs -mkdir -p /lab2/input /lab2/test1 /lab2/test2
/ lab2/test 3

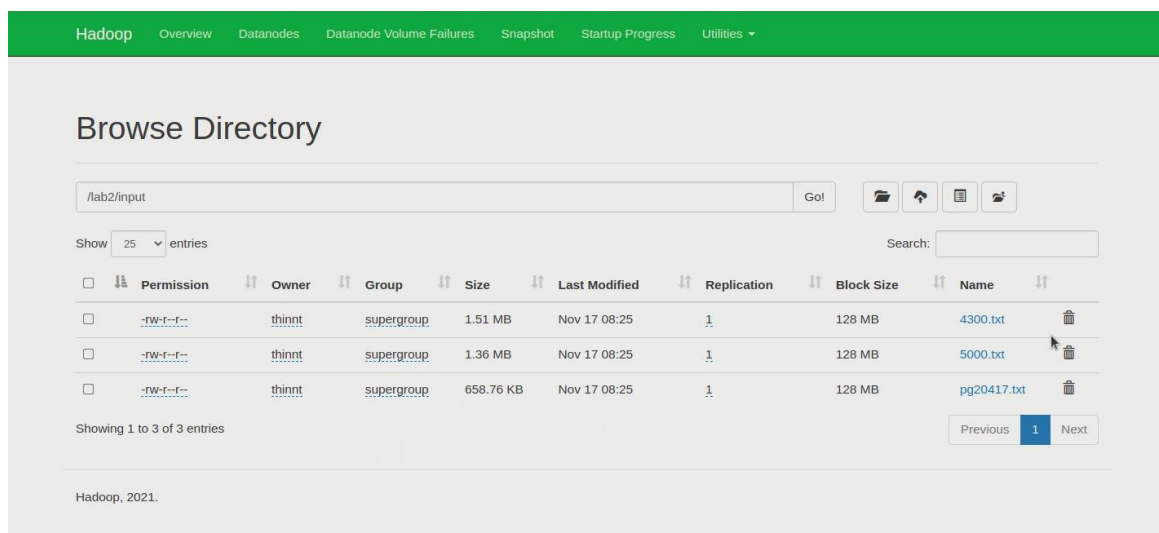
```



Hình 32. Tạo thư mục chứa input và output trên HDFS

- Copy cả 3 file text chứa nội dung của 3 cuốn sách từ local vào HDFS:

```
hadoop fs -put 4300.txt 5000.txt pg20417.txt /lab2/input
```



Hình 33. Copy 3 file input lên HDFS

2.2.1. Trường hợp đếm phân biệt hoa thường

a. Mã nguồn và giải thích

Chương trình MapReduce với MRJob sẽ được viết trong một file script bằng ngôn ngữ Python. Công việc (job) cần chạy được định nghĩa bằng một class kế thừa lớp MRJob có sẵn và override lại một số phương thức như mapper(), combiner(), reducer(). Một công việc có thể bao gồm nhiều bước (step), tại mỗi bước bao gồm mapper, combiner và reducer. Tùy vào mục đích của công việc mà sẽ thiết kế mỗi bước có đủ 3 cả mapper, combiner, reducer hoặc có ít nhất là một trong ba cái.

Mã nguồn thực hiện đếm từ có phân biệt hoa thường được trình bày và giải thích chi tiết bên dưới:


```

from mrjob.job import MRJob
import re

# class MRJWordCount cài đặt job đếm từ,
# kế thừa và override lại hàm của lớp MRJob có sẵn
class MRJWordCount(MRJob):

    # hàm mapper nhận đầu vào là cặp <key, value>,
    # giá trị key ta không cần quan tâm, còn value ở đây chính là
    # một dòng trong văn bản đầu vào,
    # mapper tách ra các từ xuất hiện ở mỗi câu
    def mapper(self, _, line):
        # loại bỏ khoảng trắng đầu và cuối dòng
        line = line.strip()
        # dùng regex tìm các từ chỉ gồm chữ cái, không gồm chữ số và kí tự khác
        words = re.findall(r"[a-zA-Z]+", line)
        # tạo đầu ra là các cặp <key, value> dạng <word, 1> cho mỗi từ trong 1
line
        for word in words:
            yield word, 1

    # nhận đầu vào sinh ra từ mapper,
    # tổng hợp lại số lần xuất hiện của các từ trên 1 line (câu),
    # làm gọn bớt đầu vào cho reducer
    def combiner(self, word, counts):
        yield word, sum(counts)

    # nhận đầu vào sinh ra từ combiner, các cặp <key, value> của cùng 1 từ
    # sẽ về cùng một reducer và tổng hợp lại số lần xuất hiện
    def reducer(self, word, counts):
        yield word, sum(counts)

if __name__ == '__main__':
    MRJWordCount.run()

```

Mã nguồn trên có tham khảo từ document của MRJob.

b. Quá trình thực thi

Testcase 1: Sách mã số 4300

Từ terminal, tại vị trí ~/Desktop/WordCount, thực thi chương trình wordcount1.py bằng lệnh:

```
python wordcount1.py -r hadoop hdfs:///lab2/input/4300.txt --
output-dir=/lab2/test1/4300_output
```

Trong câu lệnh trên:

- wordcount1.py là tên file cài đặt job cần chạy
- -r hadoop để chạy chương trình trên hadoop cluster
- hdfs:///lab2/input/4300.txt là đường dẫn đến file input trên HDFS
- --output-dir=/lab2/test1/4300_output là đường dẫn đến nơi chứa kết quả thực thi trên HDFS

Quá trình chạy mapreduce:

```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount1.py -r hadoop hdfs:///lab2/inputs/4300.txt --output-dir=/lab2/test1/4300_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar
Creating temp directory /tmp/wordcount1.thinnt.20211117.012630.257177
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012630.257177/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012630.257177/files/
Running step 1 of 1...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar749817684498003169/] [] /tmp/streamjob6824260399320789890.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0001
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0001
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0001
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0001/
Running job: job_1637112291929_0001
Job job_1637112291929_0001 running in uber mode : false
map 0% reduce 0%
map 100% reduce 0%
```

Hình 34. Chạy wordcount1.py với input là sách mã 4300

Kết quả thu được:

```
lab2 > test1 > 4300_output > E part-00000
1 "A" 690
2 "ACTUAL" 1
3 "ADONAT" 2
4 "AEROLITHS" 1
5 "AGREE" 2
6 "AGREEMENT" 1
7 "ALBERTA" 1
8 "ALEXANDER" 2
9 "ALF" 1
10 "ALL" 4
11 "AM" 1
12 "AN" 4
13 "AND" 16
14 "ANNE" 1
15 "ANNOUNCE" 1
16 "ANSWERS" 1
17 "ANY" 3
18 "ANYTHING" 1
19 "APPLEWOMAN" 1
20 "ARCHBISHOP" 2
21 "ARMAGH" 2
22 "ARTANE" 1
23 "ARTIFONI" 1
24 "ARTIUM" 1
25 "AS" 1
26 "ASCII" 2
27 "AT" 1
28 "Aaron" 2
29 "Abaft" 1
30 "Abba" 1
31 "Abbas" 2
32 "Abbey" 9
33 "Abe" 1
34 "Abeakuta" 2
35 "Abeakutic" 1
36 "Able" 1
37 "Abnegation" 1
38 "Aboard" 1
39 "About" 13
```

Hình 35. Kết quả chạy wordcount1.py với input là sách mã 4300

Testcase 2: Sách mã số 5000

Chạy lệnh thực thi:

```
python wordcount1.py -r hadoop hdfs:///lab2/input/5000.txt --output-dir=/lab2/test1/5000_output
```

Quá trình chạy mapreduce:

```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount1.py -r hadoop hdfs:///lab2/input/5000.txt --output-dir=/lab2/test1/5000_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar
Creating temp directory /tmp/wordcount1.thinnt.20211117.012733.348542
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012733.348542/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012733.348542/files/
Running step 1 of 1...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar1142597059205341062/] [] /tmp/streamjob1830498127376715754.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0002
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0002
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0002
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0002/
Running job: job_1637112291929_0002
Job job_1637112291929_0002 running in uber mode : false
  map 0% reduce 0%
  map 100% reduce 0%
```

Hình 36. Chạy wordcount1.py với input là sách mã 5000

Kết quả thu được:

```
lab2 > test1 > 5000_output > part-00000
1  "A" 556
2  "AB" 2
3  "ABOUT" 1
4  "ABOVE" 6
5  "ABSTINENCE" 1
6  "ABUTMENT" 1
7  "ACCIDENT" 1
8  "ACCIDENTS" 2
9  "ACCURACY" 1
10 "ACCURATELY" 1
11 "ACQUIRE" 1
12 "ACQUIRING" 1
13 "ACT" 1
14 "ACTION" 4
15 "AD" 1
16 "ADDED" 1
17 "ADMIRABLE" 2
18 "ADOPT" 1
19 "ADVANCED" 1
20 "ADVANCES" 1
21 "ADVANTAGE" 1
22 "ADVERSARY" 1
23 "AERIAL" 4
24 "Aeolus" 1
25 "Aesop" 1
26 "AFRICA" 1
27 "AGAIN" 1
28 "AGAINST" 2
29 "AGE" 3
30 "AGREE" 1
31 "AGREEMENT" 1
32 "AID" 1
33 "AIR" 3
34 "ALBERTI" 2
35 "ALEPO" 1
36 "ALEXIS" 2
37 "ALL" 8
38 "ALLOYING" 1
39 "ALREADY" 1
```

Hình 37. Kết quả chạy wordcount1.py với input là sách mã 5000

Testcase 3: Sách mã số 20417

Chạy lệnh thực thi:

```
python wordcount1.py -r hadoop hdfs:///lab2/input/pg20417.txt  
--output-dir=/lab2/test1/pg20417_output
```

Quá trình chạy mapreduce:

```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount1.py -r hadoop hdfs:///lab2/in  
ut/pg20417.txt --output-dir=/lab2/test1/pg20417_output  
No configs found; falling back on auto-configuration  
No configs specified for hadoop runner  
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...  
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop  
Using Hadoop version 3.3.1  
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...  
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.  
jar  
Creating temp directory /tmp/wordcount1.thinnt.20211117.012824.456622  
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012824.45662  
2/files/wd...  
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount1.thinnt.20211117.012824.456622/  
files/  
Running step 1 of 1...  
Unable to load native-hadoop library for your platform... using builtin-java classes where applica  
ble  
packageJobJar: [/tmp/hadoop-unjar5889440386904687659/] [] /tmp/streamjob127662161618086051.jar tmp  
Dir=null  
Connecting to ResourceManager at /0.0.0:8032  
Connecting to ResourceManager at /0.0.0:8032  
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0003  
Total input files to process : 1  
number of splits:2  
Submitting tokens for job: job_1637112291929_0003  
Executing with tokens: []  
resource-types.xml not found  
Unable to find 'resource-types.xml'.  
Submitted application application_1637112291929_0003  
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0003/  
Running job: job_1637112291929_0003  
Job job_1637112291929_0003 running in uber mode : false  
map 0% reduce 0%  
map 100% reduce 0%
```

Hình 38. Chạy wordcount1.py với input là sách mã 20417

Kết quả thu được:

```
lab2 > test1 > pg20417_output > E part-00000  
1 "A" 349  
2 "ABERDEEN" 1  
3 "ABOUT" 5  
4 "ABOVE" 1  
5 "ACT" 2  
6 "ACTUAL" 1  
7 "ADAMS" 2  
8 "ADAPTATION" 2  
9 "ADAPTATIONS" 3  
10 "ADAPTED" 12  
11 "AEGIR" 4  
12 "AFRICA" 4  
13 "AGES" 4  
14 "AGO" 4  
15 "AGREE" 2  
16 "AGREEMENT" 1  
17 "AIR" 3  
18 "AK" 1  
19 "ALBATROSS" 2  
20 "ALFRED" 1  
21 "ALL" 1  
22 "ALLIGATOR" 2  
23 "ALONG" 2  
24 "ALPHA" 1  
25 "ALSATIAN" 2  
26 "ALSO" 2  
27 "ALTAMIRA" 4  
28 "ALTERNATING" 2  
29 "AMOEBA" 2  
30 "AMONGST" 2  
31 "AN" 37  
32 "ANALYSING" 2  
33 "ANCESTORS" 2  
34 "AND" 98  
35 "ANDROMEDA" 2
```

Hình 39. Kết quả chạy wordcount1.py với input là sách mã 20417

2.2.2. Trường hợp đếm không phân biệt hoa thường

a. Mã nguồn và giải thích

Mã nguồn cho trường hợp đếm không phân biệt hoa thường giống với mã nguồn đếm phân biệt hoa thường, chỉ thay đổi duy nhất một dòng:

```
for word in words:
    # .lower() vì không phân biệt hoa thường
    yield word.lower(), 1
```

Với thay đổi này, output của mapper luôn có phần key là chữ dạng viết thường, sau đó reducer sẽ đếm số lượng theo chữ viết thường này.

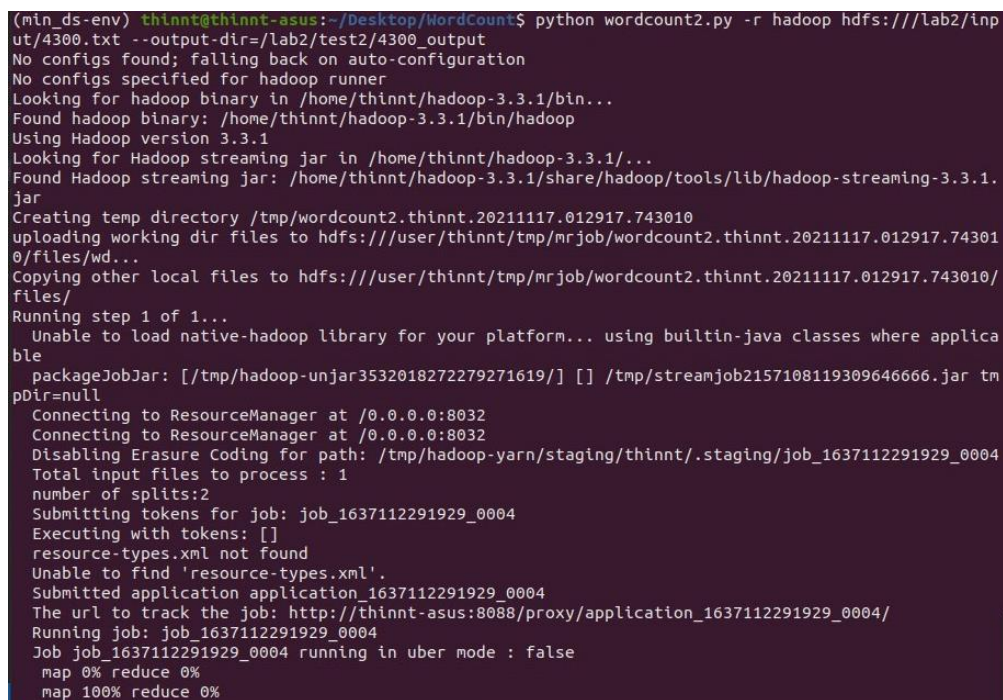
b. Quá trình thực thi

Testcase 1: Sách mã số 4300

Chạy lệnh thực thi:

```
python wordcount2.py -r hadoop hdfs:///lab2/input/4300.txt -
output-dir=/lab2/test2/4300_output
```

Quá trình chạy mapreduce:



```
(min ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount2.py -r hadoop hdfs:///lab2/inp
ut/4300.txt --output-dir=/lab2/test2/4300_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.
jar
Creating temp directory /tmp/wordcount2.thinnt.20211117.012917.743010
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.012917.74301
0/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.012917.743010/
files/
Running step 1 of 1...
Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar3532018272279271619/] [] /tmp/streamjob2157108119309646666.jar tm
pDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0004
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0004
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0004
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0004/
Running job: job_1637112291929_0004
Job job_1637112291929_0004 running in uber mode : false
  map 0% reduce 0%
  map 100% reduce 0%
```

Hình 40. Chạy wordcount2.py với input là sách mã 4300

Kết quả thu được:

```

lab2 > test2 > 4300_output > E part-00000
1  "a" 6581
2  "aaron" 2
3  "aback" 1
4  "abaft" 1
5  "abandon" 1
6  "abandoned" 7
7  "abandoning" 1
8  "abandonment" 1
9  "abasement" 2
10 "abatement" 1
11 "abattoir" 1
12 "abba" 1
13 "abbas" 2
14 "abbess" 1
15 "abbey" 13
16 "abbot" 3
17 "abbots" 1
18 "abbreviation" 1
19 "abdomen" 2
20 "abdominal" 2
21 "abe" 1
22 "abeakuta" 2
23 "abeakutic" 1
24 "aberration" 2
25 "abetting" 1
26 "abhorrence" 1
27 "abhors" 2
28 "abide" 3
29 "abigail" 1
30 "abilities" 2
31 "ability" 3
32 "abject" 1
33 "abjectly" 1
34 "abjured" 1
35 "ablation" 1
36 "able" 12

```

Hình 41. Kết quả chạy wordcount2.py với input là sách mã 4300

Testcase 2: Sách mã số 5000

Chạy lệnh thực thi:

```
python wordcount2.py -r hadoop hdfs:///lab2/input/5000.txt -
output-dir=/lab2/test2/5000_output
```

Quá trình chạy mapreduce:

```

(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount2.py -r hadoop hdfs:///lab2/in
ut/5000.txt --output-dir=/lab2/test2/5000_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.
jar
Creating temp directory /tmp/wordcount2.thinnt.20211117.013010.665999
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.013010.66599
9/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.013010.665999/
files/
Running step 1 of 1...
Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar4733185849890736610/] [] /tmp/streamjob3039131446784057958.jar tm
pDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0005
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0005
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0005
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0005/
Running job: job_1637112291929_0005
Job job_1637112291929_0005 running in uber mode : false
map 0% reduce 0%
map 100% reduce 0%

```

Hình 42. Chạy wordcount2.py với input là sách mã 5000

Kết quả thu được:

```
lab2 > thinnt > 5000_output > part-00000
1  "a" 4865
2  "aa" 1
3  "ab" 3
4  "abacho" 1
5  "abacus" 4
6  "abandon" 6
7  "abandoned" 2
8  "abandoning" 2
9  "abandons" 2
10 "abbasiden" 1
11 "abbate" 1
12 "abbey" 1
13 "abbia" 1
14 "abbiamo" 4
15 "abbozzi" 1
16 "abbracciare" 1
17 "abbreviated" 1
18 "abbreviation" 2
19 "abbreviations" 1
20 "abbreviators" 1
21 "aber" 1
22 "abila" 2
23 "abile" 1
24 "abiti" 1
25 "ablaze" 1
26 "able" 47
27 "abnimm" 1
28 "abode" 3
29 "abook" 3
30 "abortive" 1
31 "abosa" 1
32 "abound" 1
33 "abounding" 1
```

Hình 43. Kết quả chạy wordcount2.py với input là sách mã 5000

Testcase 3: Sách mã số 20417

Chạy lệnh thực thi:

```
python wordcount2.py -r hadoop hdfs:///lab2/input/pg20417.txt
-output-dir=/lab2/test2/pg20417_output
```

Quá trình chạy mapreduce:

```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount2.py -r hadoop hdfs:///lab2/inp
ut/pg20417.txt --output-dir=/lab2/test2/pg20417_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.
jar
Creating temp directory /tmp/wordcount2.thinnt.20211117.013058.904060
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.013058.90406
0/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount2.thinnt.20211117.013058.904060/
files/
Running step 1 of 1...
Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar7286178062905168949/] [] /tmp/streamjob1535836357941701731.jar tm
pDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0006
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0006
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0006
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0006/
Running job: job_1637112291929_0006
Job job_1637112291929_0006 running in uber mode : false
map 0% reduce 0%
map 100% reduce 0%
```

Hình 44. Chạy wordcount2.py với input là sách mã 20417

Kết quả thu được:

```
lab2 > test2 > pg20417_output > E part-00000
1  "a" 2834
2  "abandoned" 3
3  "abandonment" 1
4  "abbreviated" 1
5  "abbreviation" 1
6  "abdomen" 2
7  "abdominal" 1
8  "aberdeen" 1
9  "abeyance" 1
10 "abide" 1
11 "abilities" 1
12 "ability" 3
13 "able" 41
14 "aboriginal" 1
15 "about" 194
16 "above" 38
17 "abroad" 1
18 "abrupt" 1
19 "absence" 8
20 "absent" 3
21 "absolute" 4
22 "absolutely" 2
23 "absorb" 1
24 "absorbed" 7
25 "absorbing" 3
26 "absorbs" 4
27 "absorption" 3
28 "absorptive" 1
29 "abstruse" 1
30 "absurdly" 1
31 "abundance" 7
32 "abundant" 9
33 "abundantly" 2
34 "abyssal" 1
35 "abysses" 8
```

Hình 45. Kết quả chạy wordcount2.py với input là sách mã 20417

2.2.3. Tìm từ xuất hiện nhiều nhất trong tài liệu (không phân biệt hoa thường)

a. Mã nguồn và giải thích

Ở 2 trường hợp trước thì chương trình chỉ gồm 1 step bao gồm cả mapper, combiner và reducer. Tuy nhiên ở trường hợp tìm từ xuất hiện nhiều nhất, ta phải thiết kế chương trình chạy theo 2 step:

- Step 1: bao gồm mapper, combiner và reducer. Step này có nhiệm vụ đếm số lượng xuất hiện của mỗi từ, output có dạng (None, (word, count)).
- Step 2: chỉ bao gồm reducer nhận trực tiếp input từ reducer của step 1. Vì input có key đều là None, nên phần values sẽ là list các tuple dạng [(word1, count1), (word2, count2), ...]. Từ đó ta tìm ra tuple mà có phần count là lớn nhất.

Mã nguồn thực hiện tìm từ xuất hiện nhiều nhất không phân biệt hoa thường được trình bày và giải thích chi tiết bên dưới:

```
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

# pattern regex để tìm các từ chỉ gồm chữ cái, không gồm chữ số và kí tự khác
WORD_REGEX = re.compile(r"[a-zA-Z]+")

# class MRJWordCount cài đặt job đếm từ,
# kế thừa và override lại hàm của lớp MRJob có sẵn
class MRMostUsedWord(MRJob):

    def steps(self):
```



```

# Hàm trả về list các step: ở đây job gồm 2 step
# - step 1: gồm cả mapper, combiner, reducer -> đếm số lượng xuất hiện
của các từ
# - step 2: chỉ có reducer -> tìm từ xuất hiện nhiều nhất
# MRStep nhận tham số là các hàm cài đặt tương ứng
# cho mapper, combiner, reducer
return [
    MRStep(mapper=self.mapper_get_words,
            combiner=self.combiner_count_words,
            reducer=self.reducer_count_words),
    MRStep(reducer=self.reducer_find_max_word)
]

# tách ra các từ xuất hiện ở mỗi câu
def mapper_get_words(self, _, line):
    # loại bỏ khoảng trắng đầu và cuối dòng
    line = line.strip()
    # tạo đầu ra là các cặp <key, value> dạng <word, 1> cho mỗi từ trong 1
line
    for word in WORD_REGEX.findall(line):
        # .lower() vì không phân biệt hoa thường
        yield word.lower(), 1

# tổng hợp lại số lần xuất hiện của các từ trên 1 line (câu),
# làm gọn bớt đầu vào cho reducer của step 1
def combiner_count_words(self, word, counts):
    yield word, sum(counts)

# nhận đầu vào sinh ra từ combiner, các cặp <key, value> của cùng 1 từ
# sẽ về cùng một reducer và tổng hợp lại số lần xuất hiện
def reducer_count_words(self, word, counts):
    # hàm trả về <key, values> có key=None để đưa các values=(word,
num_occur)
    # về cùng một reducer ở step 2, thuận lợi cho việc tìm số lượng xuất
hiện nhiều nhất
    yield None, (word, sum(counts))

# nhận trực tiếp đầu ra của reducer ở step 1
def reducer_find_max_word(self, _, word_count_pairs):
    # word_count_pairs là list các tuple dạng (word, num_of_occurrences),
    # tìm max dựa trên phần tử thứ 2 của tuple
    yield max(word_count_pairs, key=lambda item: item[1])

if __name__ == '__main__':
    MRMostUsedWord.run()

```

b. Quá trình thực thi

Testcase 1: Sách mã số 4300

Chạy lệnh thực thi:

```
python wordcount3.py -r hadoop hdfs:///lab2/input/4300.txt -
output-dir=/lab2/test3/4300_output
```

Quá trình chạy mapreduce:

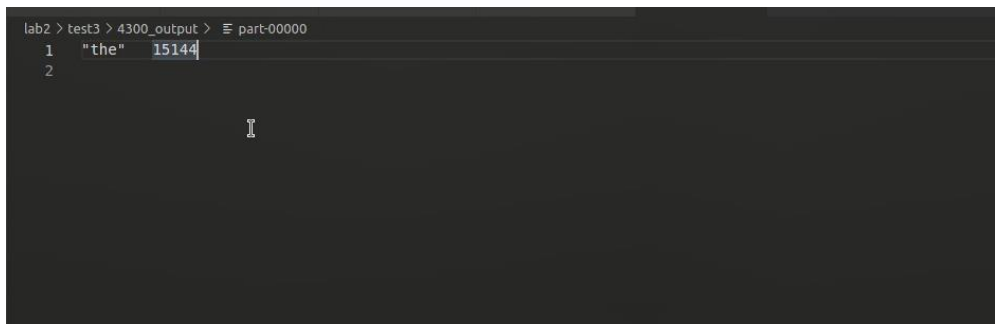
```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount3.py -r hadoop hdfs:///lab2/inputs/4300.txt --output-dir=/lab2/test3/4300_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar
Creating temp directory /tmp/wordcount3.thinnt.20211117.013150.961509
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount3.thinnt.20211117.013150.961509/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount3.thinnt.20211117.013150.961509/files/
Running step 1 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar7811205071307261432/] [] /tmp/streamjob3946956516522701832.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0007
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0007
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0007
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0007/
Running job: job_1637112291929_0007
Job job_1637112291929_0007 running in uber mode : false
  map 0% reduce 0%
  map 100% reduce 0%
  map 100% reduce 100%
```

Hình 46. Step 1 chạy wordcount3.py với input là sách mã 4300

```
thinnt@thinnt-asus: ~/Desktop/WordCount
Input split bytes=186
Map input records=33216
Map output bytes=2557319
Map output materialized bytes=544794
Map output records=271975
Merged Map outputs=2
Peak Map Physical memory (bytes)=311087104
Peak Map Virtual memory (bytes)=2596651008
Peak Reduce Physical memory (bytes)=217010176
Peak Reduce Virtual memory (bytes)=2598899712
Physical memory (bytes) snapshot=836612096
Reduce input groups=29131
Reduce input records=37933
Reduce output records=29131
Reduce shuffle bytes=544794
Shuffled Maps =2
Spilled Records=75866
Total committed heap usage (bytes)=705167360
Virtual memory (bytes) snapshot=7790551040
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Running step 2 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar6899808595315180387/] [] /tmp/streamjob8250216381431963456.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0008
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0008
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0008
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0008/
Running job: job_1637112291929_0008
Job job_1637112291929_0008 running in uber mode : false
  map 0% reduce 0%
```

Hình 47. Step 2 chạy wordcount3.py với input là sách mã 4300

Kết quả thu được:



```
lab2 > test3 > 4300_output > part-00000
1  "the" 15144
2
```

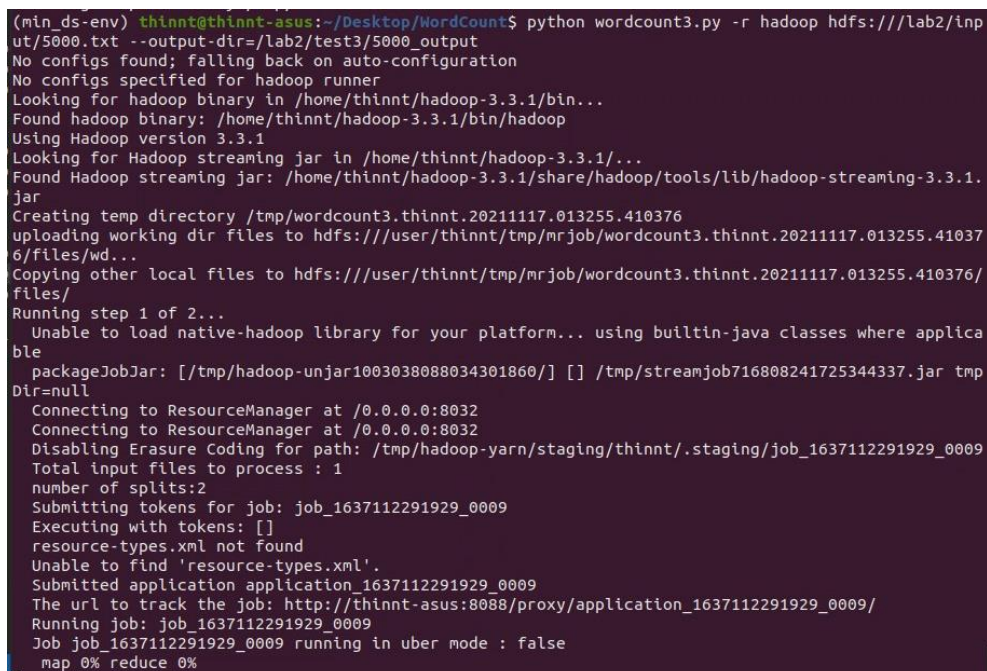
Hình 48. Kết quả chạy wordcount3.py với input là sách mã 4300

Testcase 2: Sách mã số 5000

Chạy lệnh thực thi:

```
python wordcount3.py -r hadoop hdfs:///lab2/input/5000.txt -
output-dir=/lab2/test3/5000_output
```

Quá trình chạy mapreduce:



```
(min_ds-env) thinn@thinn-asus:~/Desktop/WordCount$ python wordcount3.py -r hadoop hdfs:///lab2/inp
ut/5000.txt --output-dir=/lab2/test3/5000_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinn/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinn/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinn/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinn/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.
jar
Creating temp directory /tmp/wordcount3.thinn.20211117.013255.410376
uploading working dir files to hdfs:///user/thinn/tmp/mrjob/wordcount3.thinn.20211117.013255.41037
6/files/wd...
Copying other local files to hdfs:///user/thinn/tmp/mrjob/wordcount3.thinn.20211117.013255.410376/
files/
Running step 1 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar1003038088034301860/] [] /tmp/streamjob716808241725344337.jar tmp
Dir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinn/.staging/job_1637112291929_0009
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0009
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0009
The url to track the job: http://thinn-asus:8088/proxy/application_1637112291929_0009/
Running job: job_1637112291929_0009
Job job_1637112291929_0009 running in uber mode : false
map 0% reduce 0%
```

Hình 49. Step 1 chạy wordcount3.py với input là sách mã 5000

```
thinnt@thinnt-asus: ~/Desktop/WordCount
Input split bytes=186
Map input records=32763
Map output bytes=2302398
Map output materialized bytes=291078
Map output records=247603
Merged Map outputs=2
Peak Map Physical memory (bytes)=307748864
Peak Map Virtual memory (bytes)=2593714176
Peak Reduce Physical memory (bytes)=212889600
Peak Reduce Virtual memory (bytes)=2596421632
Physical memory (bytes) snapshot=825311232
Reduce input groups=15656
Reduce input records=20634
Reduce output records=15656
Reduce shuffle bytes=291078
Shuffled Maps =2
Spilled Records=41268
Total committed heap usage (bytes)=702545920
Virtual memory (bytes) snapshot=7782072320
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Running step 2 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar881791986966733894/] [] /tmp/streamjob936737123465757334.jar tmpD
ir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0010
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0010
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0010
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0010/
Running job: job_1637112291929_0010
Job job_1637112291929_0010 running in uber mode : false
map 0% reduce 0%
```

Hình 50. Step 2 chạy wordcount3.py với input là sách mã 5000

Kết quả thu được:

```
lab2 > test3 > 5000_output > part-00000
1  "the"  22991
2  |
```

Hình 51. Kết quả chạy wordcount3.py với input là sách mã 5000

Testcase 3: Sách mã số 20417

Chạy lệnh thực thi:

```
python wordcount3.py -r hadoop hdfs:///lab2/input/pg20417.txt
-output-dir=/lab3/test2/pg20417_output
```

Quá trình chạy mapreduce:


```
(min_ds-env) thinnt@thinnt-asus:~/Desktop/WordCount$ python wordcount3.py -r hadoop hdfs:///lab2/inputs/pg20417.txt --output-dir=/lab2/test3/pg20417_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/thinnt/hadoop-3.3.1/bin...
Found hadoop binary: /home/thinnt/hadoop-3.3.1/bin/hadoop
Using Hadoop version 3.3.1
Looking for Hadoop streaming jar in /home/thinnt/hadoop-3.3.1/...
Found Hadoop streaming jar: /home/thinnt/hadoop-3.3.1/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar
Creating temp directory /tmp/wordcount3.thinnt.20211117.013400.577951
uploading working dir files to hdfs:///user/thinnt/tmp/mrjob/wordcount3.thinnt.20211117.013400.577951/files/wd...
Copying other local files to hdfs:///user/thinnt/tmp/mrjob/wordcount3.thinnt.20211117.013400.577951/files/
Running step 1 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar980394300735986356/] [] /tmp/streamjob5368721332451225741.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0011
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0011
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0011
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0011/
Running job: job_1637112291929_0011
Job job_1637112291929_0011 running in uber mode : false
map 0% reduce 0%
```

Hình 52. Step 1 chạy wordcount3.py với input là sách mã 20417

```
thinnt@thinnt-asus: ~/Desktop/WordCount
Input split bytes=192
Map input records=12760
Map output bytes=1070142
Map output materialized bytes=189250
Map output records=111234
Merged Map outputs=2
Peak Map Physical memory (bytes)=311336960
Peak Map Virtual memory (bytes)=2593669120
Peak Reduce Physical memory (bytes)=212119552
Peak Reduce Virtual memory (bytes)=2601947136
Physical memory (bytes) snapshot=829677568
Reduce input groups=9391
Reduce input records=13199
Reduce output records=9391
Reduce shuffle bytes=189250
Shuffled Maps =2
Spilled Records=26398
Total committed heap usage (bytes)=704118784
Virtual memory (bytes) snapshot=7789232128
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Running step 2 of 2...
Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar5336754967152330778/] [] /tmp/streamjob8970067990253224502.jar tmpDir=null
Connecting to ResourceManager at /0.0.0.0:8032
Connecting to ResourceManager at /0.0.0.0:8032
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/thinnt/.staging/job_1637112291929_0012
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1637112291929_0012
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1637112291929_0012
The url to track the job: http://thinnt-asus:8088/proxy/application_1637112291929_0012/
Running job: job_1637112291929_0012
Job job_1637112291929_0012 running in uber mode : false
map 0% reduce 0%
```

Hình 53. Step 2 chạy wordcount3.py với input là sách mã 20417

Kết quả thu được:

```
lab2 > test3 > pg20417_output > part-00000
1  "the"  9334
2  |
```

Hình 54. Kết quả chạy wordcount3.py với input là sách mã 20417

3. Tài liệu tham khảo

- [1] Data test cho mức 1
https://www.kaggle.com/shashwatwork/consume-complaints-dataset-for-nlp?select=complaints_processed.csv
https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m?select=user_rating_6000_to_11000.csv
- [2] Video hướng dẫn install multiple node on Hadoop
<https://www.youtube.com/watch?v=FAly8HaYkbQ>
<https://www.youtube.com/watch?v=HFfa5wFGuS08>
- [3] MRJob Document: Fundamental
<https://mrjob.readthedocs.io/en/latest/guides/quickstart.html>
- [4] MRJob Document: Writing jobs
<https://mrjob.readthedocs.io/en/latest/guides/writing-mrjobs.html>
- [5] Lesson 11 Python MRJob installing and testing
<https://www.youtube.com/watch?v=zp1w8iM0UR4>
- [6] CS 6500: Running MapReduce Jobs with Python and MrJob
<https://www.youtube.com/watch?v=GCeb0qhnaIs>