**BÀI TẬP TRÊN LỚP**

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

**LAB 1: HDFS**

NHÓM: GIỮA CHÚNG TA

# Kết quả cài đặt

* 1. Configure the System

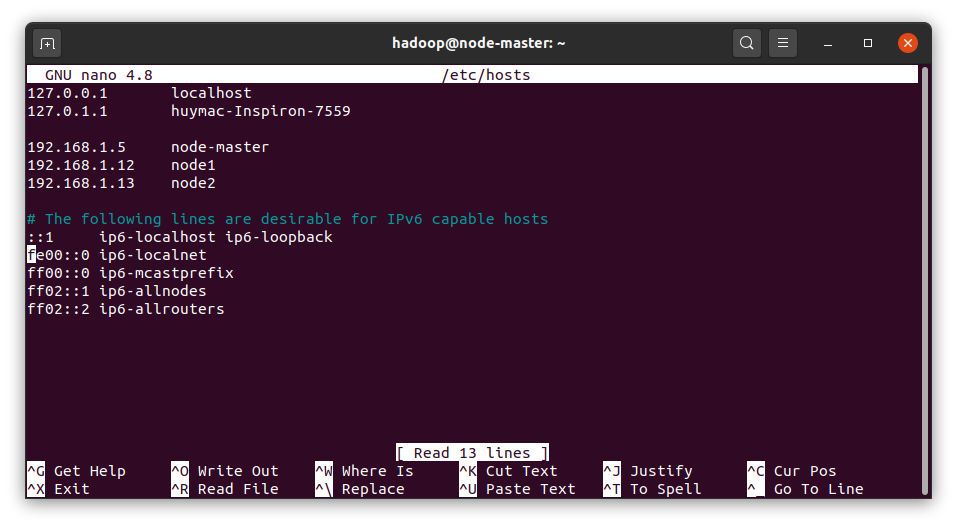


Figure :Hosts file

* 1. Set Evironment Variables

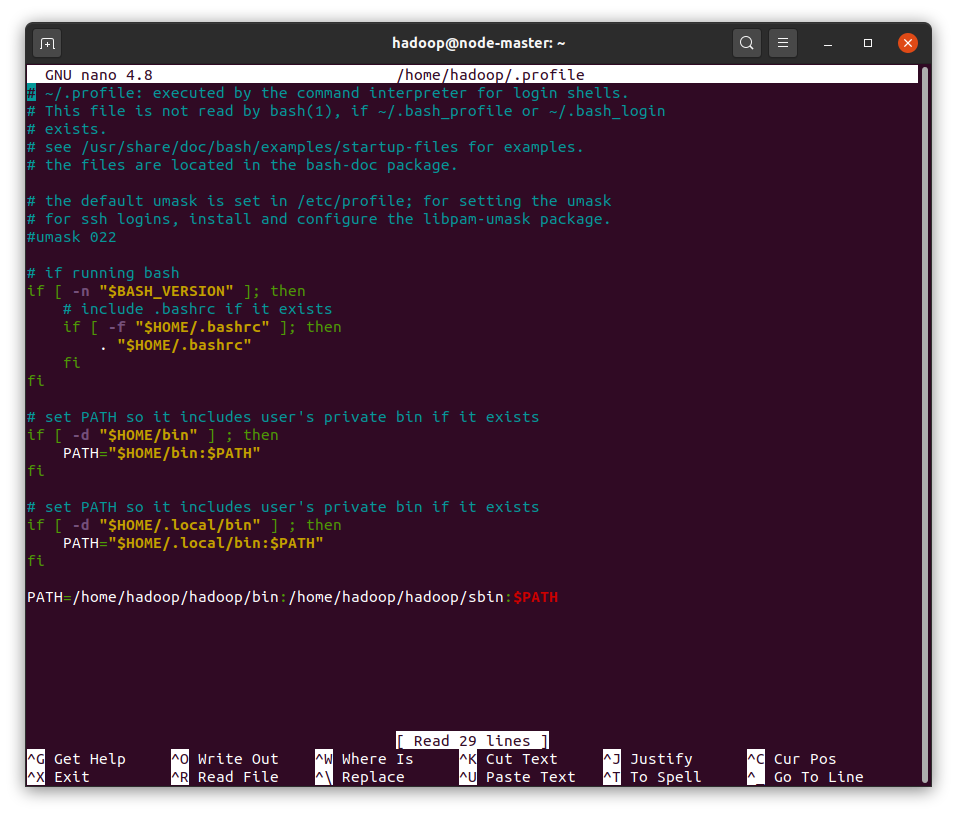


Figure :Add Hadoop binaries to PATH

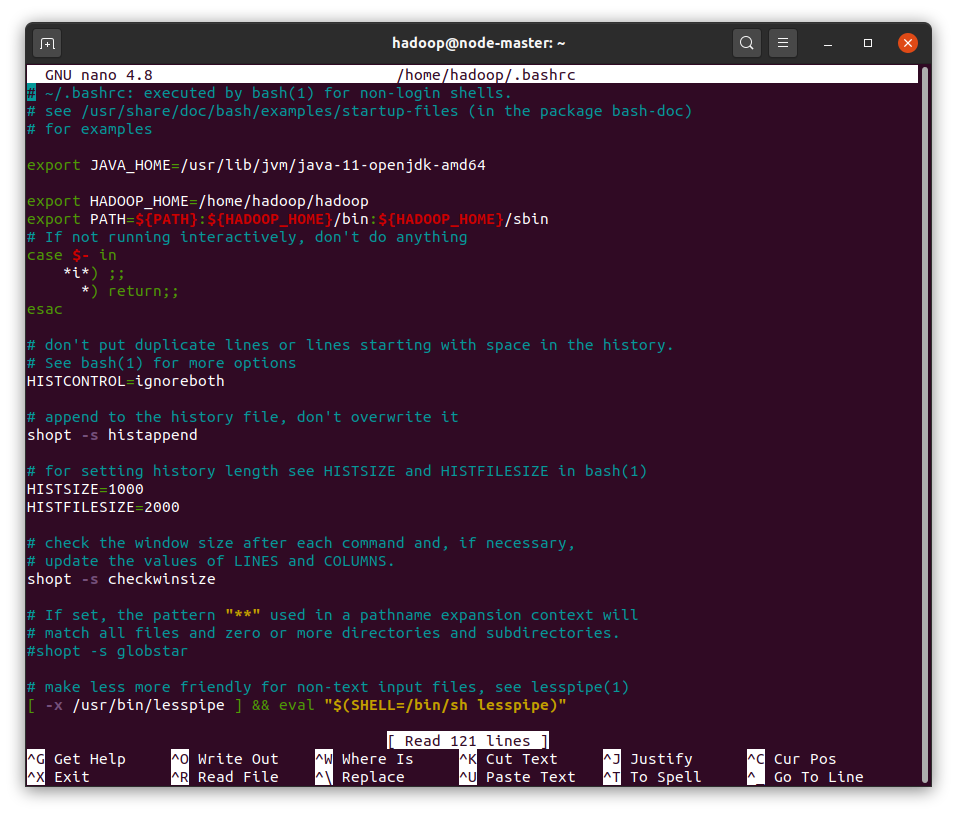


Figure : Add Hadoop to your PATH for the shell

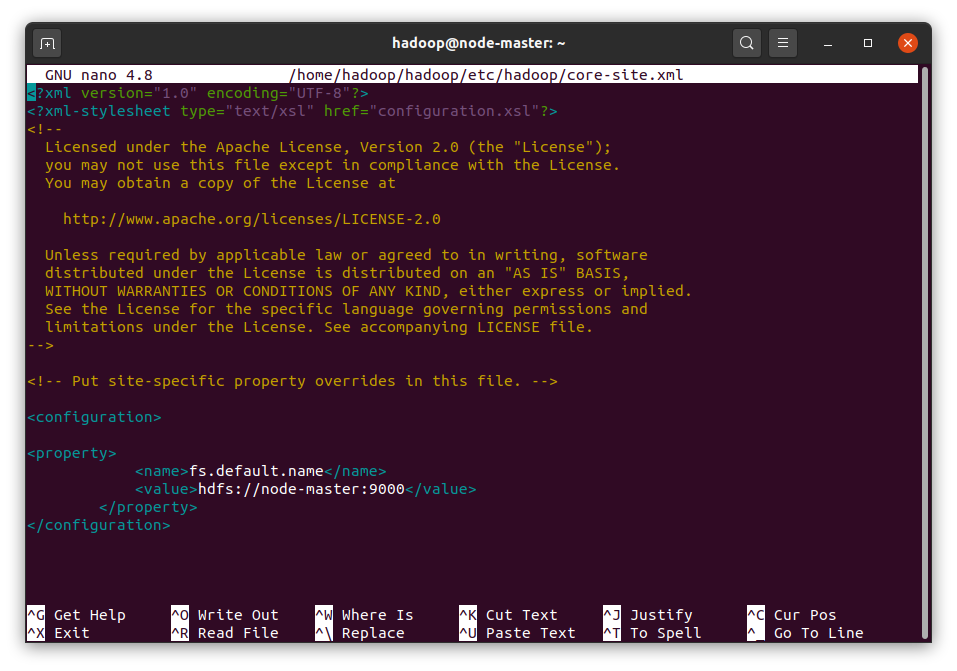


Figure : Set NameNode Location

* *Dfs.replication = 2*

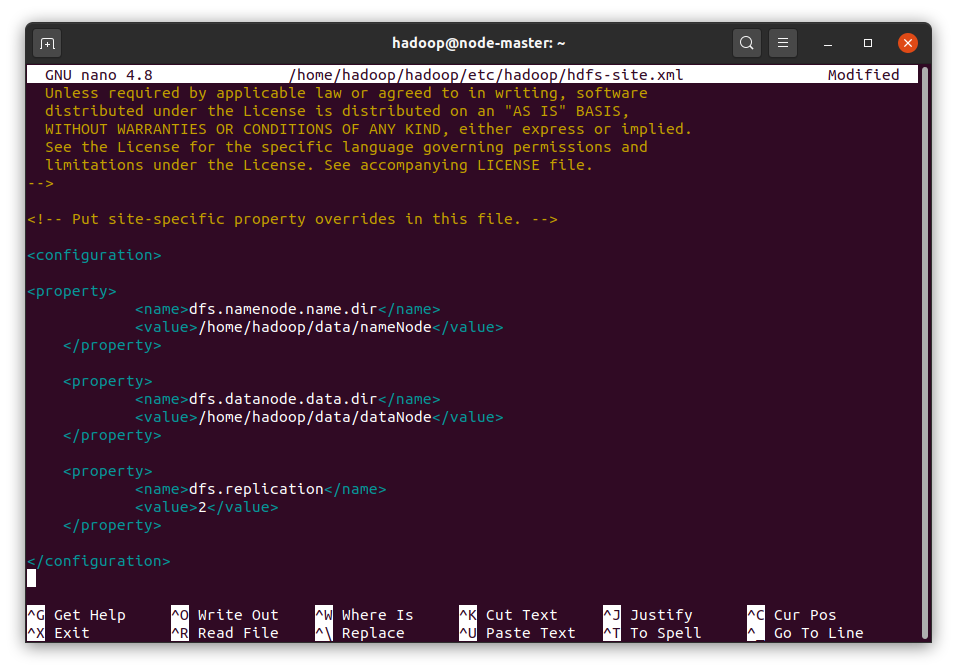


Figure : Set path for HDFSPermalink

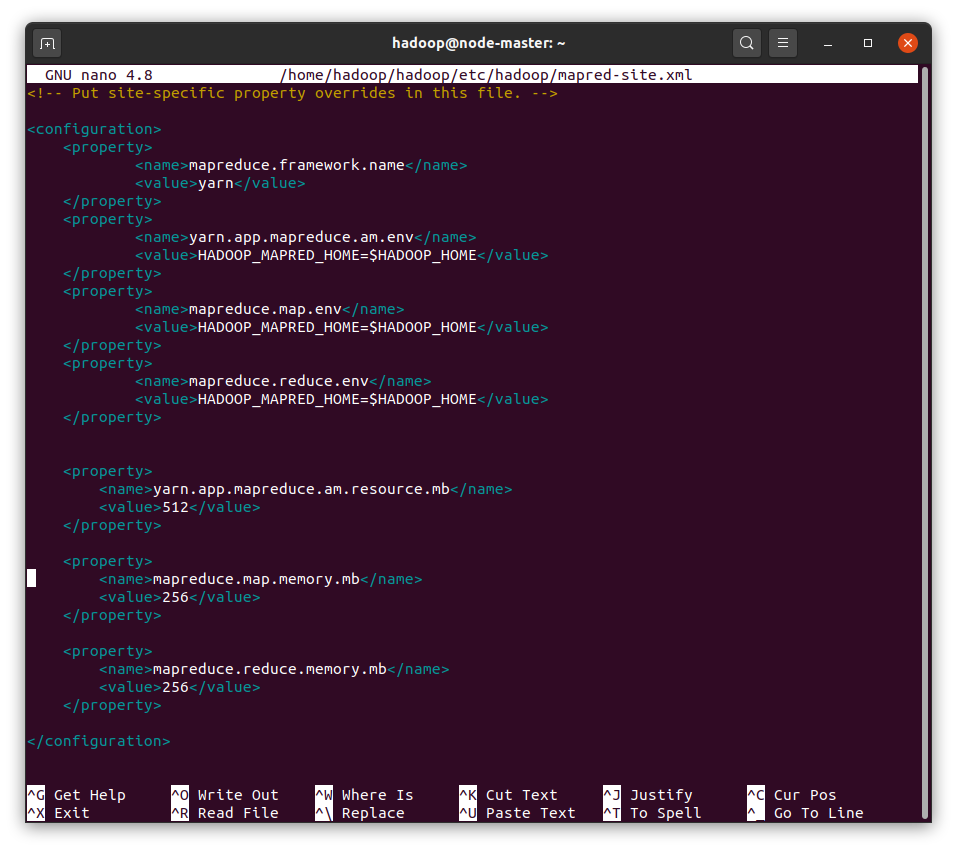


Figure : Set YARN as Job Scheduler

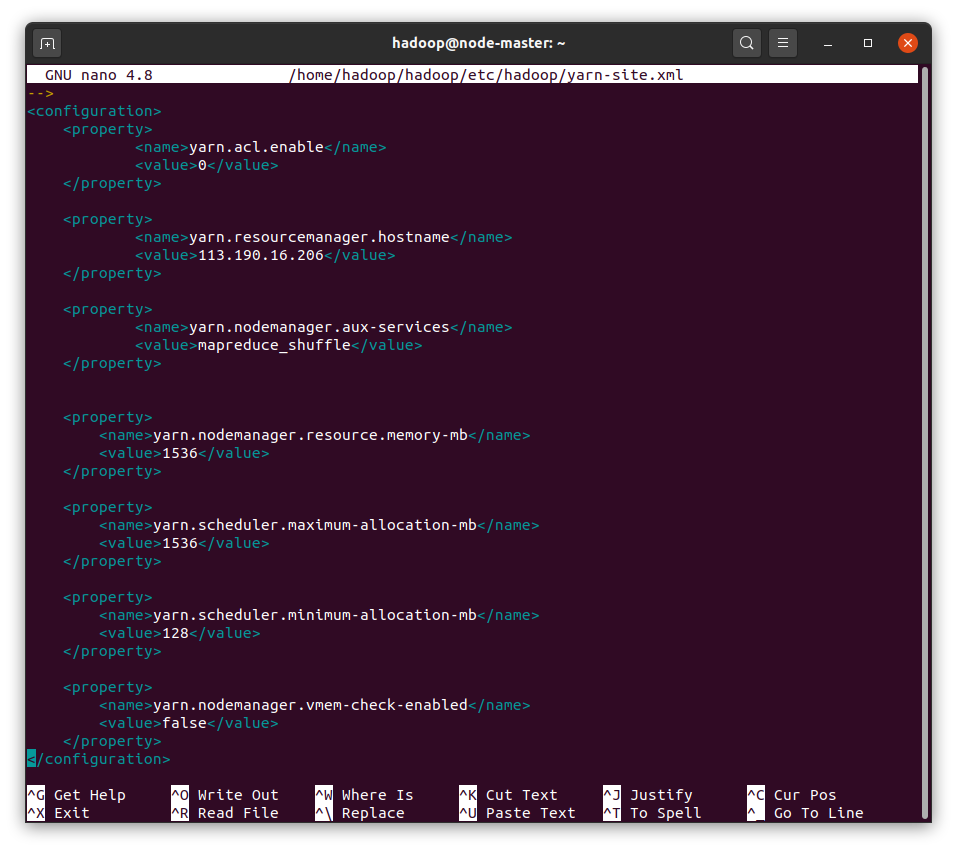


Figure : Configure YARN

* 1. Configure Memory Allocation

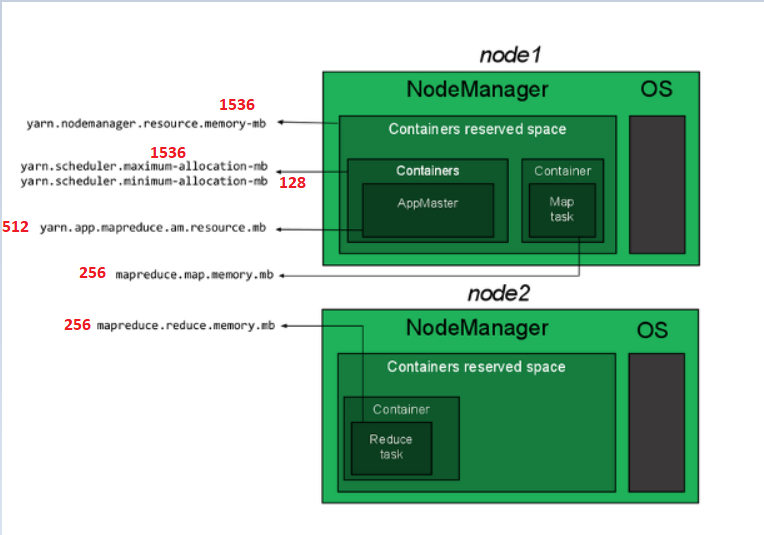


Figure : Sample config for 2GB Nodes (will change in next lab)

* 1. Duplicate Config Files On Each Node
* First, install ssh
* Generate an SSH key
* Make a new file master.pub in the /home/hadoop/.ssh. Paste your public key into
* Copy key:

cat ~/.ssh/master.pub >> ~/.ssh/authorized\_keys

* Note:
* We use this command for ssh node1, node2 without rq passwd

Ssh-copy-id –I ~/ssh/id\_rsa.pub hadoop@note-master

Ssh-copy-id –I ~/ssh/id\_rsa.pub hadoop@note1

Ssh-copy-id –I ~/ssh/id\_rsa.pub hadoop@note2

* Importance: Config username of node-master, node1, node2 is *hadoop*

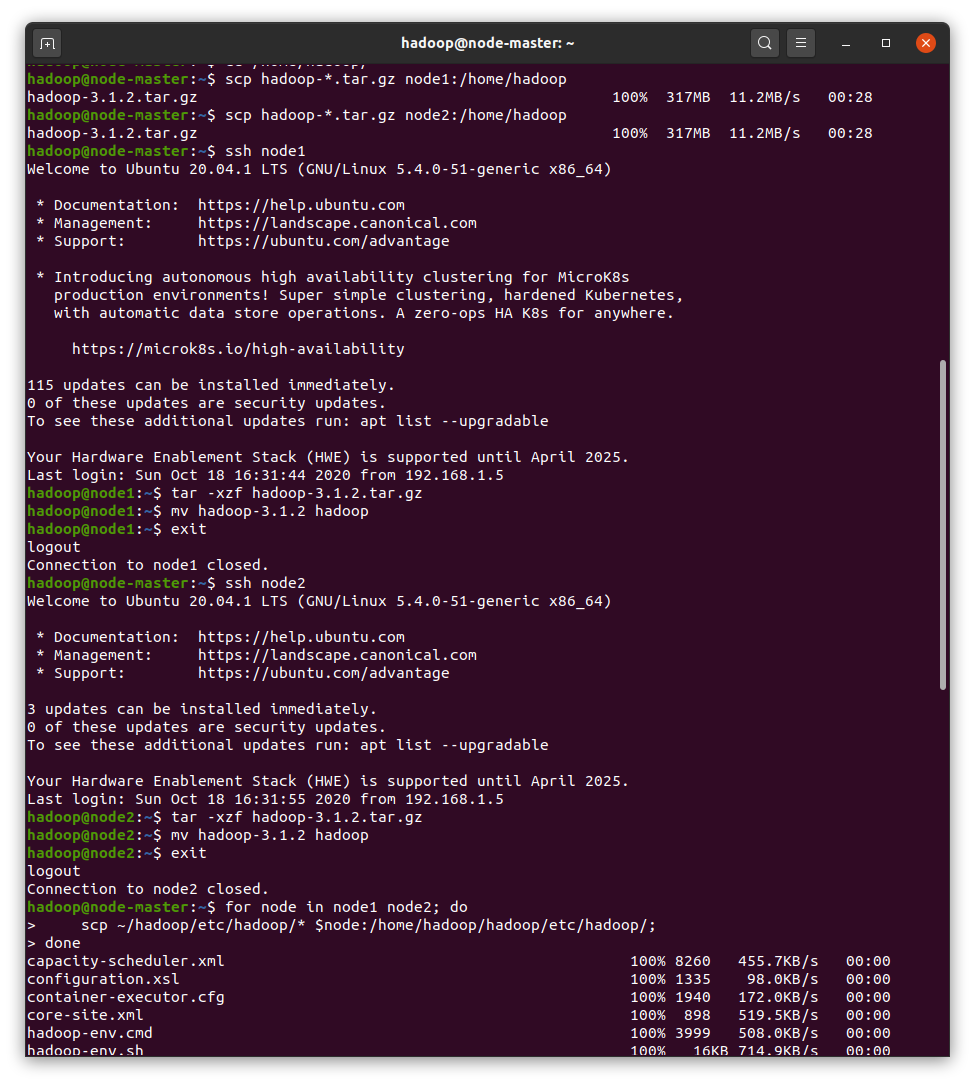


Figure : Copy files to node1, node2

* 1. Start HDFS

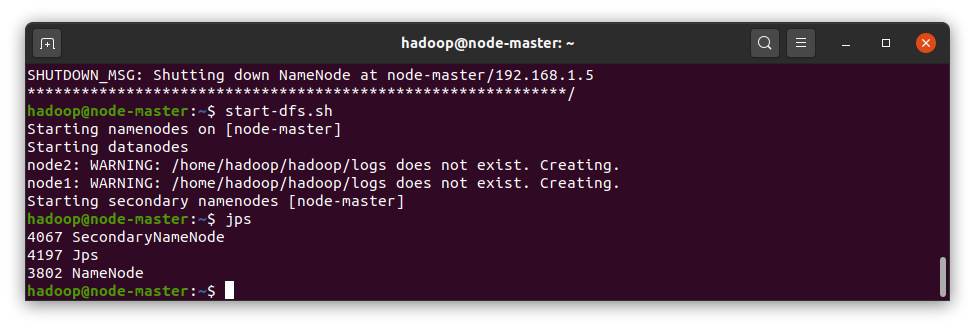


Figure : Start the HDFS by running the following script from node-master

* Or can: $ start-all.sh

# Monitor HDFS Cluter

* 1. Put Data to HDFS
* hdfs dfs -mkdir books
* use wget to download file, see here ☺: <https://speed.hetzner.de/> for what ever file size you want
* hdfs dfs -put 500MB.bin books
* hdfs dfs -put 500MB.bin books
  1. View web user interface
* <http://192.168.1.5:9870> (192.168.1.5 is my master node id)

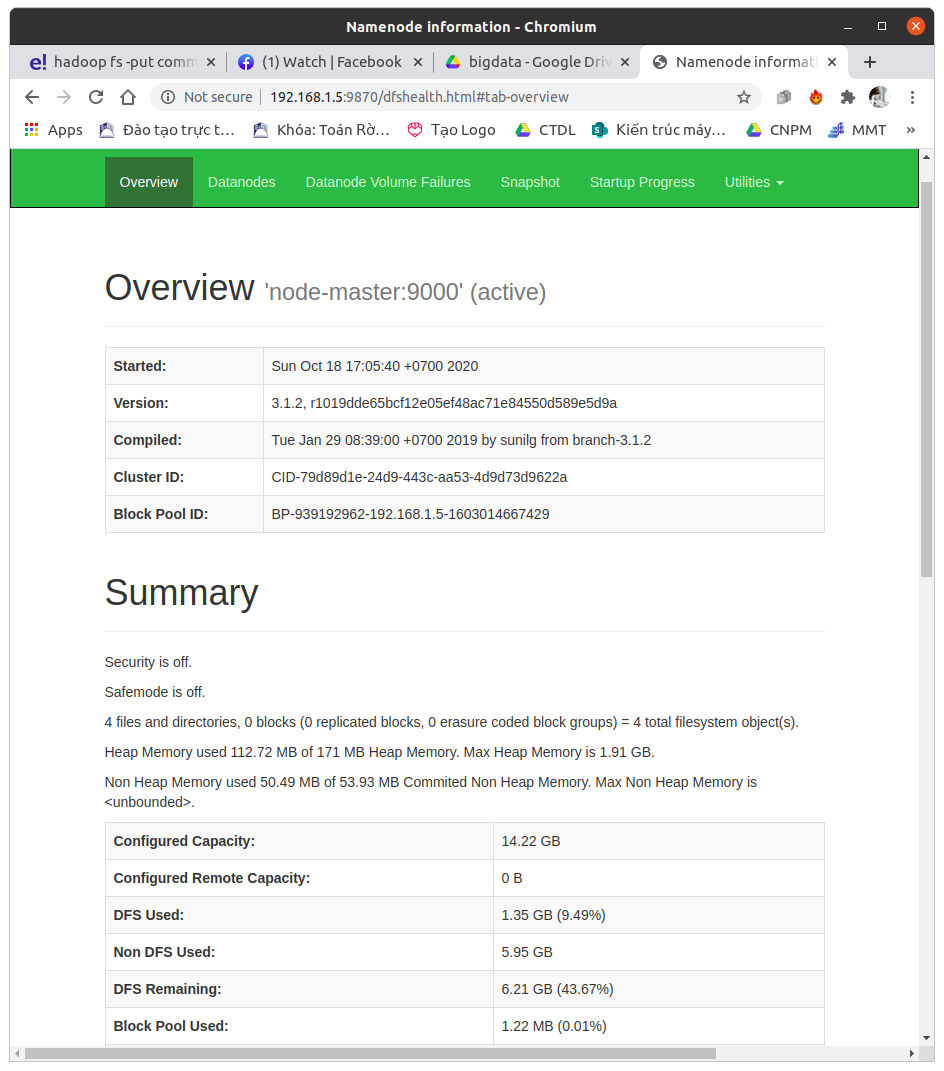


Figure : Web user interface

どうもありがとう、先生