**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**



**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

─────── \* ───────



**BÀI TẬP TRÊN LỚP**

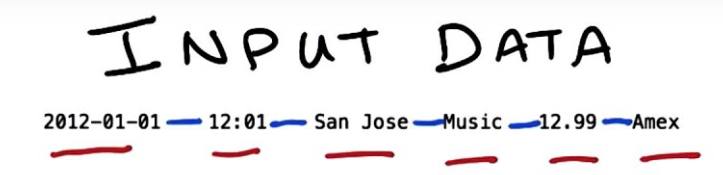
**MÔN HỌC: LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN**

**LAB 2: MAPREDUCE**

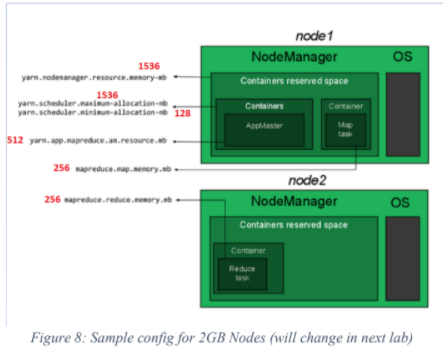
NHÓM: GIỮA CHÚNG TA

**Hà Nội, 11-2020**

**Set Up a MapReduce in Hadoop 3.2.1 Multi-Node Cluster on Ubuntu**

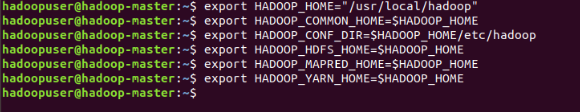


1. **Trên 3 máy master, slave1, slave2**



# 1st Step: Configure yarn

export HADOOP\_HOME="/usr/local/hadoop"  
export HADOOP\_COMMON\_HOME=$HADOOP\_HOME  
export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop  
export HADOOP\_HDFS\_HOME=$HADOOP\_HOME  
export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME  
export HADOOP\_YARN\_HOME=$HADOOP\_HOME



# 2nd Step: Configure mapred-site.xml

sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml

<configuration>

<property>

         <name>mapreduce.framework.name</name>

         <value>yarn</value>

</property>

<property>

         <name>yarn.app.mapreduce.am.env</name>

         <value>HADOOP\_MAPRED\_HOME=$HADOOP\_HOME</value>

</property>

<property>

         <name>mapreduce.map.env</name>

         <value>HADOOP\_MAPRED\_HOME=$HADOOP\_HOME</value>

</property>

<property>

         <name>mapreduce.reduce.env</name>

         <value>HADOOP\_MAPRED\_HOME=$HADOOP\_HOME</value>

</property>

<property>

     <name>yarn.app.mapreduce.am.resource.mb</name>

     <value>512</value>

</property>

<property>

     <name>mapreduce.map.memory.mb</name>

     <value>256</value>

</property>

<property>

     <name>mapreduce.reduce.memory.mb</name>

     <value>256</value>

</property>

</configuration>



# 3rd Step: Config yarn-site.xml

sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml

<configuration>

<!-- Site specific YARN configuration properties -->

<property>

         <name>yarn.acl.enable</name>

         <value>0</value>

</property>

<property>

<name>yarn.resourcemanager.hostname</name>

<value>hadoop-master</value>

</property>

 <property>

         <name>yarn.nodemanager.aux-services</name>

         <value>mapreduce\_shuffle</value>

</property>

<property>

     <name>yarn.nodemanager.resource.memory-mb</name>

     <value>1536</value>

</property>

<property>

     <name>yarn.scheduler.maximum-allocation-mb</name>

     <value>1536</value>

</property>

<property>

     <name>yarn.scheduler.minimum-allocation-mb</name>

     <value>128</value>

</property>

<property>

     <name>yarn.nodemanager.vmem-check-enabled</name>

     <value>false</value>

</property>

</configuration>



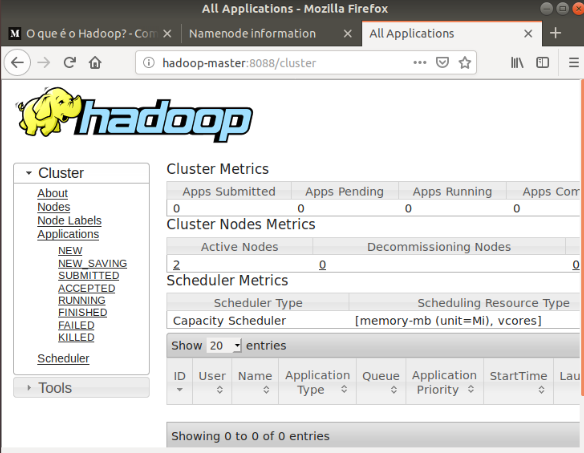
1. **Trên máy master**

# 1st Step: Start yarn

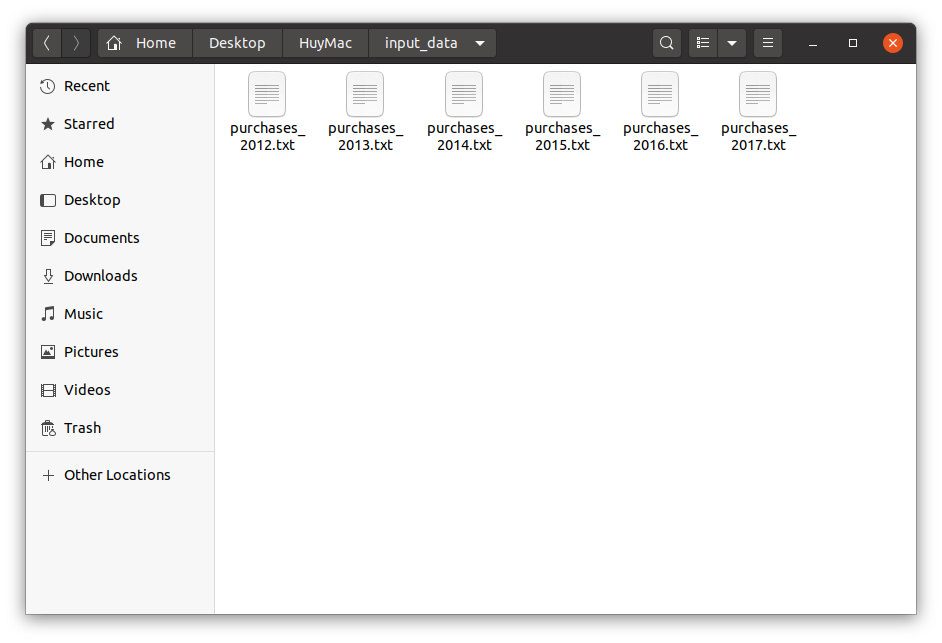
start-yarn.sh

https://lh6.googleusercontent.com/WK6xEsbcOSVxUgUVh3bGyDkFsIKOVImNHszb8zWUZr364ApPgSFUoZqSn7IQyXoa4qBAy3cpIV50yy1og2rd5w01NRy8a3N9eUrPCztE8rvJMdKgl3Ew_TL1234wh12esCw_Thhj

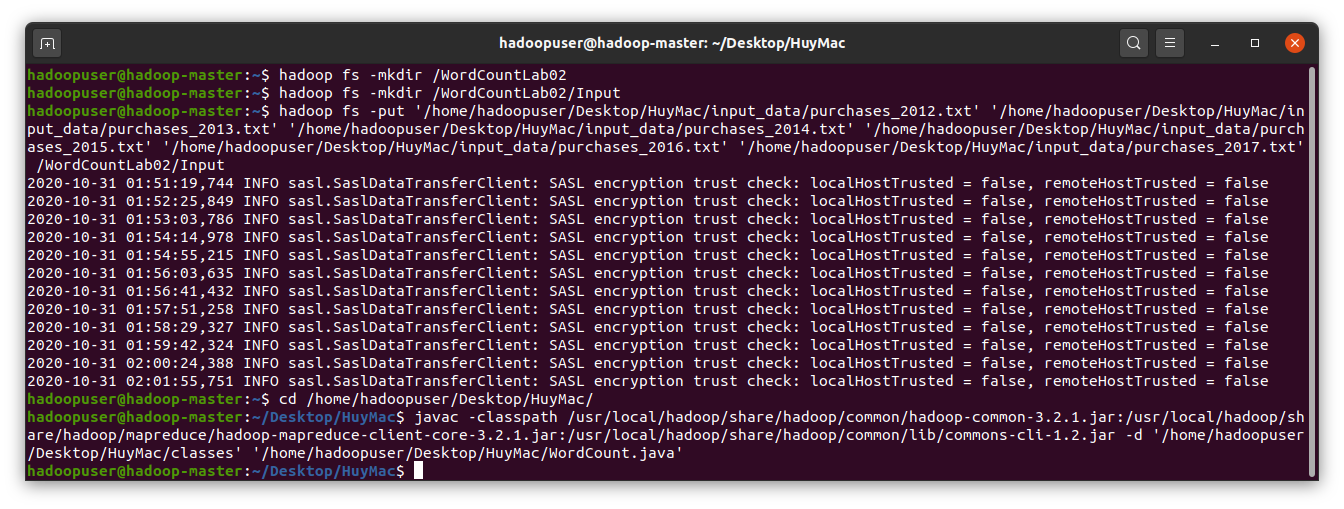
Lên web, truy cập [hadoop-master:8088/cluster](http://hadoop-master:8088/cluster)



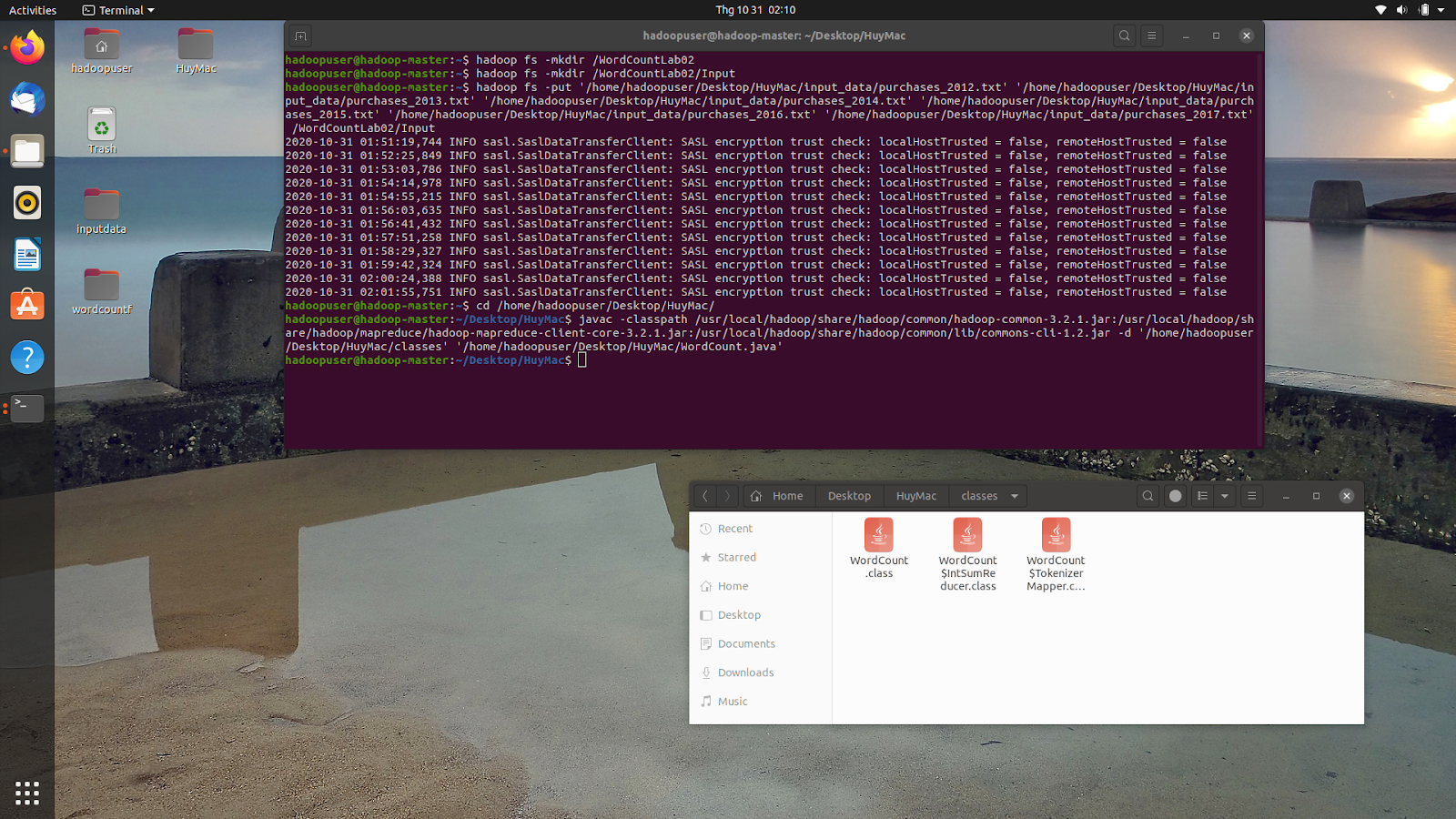
# 2nd Step: Tạo MapReduce Job

* Chuẩn bị data: Tài trên [https://github.com/MacHuy/HDFS-MultiNode/blob/main/HuyMac/input\_data/test.txt](https://github.com/MacHuy/HDFS-MultiNode/blob/main/HuyMac/input_data/test.txt%20)
* Hoặc có thể toàn bộ folder HuyMac và move nó để ở Desktop: <https://github.com/MacHuy/HDFS-MultiNode>
* 
* Updata lên HDFS folder /WordCountLab02/Input

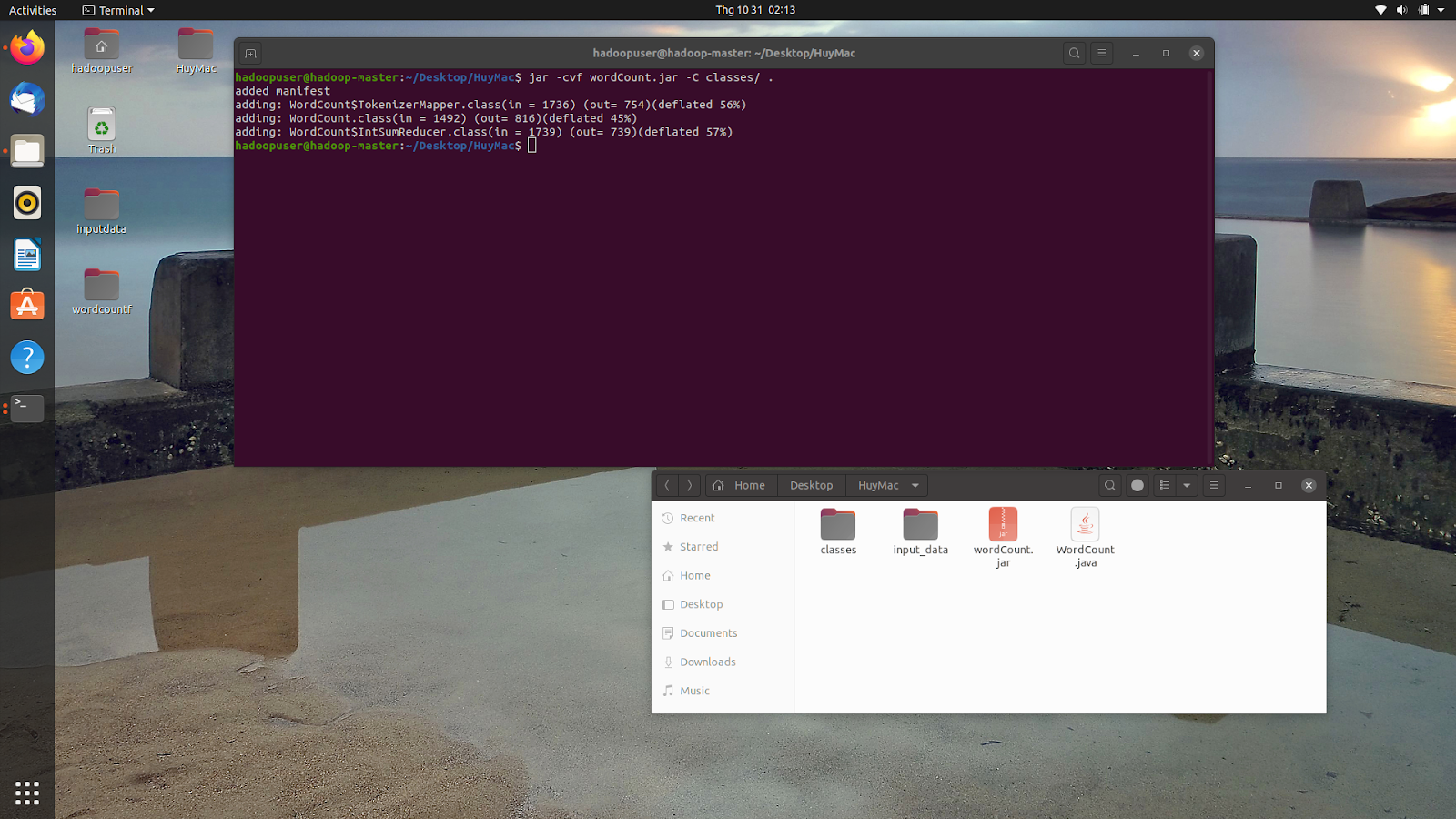
Thực thi file wordcount.java



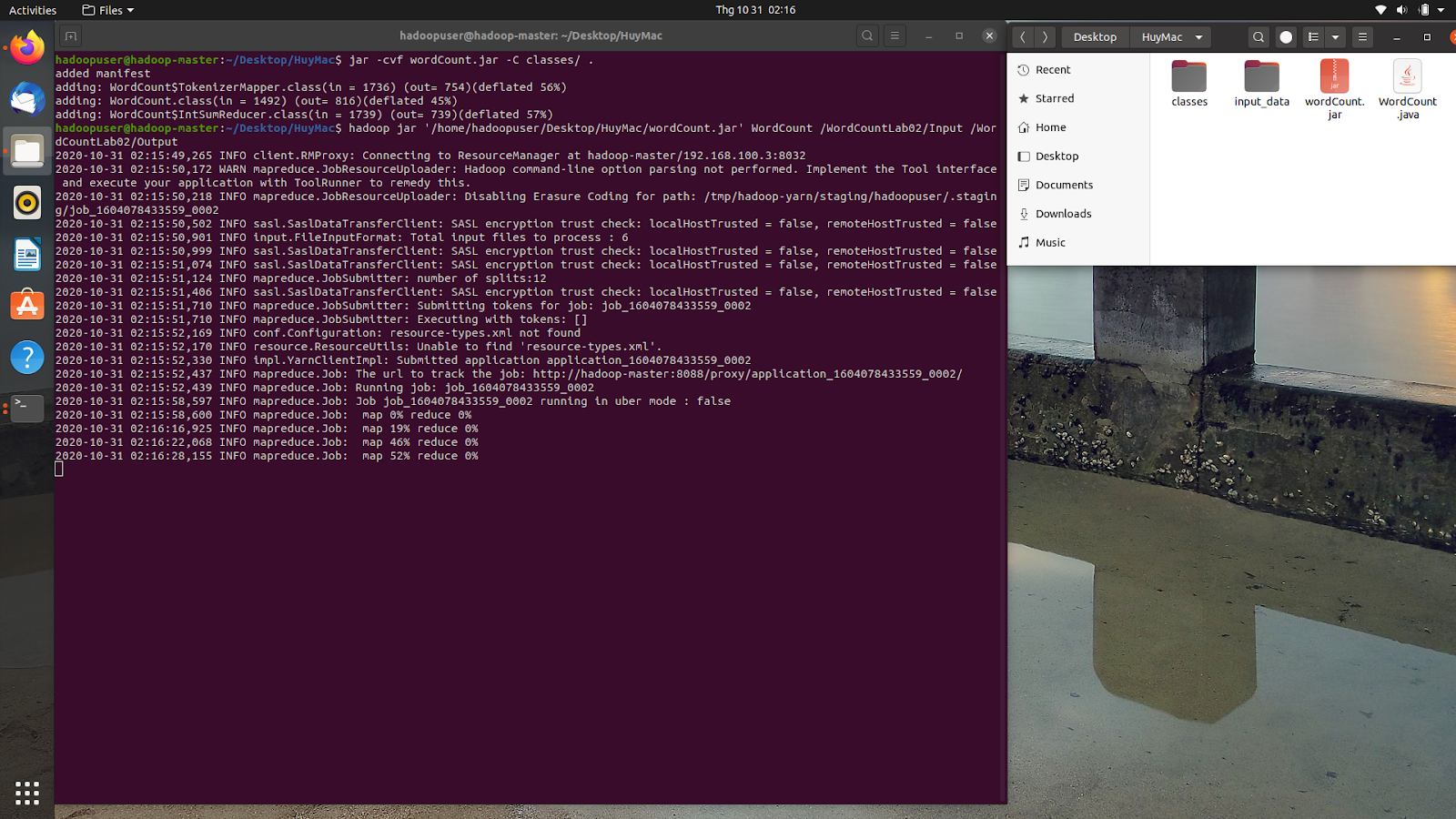
* Xuất hiện 3 file class trong folder /classes



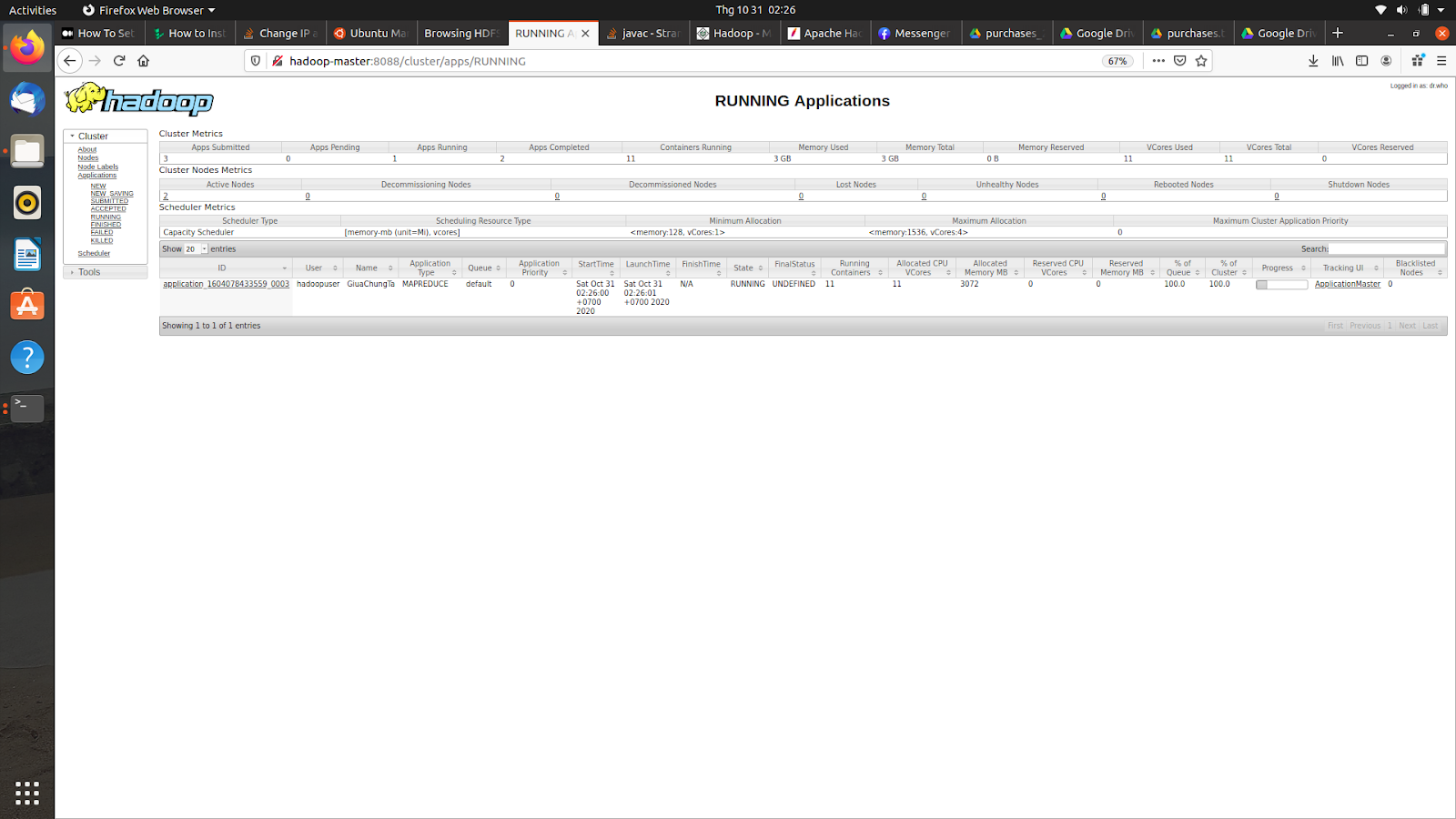
* Tạo file jar



* Sử dụng file jar thực thi chạy job mapreduce



* Lên web, check job đang chạy

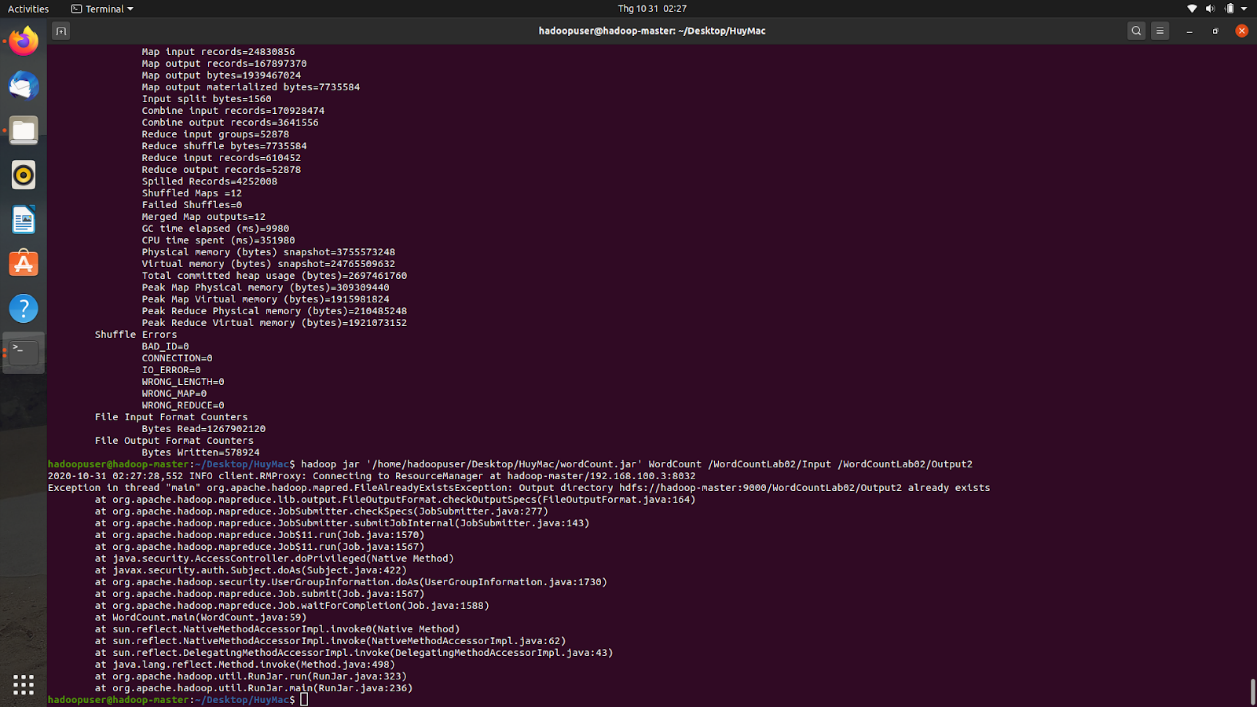


* Xem kết quả

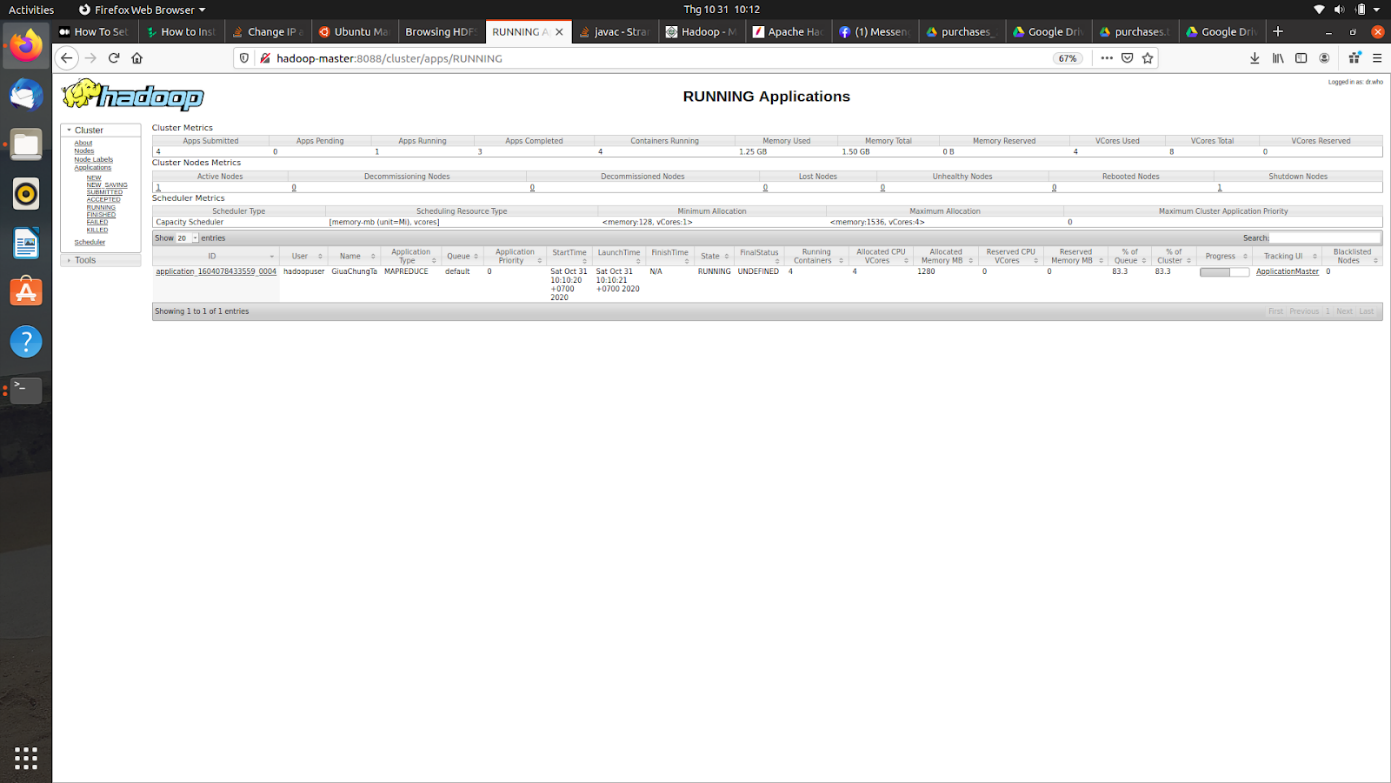
https://lh6.googleusercontent.com/xcATgNfGm6dhd7gQMfgev-WEOmrR5KsbOAS_W9dPgb9LRLusFy-1Kiw0lKTTsRCOpxkXHSWXCKeWASnAf8sLHWV7BQHoYTtzK9Yf8gsGvk5reD8BAlyIwC9KIcqf-cKC1KIjorQq



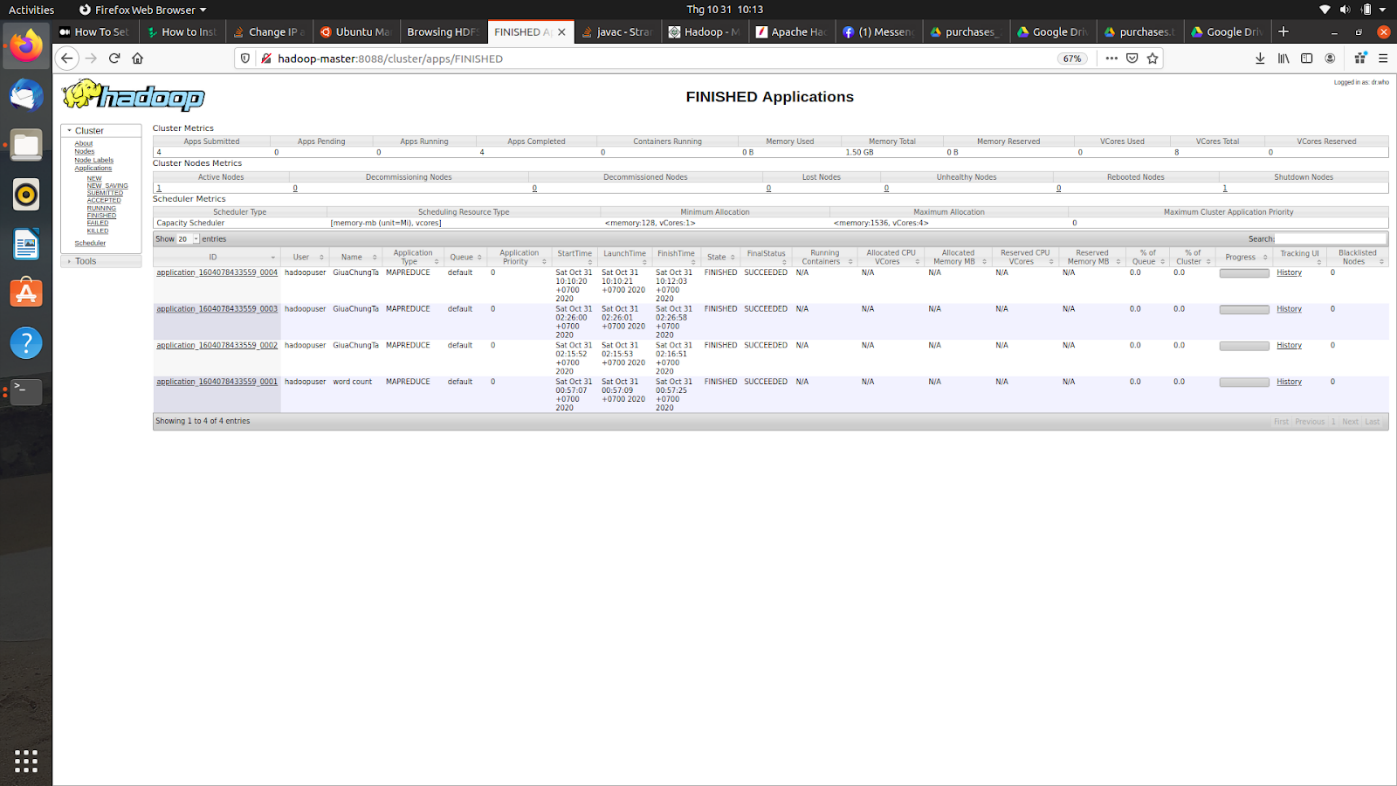
TH chạy lại job, nhớ đổi sang file output mới, folder output cũ đã có, HDFS ko cho phép overwrite



**Trường hợp chỉ có 1 datanode slave1**



* Check các job finished

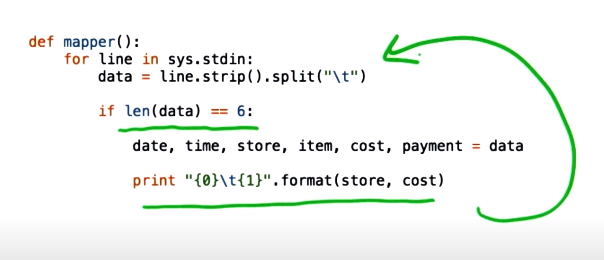


**Part2: MapReduce cho job “Tính tổng Sales/Store”**

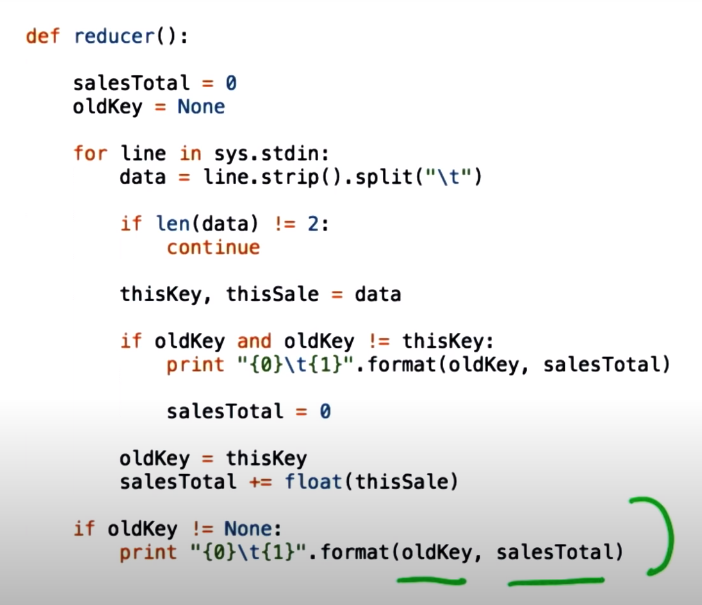
* Với dữ liệu đầu vào phần trước chọn cặp (key,value) = (store name, cost)

Sử dụng Hadoop streaming để code bằng Python

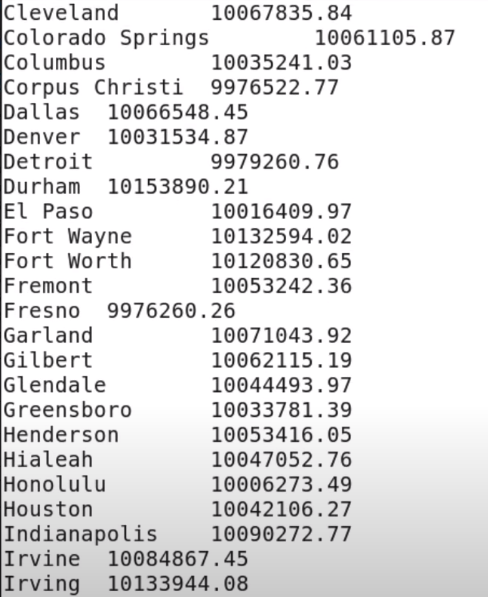
* **Mapper Code:**



* **Giữa Mapper và Reducer:** Sử dụng Shuffle and Sort
* **Reducer Code**



* **Output**



**Toàn bộ về Hadoop 3.2.1 Multi-Node Cluster and Mapreduce Job:**

<https://github.com/MacHuy/HDFS-MultiNode>

どうもありがとう、先生