Họ & Tên: Lê Thị Đoan Trang

MSSV: 20522038

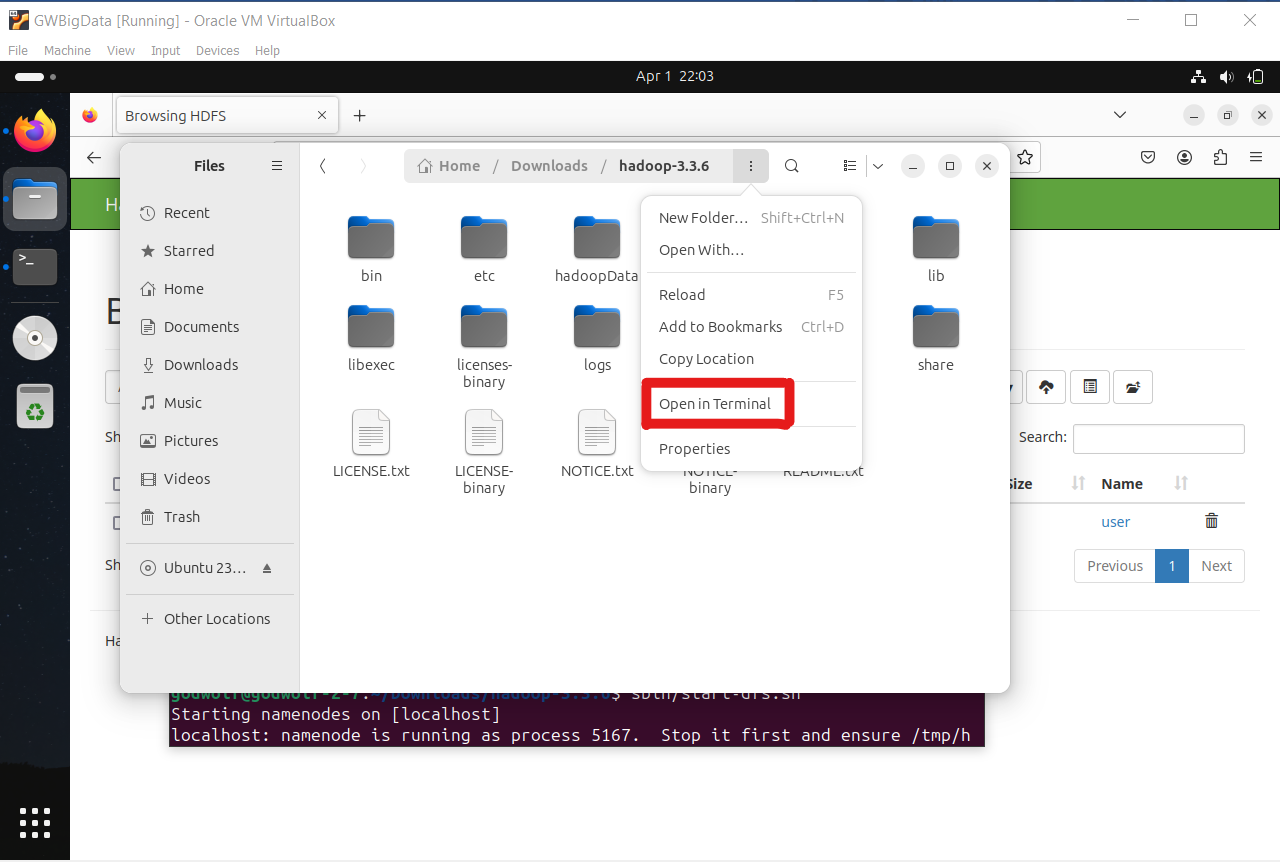
Lớp: IS405.O22.HTCL.2

**LAB 2**

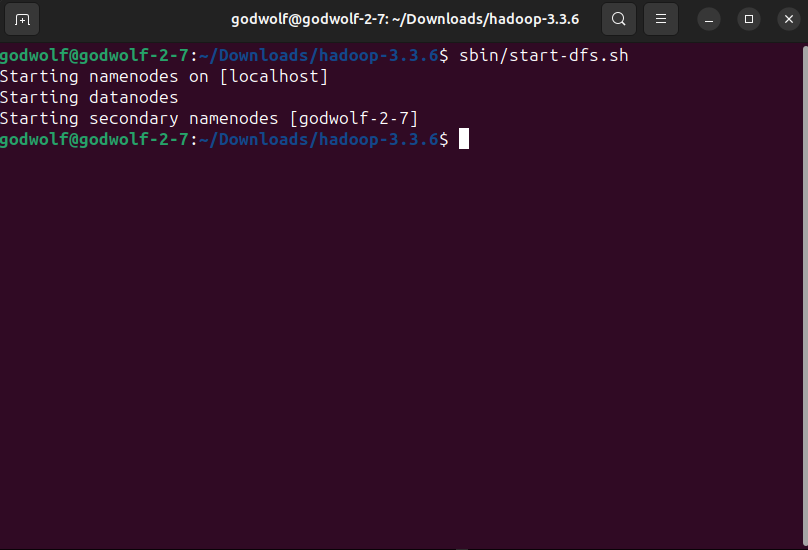
**Bài 1:** Tạo cây thư mục trong hệ thống tập tin HDFS

/PhanLop/CayQuyetDinh  
/PhanLop/Bayes  
/PhanLop/LogisticRegression  
/GomCum/kMeans  
/GomCum/DBSCAN

Bước 1: Mở thư mục hadoop bằng terminal và nhập lệnh chạy dfs

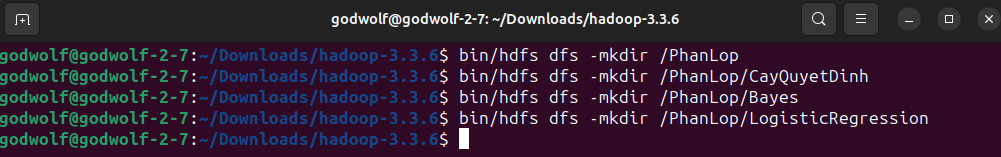


*Hình 1: Mở thư mục hadoop bằng terminal*



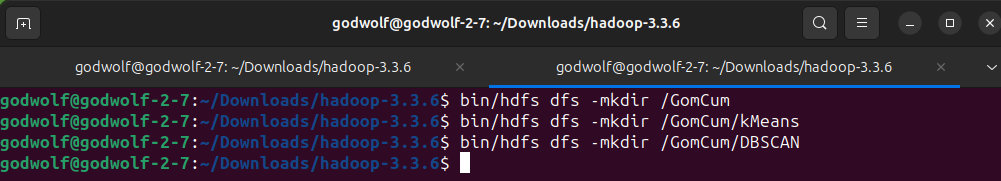
*Hình 2: Khởi chạy dfs*

Bước 2: Nhập lệnh tạo thư mục /PhanLop, sau đó trong thư mục /PhanLop tạo thêm một thư mục là CayQuyetDinh, Bayes, LogisticRegression



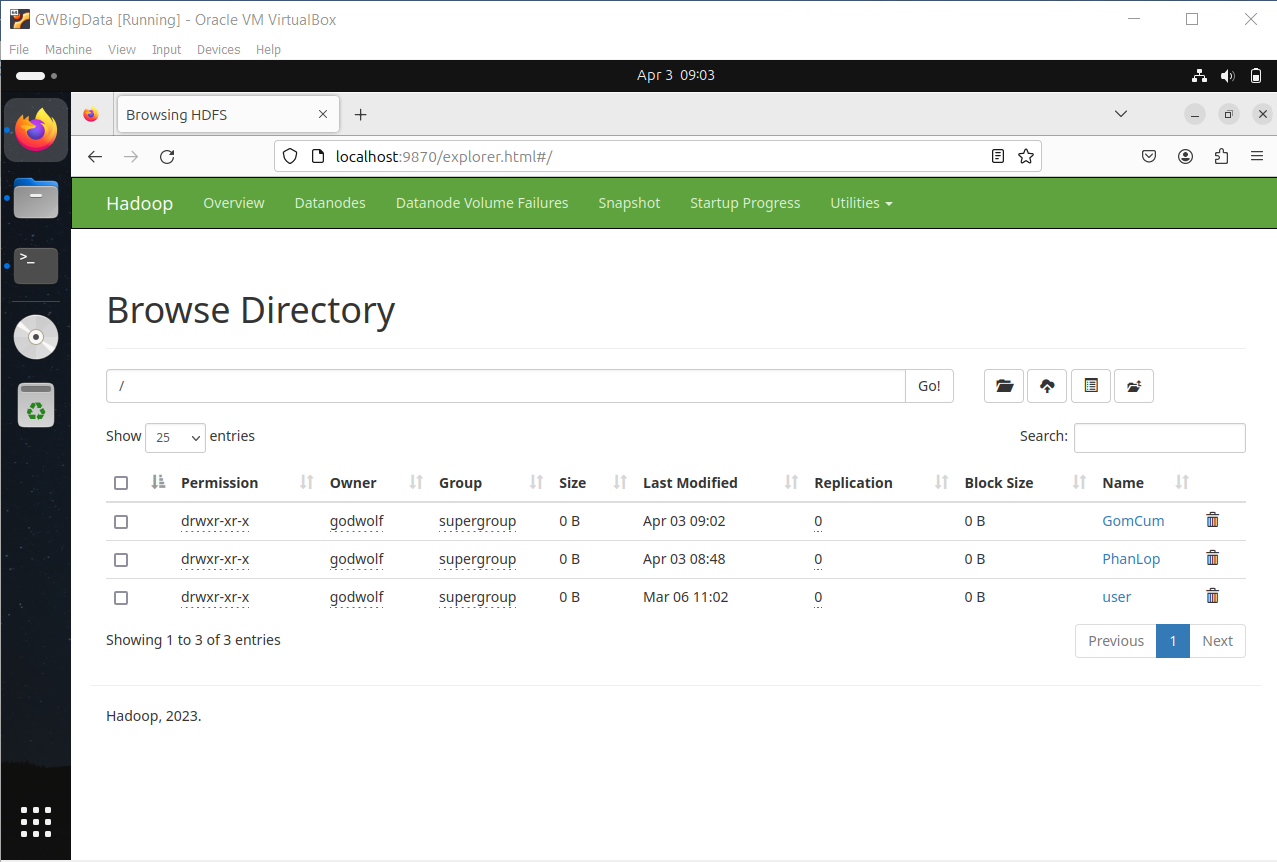
*Hình 3: Tạo cây thư mục /PhanLop/CayQuyetDinh, /PhanLop/Bayes, /PhanLop/LogisticRegression*

Bước 3: Thực hiện bước 2 tương tự với các thư mục /GomCum/kMeans, /GomCum/DBSCAN

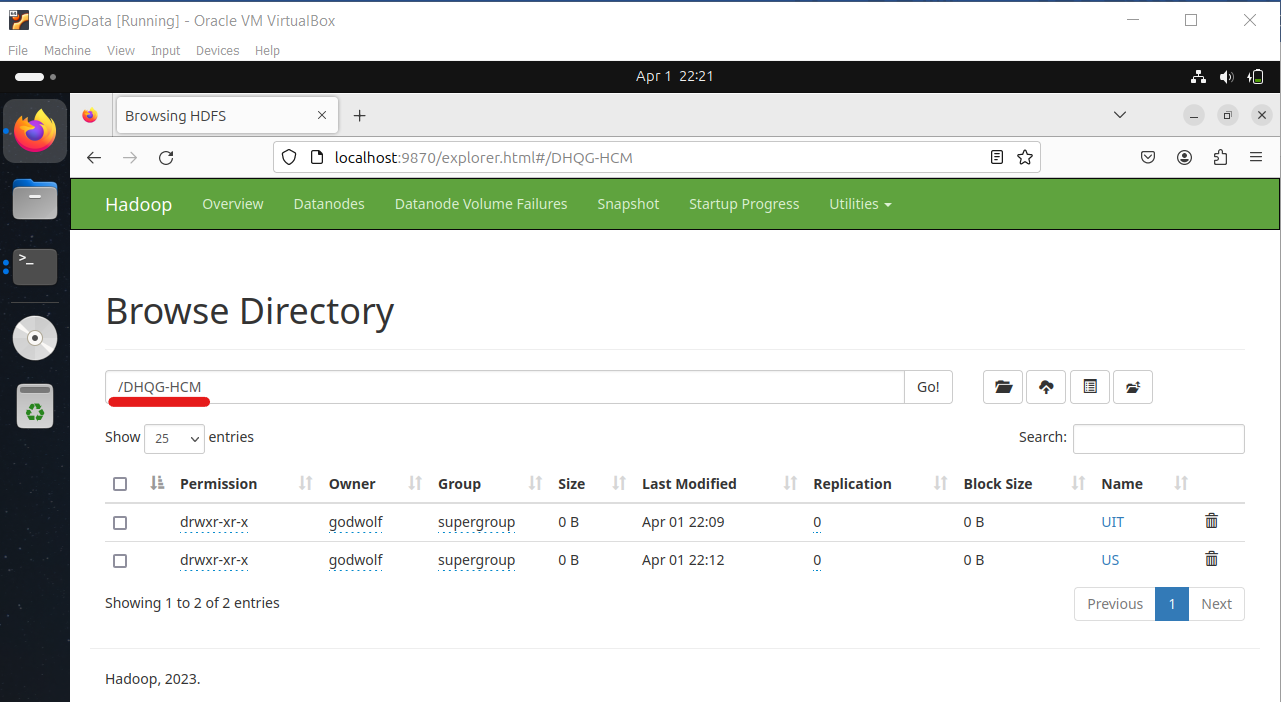


*Hình 4: Tạo cây thư mục /GomCum/kMeans, /GomCum/DBSCAN*

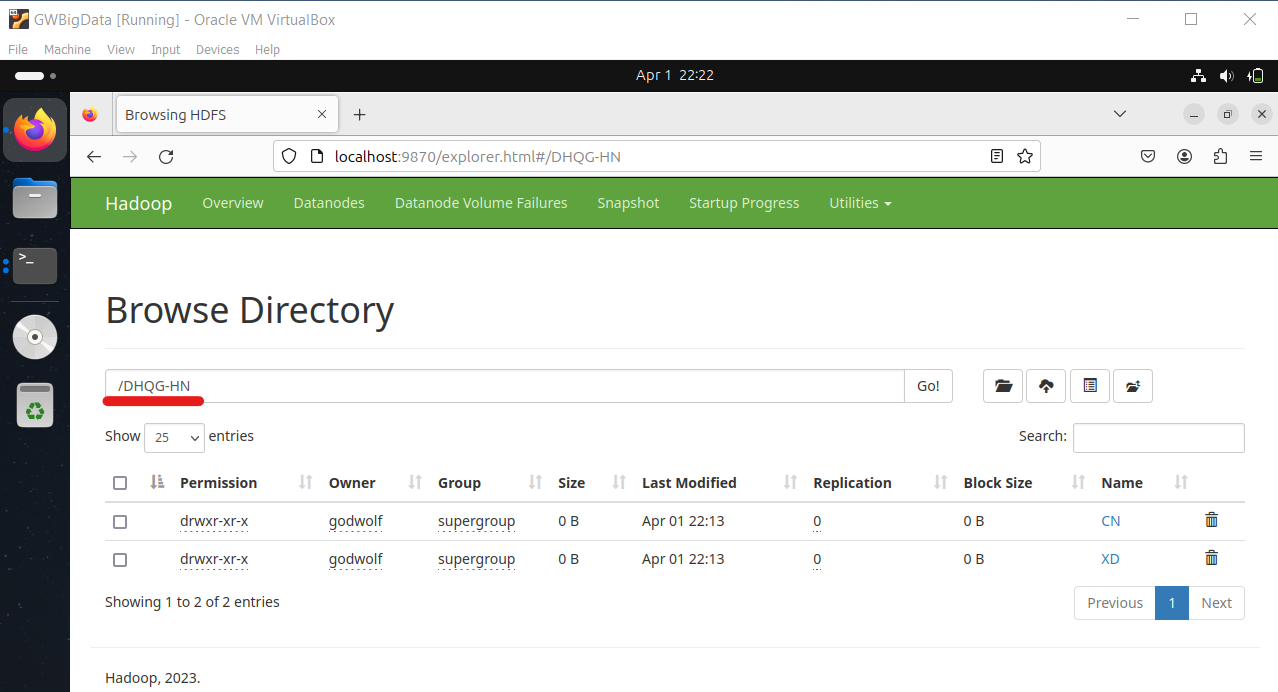
Bước 4: Vào localhost:9870 🡪 chọn Utilities 🡪 chọn Browse the file system để xem các thư mục vừa tạo



*Hình 5: Các thư mục cha*

**

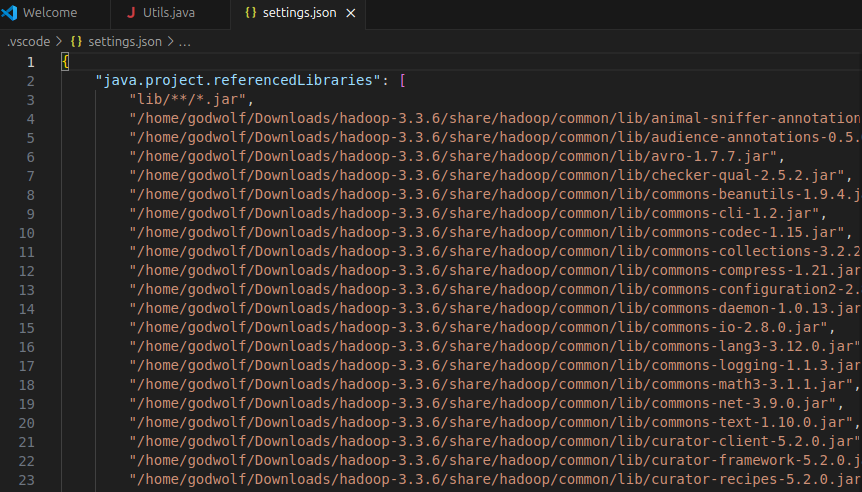
*Hình 6: Thư mục con của /DHQG-HCM*

**

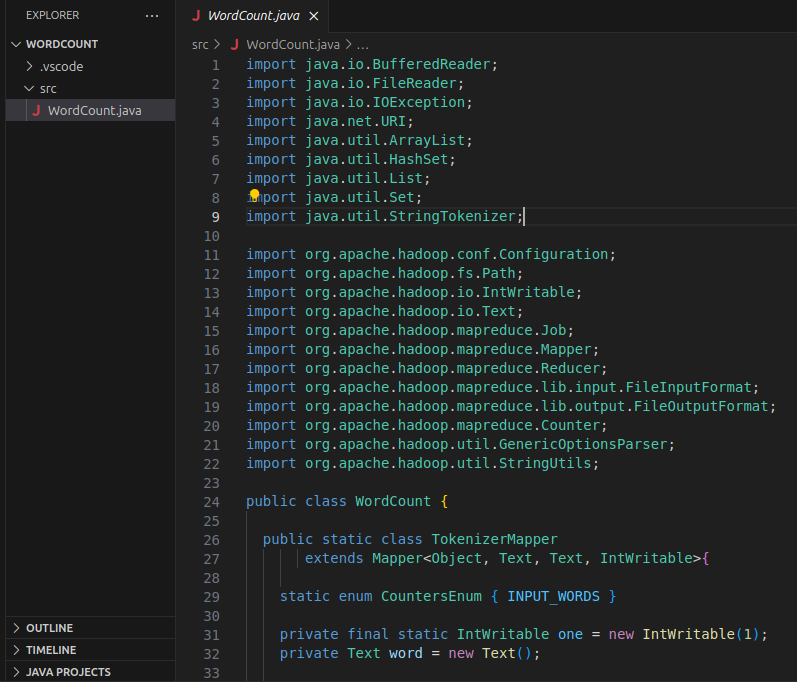
*Hình 7: Thư mục con của /DHQG-HN*

**Bài 2:** Tạo project WordCount bằng ngôn ngữ Java, thêm các thư viện cần thiết của Apache Hadoop

Bước 1: Tạo Java project mới, thêm các thư viện cần thiết



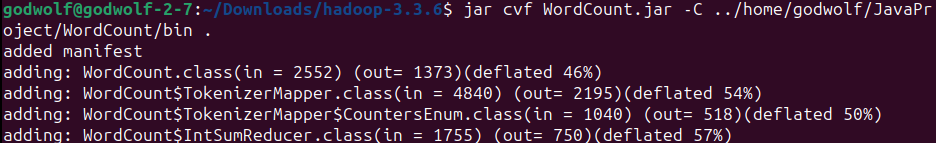
Bước 2: Thêm code chương trình WordCount v2.0



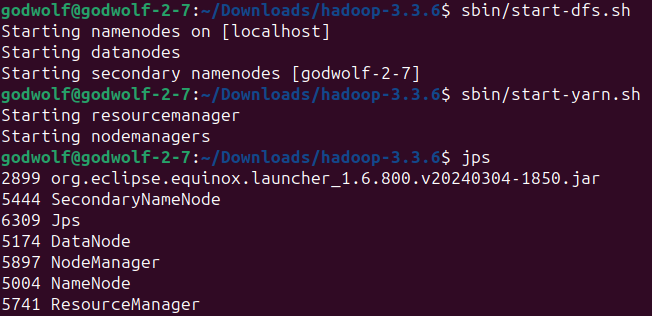
Bước 3: Nhập lệnh compile code java cho chương trình WordCount



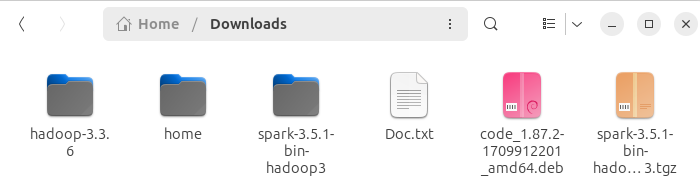
Bước 4: Tạo file jar tên WordCount chứa các class files:



Bước 5: Nhập lệnh chạy dfs và yarn

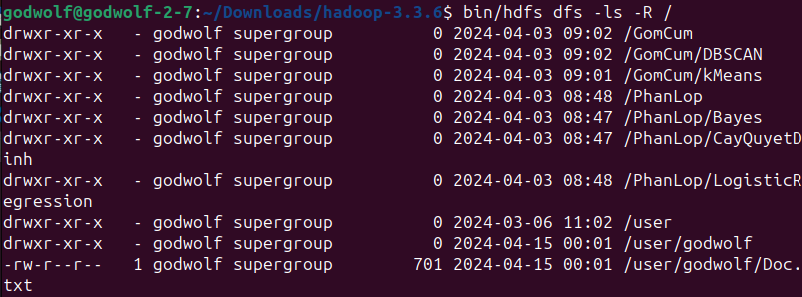


Bước 6: Tạo file Doc.txt trong folder Downloads. Nhập lệnh copy file doc từ máy chủ sang hdfs



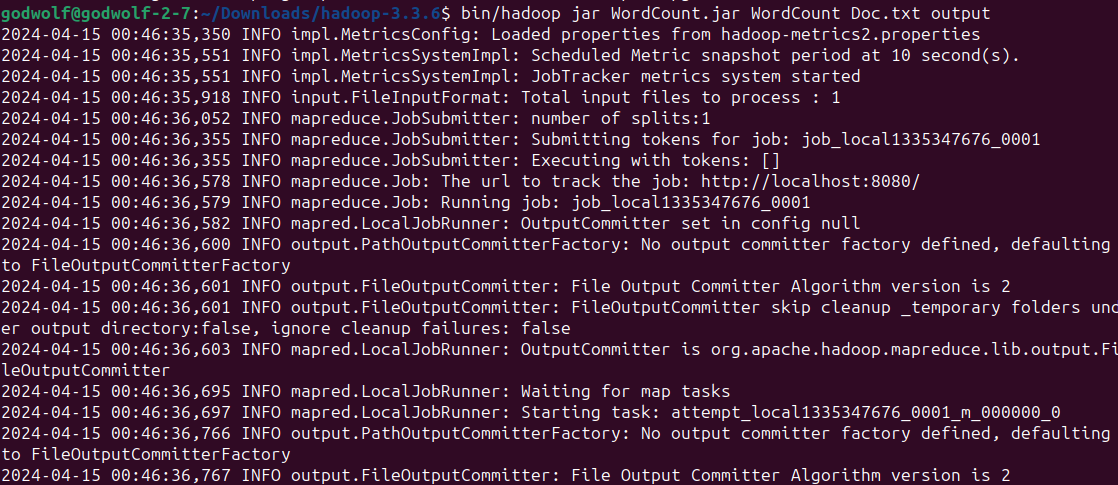


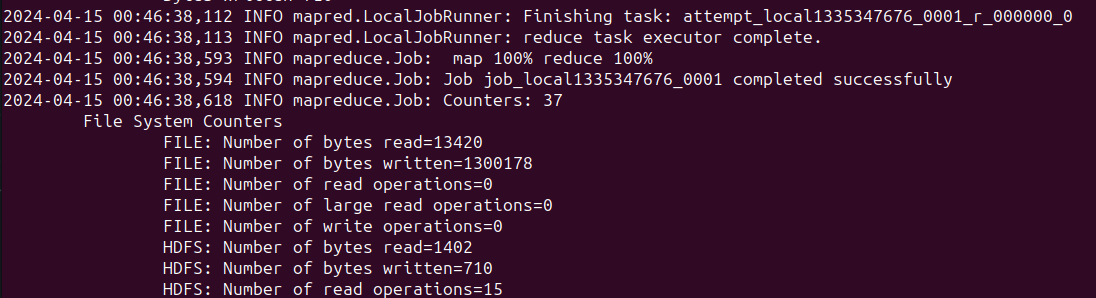
Bước 7: Nhập lệnh hiển thị danh sách file trên hdfs kiểm tra xem có file Doc không

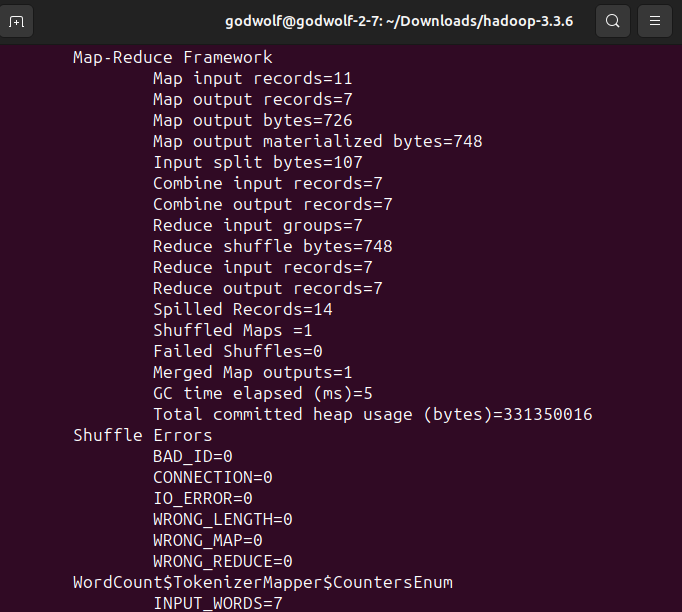


Bước 8: Xóa các output đang tồn tại và chạy chương trình

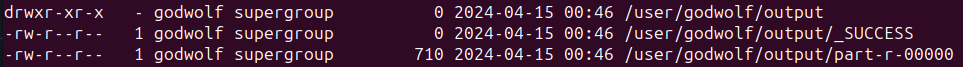






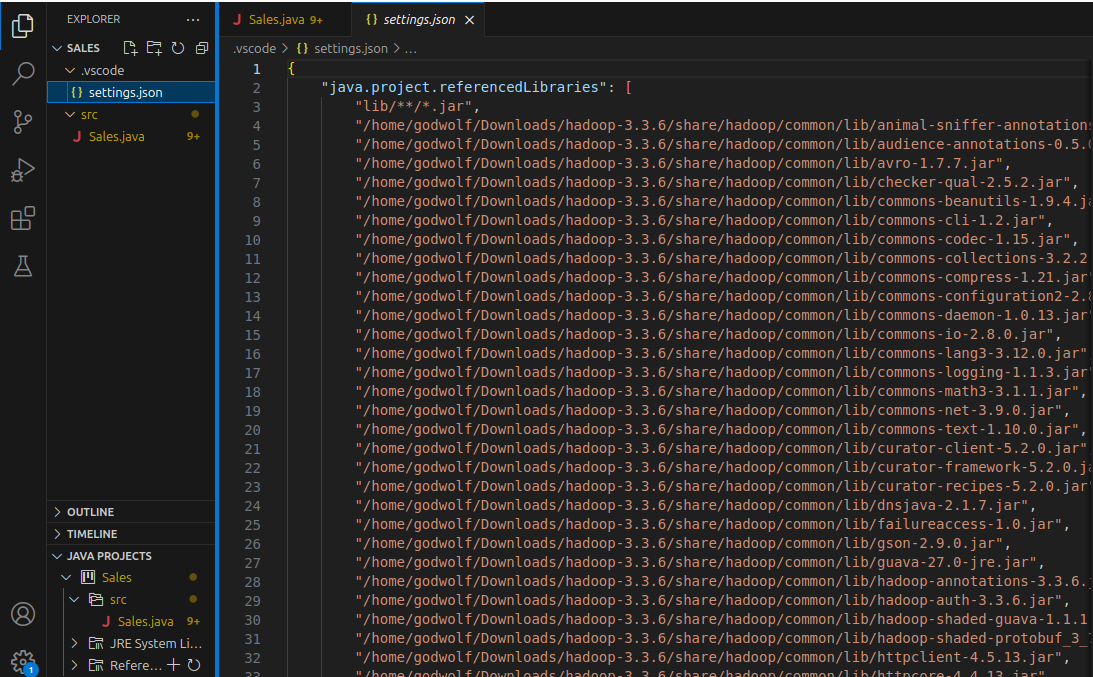


Bước 9: Kiểm tra output trên hdfs

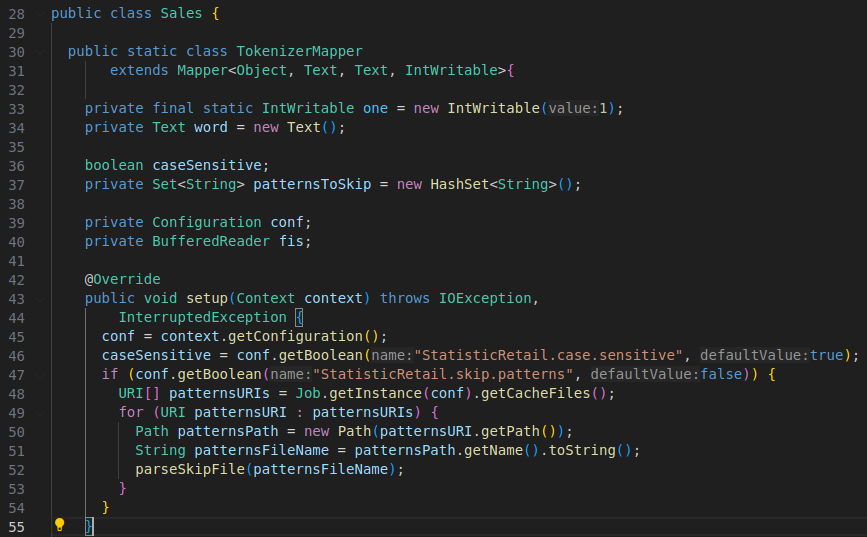


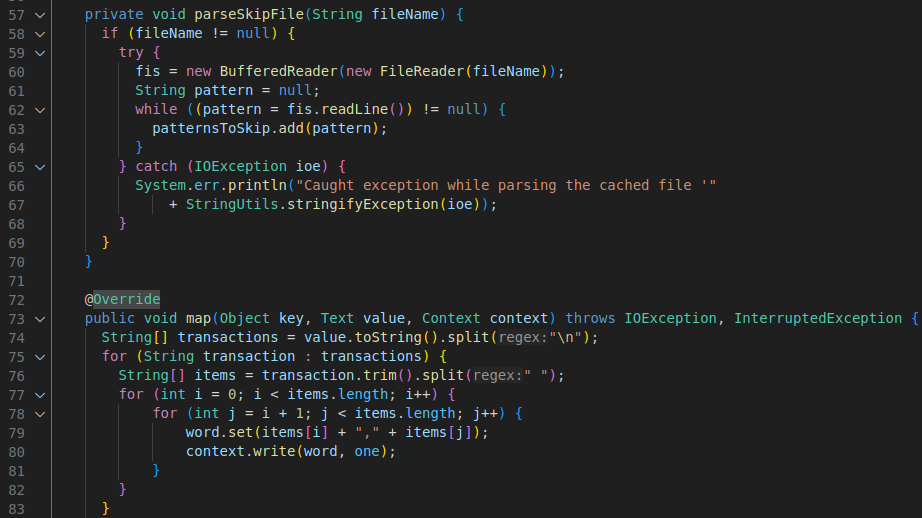
**Bài 3: b**

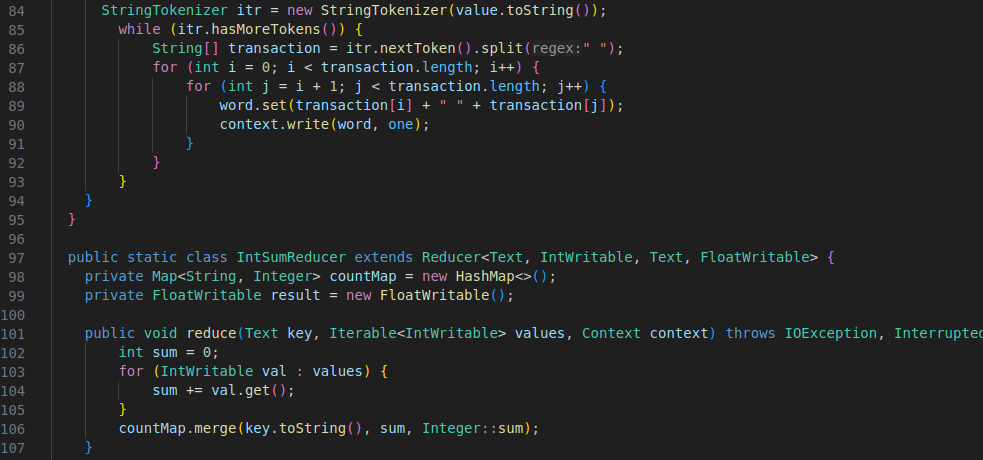
Bước 1: Tạo project mới và thêm các thư viện cần thiết

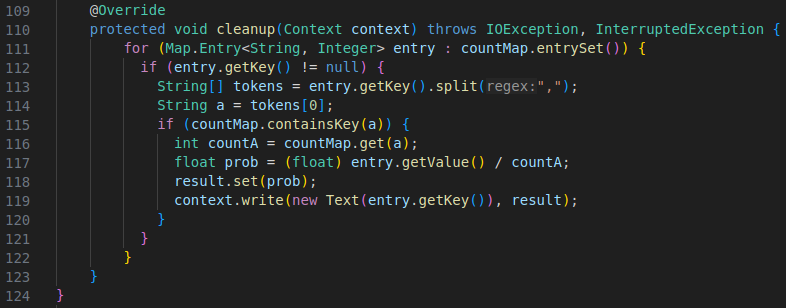


Bước 2: Viết chương trình MapReduce có chức năng theo yêu cầu đề 3b





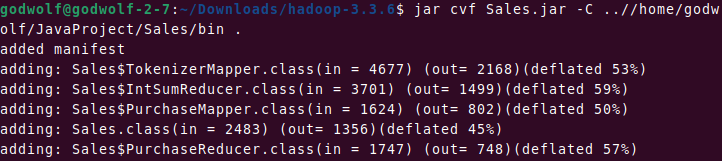




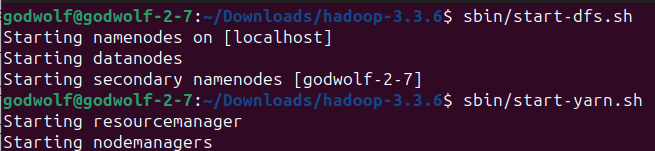
Bước 3: Nhập lệnh compile code java cho chương trình



Bước 4: Tạo file jar tên Sales chứa các class files:



Bước 5: Nhập lệnh chạy dfs và yarn



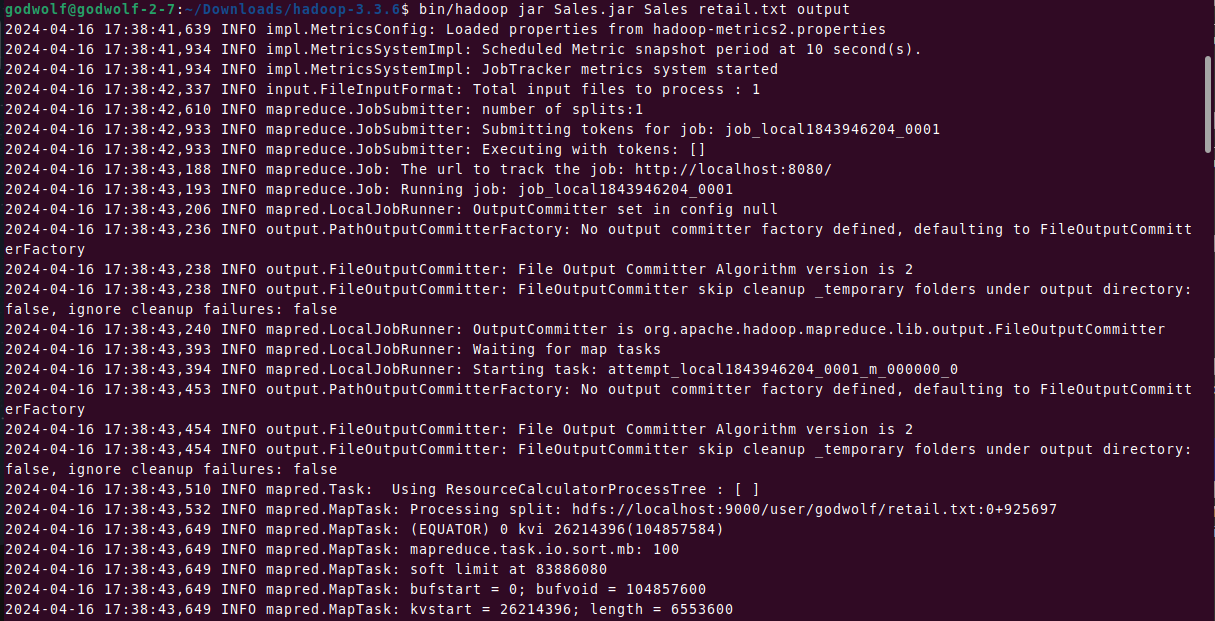
Bước 6: Tải dữ liệu retail.dat và lưu trữ vào hdfs

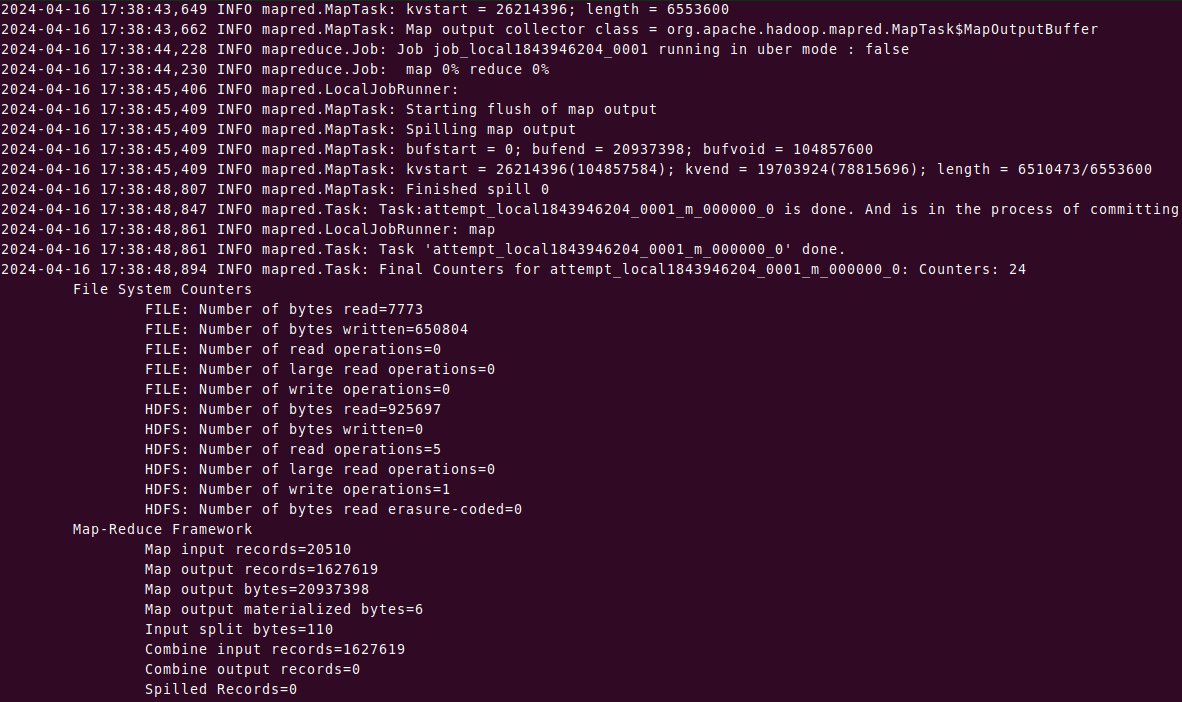


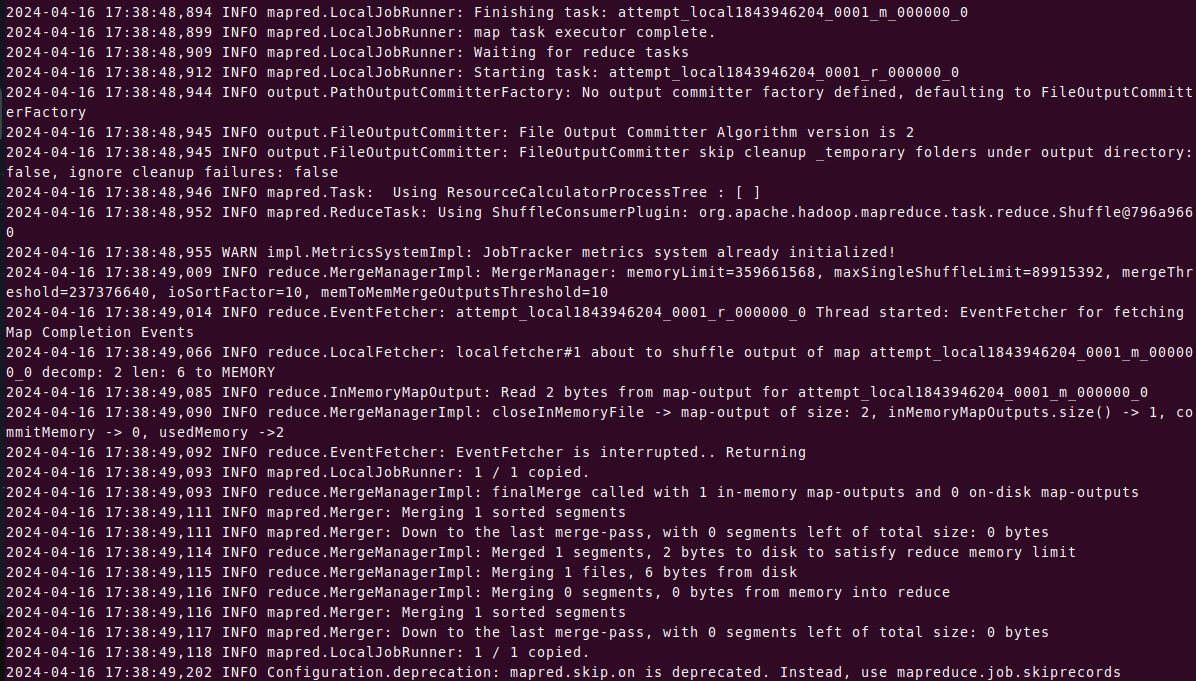
Bước 7: Nhập lệnh hiển thị danh sách file trên hdfs kiểm tra xem có file dat vừa lưu không

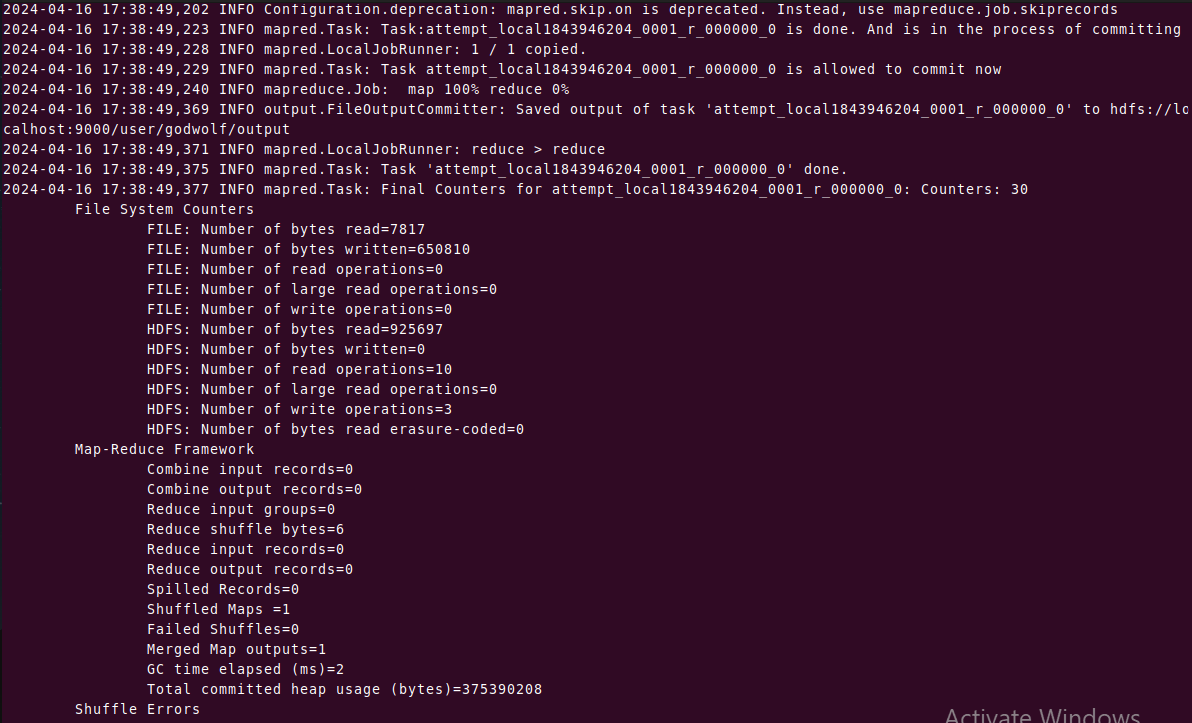
$bin/hdfs dfs -ls -R /

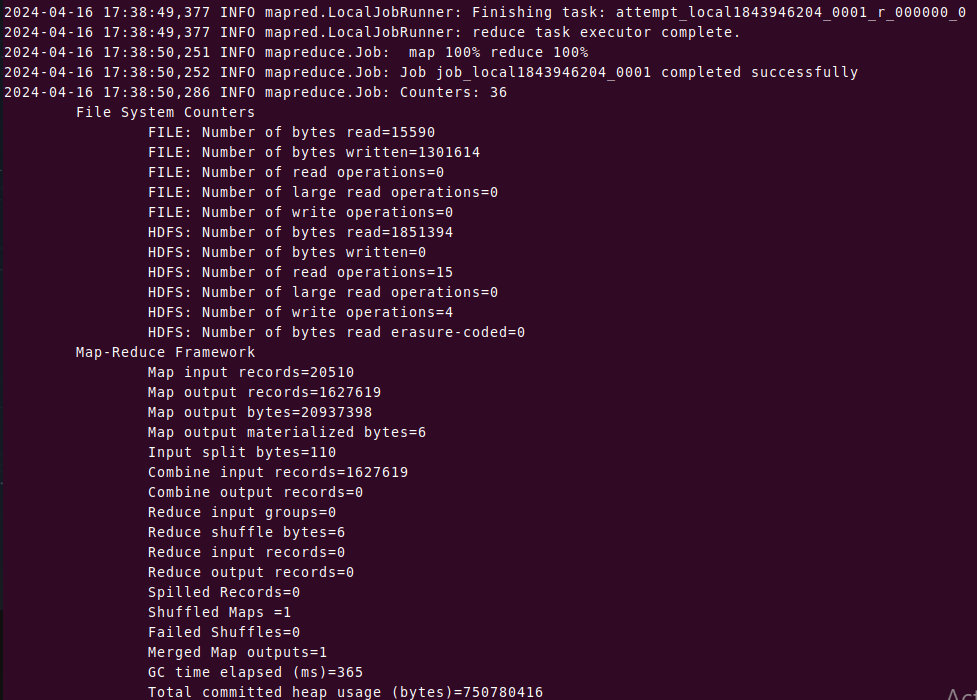


Bước 8: Chạy chương trình 









Bước 9: Kiểm tra output trên hdfs

