Môn học: Dữ liệu lớn - IS405.P23

Tên: Nguyễn Hạ Phương

MSSV: 22521166

BÀI THỰC HÀNH SỐ 2

HDFS COMMANDS VÀ LẬP TRÌNH TRÊN HADOOP

MỤC LỤC

[**1.** **Bài tập 1** 2](#_Toc195124333)

[**2.** **Bài tập 2** 4](#_Toc195124334)

1. **Bài tập 1**

* **Bước 1:** Mở teminal trong thư mục hadoop và nhập lệnh khởi chạy dfs.

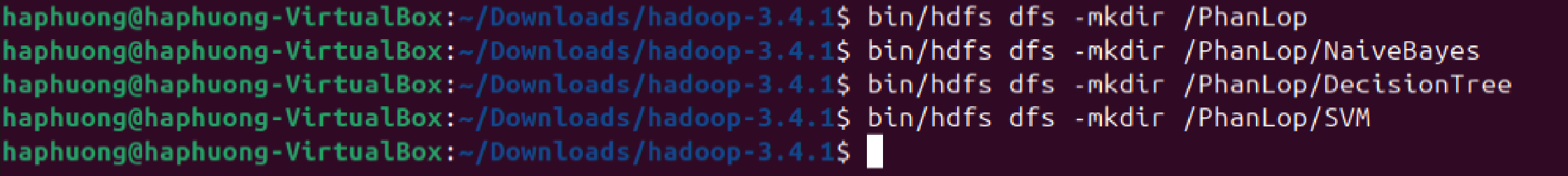
A computer screen shot of a program

AI-generated content may be incorrect.

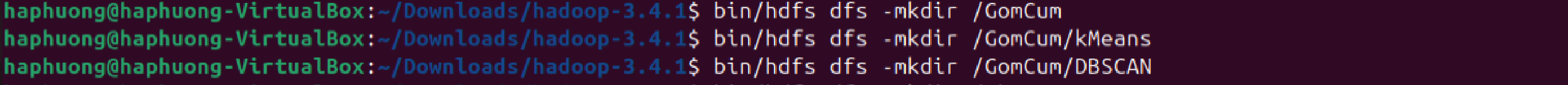
A computer screen with text

AI-generated content may be incorrect.

* **Bước 2:** Nhập lệnh tạo thư mục /PhanLop, sau đó trong thư mục /PhanLop tạo thêm các thư mục là /NaiveBayes, /DecisionTree, /SVM.



* **Bước 3:** Thực hiện tương tự bước 2, nhập lệnh tạo thư mục /GomCum, trong thư mục /GomCum tạo thêm các thư mục /kMeans và /DBSCAN.



* **Bước 4:** Vào localhost:9870/, chọn *Utilities*, chọn *Browse the file system* để xem các thư mục vừa tạo.

A screenshot of a computer

AI-generated content may be incorrect.

Các thư mục con trong thư mục /PhanLop.

A screenshot of a computer

AI-generated content may be incorrect.

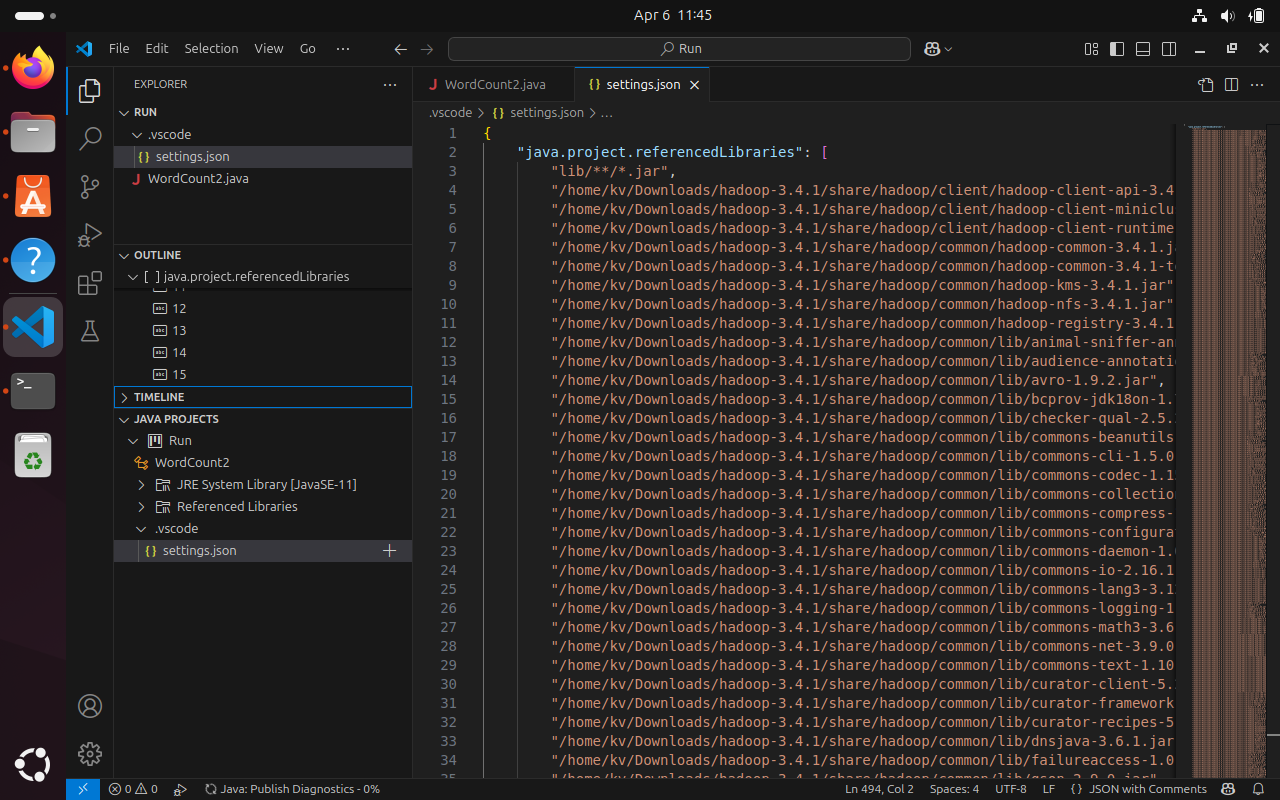
Các thư mục con trong thư mục /GomCum.

A screenshot of a computer

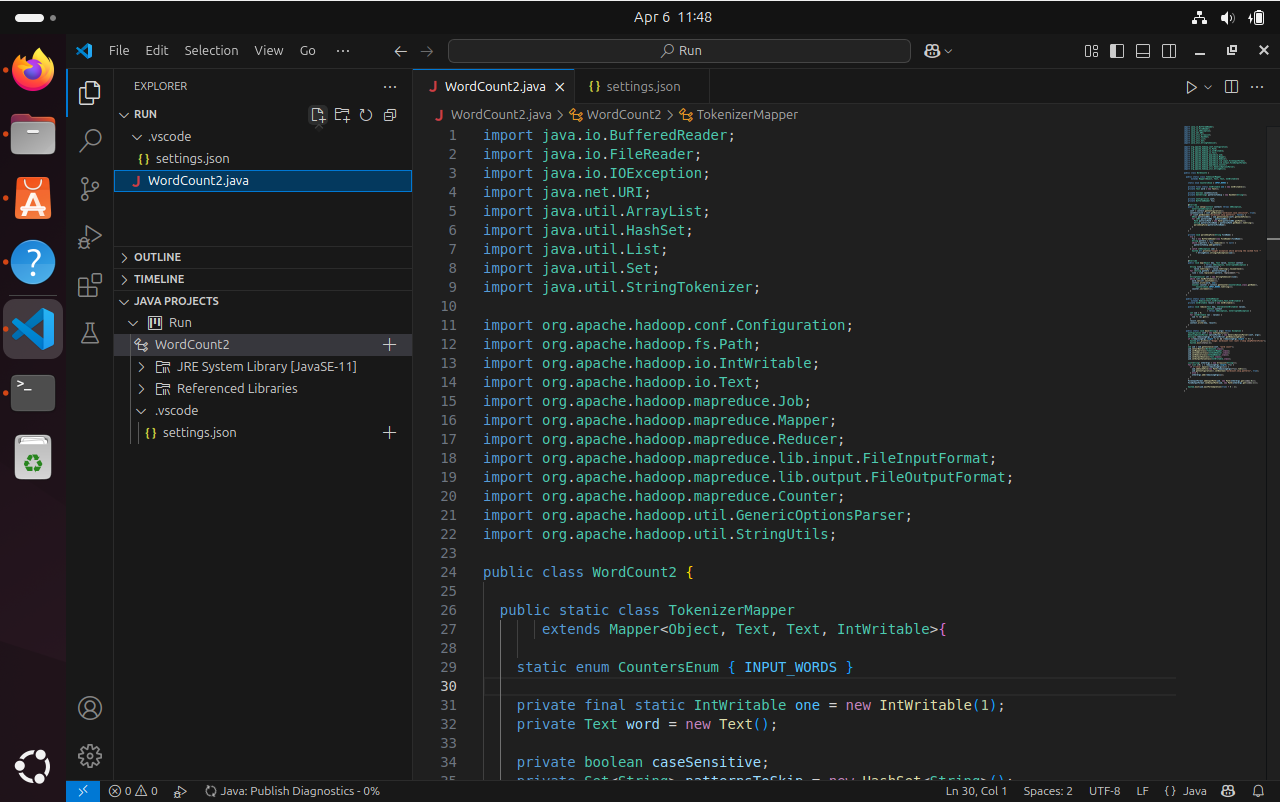
AI-generated content may be incorrect.

1. **Bài tập 2**

* **Bước 1:** Tạo project Java mới và thêm các thư viện cần thiết.



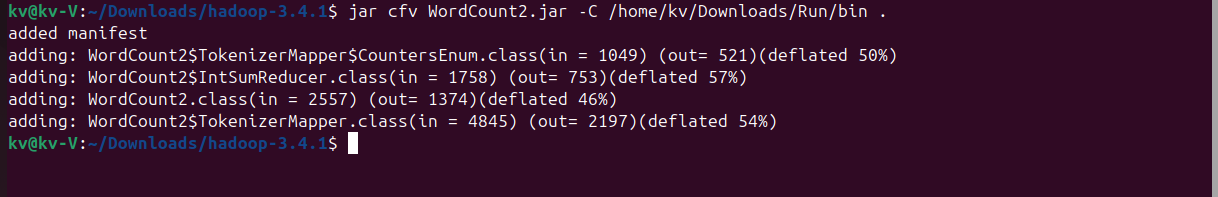
* **Bước 2:** Thêm code vào file WordCount.java từ nguồn WordCount v2.0.



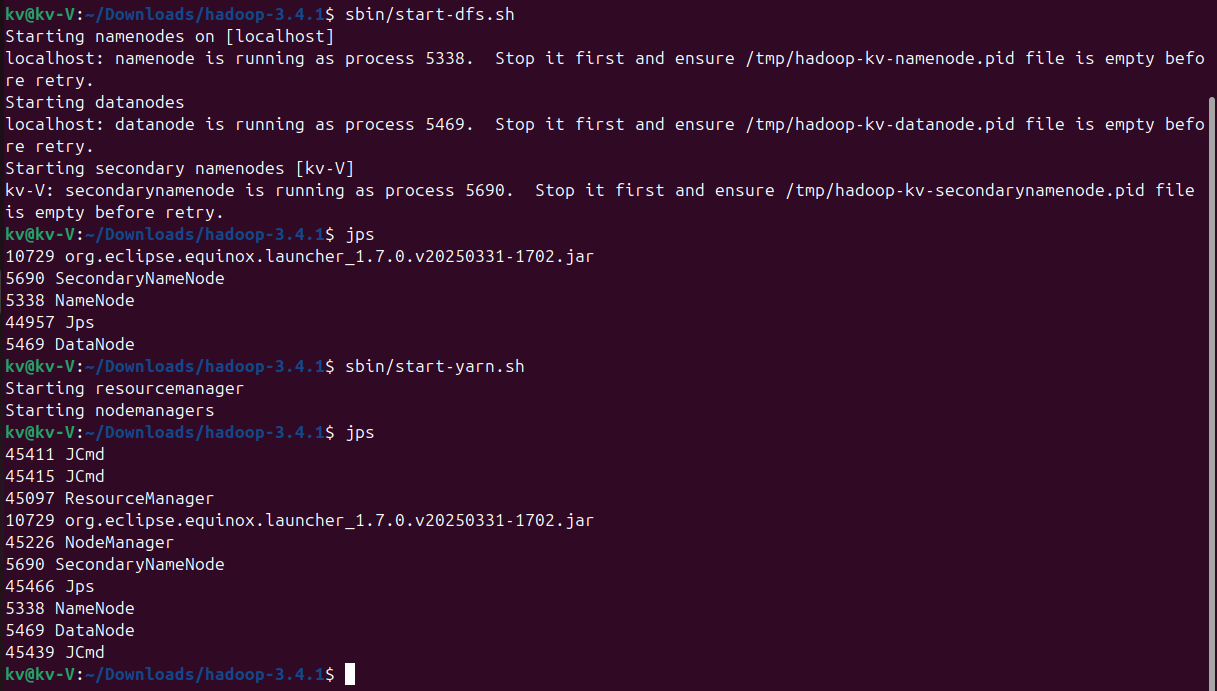
* **Bước 3:** Nhập lệnh biên dịch code java cho chương trình.



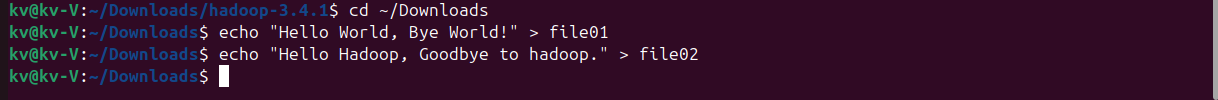
* **Bước 4:** Tạo file jar tên WordCount2 chứa các class files.



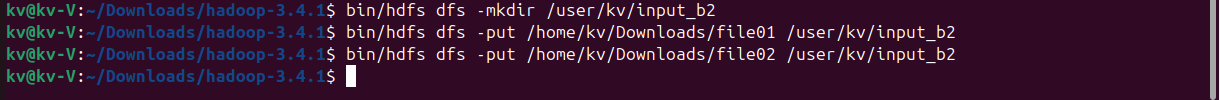
* **Bước 5:** Nhập lệnh chạy dfs và yarn.



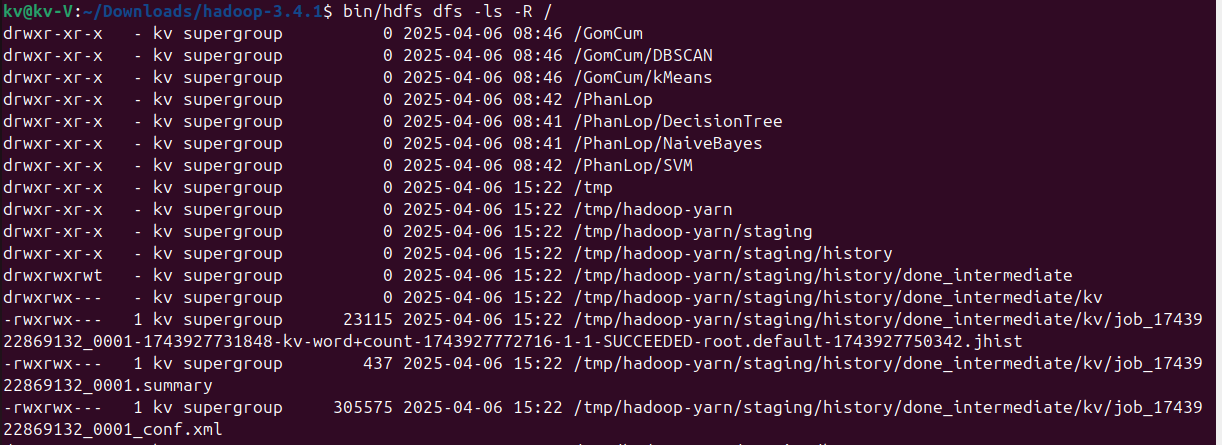
* **Bước 6:** Tạo các file text chứa dữ liệu (nếu chưa có).

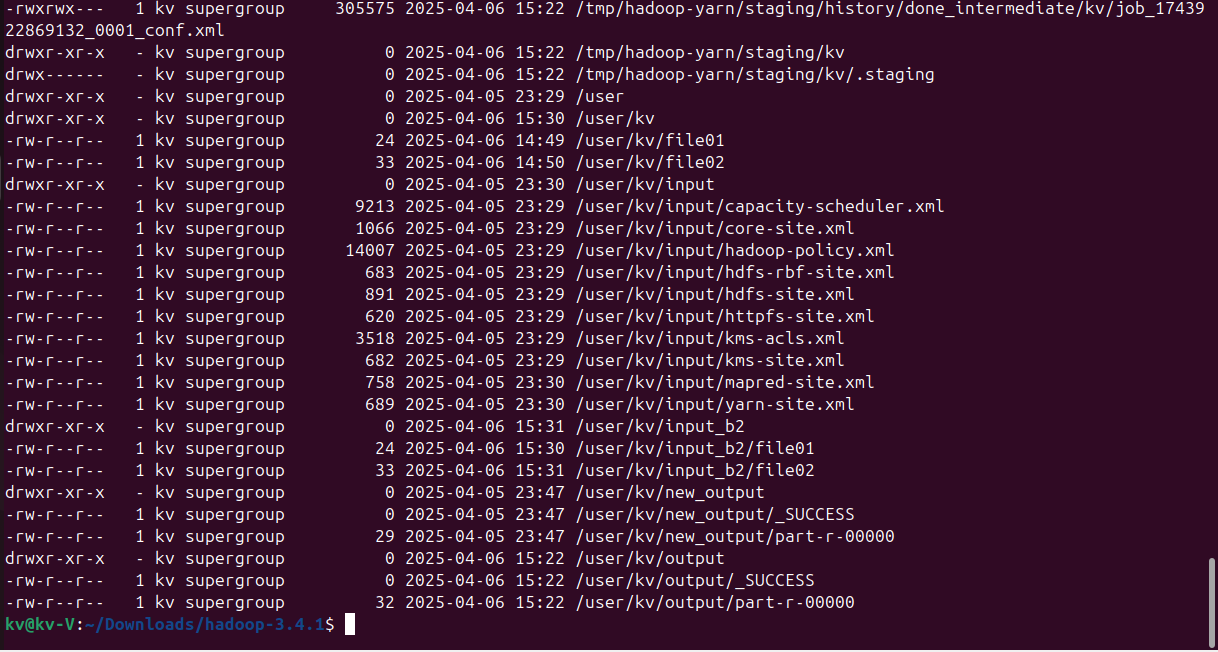


* **Bước 7:** Nhập lệnh copy file text từ máy chủ sang hdfs.

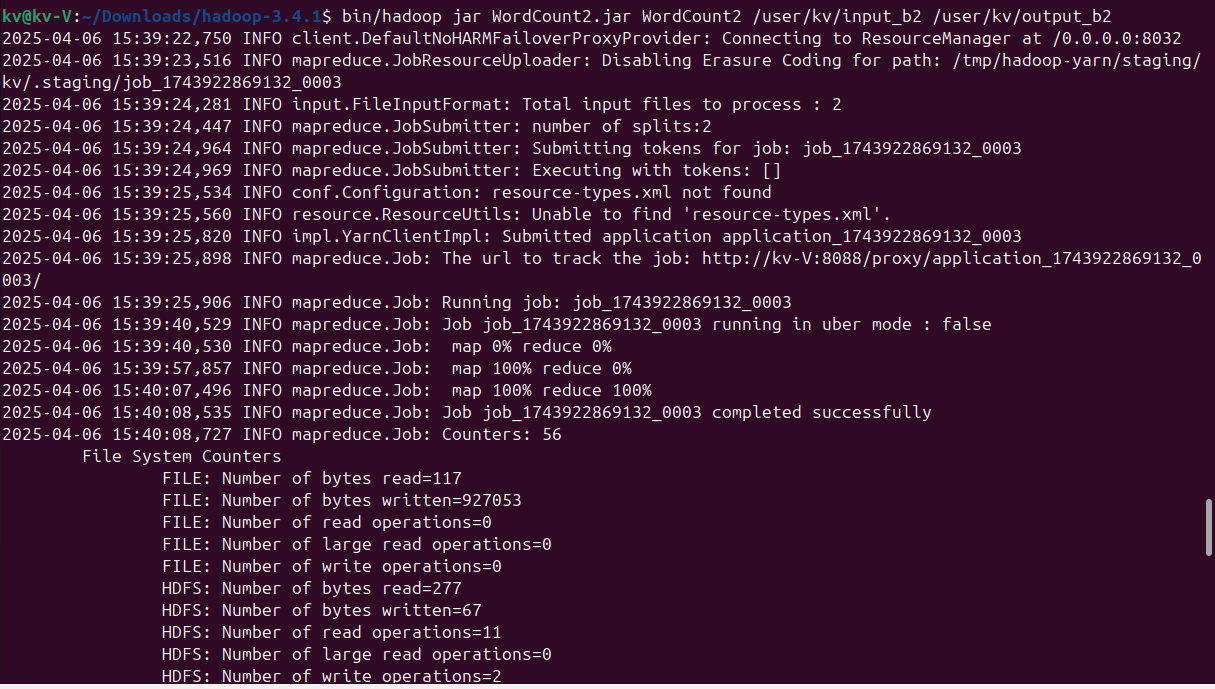


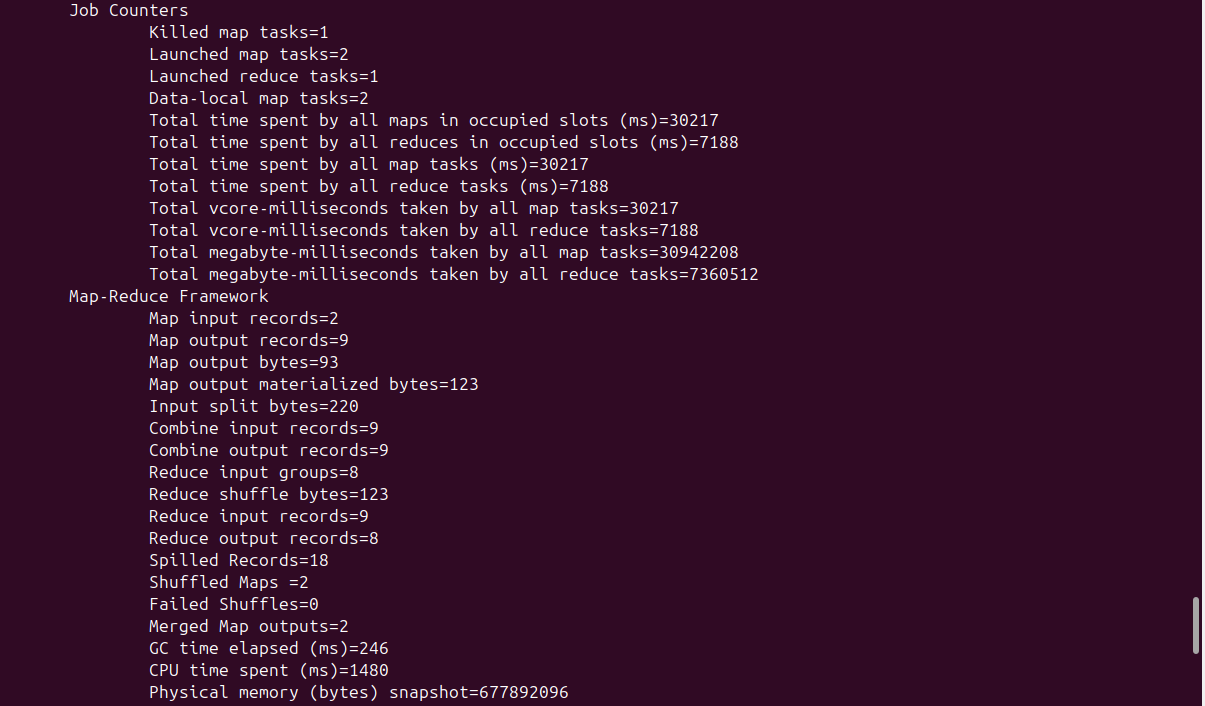
* **Bước 8:** Nhập lệnh hiển thị danh sách file trên hdfs kiểm tra xem có file text đã put lên hay không.

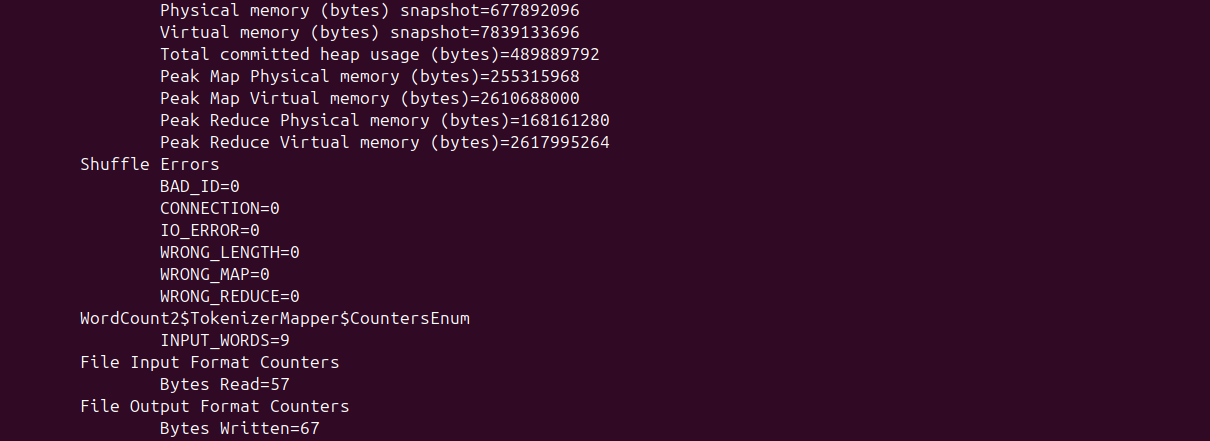




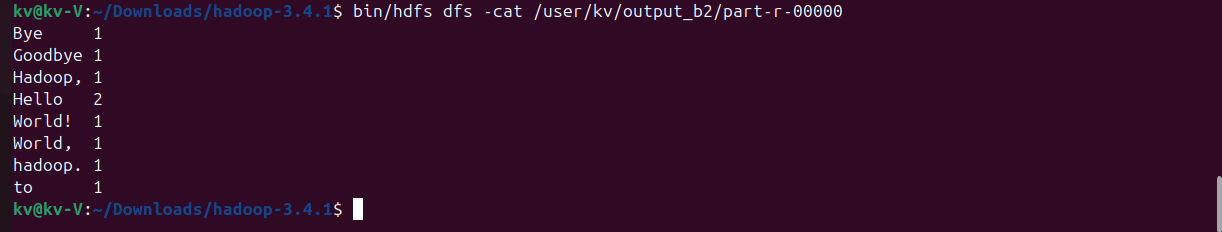
* **Bước 9:** Nhập lệnh để thực thi chương trình với dữ liệu đầu vào là 2 file text file01 và file02 trong thư mục input\_b2 trên hdfs.



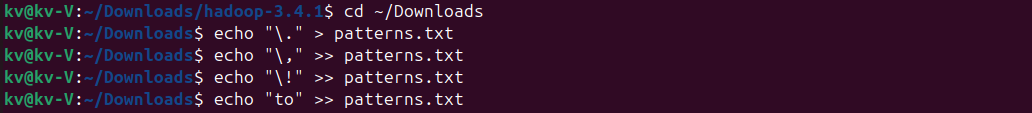




* **Bước 10:** Nhập lệnh để kiểm tra kết quả chương trình đã thực thi.



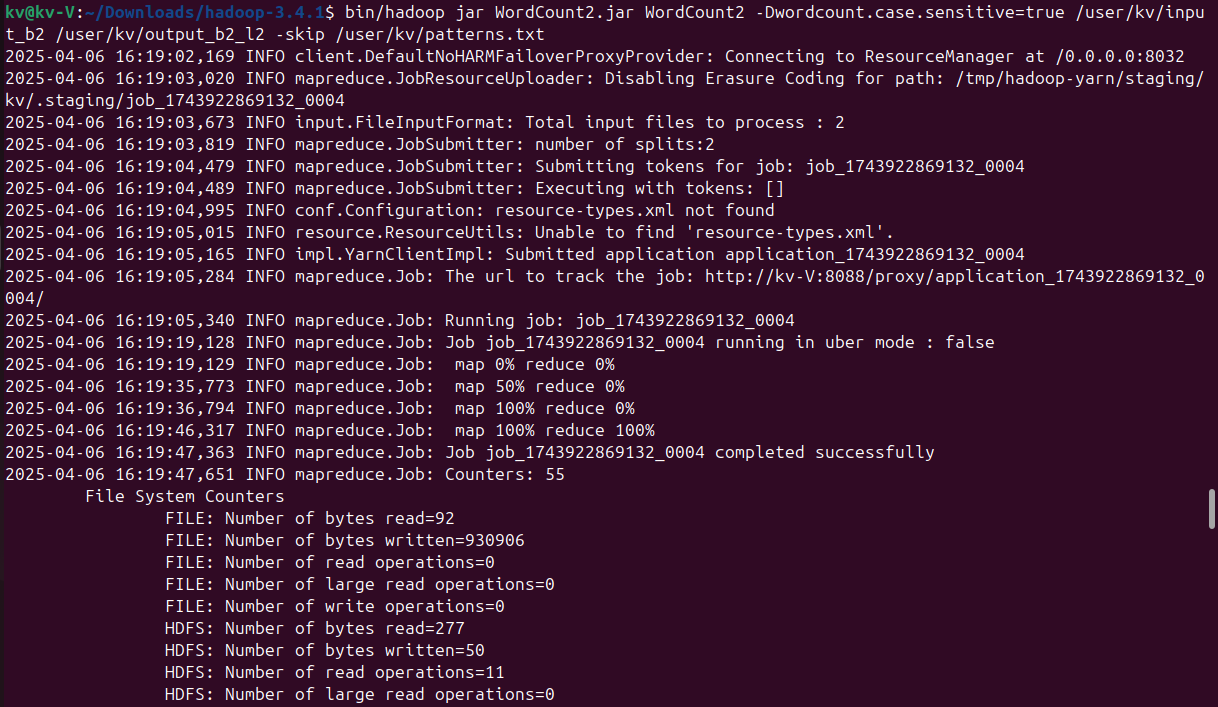
* **Bước 11:** Tạo tệp patterns.txt - một tệp liệt kê các từ sẽ bị bỏ qua thông qua DistributedCache.

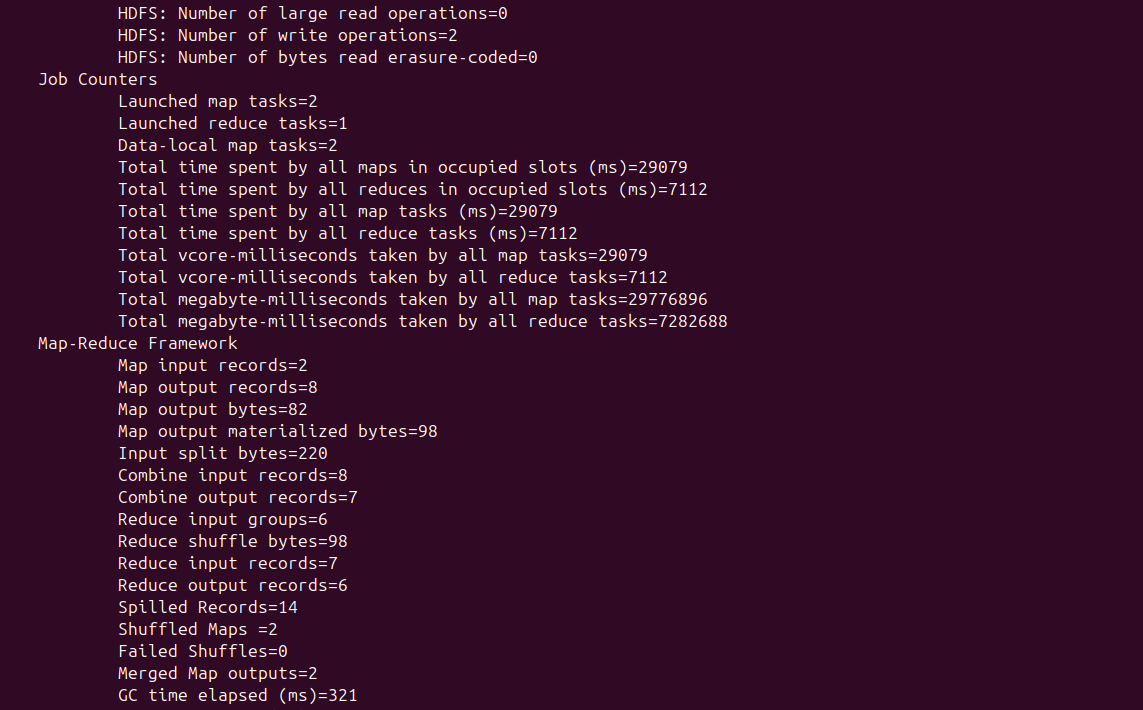


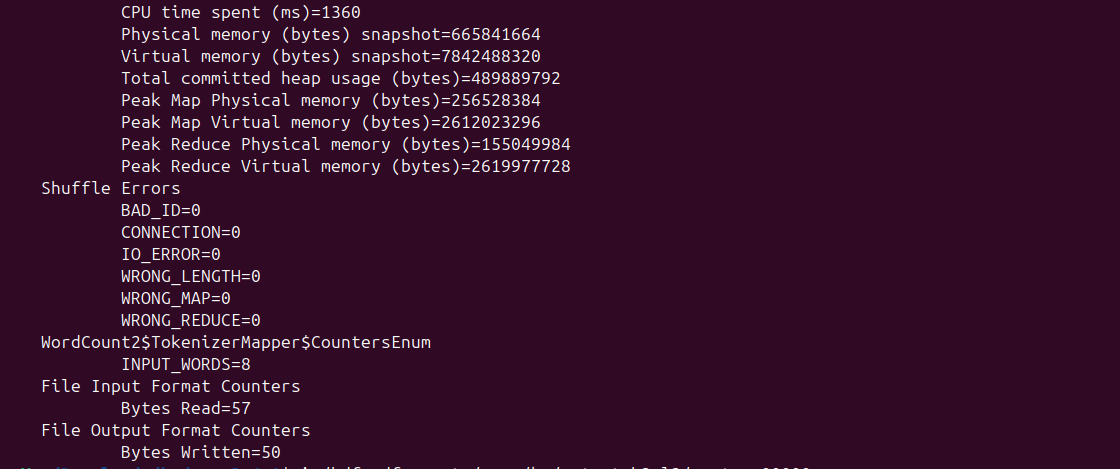
* **Bước 12:** Nhập lệnh copy patterns.txt từ máy chủ sang hdfs.



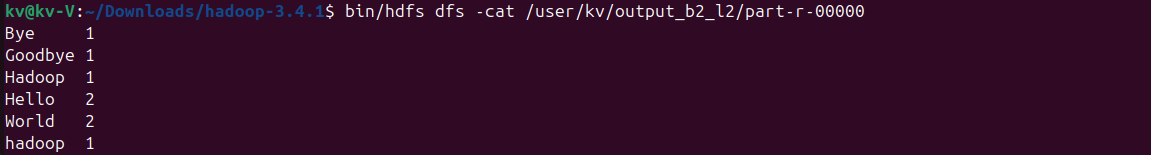
* **Bước 13:** Thực thi lại chương trình với tùy chọn bỏ qua một số ký tự đã liệt kê trong file patterns.txt (có phân biệt chữ hoa chữ thường).



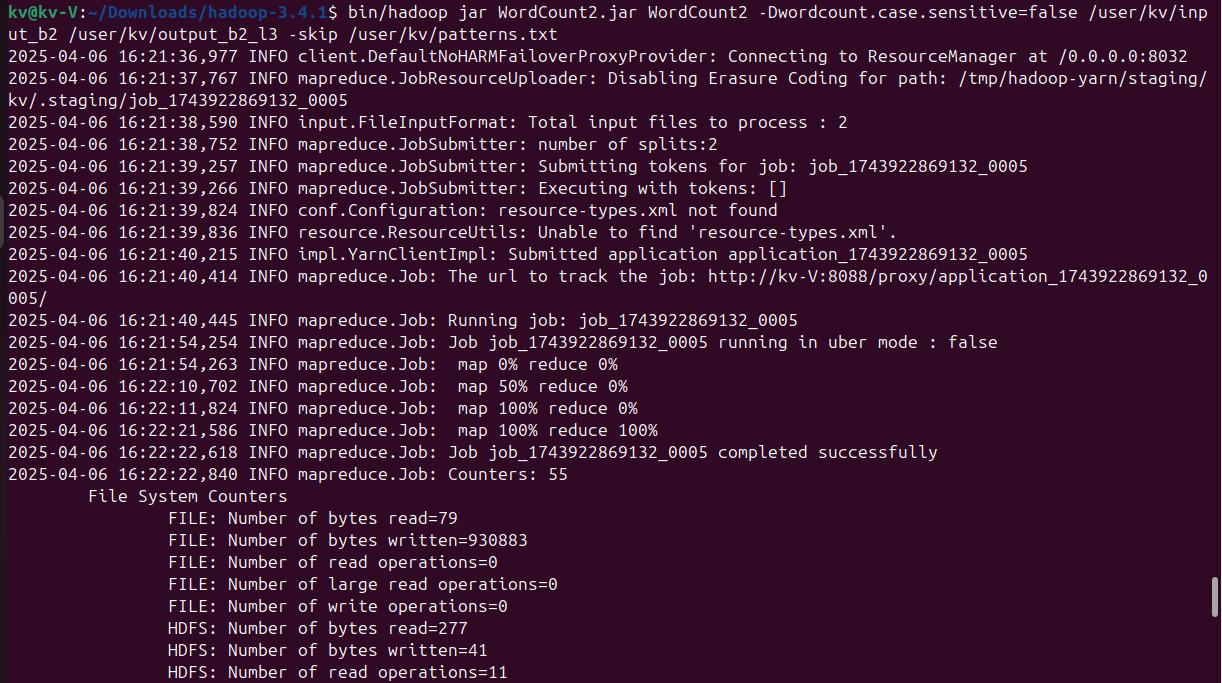


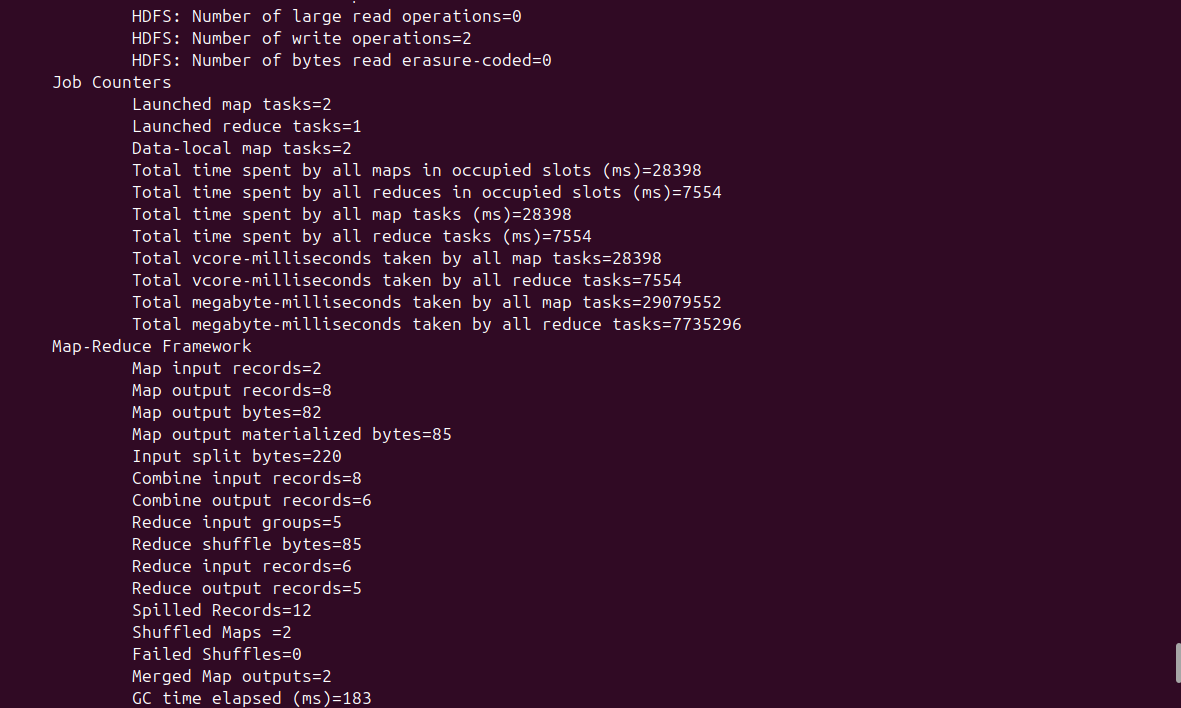


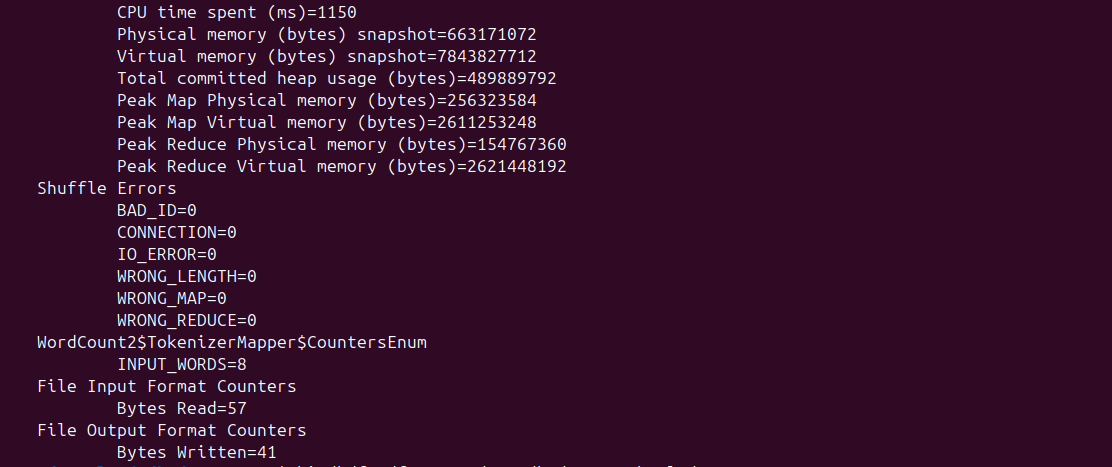
* **Bước 14:** Nhập câu lệnh kiểm tra kết quả chương trình vừa thực thi.



* **Bước 15:** Thực thi lại chương trình với tùy chọn bỏ qua một số ký tự đã liệt kê trong file patterns.txt (không phân biệt chữ hoa chữ thường).







* **Bước 16:** Nhập câu lệnh kiểm tra kết quả chương trình vừa thực thi.

