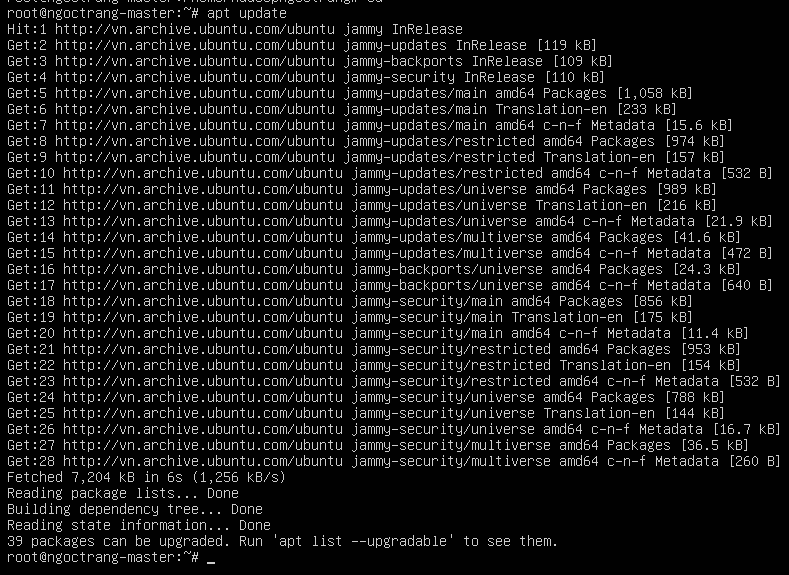
Trần Thị Ngọc Trang

MSSV: 21133109

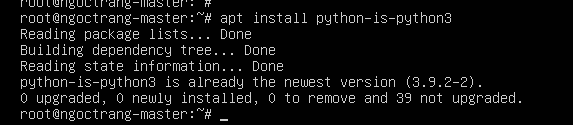
**HADOOP STREAMING**

1. **Install Python**

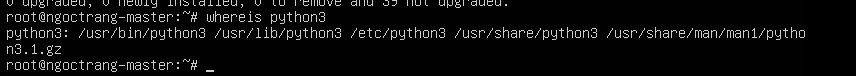
# apt update



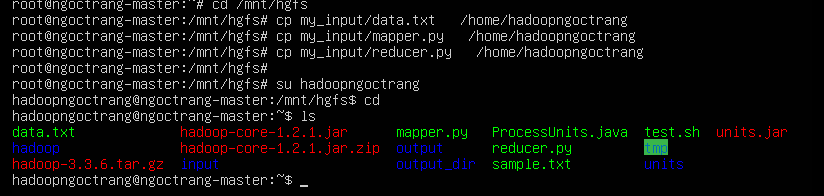
# apt install python-is-python3



# whereis python3

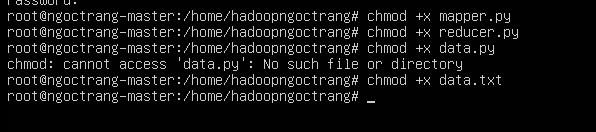


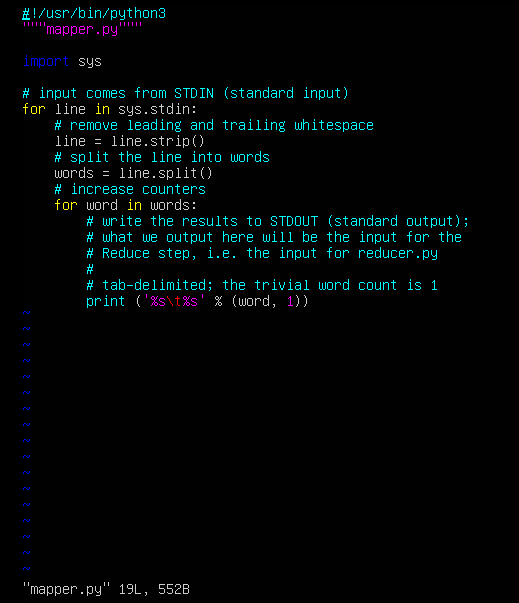
1. **Example Using Python**



* 1. **Mapper Phase Code**

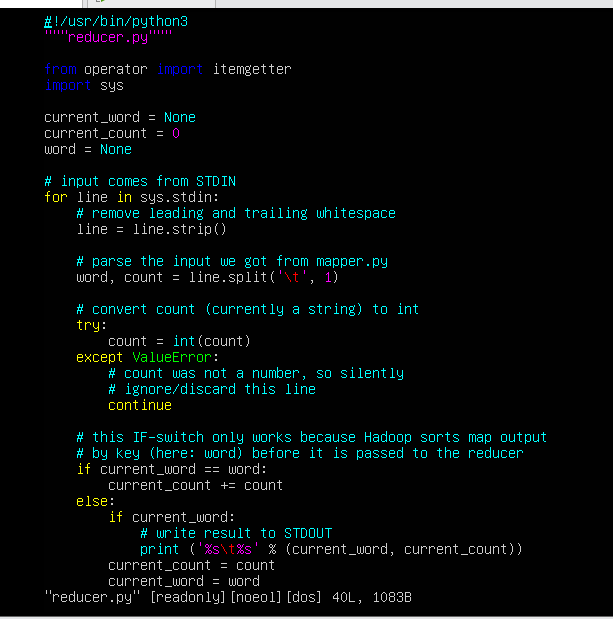
File mapper.py, chmod mapper.py



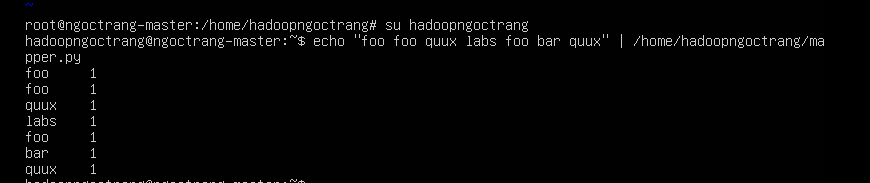


* 1. **Reducer Phase Code**

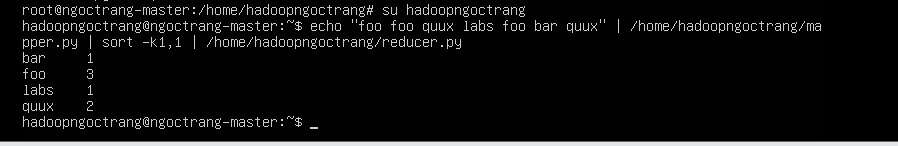
File reducer.py, chmod +x reducer.py



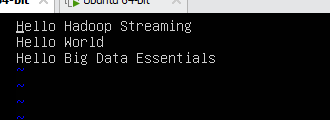
1. **Thực thi chương trình WordCount trên thư mục cục bộ**

$ echo "foo foo quux labs foo bar quux" | /home/hadoopngoctrang/mapper.py ****

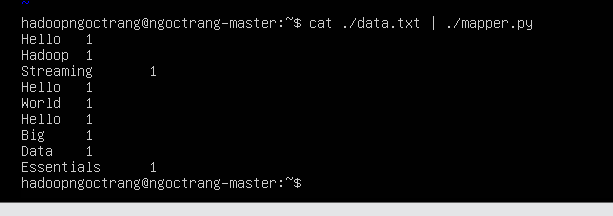
echo "foo foo quux labs foo bar quux" | /home/hadoopngoctrang/mapper.py | sort -k1,1 | /home/hadoopngoctrang/reducer.py



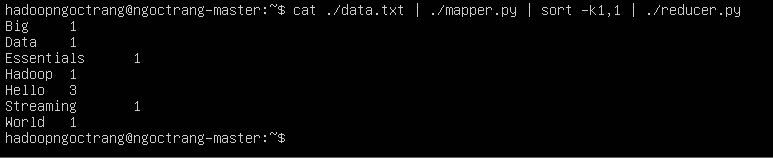
Tạo file data.txt chứa dữ liệu



cat ./data.txt | ./mapper.py



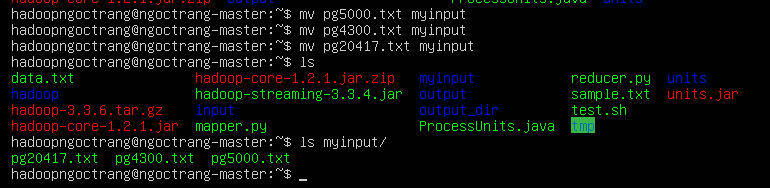
cat ./data.txt | ./mapper.py | sort -k1,1 | ./reducer.py



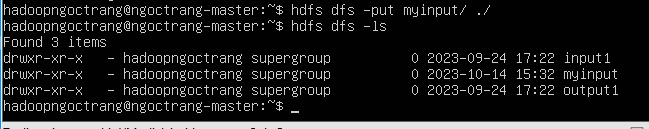
1. **Thực thi chương trình WordCount trên HDFS**

Tạo thư mục myinput chứa dữ liệu

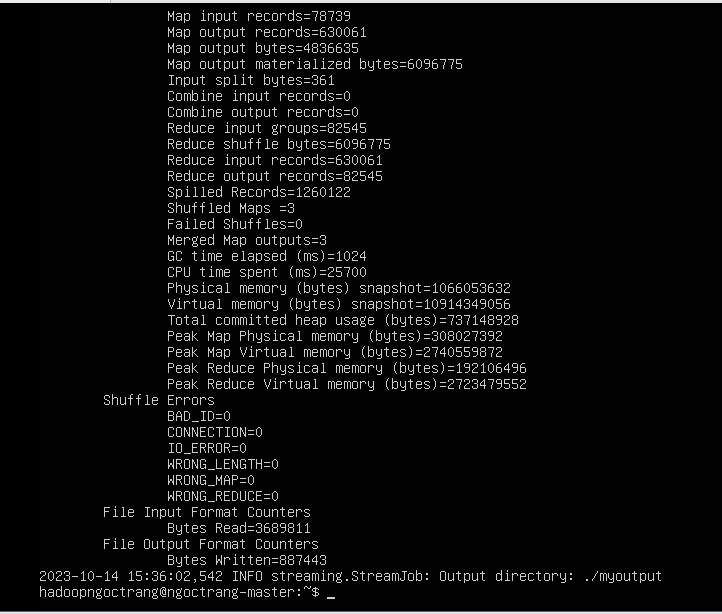
****

****

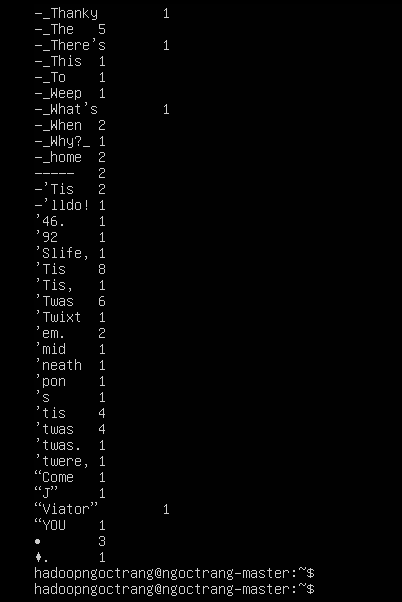
Copy thư mục myinput vào HDFS

****

hadoop jar hadoop-streaming-3.3.4.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input ./myinput -output ./myoutput

****

hdfs dfs -cat ./myoutput/part-00000

****