**Báo cáo Lưu Trữ Và Xử Lý Dữ Liệu Lớn**

**Lab4:Spark**

**Nhóm**:LoveData

**Thành viên**:

Phương Trung Kiên 20183776

Lê Nguyễn Thành Chung 20183696

Đào Trọng Hiếu 20183739

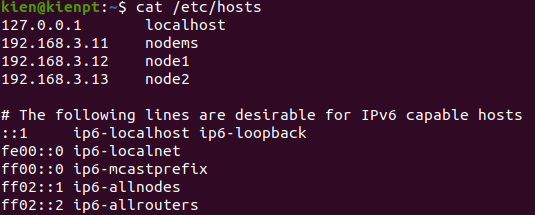
HIM PHON 20170283

1. **Yêu cầu:**

cài đặt cụm Spark (có thể dùng docker hoặc cài trực tiếp trên máy). Yêu cầu cài đặt trên 3 máy, có chụp hình minh chứng, và chạy chương trình wordcount (đọc dữ liệu từ HDFS).

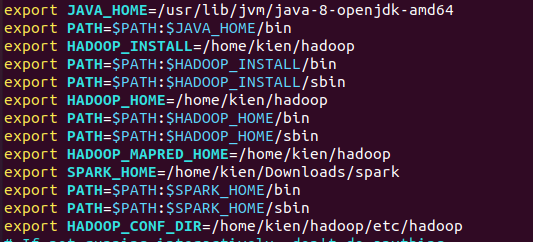
1. **Mô tả cài đặt:**

-Sử dụng 3 máy ảo từ 2 bài thực hành trước với các địa chỉ IP:



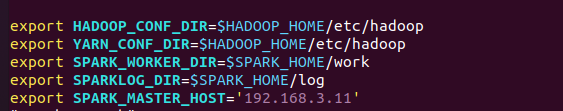
-Cài đặt spark trên từng máy ảo.

-Cài đường dẫn trong file .bashrc:

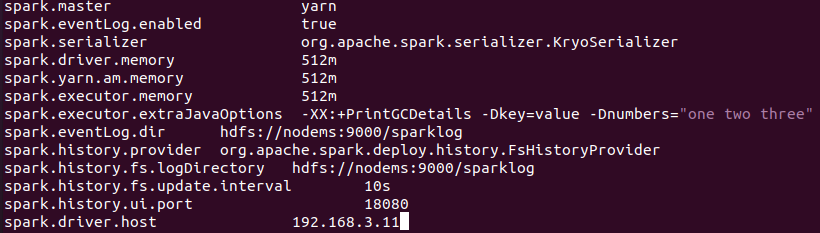


-Cấu hình các file của spark trong spark/conf:

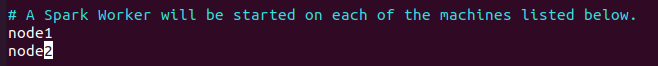
* File spark-env.sh:



* File spark-default.conf:



* File slaves:

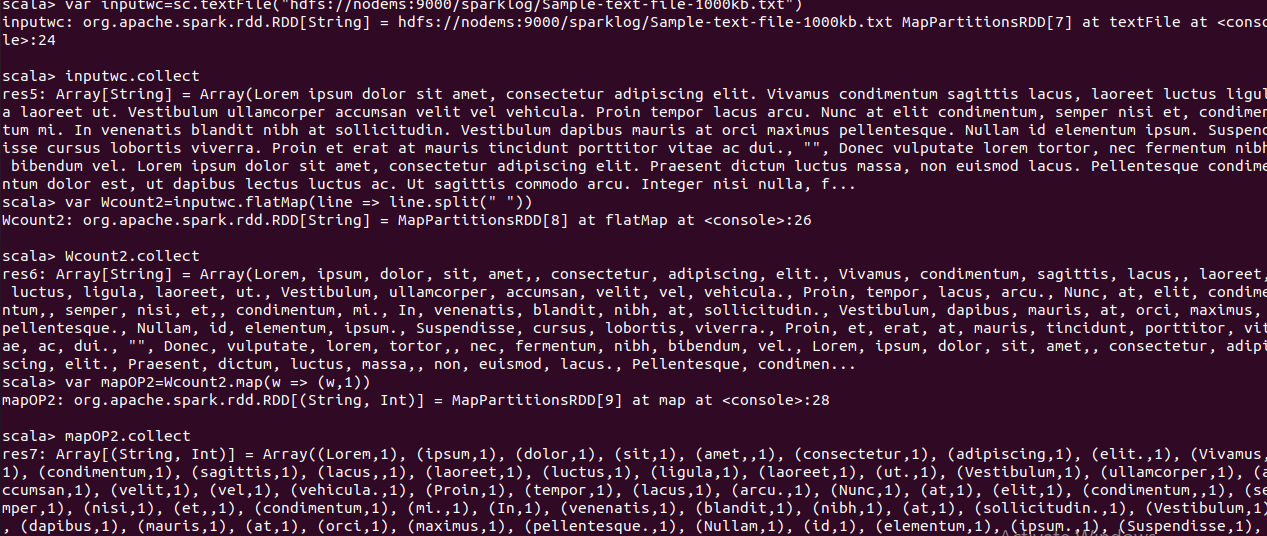


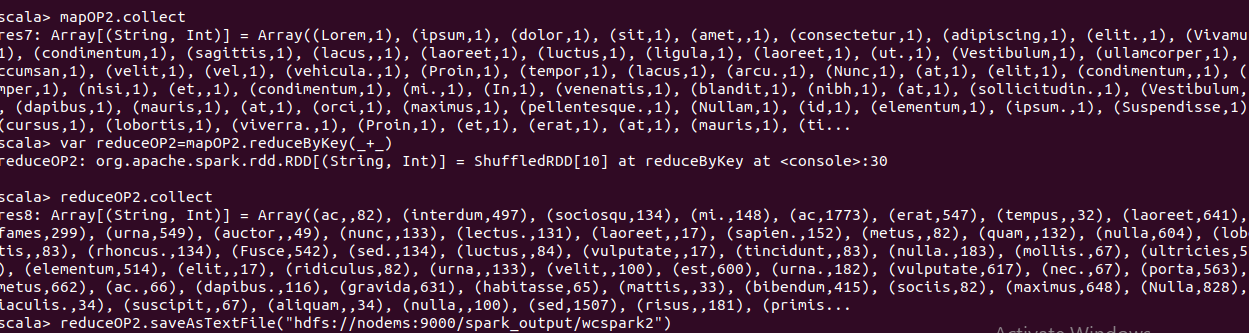
-Khởi tạo spark tới Yarn với chế độ Cluster:





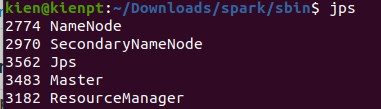
-Chạy chương trình WordCount trên scala sau khi thao tác lệnh spark-shell:

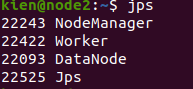


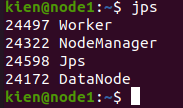


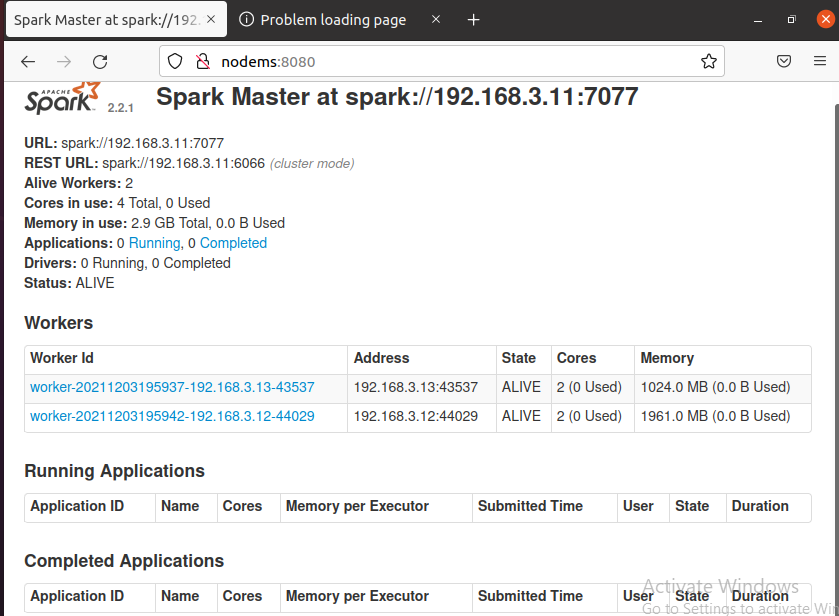
1. **Kết quả:**

-Kết quả thu được sau khi chạy thành công spark trên cụm máy ảo :

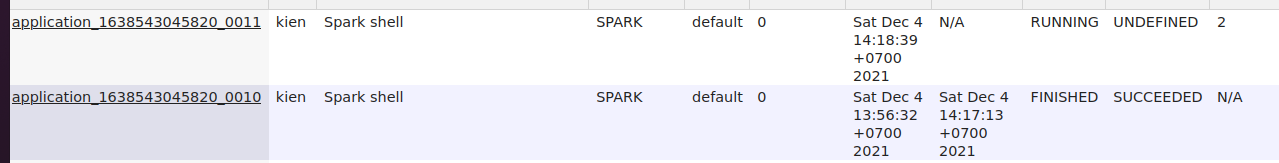


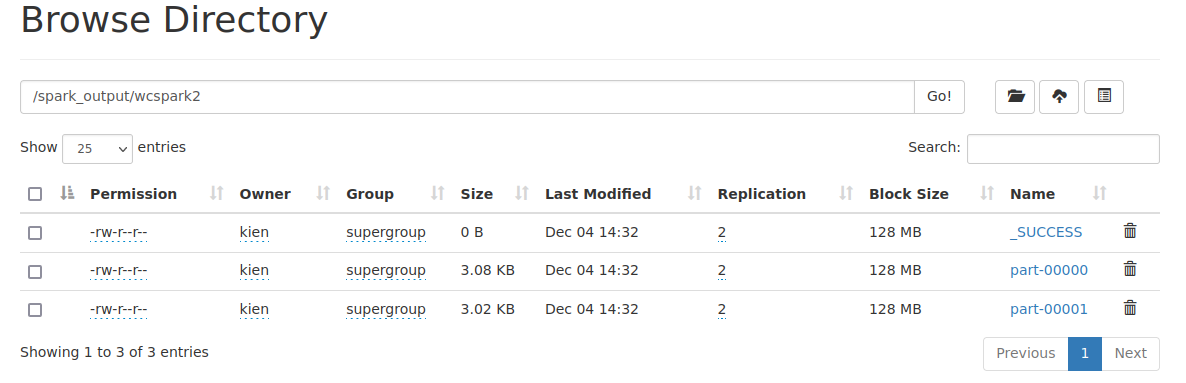




-

-Kết quả khi chạy spark-shell,job sẽ được hiển thị:



-Kết quả khi chạy được chương trình wordcount:

-Dữ liệu thu được:

