

Nhập Môn Dữ Liệu Lớn

LAP01: Introduction to Hadoop Ecosystem

I. Thông tin sinh viên

- Họ và tên: Phan Bá Đức
- MSSV: 22120071
- Lớp: CQ2022/21

II. Các bước thực hiện cài đặt Hadoop

- **Lệnh:** sudo apt update
 - Cập nhật các gói nâng cấp phần mềm có sẵn trong hệ thống

```
[~] base ➤ ducphan ➤ sudo apt update

Hit:1 http://vn.archive.ubuntu.com/ubuntu noble InRelease
Hit:2 http://vn.archive.ubuntu.com/ubuntu noble-updates InRelease
Hit:3 http://vn.archive.ubuntu.com/ubuntu noble-backports InRelease
Hit:4 https://dl.google.com/linux/chrome/deb stable InRelease
Hit:5 http://security.ubuntu.com/ubuntu noble-security InRelease
Hit:6 https://packages.microsoft.com/repos/vscode stable InRelease
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
1 package can be upgraded. Run 'apt list --upgradable' to see it.
N: Skipping acquire of configured file 'main/binary-i386/Packages' as repository
 'https://packages.microsoft.com/repos/vscode stable InRelease' doesn't support
 architecture 'i386'
```

- **Lệnh:** sudo apt install openjdk-11-jdk
 - Cài đặt OpenJDK 11 (Java Development Kit), một bộ công cụ phát triển Java cần thiết để chạy các ứng dụng Java, bao gồm Hadoop.

```
[~] base ➤ ducphan ➤ sudo apt install openjdk-11-jdk
```

- **Lệnh:** java -version
 - Kiểm tra phiên bản Java hiện tại được cài đặt trên hệ thống. Lệnh này hiển thị thông tin về phiên bản Java và môi trường runtime.

```
[~] base ➤ ducphan ➤ java -version

openjdk version "11.0.26" 2025-01-21
OpenJDK Runtime Environment (build 11.0.26+4-post-Ubuntu-1ubuntu124.04)
OpenJDK 64-Bit Server VM (build 11.0.26+4-post-Ubuntu-1ubuntu124.04, mixed mode,
sharing)
```

Tải và cài đặt Hadoop phiên bản 3.4.1

- **Lệnh:** wget
<https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz>
 - Tải xuống tệp nén Hadoop 3.4.1 từ trang chủ Apache Hadoop bằng công cụ wget. Tệp này là bản phân phối Hadoop dưới dạng .tar.gz, chứa toàn bộ mã nguồn và các tệp cần thiết để cài đặt Hadoop.

```
[~] base > ducphan > wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
```

- **Lệnh:** tar -xzvf hadoop-3.4.1.tar.gz
 - Giải nén tệp hadoop-3.4.1.tar.gz bằng lệnh tar. Các tùy chọn được sử dụng:
 - -x: Giải nén tệp.
 - -z: Giải nén tệp nén bằng gzip (.gz).
 - -v: Hiển thị chi tiết các tệp đang được giải nén (verbose).
 - -f: Chỉ định tên tệp cần giải nén.

```
[~] base > ducphan > tar -xzvf hadoop-3.4.1.tar.gz
```

- **Lệnh:** sudo hadoop-3.4.1 /usr/local/had
 - Di chuyển thư mục Hadoop vào thư mục /usr/local/

```
[~] base > ducphan > sudo mv hadoop-3.4.1 /usr/local/hadoop  
[sudo] password for ducphan: [REDACTED]
```

Cấu hình hadoop

a. core-site.xml:

- Lệnh: nano \$HADOOP_HOME/etc/hadoop/core-site.xml
 - Mở tệp /etc/hadoop/core-site.xml và thêm thông tin về hệ thống tệp HDFS.

```
nano $HADOOP_HOME/etc/hadoop/core-site.xml
GNU nano 7.2
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

b. hdfs-site.xml:

- Lệnh: nano \$HADOOP_HOME/etc/hadoop/hdfs-site.xml
 - Mở tệp /etc/hadoop/hdfs-site.xml và thêm thông tin cấu hình cho HDFS

```
nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml *
GNU nano 7.2
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.name.dir</name>
<value>/usr/local/hadoop/hdfs/name</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/usr/local/hadoop/hdfs/data</value>
</property>
</configuration>
```

c. mapred-site.xml

- Lệnh: nano \$HADOOP_HOME/etc/hadoop/mapred-site.xml
 - Mở tệp /etc/hadoop/mapred-site.xml và thêm cấu hình cho MapReduce

The screenshot shows a terminal window with the title "nano \$HADOOP_HOME/etc/hadoop/mapred-site.xml". The file content is as follows:

```
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.

-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

The bottom of the screen shows the nano editor's command bar with various keyboard shortcuts.

d. yarn-site.xml

- Lệnh: nano \$HADOOP_HOME/etc/hadoop/yarn-site.xml
 - Mở tệp /etc/hadoop/yarn-site.xml và thêm cấu hình cho YARN.

```
GNU nano 7.2      /usr/local/hadoop/etc/hadoop/yarn-site.xml *
```

```
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
</configuration>
```

Key Bindings:

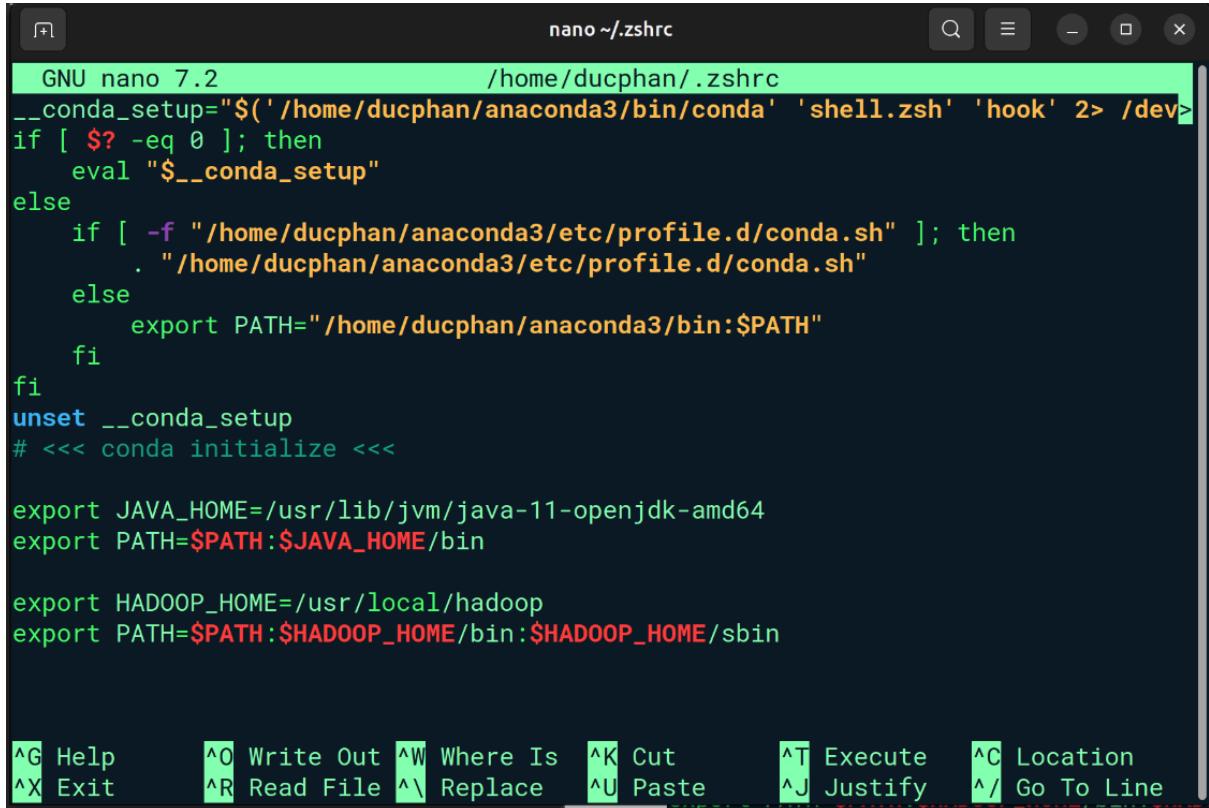
- ^G Help
- ^O Write Out
- ^W Where Is
- ^K Cut
- ^T Execute
- ^C Location
- ^X Exit
- ^R Read File
- ^\\ Replace
- ^U Paste
- ^J Justify
- ^/ Go To Line

Thao tác trên .zshrc

Lý do sử dụng file .zshrc thay vì .bashrc là vì em đang dùng framework Oh My Zsh giúp tùy chỉnh giao diện terminal. Tức là hệ thống đang sử dụng zsh làm cell mặc định

- **Lệnh:** nano ~/.zshrc
 - Mở tệp .zshrc trong trình soạn thảo văn bản nano. Tệp .zshrc là tệp cấu hình cho Zsh shell, chứa các thiết lập và biến môi trường được áp dụng mỗi khi mở terminal sử dụng Zsh. Lệnh này cho phép chỉnh sửa hoặc thêm các cấu hình vào tệp này.
- Thêm các dòng lệnh sau vào file .zshrc
 1. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
 - > Đặt biến môi trường JAVA_HOME để trỏ đến thư mục cài đặt Java (OpenJDK 11). Hadoop cần biết vị trí của Java để hoạt động.
 2. export PATH=\$PATH:\$JAVA_HOME/bin
 - > Thêm thư mục chứa các lệnh thực thi của Java (bin) vào biến môi trường PATH, giúp bạn có thể chạy các lệnh Java từ bất kỳ đâu trong terminal.
 3. export HADOOP_HOME=/usr/local/hadoop
 - > Đặt biến môi trường HADOOP_HOME để trỏ đến thư mục cài đặt Hadoop.

4. `export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin`
 -> Thêm thư mục chứa các lệnh thực thi của Hadoop (bin và sbin) vào biến môi trường PATH, giúp bạn có thể chạy các lệnh Hadoop từ bất kỳ đâu trong terminal.



```

GNU nano 7.2                               /home/ducphan/.zshrc
__conda_setup="$('/home/ducphan/anaconda3/bin/conda' 'shell.zsh' 'hook' 2> /dev>
if [ $? -eq 0 ]; then
    eval "$__conda_setup"
else
    if [ -f "/home/ducphan/anaconda3/etc/profile.d/conda.sh" ]; then
        . "/home/ducphan/anaconda3/etc/profile.d/conda.sh"
    else
        export PATH="/home/ducphan/anaconda3/bin:$PATH"
    fi
fi
unset __conda_setup
# <<< conda initialize <<<

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export PATH=$PATH:$JAVA_HOME/bin

export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

^G Help      ^O Write Out  ^W Where Is  ^K Cut      ^T Execute   ^C Location
^X Exit      ^R Read File  ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line

```

- **Lệnh:** `source ~/.zshrc`
 - Lệnh này thực thi lại tệp .zshrc, cập nhật các biến môi trường và thiết lập mới được thêm vào.



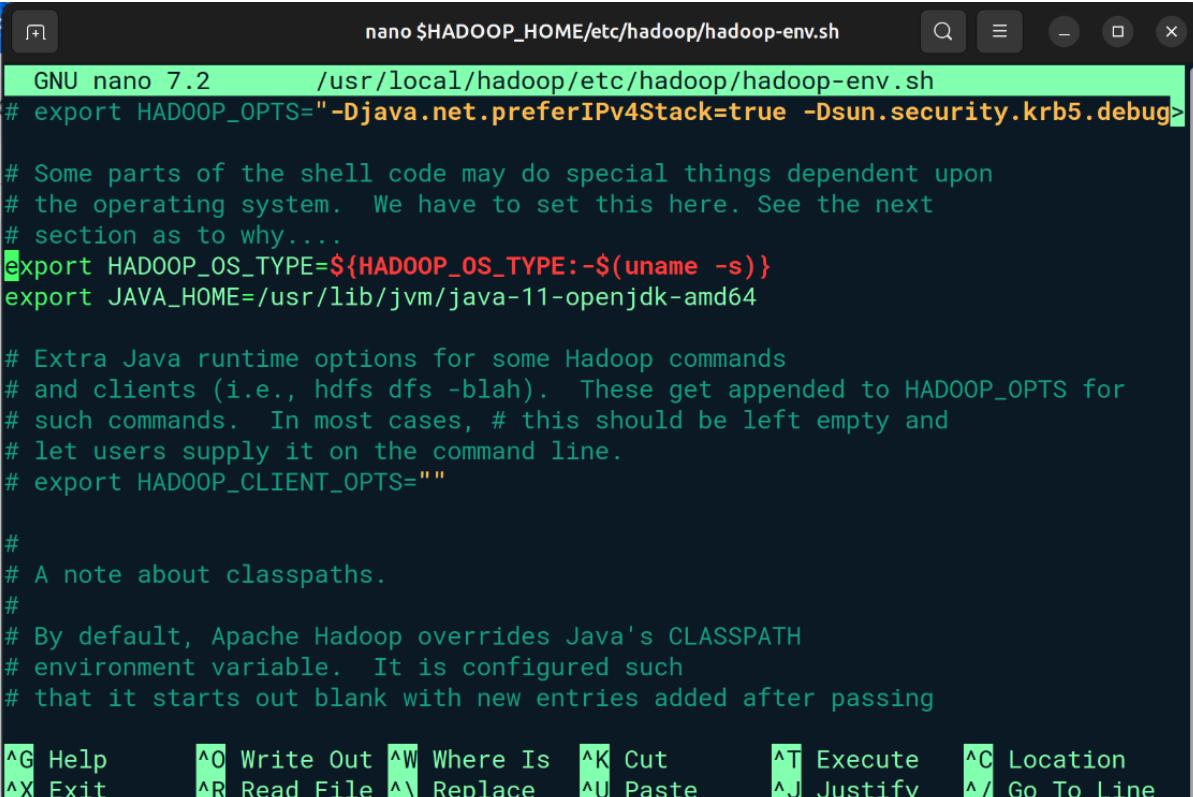
Tiến hành format cho HDFS filesystem

- **Lệnh:** `$HADOOP_HOME/bin/hdfs namenode -format`
 - Mục đích:
 - Định dạng NameNode: Lệnh này khởi tạo các cấu trúc dữ liệu cần thiết cho NameNode, bao gồm metadata của HDFS. Đây là bước bắt buộc trước khi khởi động Hadoop lần đầu tiên.
 - Tạo thư mục logs: Nếu thư mục logs chưa tồn tại, hệ thống sẽ tự động tạo ra để lưu trữ các tệp log.

```
[~] base ➤ ducphan ➤ $SHADOOP_HOME/bin/hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
2025-03-20 13:10:41,782 INFO namenode.NameNode: STARTUP_MSG:
/*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = Luminous/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.4.1
```

Tiến hành set JAVA_HOME vào trong cấu hình Hadoop:

- **Lệnh:** export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
 - Mục đích
 - Xác định vị trí Java: Hadoop yêu cầu biến JAVA_HOME để biết vị trí cài đặt Java, vì Hadoop được viết bằng Java và cần Java để chạy.
 - Đảm bảo tính tương thích: Việc thiết lập JAVA_HOME giúp Hadoop sử dụng đúng phiên bản Java được cài đặt trên hệ thống.



```
GNU nano 7.2      /usr/local/hadoop/etc/hadoop/hadoop-env.sh
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -Dsun.security.krb5.debug>

# Some parts of the shell code may do special things dependent upon
# the operating system. We have to set this here. See the next
# section as to why....
export HADOOP_OS_TYPE=${HADOOP_OS_TYPE:-$(uname -s)}
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64

# Extra Java runtime options for some Hadoop commands
# and clients (i.e., hdfs dfs -blah). These get appended to HADOOP_OPTS for
# such commands. In most cases, # this should be left empty and
# let users supply it on the command line.
# export HADOOP_CLIENT_OPTS=""

#
# A note about classpaths.
#
# By default, Apache Hadoop overrides Java's CLASSPATH
# environment variable. It is configured such
# that it starts out blank with new entries added after passing
```

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
 ^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

Cấu hình SSH

- **Lệnh:** ssh-keygen -t rsa -P " -f ~/.ssh/id_rsa
 - Tạo một cặp khóa SSH (public key và private key) để thiết lập kết nối SSH không cần mật khẩu giữa các máy trong cụm Hadoop. Cụ thể:

- ssh-keygen: Lệnh dùng để tạo cặp khóa SSH.
- -t rsa: Chỉ định loại khóa được tạo là RSA (một thuật toán mã hóa phổ biến).
- -P "": Đặt passphrase (mật khẩu bảo vệ khóa) là rỗng (""), giúp kết nối SSH không yêu cầu nhập passphrase.
- -f ~/.ssh/id_rsa: Chỉ định đường dẫn và tên tệp lưu khóa. Ở đây, khóa private sẽ được lưu tại ~/.ssh/id_rsa và khóa public tại ~/.ssh/id_rsa.pub.
- **Ghi chú:**
 - Trước đó em đã cài SSH server nên không có hình chụp màn hình
 - Trong trường hợp này, tệp khóa đã tồn tại (/home/ducphan/.ssh/id_rsa), và đã chọn ghi đè (y) để tạo lại khóa mới.
- **Lệnh:** cat \$HOME/.ssh/id_rsa.pub >> \$HOME/.ssh/authorized_keys
 - Thêm khóa công khai (public key) từ tệp id_rsa.pub vào tệp authorized_keys
 - Mục đích:
 - Tệp authorized_keys chứa danh sách các khóa công khai được phép kết nối SSH đến máy hiện tại mà không cần nhập mật khẩu.
 - Bằng cách thêm khóa công khai vào authorized_keys, sẽ cho phép máy hiện tại kết nối SSH đến chính nó (hoặc các máy khác nếu sao chép authorized_keys sang máy đó) mà không cần mật khẩu.
- **Lệnh:** chmod 0600 ~/.ssh/authorized_keys
 - Thay đổi quyền truy cập (permissions) của tệp authorized_keys thành 0600. Cụ thể
 - 6 (chủ sở hữu): Đọc và ghi (rw-).
 - 0 (nhóm): Không có quyền (---).
 - 0 (người dùng khác): Không có quyền (---).

Kiểm tra xem kết nối SSH không cần mật khẩu đã được thiết lập thành công hay chưa.

- **Lệnh:** ssh localhost
 - Kết nối SSH đến máy cục bộ (localhost) sử dụng giao thức SSH.
- Thông báo xác thực: khi kết nối lần đầu tiên, hệ thống sẽ hiển thị thông báo xác thực máy chủ:
 - Điều này có nghĩa là máy chủ localhost chưa được lưu trong danh sách các máy chủ đã biết (~/.ssh/known_hosts).
- **Xác nhận kết nối**

- Nhập yes để chấp nhận và thêm máy chủ localhost vào danh sách known_hosts.
- Sau khi xác nhận, kết nối SSH được thiết lập thành công, và thấy thông báo chào mừng từ hệ thống

```
[~] x base ➤ ducphan ➤ ssh localhost

The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:g/JyuqG4PkJ823P6NdPSubq5hJez6if4eB774nKpDMA.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 24.04.2 LTS (GNU/Linux 6.8.0-55-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

8 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
```

Kiểm tra, chạy các dòng lệnh sau

- start-dfs.sh: Khởi động các dịch vụ liên quan đến HDFS (Hadoop Distributed File System).
- start-yarn.sh: Khởi động các dịch vụ liên quan đến YARN (Yet Another Resource Negotiator), framework quản lý tài nguyên và thực thi các ứng dụng trên Hadoop.
- jps: Kiểm tra các tiến trình Java đang chạy trên máy, bao gồm các dịch vụ Hadoop

Ghi chú: Từ phần này trở đi, em vừa cài lại giao diện cho terminal

```

ducphan@Luminous:~          base 13:36:12
start-dfs.sh
start-yarn.sh

Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [Luminous]
Starting resourcemanager
Starting nodemanagers

jps
7698 NodeManager
6806 NameNode
7547 ResourceManager
7259 SecondaryNameNode
8123 Jps

```

Tiến hành xác thực lại HDFS và YARN bằng web UI:

- HDFS Web UI: <http://localhost:9870>

The screenshot shows the HDFS Web UI Overview page for 'localhost:9000' (✓active). The page displays the following information:

Started:	Fri Mar 21 14:25:52 +0700 2025
Version:	3.4.1, r4d7825309348956336b8f06a08322b78422849b1
Compiled:	Wed Oct 09 21:57:00 +0700 2024 by mthakur from branch-3.4.1
Cluster ID:	CID-da5a3267-315a-4a8d-9385-f5ded054247d
Block Pool ID:	BP-402673187-127.0.1.1-1742451042491

Summary

Security is off.
Safemode is off.

4 files and directories, 1 blocks (1 replicated blocks, 0 erasure coded block groups) = 5 total filesystem object(s).

Heap Memory used 181.72 MB of 298 MB Heap Memory. Max Heap Memory is 3.83 GB.

Non Heap Memory used 61.98 MB of 63.75 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	32.56 GB
----------------------	----------

- YARN Web UI: <http://localhost:8088>

The screenshot shows the Hadoop Web UI interface at localhost:8088/cluster. The left sidebar has a tree view with 'Cluster' expanded, showing 'About', 'Nodes', 'Node Labels', 'Applications' (with sub-options: NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), 'Scheduler', and 'Tools'. The main content area displays 'Cluster Metrics' with zero values for all metrics: Apps Submitted (0), Apps Pending (0), Apps Running (0), Apps Completed (0), and Containers Running (0). Below this is the 'Cluster Nodes Metrics' section, which shows 1 Active Node and 0 Decommissioning Nodes. Under 'Scheduler Metrics', it indicates the Capacity Scheduler is active, using [memory-mb (unit=Mi), vcores] as the scheduling resource type, with a minimum allocation of <memory:1024, vCores:1>. A table titled 'Show 20 entries' lists application details like ID, User, Name, Application Type, Application Tags, Queue, Application Priority, Start Time, and Launch Time, but shows 0 entries.

Như vậy ta đã hoàn thành xong việc cài đặt Hadoop.

III. HDFS Operations

1. Tạo thư mục với đường dẫn là /hcmus trên HDFS

- **Lệnh:** hdfs dfs -mkdir /hcmus

A terminal window titled 'ducphan@Luminous:~' shows the command 'hdfs dfs -mkdir /hcmus' being run. The output indicates success with a green checkmark icon and the text 'base 13:39:13'.

2. Tạo user với tên là khtn_<mssv> với mssv là 22120071

- **Lệnh:** sudo adduser khtn_22120071

```
ducphan@Luminous:~
```

```
sudo adduser khtn_22120071
```

```
info: Adding user `khtn_22120071' ...
info: Selecting UID/GID from range 1000 to 59999 ...
info: Adding new group `khtn_22120071' (1001) ...
info: Adding new user `khtn_22120071' (1001) with group `khtn_22120071 (1001)' ...
info: Creating home directory `/home/khtn_22120071' ...
info: Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for khtn_22120071
Enter the new value, or press ENTER for the default
      Full Name []: Phan Ba Duc
      Room Number []:
      Work Phone []:
      Home Phone []:
      Other []:
Is the information correct? [Y/n] y
info: Adding new user `khtn_22120071' to supplemental / extra groups `users' ...
info: Adding user `khtn_22120071' to group `users' ...
```

3. Tạo folder phụ /hcmus/<mssv> với mssv là 22120071 và upload 1 file lên

• Lệnh

- hdfs dfs -mkdir /hcmus/22120071: tạo thư mục con 22120071 trong thư mục hcmus
- echo "Hello Phan Ba Duc" > Test.txt: tạo file Test.txt với nội dung là "Hello Phan Ba Duc"
- hdfs dfs -put Test.txt /hcmus/22120071/: tải tệp Test.txt lên thư mục 22120071

```
hdfs dfs -mkdir /hcmus/22120071
```

```
echo "Hello Phan Ba Duc" > Test.txt
```

```
hdfs dfs -put Test.txt /hcmus/22120071/
```

4. Chmod 744 /hcmus/<mssv> và đặt quyền sở hữu thư mục con cho user khtn_<mssv>

```

[✓] base ✘ 13:58:17 ⓘ
└─ hdfs dfs -chmod 744 /hcmus/22120071
[✓] base ✘ 13:59:17 ⓘ
└─ hdfs dfs -chown khtn_22120071 /hcmus/22120071

```

5. Chạy file jar đính kèm có tên là hadoop-test.jar.

```

[✓] base ✘ 14:00:18 ⓘ
└─ java -jar /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/hadoop-test.jar 9000 /hcmus/22120071
Trying to read /hcmus/22120071
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info
.
Found hdfs://localhost:9000/hcmus/22120071/Test.txt
Ensure the permission is set to 744

```

Lỗi : Lỗi không có quyền truy cập trên file Test.txt (lệnh phía trên chỉ cấp lệnh tuy cập vào thư mục)

Khắc phục: Chạy lệnh sau để cấp quyền cho toàn bộ thư mục và tất cả file bên trong

- hdfs dfs -chmod -R 744 /hcmus/22120071
- hdfs dfs -chown -R khtn_22120071 /hcmus/22120071

Chạy lại file jar (sử dụng user mặc định là ducphan)

```

[✓] base ✘ 14:08:13 ⓘ
└─ java -jar /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/hadoop-test.jar 9000 /hcmus/22120071
Trying to read /hcmus/22120071
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Found hdfs://localhost:9000/hcmus/22120071/Test.txt
Your student ID: 22120071 (ensure it matches your student ID)
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
File written at /home/ducphan/22120071_verification.txt

```

Kết quả: Có kết quả đã được ghi vào 22120071_verification.txt

Tuy nhiên, phương pháp thức nhất để lấy địa chỉ MAC thất bại nhưng đã sử dụng phương pháp thay thế

Thử chạy bằng user khtn_22120071

```

[✓] base ✘ 14:08:24 ⓘ
└─ su - khtn_22120071
Password:
khtn_22120071@Luminous:~$ java -jar /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/hadoop-test.jar 9000 /hcmus/22120071
Error: Unable to access jarfile /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/hadoop-test.jar

```

Lỗi: file hadoop-test.jar đang ở thư mục /home của user ducphan nên khtn_22120071 không có quyền truy cập

Cách xử lý: sao chép tệp jar vào thư mục /home của khtn_22120071 và cấp quyền truy cập file

```
khtn_22120071@Luminous:~$ exit
logout
[sudo] password for ducphan:
```

Chạy lại file hadoop-test.jar

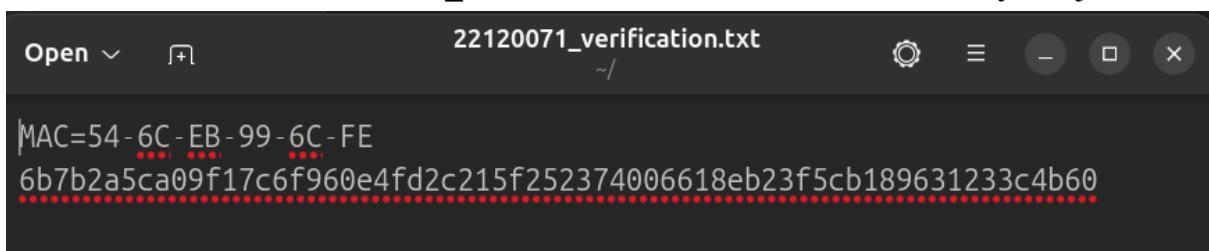
```
khtn_22120071@Luminous:~$ java -jar ~/hadoop-test.jar 9000 /hcmus/22120071
Trying to read /hcmus/22120071
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Found hdfs://localhost:9000/hcmus/22120071/Test.txt
Your student ID: 22120071 (ensure it matches your student ID)
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
File written at /home/khtn_22120071/22120071_verification.txt
khtn_22120071@Luminous:~$ ls
22120071_verification.txt  hadoop-test.jar
khtn_22120071@Luminous:~$ cat 22120071_verification.txt
MAC=54-6C-EB-99-6C-FE
6b7b2a5ca09f17c6f960e4fd2c215f252374006618eb23f5cb189631233c4b60khtn_22120071@Luminous:~$
```

Kết quả tương tự khi thực hiện với user ducphan

Cũng ra file 22120071_verification.txt và phương pháp thứ nhất để lấy địa chỉ MAC bị lỗi, nên đã sử dụng phương pháp thay thế

- **Kiểm tra phương pháp thay thế có thành công không**

- Mở file 22120071_verification.txt được tạo ra sau khi chạy file jar



```
Open  22120071_verification.txt  ~/
MAC=54-6C-EB-99-6C-FE
6b7b2a5ca09f17c6f960e4fd2c215f252374006618eb23f5cb189631233c4b60
```

Ta thấy địa chỉ MAC thu được như trong hình => Kiểm tra xem có phải là địa chỉ MAC thật của máy không

```
ip link

1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT group default
    qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: wlp0s20f3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP mode DORMANT
    group default qlen 1000
    link/ether 54:6c:eb:99:6c:fe brd ff:ff:ff:ff:ff:ff
```

Như vậy địa chỉ MAC trong 22120071_verification.txt là đúng => Thành công

- **Tìm hiểu lý do vì sao “First method” thất bại**

- Nếu Java vẫn không thể lấy MAC vì bị hạn chế quyền cấp thấp (raw socket), bạn có thể cấp quyền cho java bằng setcap và chạy lại file jar

```
sudo setcap cap_net_raw+ep $(readlink -f $(which java))

java -jar /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/hadoop-test.jar 9000 /hcmus/22120071

Trying to read /hcmus/22120071
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Found hdfs://localhost:9000/hcmus/22120071/Test.txt
Your student ID: 22120071 (ensure it matches your student ID)
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
File written at /home/ducphan/22120071_verification.txt
```

“The first method to get MAC address is failed: Could not get network interface” vẫn xảy ra => **Điều này cho thấy vấn đề không phải do quyền, mà có thể do Java không chọn đúng interface để lấy MAC.**

- Xác nhận xem interface wlp0s20f3 có được Java nhận diện không bằng đoạn code sau

```
GNU nano 7.2          CheckMAC.java
import java.net.*;
import java.util.*;

public class CheckMAC {
    public static void main(String[] args) throws Exception {
        Enumeration<NetworkInterface> interfaces = NetworkInterface.getNetworkInterfaces();
        while (interfaces.hasMoreElements()) {
            NetworkInterface ni = interfaces.nextElement();
            System.out.println("Interface: " + ni.getName());

            try {
                byte[] mac = ni.getHardwareAddress();
                if (mac != null) {
                    System.out.print(" MAC: ");
                    for (int i = 0; i < mac.length; i++) {
                        System.out.format("%02X%s", mac[i], (i < mac.length - 1) ? "-" : "\n");
                    }
                    System.out.println();
                } else {
                    System.out.println(" MAC: null");
                }
            }
        }
    }
}
```

[Read 27 lines]

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

```
nano CheckMAC.java
javac CheckMAC.java
java CheckMAC

Interface: wlp0s20f3
MAC: 54-6C-EB-99-6C-FE
Interface: lo
MAC: null
```

=> Java đã truy cập được interface wlp0s20f3 và lấy MAC thành công

→ Điều này có là do chương trình bên trong .jar đang cố lấy MAC từ interface mặc định, có thể là lo hoặc null, thay vì wlp0s20f3.

IV. Warm up with word count:

1. Tạo file word_count.py, triển hành code và cấp quyền thực thi

```
~/3/B/N/NMDLL_Lab_1/22120071/src/WordCount
nano word_count.py
chmod +x word_count.py
```

```
GNU nano 7.2                               word_count.py *
#!/usr/bin/env python3
from mrjob.job import MRJob
import re

class LetterCount(MRJob):
    valid_letters = set("afjghcmus")

    def mapper(self, _, line):
        pattern = r'[a-zA-Z]+'
        words = re.findall(pattern, line)

        for word in words:
            if word:
                first_char = word[0].lower()
                if first_char in self.valid_letters:
                    yield first_char, 1

    def combiner(self, letter, counts):
        """Combiner function: tổng hợp cục bộ để giảm lượng dữ liệu truyền qua mạng"""
        yield letter, sum(counts)

    def reducer(self, letter, counts):
        yield letter, sum(counts)

if __name__ == '__main__':
    LetterCount.run()
```

2. Đưa file words.txt lên hdfs

```
26s ✘ base ✨ 08:55:35
hdbs dfs -put /home/ducphan/3rd/Big_data/NMDLL_Lab_1/NMDLL_Lab_1/words.txt /hcmus/words.txt
```

3. Tiến hành chạy MapReduce Job:

```

ducphan@Luminous:~/3rd/Big_data/NMDLL_Lab_1/22120071/src/WordCount
python word_count.py -r hadoop hdfs:///hcmus/words.txt -o hdfs:///hcmus/wordcount_output
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/local/hadoop/bin...
Found hadoop binary: /usr/local/hadoop/bin/hadoop
Using Hadoop version 3.4.1
Looking for Hadoop streaming jar in /usr/local/hadoop...
Found Hadoop streaming jar: /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar
Creating temp directory /tmp/word_count.ducphan.20250328.140531.635312
uploading working dir files to hdfs:///user/ducphan/tmp/mrjob/word_count.ducphan.20250328.140531.635312/files/...
Copying other local files to hdfs:///user/ducphan/tmp/mrjob/word_count.ducphan.20250328.140531.635312/files/
Running step 1 of 1...
  packageJobJar: [/tmp/hadoop-unjar6449287807773093982/] [] /tmp/streamjob2233238202776947108.jar
  tmpDir=null
  Connecting to ResourceManager at /0.0.0.0:8032
  Connecting to ResourceManager at /0.0.0.0:8032
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ducphan/.staging/job_1743169893056_0002
Total input files to process : 1
number of splits:2
Submitting tokens for job: job_1743169893056_0002
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1743169893056_0002
The url to track the job: http://Luminous:8088/proxy/application_1743169893056_0002/
Running job: job_1743169893056_0002

```

Kiểm tra file kết quả tạo ra và nội dung bên trong

```

~/3/B/N/NMDLL_Lab_1/22120071/src/WordCount
hdfs dfs -ls /hcmus
Found 3 items
drwxr--r--  - khtn_22120071 supergroup      0 2025-03-24 13:56 /hcmus/22120071
drwxr-xr-x  - ducphan       supergroup      0 2025-03-28 21:06 /hcmus/wordcount_output
-rw-r--r--  1 ducphan       supergroup 4862985 2025-03-24 15:16 /hcmus/words.txt

~/3/B/N/NMDLL_Lab_1/22120071/src/WordCount
hdfs dfs -cat /hcmus/wordcount_output/part-*
``a''    32921
``c''    42817
``f''    18793
``g''    16002
``h''    20911
``j''    4530
``m''    27239
``s''    59567
``u''    24301

~/3/B/N/NMDLL_Lab_1/22120071/src/WordCount

```

Xuất kết quả ra file result.txt

```

~/3/B/N/NMDLL_Lab_1/22120071/src/WordCount
hdfs dfs -getmerge /hcmus/wordcount_output result.txt

```