

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN
□□□□

**ỨNG DỤNG HADOOP THỰC HIỆN XÂY DỰNG HỆ
THỐNG PHÂN TÍCH CẠNH TRANH THỊ TRƯỜNG
DỰA TRÊN DỮ LIỆU GIÁ VÀ KHUYẾN MÃI SẢN
PHẨM LAPTOP**

SVTH: Phan Trọng Phú - 23133056

GVHD: Trần Quang Khải

HK1, NĂM 2025

1. PHẦN MỞ ĐẦU

1. Cài đặt các file cấu hình

1. Cài hadoop: wget <https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz>
2. Cài jdk8: sudo apt install openjdk-8-jdk -y
3. Cài ssh : sudo apt install openssh-server -y
4. Cài đặt mysql : sudo apt install mysql-server -y
5. Cài đặt JDBC: wget <https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz>

2. Minh chứng các phần cài đặt :

1. Hadoop

1.1 Lệnh cài đặt

```
phu@DESKTOP-04NCH03:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
--2025-10-13 15:52:56-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 730107476 (696M) [application/x-gzip]
Saving to: 'hadoop-3.3.6.tar.gz.1'

hadoop-3.3.6.tar.gz.1 75%[=====>] 527.85M 5.86MB/s eta 26s
```

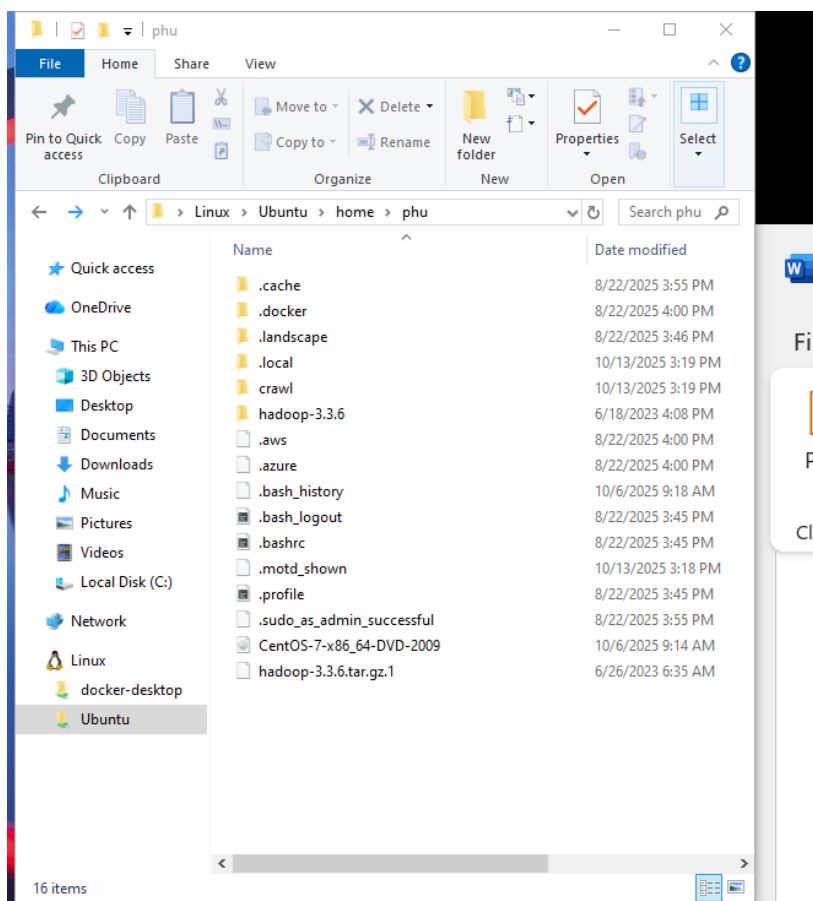
1.2 Minh chứng hoàn thành

```
phu@DESKTOP-04NCH03:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
--2025-10-13 15:52:56-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 730107476 (696M) [application/x-gzip]
Saving to: 'hadoop-3.3.6.tar.gz.1'

hadoop-3.3.6.tar.gz.1 100%[=====>] 696.28M 8.40MB/s in 1m 57s
2025-10-13 15:55:13 (5.93 MB/s) - 'hadoop-3.3.6.tar.gz.1' saved [730107476/730107476]

phu@DESKTOP-04NCH03:~$
```

Khi giải nén xong ta sẽ thấy được folder của hadoop



1.3 Kiểm tra version của hadoop

```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ hadoop version  
Hadoop 3.3.6  
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c  
Compiled by ubuntu on 2023-06-18T08:22Z  
Compiled on platform linux-x86_64  
Compiled with protoc 3.7.1  
From source with checksum 5652179ad55f76cb287d9c633bb53bbd  
This command was run using /home/phu/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar  
phu@DESKTOP-O4NCH03:~$
```

2. Jdk 8

2.1 Lệnh cài đặt

```
phu@DESKTOP-04NCH03:~$ sudo apt install openjdk-8-jdk -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openjdk-8-jdk is already the newest version (8u462-ga~us1-0ubuntu2~24.04.2).
The following packages were automatically installed and are no longer required:
  libdrm-nouveau2 libdrm-radeon1 libgl1-amber-dri libglapi-mesa libllvm17t64
  libxcb-dri2-0
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 28 not upgraded.
phu@DESKTOP-04NCH03:~$
```

2.2 Minh chứng hoàn thành

```
phu@DESKTOP-04NCH03:~$ sudo apt install openjdk-8-jdk -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openjdk-8-jdk is already the newest version (8u462-ga~us1-0ubuntu2~24.04.2).
The following packages were automatically installed and are no longer required:
  libdrm-nouveau2 libdrm-radeon1 libgl1-amber-dri libglapi-mesa libllvm17t64
  libxcb-dri2-0
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 28 not upgraded.
phu@DESKTOP-04NCH03:~$
```

2.3 Kiểm tra version của java

```
phu@DESKTOP-04NCH03:~$ java -version
openjdk version "1.8.0_462"
OpenJDK Runtime Environment (build 1.8.0_462-8u462-ga~us1-0ubuntu2~24.04.2-b08)
OpenJDK 64-Bit Server VM (build 25.462-b08, mixed mode)
phu@DESKTOP-04NCH03:~$
```

3. SSH

3.1 Lệnh cài ssh

```
phu@DESKTOP-04NCH03:~$ sudo apt install openssh-server -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libdrm-nouveau2 libdrm-radeon1 libgl1-amber-dri libglapi-mesa libllvm17t64
  libxcb-dri2-0
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  libwrap0 ncurses-term openssh-client openssh-sftp-server ssh-import-id
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard ufw
The following NEW packages will be installed:
  libwrap0 ncurses-term openssh-server openssh-sftp-server ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 5 newly installed, 0 to remove and 28 not upgraded.
Need to get 1786 kB of archives.
After this operation, 6853 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu noble-updates/main amd64 openssh-client amd64 1:
9.6p1-3ubuntu13.14 [906 kB]
17% [1 openssh-client 382 kB/906 kB 42%]
```

3.2 Minh chứng hoàn thành

```
phu@DESKTOP-04NCH03: ~
Fetched 1786 kB in 5s (369 kB/s)
Preconfiguring packages ...
(Reading database ... 42909 files and directories currently installed.)
Preparing to unpack .../0-openssh-client_1%3a9.6p1-3ubuntu13.14_amd64.deb ...
Unpacking openssh-client (1:9.6p1-3ubuntu13.14) over (1:9.6p1-3ubuntu13.13) ...
Selecting previously unselected package openssh-sftp-server.
Preparing to unpack .../1-openssh-sftp-server_1%3a9.6p1-3ubuntu13.14_amd64.deb ...
Unpacking openssh-sftp-server (1:9.6p1-3ubuntu13.14) ...
Selecting previously unselected package libwrap0:amd64.
Preparing to unpack .../2-libwrap0_7.6.q-33_amd64.deb ...
Unpacking libwrap0:amd64 (7.6.q-33) ...
Selecting previously unselected package openssh-server.
Preparing to unpack .../3-openssh-server_1%3a9.6p1-3ubuntu13.14_amd64.deb ...
Unpacking openssh-server (1:9.6p1-3ubuntu13.14) ...
Selecting previously unselected package ncurses-term.
Preparing to unpack .../4-ncurses-term_6.4+20240113-1ubuntu2_all.deb ...
Unpacking ncurses-term (6.4+20240113-1ubuntu2) ...
Selecting previously unselected package ssh-import-id.
Preparing to unpack .../5-ssh-import-id_5.11-0ubuntu2.24.04.1_all.deb ...
Unpacking ssh-import-id (5.11-0ubuntu2.24.04.1) ...
Setting up openssh-client (1:9.6p1-3ubuntu13.14) ...
Setting up ssh-import-id (5.11-0ubuntu2.24.04.1) ...
Setting up libwrap0:amd64 (7.6.q-33) ...
Setting up ncurses-term (6.4+20240113-1ubuntu2) ...
Setting up openssh-sftp-server (1:9.6p1-3ubuntu13.14) ...
Setting up openssh-server (1:9.6p1-3ubuntu13.14) ...

Creating config file /etc/ssh/sshd_config with new version
Creating SSH2 RSA key; this may take some time ...
3072 SHA256:H9RnIbTV+D/4hAruvT8excQDD6J3SaSThYV0IKgIok root@DESKTOP-04NCH03 (RSA)
Creating SSH2 ECDSA key; this may take some time ...
256 SHA256:h1w1YavBs7MlyprdB8RbpdYZqjllsIj17KQhM/5u4NB4 root@DESKTOP-04NCH03 (ECDSA)
Creating SSH2 ED25519 key; this may take some time ...
256 SHA256:qGWST+3P+ZALJ5hg5wOb8QW8rf+4vCgcPLI+Sn3K0zA root@DESKTOP-04NCH03 (ED25519)
Created symlink /etc/systemd/system/sockets.target.wants/ssh.socket → /usr/lib/systemd/system/ssh.socket.
Created symlink /etc/systemd/system/ssh.service.requires/ssh.socket → /usr/lib/systemd/system/ssh.socket.
Processing triggers for man-db (2.12.0-4build2) ...
Processing triggers for libc-bin (2.39-0ubuntu8.6) ...
phu@DESKTOP-04NCH03:~$
```

3.3 Kiểm tra version của ssh

```
phu@DESKTOP-04NCH03:~$ ssh -V
OpenSSH_9.6p1 Ubuntu-3ubuntu13.14, OpenSSL 3.0.13 30 Jan 2024
phu@DESKTOP-04NCH03:~$
```

4. Mysql

4.1 Lệnh cài đặt

```
phu@DESKTOP-04NCH03:~$ sudo apt install mysql-server -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libdrm-nouveau2 libdrm-radeon1 libgl1-amber-dri libglapi-mesa libllvm17t64
  libxcb-dri2-0
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  libaio1t64 libcgi-fast-perl libcgi-pm-perl libclone-perl libencode-locale-perl
  libevent-pthreads-2.1-7t64 libfcgi-bin libfcgi-perl libfcgi0t64 libhtml-parser-perl
  libhtml-tagset-perl libhtml-template-perl libhttp-date-perl libhttp-message-perl
  libio-html-perl liblwp-mediatypes-perl libmecab2 libnuma1 libprotobuf-lite32t64
  libtimedate-perl liburi-perl mecab-ipadic mecab-ipadic-utf8 mecab-utils
  mysql-client-8.0 mysql-client-core-8.0 mysql-common mysql-server-8.0
  mysql-server-core-8.0
Suggested packages:
  libdata-dump-perl libipc-sharedcache-perl libio-compress-brotli-perl
  libbusiness-isbn-perl libregexp-ipv6-perl libwww-perl mailx tinyca
The following NEW packages will be installed:
  libaio1t64 libcgi-fast-perl libcgi-pm-perl libclone-perl libencode-locale-perl
  libevent-pthreads-2.1-7t64 libfcgi-bin libfcgi-perl libfcgi0t64 libhtml-parser-perl
  libhtml-tagset-perl libhtml-template-perl libhttp-date-perl libhttp-message-perl
  libio-html-perl liblwp-mediatypes-perl libmecab2 libnuma1 libprotobuf-lite32t64
  libtimedate-perl liburi-perl mecab-ipadic mecab-ipadic-utf8 mecab-utils
  mysql-client-8.0 mysql-client-core-8.0 mysql-common mysql-server mysql-server-8.0
  mysql-server-core-8.0
0 upgraded, 30 newly installed, 0 to remove and 28 not upgraded.
Need to get 29.6 MB of archives.
After this operation, 243 MB of additional disk space will be used.
```

4.2 Minh chứng hoàn thành

```
phu@DESKTOP-O4NCH03: ~  
reading /usr/share/mecab/dic/ipadic/Adverb.csv ... 3032  
reading /usr/share/mecab/dic/ipadic/Noun.nai.csv ... 42  
reading /usr/share/mecab/dic/ipadic/Noun.demonst.csv ... 120  
reading /usr/share/mecab/dic/ipadic/Noun.proper.csv ... 27328  
reading /usr/share/mecab/dic/ipadic/Noun.place.csv ... 72999  
reading /usr/share/mecab/dic/ipadic/Others.csv ... 2  
reading /usr/share/mecab/dic/ipadic/Symbol.csv ... 208  
reading /usr/share/mecab/dic/ipadic/Filler.csv ... 19  
reading /usr/share/mecab/dic/ipadic/Interjection.csv ... 252  
reading /usr/share/mecab/dic/ipadic/Verb.csv ... 130750  
reading /usr/share/mecab/dic/ipadic/Postp-col.csv ... 91  
reading /usr/share/mecab/dic/ipadic/Adj.csv ... 27210  
reading /usr/share/mecab/dic/ipadic/Noun.name.csv ... 34202  
reading /usr/share/mecab/dic/ipadic/Postp.csv ... 146  
reading /usr/share/mecab/dic/ipadic/Conjunction.csv ... 171  
reading /usr/share/mecab/dic/ipadic/Prefix.csv ... 221  
reading /usr/share/mecab/dic/ipadic/Noun.org.csv ... 16668  
reading /usr/share/mecab/dic/ipadic/Noun.adverbal.csv ... 795  
emitting double-array: 100% |#####|  
reading /usr/share/mecab/dic/ipadic/matrix.def ... 1316x1316  
emitting matrix : 100% |#####|  
  
done!  
update-alternatives: using /var/lib/mecab/dic/ipadic-utf8 to provide /var/lib/mecab/dic/  
/debian (mecab-dictionary) in auto mode  
Setting up libhtml-parser-perl:amd64 (3.81-1build3) ...#####.....]  
Setting up libhttp-message-perl (6.45-1ubuntu1) ...#####.....]  
Setting up mysql-server-8.0 (8.0.43-0ubuntu0.24.04.2) ...#####.....]  
update-alternatives: using /etc/mysql/mysql.cnf to provide /etc/mysql/my.cnf (my.cnf) i  
n auto mode  
Renaming removed key_buffer and myisam-recover options (if present)  
mysqld will log errors to /var/log/mysql/error.log  
mysqld is running as pid 318331  
Created symlink /etc/systemd/system/multi-user.target.wants/mysql.service → /usr/lib/sy  
stemd/system/mysql.service.  
Setting up libcgi-pm-perl (4.63-1) ...#####.....]  
Setting up libhtml-template-perl (2.97-2) ...#####.....]  
Setting up mysql-server (8.0.43-0ubuntu0.24.04.2) ...#####.....]  
Setting up libcgi-fast-perl (1:2.17-1) ...#####.....]  
Processing triggers for man-db (2.12.0-4build2) ...#####.....]  
Processing triggers for libc-bin (2.39-0ubuntu8.6) ...  
phu@DESKTOP-O4NCH03:~$
```

4.3 Kiểm tra version


```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ hadoop version  
Hadoop 3.3.6  
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c  
Compiled by ubuntu on 2023-06-18T08:22Z  
Compiled on platform linux-x86_64  
Compiled with protoc 3.7.1  
From source with checksum 5652179ad55f76cb287d9c633bb53bbd  
This command was run using /home/phu/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar  
phu@DESKTOP-O4NCH03:~$ java -version  
openjdk version "1.8.0_462"  
OpenJDK Runtime Environment (build 1.8.0_462-8u462-ga~us1-0ubuntu2~24.04.2-b08)  
OpenJDK 64-Bit Server VM (build 25.462-b08, mixed mode)  
phu@DESKTOP-O4NCH03:~$ mysql --version  
mysql Ver 8.0.43-0ubuntu0.24.04.2 for Linux on x86_64 ((Ubuntu))  
phu@DESKTOP-O4NCH03:~$
```

5. JDBC

5.1 Lệnh cài đặt

```
phu@DESKTOP-O4NCH03:~$ wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz  
--2025-10-13 16:05:47-- https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz  
Resolving dev.mysql.com (dev.mysql.com)... 184.85.112.229, 2600:1417:4400:8ac::2e31, 2600:1417:4400:89a::2e31  
Connecting to dev.mysql.com (dev.mysql.com)|184.85.112.229|:443... connected.  
HTTP request sent, awaiting response... 302 Moved Temporarily  
Location: https://cdn.mysql.com//archives/mysql-connector-java-8.0/mysql-connector-j-8.0.33.tar.gz [following]  
--2025-10-13 16:05:48-- https://cdn.mysql.com//archives/mysql-connector-java-8.0/mysql-connector-j-8.0.33.tar.gz  
Resolving cdn.mysql.com (cdn.mysql.com)... 23.7.220.59, 2600:1417:4400:8b6::1d68, 2600:1417:4400:8ae::1d68  
Connecting to cdn.mysql.com (cdn.mysql.com)|23.7.220.59|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 4236147 (4.0M) [application/x-tar-gz]  
Saving to: 'mysql-connector-j-8.0.33.tar.gz'  
  
mysql-connector-j-8.0 100%[=====>] 4.04M 3.73MB/s in 1.1s  
  
2025-10-13 16:05:49 (3.73 MB/s) - 'mysql-connector-j-8.0.33.tar.gz' saved [4236147/4236147]  
  
phu@DESKTOP-O4NCH03:~$
```

5.2 Minh chứng hoàn thành


```
phu@DESKTOP-04NCH03:~$ wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz
--2025-10-13 16:05:47-- https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz
Resolving dev.mysql.com (dev.mysql.com)... 184.85.112.229, 2600:1417:4400:8ac::2e31, 2600:1417:4400:89a::2e31
Connecting to dev.mysql.com (dev.mysql.com)|184.85.112.229|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://cdn.mysql.com//archives/mysql-connector-java-8.0/mysql-connector-j-8.0.33.tar.gz [following]
--2025-10-13 16:05:48-- https://cdn.mysql.com//archives/mysql-connector-java-8.0/mysql-connector-j-8.0.33.tar.gz
Resolving cdn.mysql.com (cdn.mysql.com)... 23.7.220.59, 2600:1417:4400:8b6::1d68, 2600:1417:4400:8ae::1d68
Connecting to cdn.mysql.com (cdn.mysql.com)|23.7.220.59|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4236147 (4.0M) [application/x-tar-gz]
Saving to: 'mysql-connector-j-8.0.33.tar.gz'

mysql-connector-j-8.0 100%[=====>] 4.04M 3.73MB/s in 1.1s

2025-10-13 16:05:49 (3.73 MB/s) - 'mysql-connector-j-8.0.33.tar.gz' saved [4236147/4236147]

phu@DESKTOP-04NCH03:~$
```

6. SQOOP

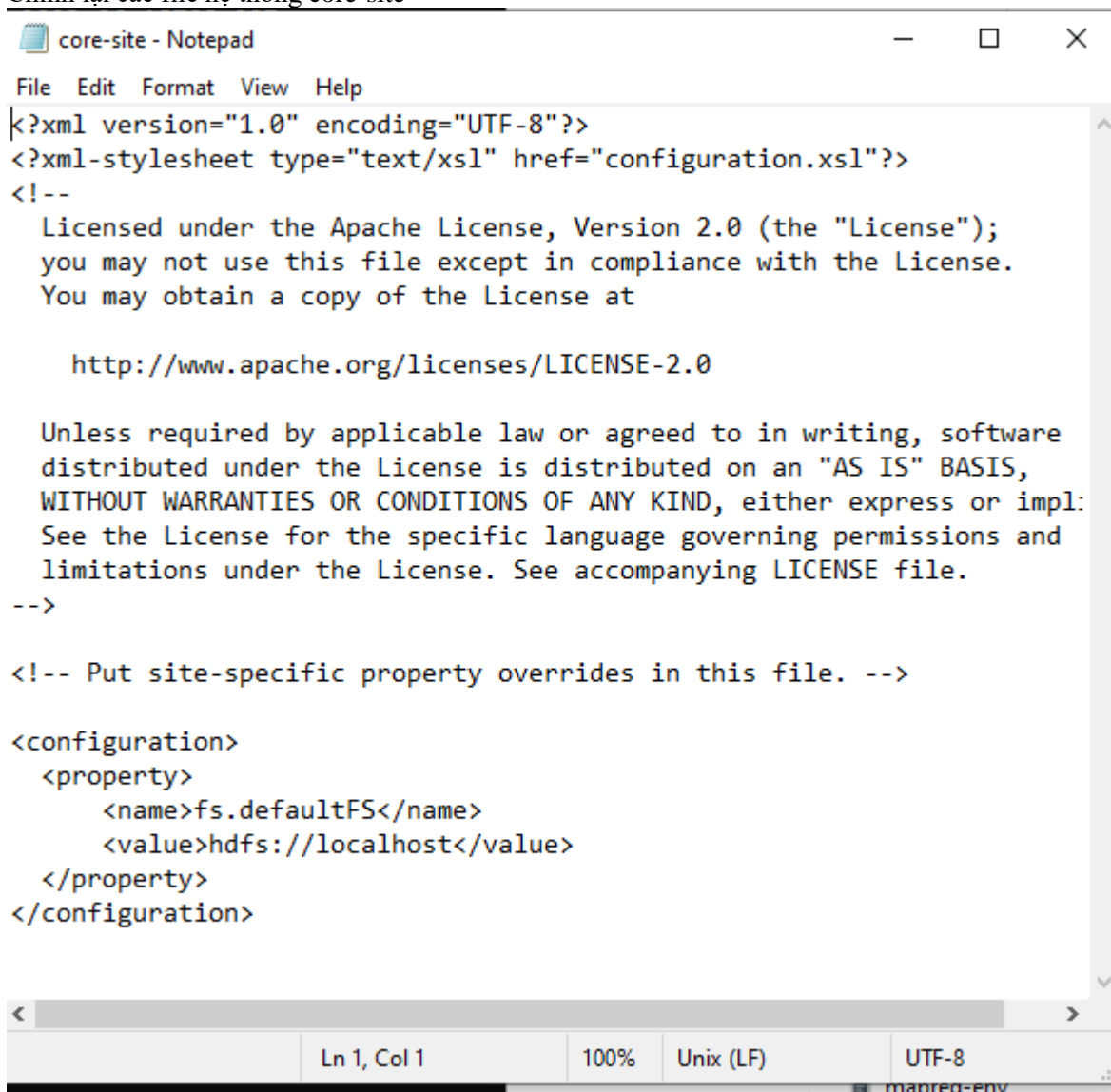
71. Cài đặt và minh chứng thành công

```
phu@DESKTOP-04NCH03:~$ wget https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
--2025-10-13 16:50:57-- https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17953604 (17M) [application/x-gzip]
Saving to: 'sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz'

sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz 7%[>] 1.21M 149KB/s eta 2m 7s
```

7.2 Kiểm tra version

Phân cấu hình hadoop
Chỉnh lại các file hệ thống core-site



```
<?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

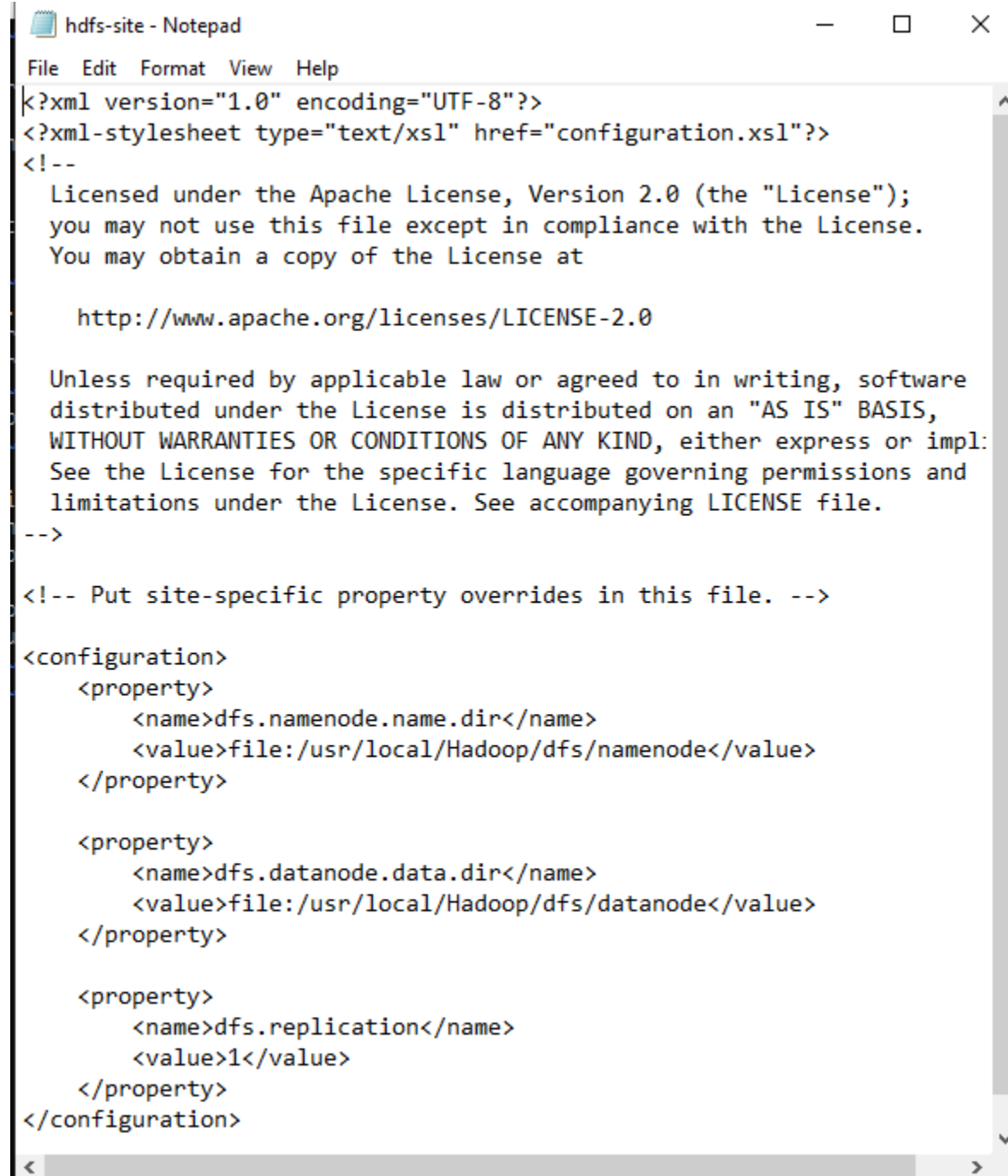
    http://www.apache.org/licenses/LICENSE-2.0

  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 impl:
  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</value>
  </property>
</configuration>
```

File hdfs-site



```
<?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.
-->

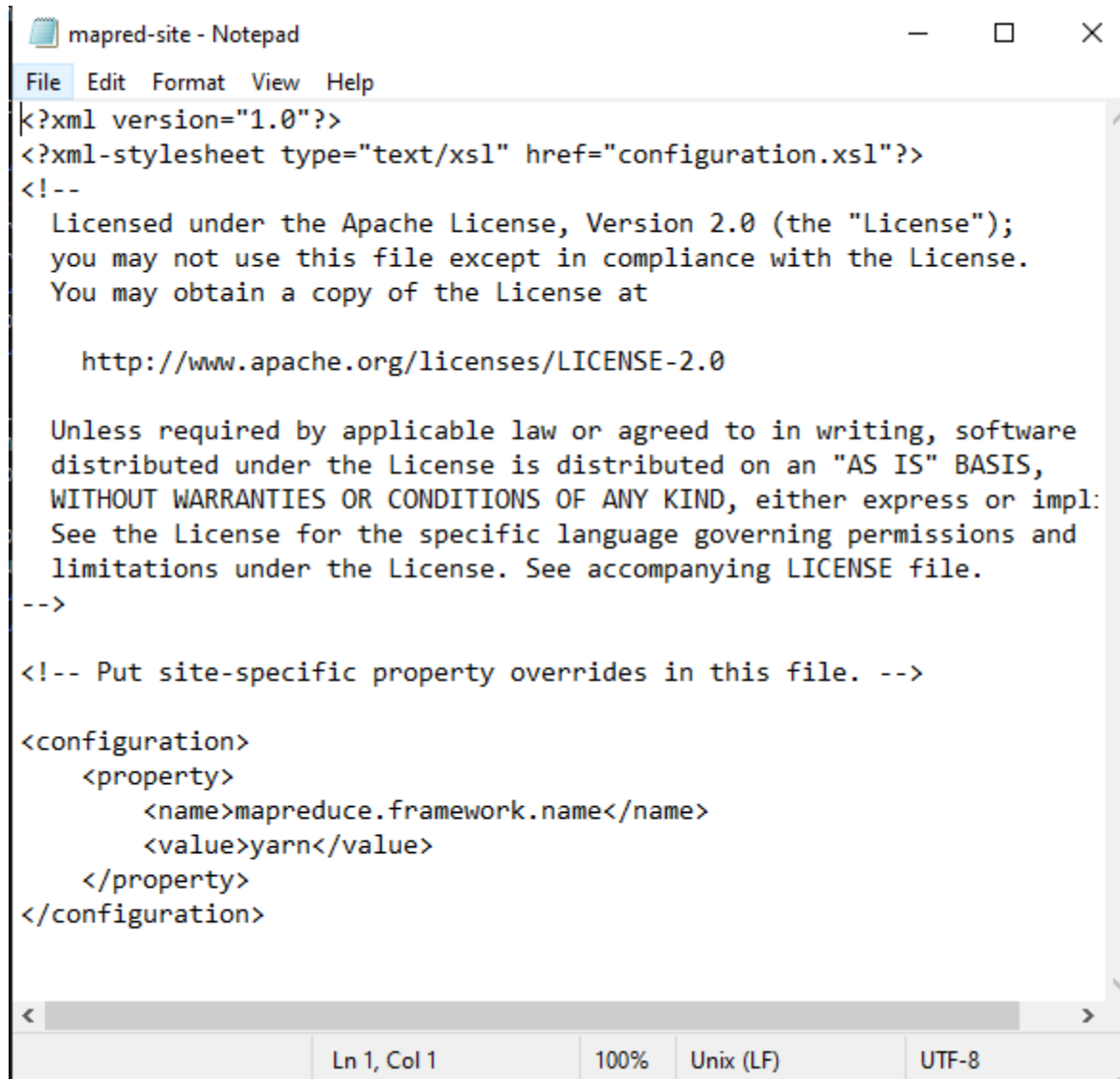
<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/usr/local/Hadoop/dfs/namenode</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/usr/local/Hadoop/dfs/datanode</value>
  </property>

  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

Chỉnh lại file mapred-site



```
<?xml version="1.0"?>
<?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

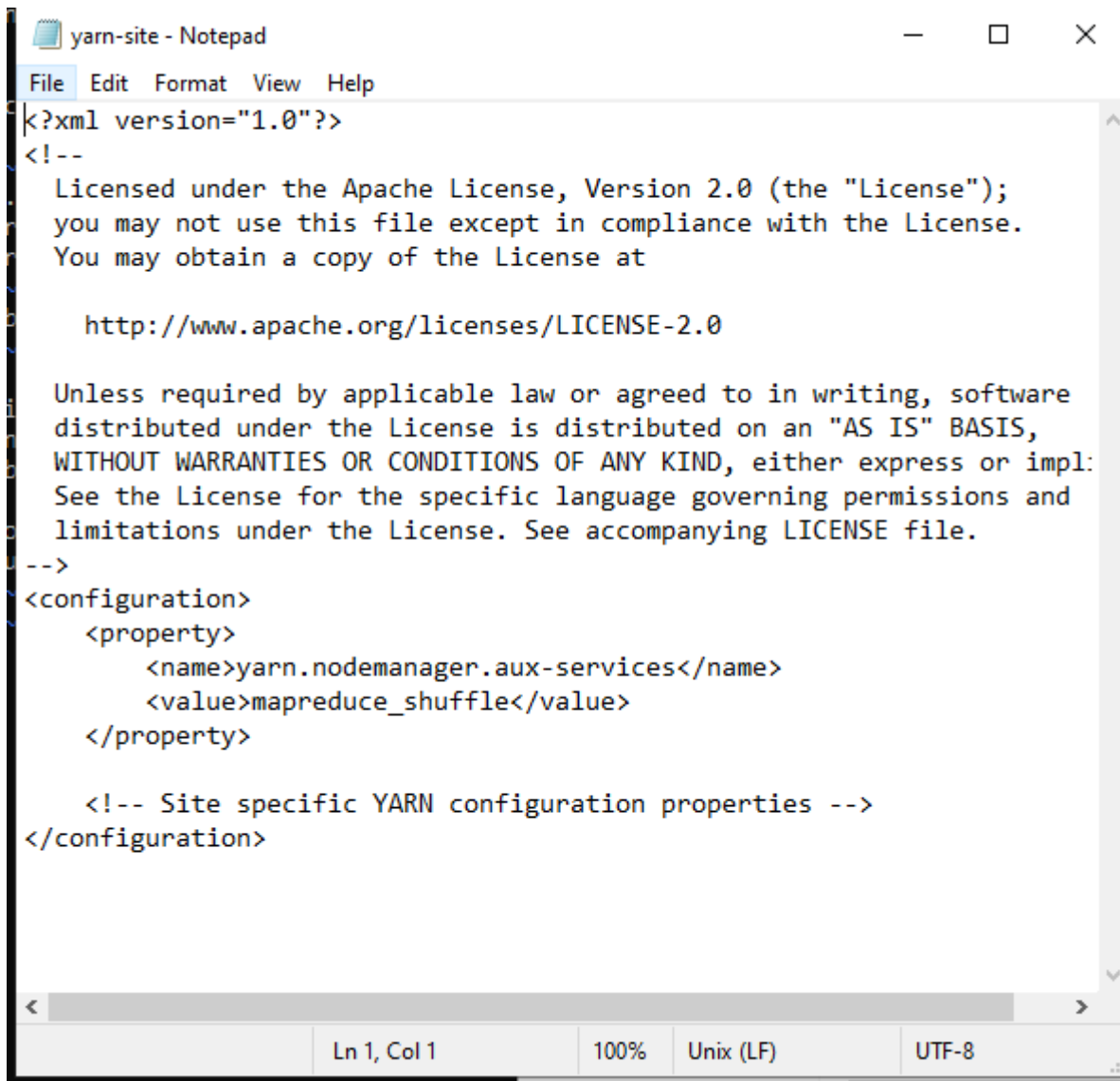
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or impl:
  See the License for the specific language governing permissions and
  limitations under the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

Ln 1, Col 1 100% Unix (LF) UTF-8

Chỉnh lại file yarn-site



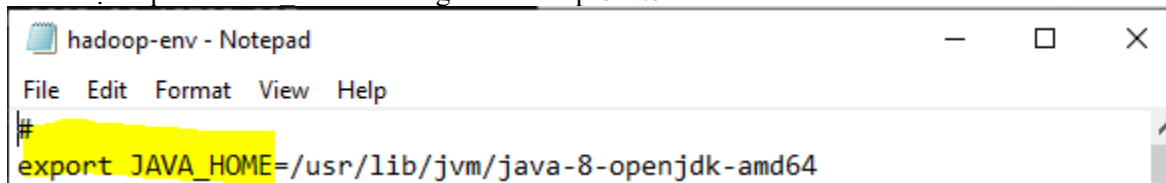
```
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    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 impl:
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

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

Chỉnh lại export JAVA_HOME trong file hadoop-env.sh



```
#
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

Cấu hình ssh

1. Tạo keygen không mật khẩu

```
phu@DESKTOP-04NCH03:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/phu/.ssh'.
Your identification has been saved in /home/phu/.ssh/id_rsa
Your public key has been saved in /home/phu/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:Q4jSYoyT9rUe57P7yyslnCaE5sTAfIGc/9D1B+ECWcY phu@DESKTOP-04NCH03
The key's randomart image is:
+---[RSA 3072]---+
|+ 0...+0  ..|
|@ 0 ooE..|
|.X =.O.O..|
|.O.X.O....|
|+.+O..S .|
|..O+= O|
|.ooo|
|. +|
|O+=O|
+----[SHA256]-----+
phu@DESKTOP-04NCH03:~$
```

2. Thêm khóa công khai vừa tạo

```
+----[SHA256]-----+
phu@DESKTOP-04NCH03:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
phu@DESKTOP-04NCH03:~$
```

3. Đặt quyền truy cập

```
phu@DESKTOP-04NCH03:~$ chmod 0600 ~/.ssh/authorized_keys
phu@DESKTOP-04NCH03:~$
```

4. Kiểm tra kết nối


```
phu@DESKTOP-04NCH03: ~  
phu@DESKTOP-04NCH03:~$ ssh localhost  
The authenticity of host 'localhost (127.0.0.1)' can't be established.  
ED25519 key fingerprint is SHA256:qGWST+3P+ZALJ5hgSwOb8QW8rf+4vCgcPlI+Sn3KOzA.  
This key is not known by any other names.  
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes  
Please type 'yes', 'no' or the fingerprint: yes  
Please type 'yes', 'no' or the fingerprint: yes  
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.  
Welcome to Ubuntu 24.04.3 LTS (GNU/Linux 6.6.87.2-microsoft-standard-WSL2 x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:       https://ubuntu.com/pro  
  
System information as of Mon Oct 13 16:26:00 +07 2025  
  
System load:  0.0          Processes:            36  
Usage of /:   0.5% of 1006.85GB  Users logged in:     2  
Memory usage: 20%          IPv4 address for eth0: 172.30.212.45  
Swap usage:   0%  
  
* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s  
just raised the bar for easy, resilient and secure K8s cluster deployment.  
  
https://ubuntu.com/engage/secure-kubernetes-at-the-edge  
Welcome to Ubuntu 24.04.3 LTS (GNU/Linux 6.6.87.2-microsoft-standard-WSL2 x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:       https://ubuntu.com/pro  
  
System information as of Mon Oct 13 15:42:48 +07 2025  
  
System load:  0.87          Processes:            37  
Usage of /:   0.5% of 1006.85GB  Users logged in:     1  
Memory usage: 13%          IPv4 address for eth0: 172.30.212.45  
Swap usage:   0%  
  
* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s  
just raised the bar for easy, resilient and secure K8s cluster deployment.  
  
https://ubuntu.com/engage/secure-kubernetes-at-the-edge  
phu@DESKTOP-04NCH03:~$
```

Định dạng hdfs format và tiến hành cài đặt

```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ hdfs namenode -format  
WARNING: /home/phu/hadoop/logs does not exist. Creating.  
2025-10-13 16:26:44,587 INFO namenode.NameNode: STARTUP_MSG:  
/*****  
STARTUP_MSG: Starting NameNode  
STARTUP_MSG:   host = DESKTOP-O4NCH03.localdomain/127.0.1.1  
STARTUP_MSG:   args = [-format]  
STARTUP_MSG:   version = 3.3.6  
STARTUP_MSG:   classpath = /home/phu/hadoop/etc/hadoop:/home/phu/hadoop/share/hadoop/co  
mmon/lib/animal-sniffer-annotations-1.17.jar:/home/phu/hadoop/share/hadoop/common/lib/h  
adoop-shaded-guava-1.1.1.jar:/home/phu/hadoop/share/hadoop/common/lib/netty-codec-socks  
-4.1.89.Final.jar:/home/phu/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/home/phu/  
hadoop/share/hadoop/common/lib/netty-transport-rxtx-4.1.89.Final.jar:/home/phu/hadoop/s  
hare/hadoop/common/lib/netty-transport-native-kqueue-4.1.89.Final-osx-x86_64.jar:/home/  
phu/hadoop/share/hadoop/common/lib/commons-compress-1.21.jar:/home/phu/hadoop/share/had  
oop/common/lib/jsr311-api-1.1.1.jar:/home/phu/hadoop/share/hadoop/common/lib/jettison-1  
.5.4.jar:/home/phu/hadoop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/home/p  
hu/hadoop/share/hadoop/common/lib/kerb-common-1.0.1.jar:/home/phu/hadoop/share/hadoop/c  
ommon/lib/kerb-core-1.0.1.jar:/home/phu/hadoop/share/hadoop/common/lib/slf4j-api-1.7.36  
.jar:/home/phu/hadoop/share/hadoop/common/lib/j2objc-annotations-1.1.jar:/home/phu/hado  
op/share/hadoop/common/lib/commons-beanutils-1.9.4.jar:/home/phu/hadoop/share/hadoop/co  
mmon/lib/netty-common-4.1.89.Final.jar:/home/phu/hadoop/share/hadoop/common/lib/commons  
-math3-3.1.1.jar:/home/phu/hadoop/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/home/p  
hu/hadoop/share/hadoop/common/lib/netty-codec-memcache-4.1.89.Final.jar:/home/phu/hadoo  
p/share/hadoop/common/lib/netty-transport-udt-4.1.89.Final.jar:/home/phu/hadoop/share/h  
adoop/common/lib/jul-to-slf4j-1.7.36.jar:/home/phu/hadoop/share/hadoop/common/lib/jaxb-  
api-2.2.11.jar:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar:/home  
phu/hadoop/share/hadoop/common/lib/netty-codec-http2-4.1.89.Final.jar:/home/phu/hadoop  
/share/hadoop/common/lib/httpcore-4.4.13.jar:/home/phu/hadoop/share/hadoop/common/lib/j  
ackson-core-2.12.7.jar:/home/phu/hadoop/share/hadoop/common/lib/jakarta.activation-api  
-1.2.1.jar:/home/phu/hadoop/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/ph  
u/hadoop/share/hadoop/common/lib/netty-resolver-dns-native-macos-4.1.89.Final-osx-x86_64  
.jar:/home/phu/hadoop/share/hadoop/common/lib/jersey-core-1.19.4.jar:/home/phu/hadoop/s  
hare/hadoop/common/lib/kerby-config-1.0.1.jar:/home/phu/hadoop/share/hadoop/common/lib/  
commons-text-1.10.0.jar:/home/phu/hadoop/share/hadoop/common/lib/jetty-server-9.4.51.v2  
0230217.jar:/home/phu/hadoop/share/hadoop/common/lib/nimbus-jose-jwt-9.8.1.jar:/home/p  
hu/hadoop/share/hadoop/common/lib/checker-qual-2.5.2.jar:/home/phu/hadoop/share/hadoop/c  
ommon/lib/netty-codec-4.1.89.Final.jar:/home/phu/hadoop/share/hadoop/common/lib/zookeep  
er-jute-3.6.3.jar:/home/phu/hadoop/share/hadoop/common/lib/hadoop-shaded-protobuf_3_7-1  
.1.1.jar:/home/phu/hadoop/share/hadoop/common/lib/audience-annotations-0.5.0.jar:/home/  
phu/hadoop/share/hadoop/common/lib/jsp-api-2.1.jar:/home/phu/hadoop/share/hadoop/common  
/lib/guava-27.0-jre.jar:/home/phu/hadoop/share/hadoop/common/lib/jersey-json-1.20.jar:/  
home/phu/hadoop/share/hadoop/common/lib/jetty-http-9.4.51.v20230217.jar:/home/phu/hadoo
```

Kiểm tra cấu hình hệ thống

```
phu@DESKTOP-04NCH03:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as phu in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [DESKTOP-04NCH03]
Starting resourcemanager
Starting nodemanagers
phu@DESKTOP-04NCH03:~$ jps
325477 Jps
325127 NodeManager
324817 SecondaryNameNode
324472 NameNode
324632 DataNode
325003 ResourceManager
phu@DESKTOP-04NCH03:~$
```

Phần cấu hình SQOOP
Cài đặt như phần trên
Đặt cấu hình trên bashrc

```
.bashrc - Notepad
File Edit Format View Help
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

#java en
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

#hadoop en
export HADOOP_HOME=/home/phu/hadoop
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME

#add hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

export SQOOP_HOME=/opt/sqoop
export PATH=$PATH:$SQOOP_HOME/bin|
< | >
Ln 134, Col 34    100%    Unix (LF)    UTF-8
```

Chỉnh sửa các cấu hình trong mysql

```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ source ~/.bashrc  
phu@DESKTOP-O4NCH03:~$ sudo mysql_secure_installation  
  
Securing the MySQL server deployment.  
  
Connecting to MySQL using a blank password.  
  
VALIDATE PASSWORD COMPONENT can be used to test passwords  
and improve security. It checks the strength of password  
and allows the users to set only those passwords which are  
secure enough. Would you like to setup VALIDATE PASSWORD component?  
  
Press y|Y for Yes, any other key for No: y  
  
There are three levels of password validation policy:  
  
LOW      Length >= 8  
MEDIUM  Length >= 8, numeric, mixed case, and special characters  
STRONG  Length >= 8, numeric, mixed case, special characters and dictionary  
         file  
  
Please enter 0 = LOW, 1 = MEDIUM and 2 = STRONG: 0  
  
Skipping password set for root as authentication with auth_socket is used by default.  
If you would like to use password authentication instead, this can be done with the "AL  
TER_USER" command.  
See https://dev.mysql.com/doc/refman/8.0/en/alter-user.html#alter-user-password-managem  
ent for more information.  
  
By default, a MySQL installation has an anonymous user,  
allowing anyone to log into MySQL without having to have  
a user account created for them. This is intended only for  
testing, and to make the installation go a bit smoother.  
You should remove them before moving into a production  
environment.  
  
Remove anonymous users? (Press y|Y for Yes, any other key for No) : y  
Success.  
  
Normally, root should only be allowed to connect from  
'localhost'. This ensures that someone cannot guess at  
the root password from the network.
```

Chỉnh lại lắng nghe mọi port

```
phu@DESKTOP-O4NCH03: ~  
GNU nano 7.2 /etc/mysql/mysql.conf.d/mysqld.cnf *  
#  
# The MySQL database server configuration file.  
#  
# One can use all long options that the program supports.  
# Run program with --help to get a list of available options and with  
# --print-defaults to see which it would actually understand and use.  
#  
# For explanations see  
# http://dev.mysql.com/doc/mysql/en/server-system-variables.html  
#  
# Here is entries for some specific programs  
# The following values assume you have at least 32M ram  
[mysqld]  
#  
# * Basic Settings  
#  
user                = mysql  
# pid-file           = /var/run/mysqld/mysqld.pid  
# socket              = /var/run/mysqld/mysqld.sock  
# port                = 3306  
# datadir             = /var/lib/mysql  
#  
# If MySQL is running as a replication slave, this should be  
# changed. Ref https://dev.mysql.com/doc/refman/8.0/en/server-system-variables.html#sy  
# tmpdir              = /tmp  
#  
# Instead of skip-networking the default is now to listen only on  
# localhost which is more compatible and is not less secure.  
bind-address         = 0.0.0.0  
mysqlx-bind-address  = 0.0.0.0  
#  
# * Fine Tuning  
#  
key_buffer_size      = 16M  
# max_allowed_packet  = 64M  
# thread_stack        = 256K  
#  
^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
```

Di chuyển thư mục đó vào trong lib


```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ ls  
CentOS-7-x86_64-DVD-2009.iso  mysql-connector-j-8.0.33.tar.gz  
crawl                        sqoop-1.4.7.bin__hadoop-2.6.0  
hadoop                        sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz  
hadoop-3.3.6.tar.gz.1  
phu@DESKTOP-O4NCH03:~$ sudo mv sqoop-1.4.7.bin__hadoop-2.6.0 /usr/lib/sqoop
```

Chỉnh file cấu hình SQOOP

```

phu@DESKTOP-04NCH03: /opt/sqoop/conf
GNU nano 7.2 sqoop-env.sh *
# Set the path to where you have installed Java
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Set the path to where you have installed Hadoop
export HADOOP_COMMON_HOME=/home/phu/hadoop

# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You 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.

# included in all the hadoop scripts with source command
# should not be executable directly
# also should not be passed any arguments, since we need original $*

# Set Hadoop-specific environment variables here.

#Set path to where bin/hadoop is available
#export HADOOP_COMMON_HOME=

#Set path to where hadoop-*-core.jar is available
#export HADOOP_MAPRED_HOME=

#set the path to where bin/hbase is available
#export HBASE_HOME=

#Set the path to where bin/hive is available
#export HIVE_HOME=

^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

```

Copy cấu hình JDBC vào liv của SQOOP

```

phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$ sudo cp mysql-connector-j-8.0.33/mysql-connector-j-8.0.33.jar /o
pt/sqoop/lib/
phu@DESKTOP-04NCH03:~$

```

Tải coomon lang

```
phu@DESKTOP-04NCH03:~$ wget https://repo1.maven.org/maven2/commons-lang/commons-lang/2.6/commons-lang-2.6.jar
--2025-10-13 17:18:55-- https://repo1.maven.org/maven2/commons-lang/commons-lang/2.6/commons-lang-2.6.jar
Resolving repo1.maven.org (repo1.maven.org)... 104.18.18.12, 104.18.19.12, 2606:4700::6812:130c, ...
Connecting to repo1.maven.org (repo1.maven.org)|104.18.18.12|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 284220 (278K) [application/java-archive]
Saving to: 'commons-lang-2.6.jar'

commons-lang-2.6.jar 100%[=====>] 277.56K 1.09MB/s in 0.2s

2025-10-13 17:18:56 (1.09 MB/s) - 'commons-lang-2.6.jar' saved [284220/284220]

phu@DESKTOP-04NCH03:~$
```

Copy jarr vào lib

```
cp: cannot stat '/home/phu/downloads/commons-lang-2.6.jar': No such file or directory
phu@DESKTOP-04NCH03:~$ ls
CentOS-7-x86_64-DVD-2009.iso  hadoop-3.3.6.tar.gz.1
commons-lang-2.6.jar          mysql-connector-j-8.0.33
crawl                        mysql-connector-j-8.0.33.tar.gz
hadoop                      sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
phu@DESKTOP-04NCH03:~$ sudo cp commons-lang-2.6.jar /opt/sqoop/lib/
phu@DESKTOP-04NCH03:~$
```

PHẦN VỀ THỰC HIỆN CẤU HÌNH PIG

Thực hiện cài đặt

```
phu@DESKTOP-04NCH03: ~/crawl
phu@DESKTOP-04NCH03:~/crawl$
phu@DESKTOP-04NCH03:~/crawl$ wget https://archive.apache.org/dist/pig/pig-0.17.0/pig-0.17.0.tar.gz
--2025-10-13 17:47:50-- https://archive.apache.org/dist/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'

pig-0.17.0.tar.gz      0%[ ] 523.32K 130KB/s eta 28m 18s
```

Di chuyển và thực hiện việc cấp quyền

```
phu@DESKTOP-04NCH03:~/crawl$ sudo mv pig-0.17.0 /opt/pig
[sudo] password for phu:
phu@DESKTOP-04NCH03:~/crawl$ sudo chown -R $USER:$USER /opt/pig
```

Chỉnh lại file bashrc

```
phu@
[sudo] export PIG_HOME=/opt/pig
phu@ export PATH=$PATH:$PIG_HOME/bin|
phu@ <
cat:
phu@
```

Ln 137, Col 32	100%	Unix (LF)	UTF-8
----------------	------	-----------	-------

Sao chép file cấu hình mẫu sang file cấu hình hiện tại

```
phu@DESKTOP-04NCH03: /opt/pig/conf
GNU nano 7.2 /opt/pig/conf/pig-env.sh *
# Thiết lập đường dẫn Hadoop để Pig có thể tìm thấy các thư viện cần thiết
export HADOOP_HOME=/home/phu/hadoop
```

Cấp quyền thực thi

```
phu@DESKTOP-04NCH03: /opt/pig/conf$ sudo nano /opt/pig/conf/pig-env.sh
phu@DESKTOP-04NCH03: /opt/pig/conf$ sudo chmod +x /opt/pig/conf/pig-env.sh
phu@DESKTOP-04NCH03: /opt/pig/conf$
```

Kiểm tra version của pig

```
phu@DESKTOP-04NCH03: /opt/pig/conf$ pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
phu@DESKTOP-04NCH03: /opt/pig/conf$
```

PHẦN VỀ CẤU HÌNH HIVE

1 Tải hive : wget <https://archive.apache.org/dist/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz>

```
phu@DESKTOP-04NCH03: /opt/pig/conf
phu@DESKTOP-04NCH03: /opt/pig/conf$ wget https://archive.apache.org/dist/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz
--2025-10-13 18:20:15-- https://archive.apache.org/dist/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 326940667 (312M) [application/x-gzip]
Saving to: 'apache-hive-3.1.3-bin.tar.gz'

apache-hive-3.1.3-bin.tar.gz 100%[=====>] 311.79M 219KB/s in 31m 29s

2025-10-13 18:51:46 (169 KB/s) - 'apache-hive-3.1.3-bin.tar.gz' saved [326940667/326940667]

phu@DESKTOP-04NCH03: /opt/pig/conf$
```

3. Giải nén và thực hiện di chuyển

```
apache-hive-3.1.3-bin/hcatalog/share/webhcat/ivr/lib/wadl-resourcedoc-doclet-1.4.jar
apache-hive-3.1.3-bin/hcatalog/share/webhcat/ivr/lib/xercesImpl-2.9.1.jar
apache-hive-3.1.3-bin/hcatalog/share/webhcat/ivr/lib/xml-apis-1.3.04.jar
apache-hive-3.1.3-bin/hcatalog/share/webhcat/ivr/lib/commons-exec-1.1.jar
apache-hive-3.1.3-bin/hcatalog/share/webhcat/java-client/hive-webhcat-java-client-3.1.3.jar
phu@DESKTOP-04NCH03: /opt/pig/conf$ sudo mv apache-hive-3.1.3-bin /opt/hive
[sudo] password for phu:
phu@DESKTOP-04NCH03: /opt/pig/conf$
```

Thực hiện xóa guava cũ và thay guava mới

```
phu@DESKTOP-04NCH03: /opt/pig/conf$ sudo mv apache-hive-3.1.3-bin /opt/hive
[sudo] password for phu:
phu@DESKTOP-04NCH03: /opt/pig/conf$ sudo chown -R $USER:$USER /opt/hive
phu@DESKTOP-04NCH03: /opt/pig/conf$ source ~/.bashrc
phu@DESKTOP-04NCH03: /opt/pig/conf$ cd $HIVE_HOME
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$
phu@DESKTOP-04NCH03: /opt/hive$ rm /opt/hive/lib/guava-19.0.jar
phu@DESKTOP-04NCH03: /opt/hive$ cp /home/phu/hadoop/share/hadoop/hdfs/lib/guava-27.0-jre.jar /opt/hive/lib/
phu@DESKTOP-04NCH03: /opt/hive$
```

4. Cấu hình metastore

```
phu@DESKTOP-04NCH03: /opt/hive$ sudo cp /opt/sqoop/lib/mysql-connector-j-8.0.33.jar /opt/hive/lib/
phu@DESKTOP-04NCH03: /opt/hive$
```

Tạo cấu hình metastore

```
phu@DESKTOP-04NCH03:/opt/hive$ mysql -u root -p'Nhoc2207@' -e "CREATE DATABASE IF NOT EXISTS hive_metastore;"
mysql: [Warning] Using a password on the command line interface can be insecure.
phu@DESKTOP-04NCH03:/opt/hive$
```

Tạo file cấu hình hive-site.xml

```
phu@DESKTOP-04NCH03: /opt/hive/conf
GNU nano 7.2 /opt/hive/conf/hive-site.xml *
<configuration>
  <property>
    <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:mysql://localhost/hive_metastore?createDatabaseIfNotExist=true&useSSL=false&autoReconnect=true</value>
    <description>JDBC connect string for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionDriverName</name>
    <value>com.mysql.cj.jdbc.Driver</value>
    <description>Driver class name for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionUserName</name>
    <value>root</value>
    <description>Username for the database</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionPassword</name>
    <value>Nhoc2207@</value>
    <description>Password for the database</description>
  </property>
  <property>
    <name>hive.metastore.warehouse.dir</name>
    <value>/user/hive/warehouse</value>
    <description>location of default database for the warehouse</description>
  </property>
</configuration>
```

Nội dung :

```
<configuration>
```

```
  <property>
```

```
    <name>javax.jdo.option.ConnectionURL</name>
```

```
    <value>jdbc:mysql://localhost/hive_metastore?createDatabaseIfNotExist=true&useSSL=false&
    &autoReconnect=true</value>
```

```
    <description>JDBC connect string for a JDBC metastore</description>
```

```
  </property>
```

```
  <property>
```

```
    <name>javax.jdo.option.ConnectionDriverName</name>
```

```
    <value>com.mysql.cj.jdbc.Driver</value>
```

```
    <description>Driver class name for a JDBC metastore</description>
```

```
  </property>
```

```
  <property>
```

```
    <name>javax.jdo.option.ConnectionUserName</name>
```

```
    <value>root</value>
```

```
    <description>Username for the database</description>
```

```
  </property>
```

```
  <property>
```

```
    <name>javax.jdo.option.ConnectionPassword</name>
```

```
    <value>Nhoc2207@</value>
```

```
    <description>Password for the database</description>
```

```
  </property>
```

```
  </property>
```

```
<name>hive.metastore.warehouse.dir</name>
<value>/user/hive/warehouse</value>
<description>location of default database for the warehouse</description>
</property>
</configuration>
```

Khởi tạo metastore

```
phu@DESKTOP-04NCH03:/opt/hive/conf$ nano /opt/hive/conf/hive-site.xml
phu@DESKTOP-04NCH03:/opt/hive/conf$ schematool -dbType mysql -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
```

Kết quả

```
Initialization script completed
schemaTool completed
phu@DESKTOP-04NCH03:/opt/hive/conf$
```

Tạo thư mục warehouse trên hdfs

```
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -mkdir -p /user/hive/warehouse
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -mkdir /tmp/hive
mkdir: `hdfs://localhost/tmp': No such file or directory
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod g+w /user/hive/warehouse
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod g+w /tmp/hive
chmod: `/tmp/hive': No such file or directory
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -mkdir -p /tmp/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -mkdir /tmp/hive
mkdir: `/tmp/hive': File exists
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod g+w /user/hive/warehouse
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod g+w /tmp/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$
```

Đảm bảo trên thư mục

```
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod 777 /tmp
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod -R 777 /user/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chown -R $USER:$USER /user/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$
```



```
phu@DESKTOP-04NCH03:/opt/hive/conf$ # Đảm bảo /tmp có quyền 777 (Read, Write, Execute cho mọi người)
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod 777 /tmp
tmp/hive
hdfs dfs -chmod 777 /tmp/hive

# Cấp quyền 777 đệ quy cho thư mục Hive Warehouse
hdfs dfs -chmod -R 777 /user/hivephu@DESKTOP-04NCH03:/opt/hive/conf$
phu@DESKTOP-04NCH03:/opt/hive/conf$ # Cấp quyền 777 cho thư mục /tmp/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod 777 /tmp/hive

phu@DESKTOP-04NCH03:/opt/hive/conf$
phu@DESKTOP-04NCH03:/opt/hive/conf$ # Cấp quyền 777 đệ quy cho thư mục Hive Warehouse
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -chmod -R 777 /user/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -ls /tmp
Found 1 items
drwxrwxrwx - phu supergroup 0 2025-10-13 19:01 /tmp/hive
phu@DESKTOP-04NCH03:/opt/hive/conf$ hdfs dfs -ls /user/hive
Found 1 items
drwxrwxrwx - phu phu 0 2025-10-13 19:00 /user/hive/warehouse
phu@DESKTOP-04NCH03:/opt/hive/conf$
```

Kiểm tra cấu hình hive

```
phu@DESKTOP-04NCH03:/opt/hive/conf$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 6e68fd46-114e-4502-8ce3-1ae3af7fb621

Logging initialized using configuration in jar:file:/opt/hive/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true
Hive Session ID = 6b904faf-d6f4-4798-9fbf-5176180be126
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

Và kết quả khi show database

```
phu@DESKTOP-04NCH03:/opt/hive/conf$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 6e68fd46-114e-4502-8ce3-1ae3af7fb621

Logging initialized using configuration in jar:file:/opt/hive/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true
Hive Session ID = 6b904faf-d6f4-4798-9fbf-5176180be126
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show databases;
OK
default
Time taken: 0.466 seconds, Fetched: 1 row(s)
hive>
```

PHẦN VỀ THỰC HIỆN APACHE DRILL

```
phu@DESKTOP-04NCH03:~$ wget https://d1cdn.apache.org/drill/1.21.2/apache-drill-1.21.2.tar.gz
--2025-10-13 19:12:15-- https://d1cdn.apache.org/drill/1.21.2/apache-drill-1.21.2.tar.gz
Resolving d1cdn.apache.org (d1cdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to d1cdn.apache.org (d1cdn.apache.org)[151.101.2.132]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 457844959 (437M) [application/x-gzip]
Saving to: 'apache-drill-1.21.2.tar.gz'

apache-drill-1.21.2.tar.gz      100%[=====>] 436.63M  5.23MB/s   in 93s

2025-10-13 19:14:24 (4.69 MB/s) - 'apache-drill-1.21.2.tar.gz' saved [457844959/457844959]

phu@DESKTOP-04NCH03:~$ tar -xzf apache-drill-1.21.2.tar.gz
apache-drill-1.21.2/KEYS
apache-drill-1.21.2/LICENSE
```

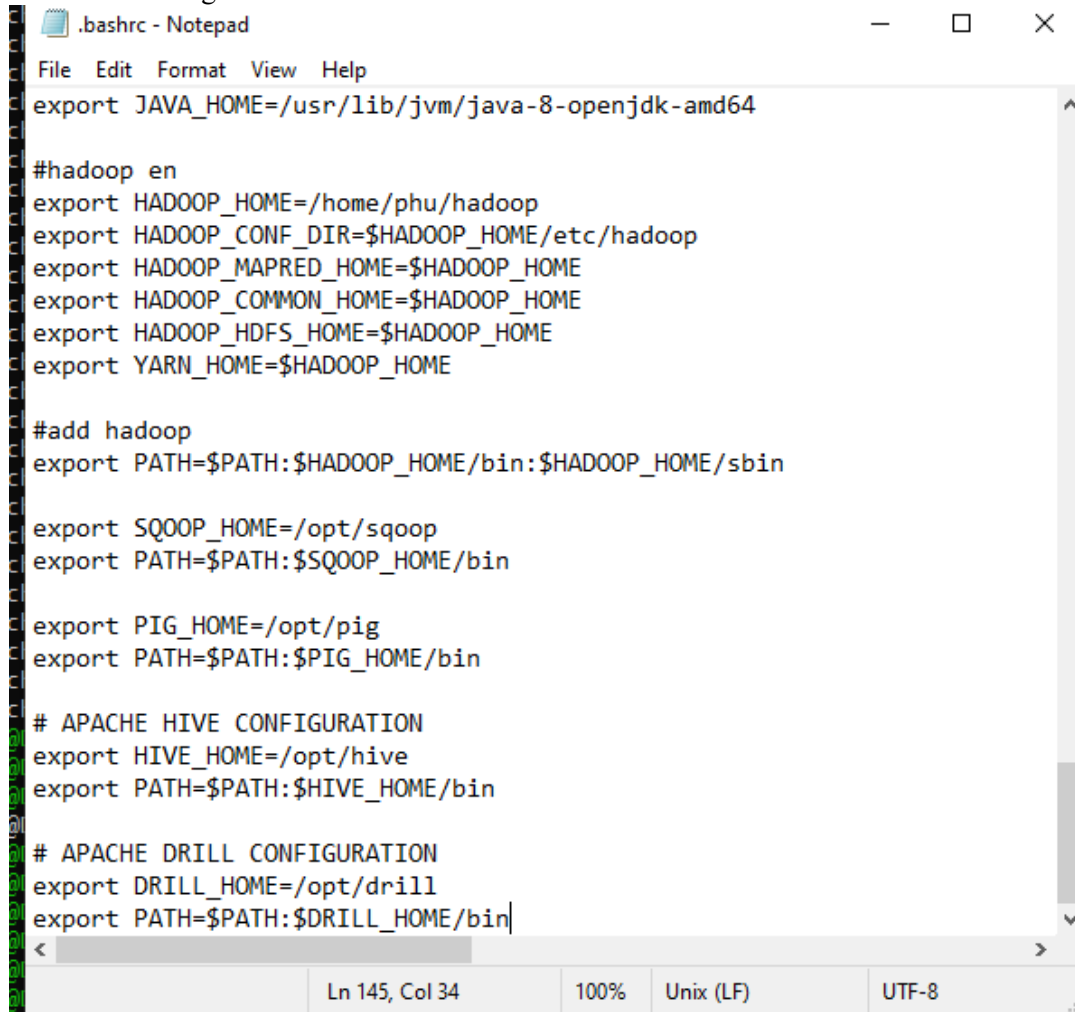
Di chuyển vào file thư mục

```
phu@DESKTOP-04NCH03:~$ sudo mv apache-drill-1.21.2 /opt/drill
[sudo] password for phu:
phu@DESKTOP-04NCH03:~$
```

Cấp quyền

```
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/drill
phu@DESKTOP-04NCH03:~$
```

Chỉnh thêm trong file cấu hình bashrc



```
.bashrc - Notepad
File Edit Format View Help
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

#hadoop en
export HADOOP_HOME=/home/phu/hadoop
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME

#add hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

export SQOOP_HOME=/opt/sqoop
export PATH=$PATH:$SQOOP_HOME/bin

export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin

# APACHE HIVE CONFIGURATION
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin

# APACHE DRILL CONFIGURATION
export DRILL_HOME=/opt/drill
export PATH=$PATH:$DRILL_HOME/bin
```

Cấp quyền và cấu hình tích hợp các dịch vụ hadoop vào trong drill

```
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$ sudo mv apache-drill-1.21.2 /opt/drill  
[sudo] password for phu:  
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/drill  
phu@DESKTOP-04NCH03:~$ source ~/.bashrc  
phu@DESKTOP-04NCH03:~$ mkdir -p /opt/drill/conf/hadoop  
phu@DESKTOP-04NCH03:~$ cp /home/phu/hadoop/etc/hadoop/core-site.xml /opt/drill/conf/hadoop/  
phu@DESKTOP-04NCH03:~$ cp /home/phu/hadoop/etc/hadoop/hdfs-site.xml /opt/drill/conf/hadoop/  
phu@DESKTOP-04NCH03:~$
```

Cài đặt zookeeper

```
phu@DESKTOP-04NCH03: ~  
phu@DESKTOP-04NCH03:~$ wget https://archive.apache.org/dist/zookeeper/zookeeper-3.6.4/apache-zookeeper-3.6.4-bin.tar.gz  
--2025-10-13 20:16:24-- https://archive.apache.org/dist/zookeeper/zookeeper-3.6.4/apache-zookeeper-3.6.4-bin.tar.gz  
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2  
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 12653363 (12M) [application/x-gzip]  
Saving to: 'apache-zookeeper-3.6.4-bin.tar.gz'  
  
apache-zookeeper-3.6.4-bin.tar 100%[=====] 12.07M 170KB/s in 2m 11s  
2025-10-13 20:18:36 (94.6 KB/s) - 'apache-zookeeper-3.6.4-bin.tar.gz' saved [12653363/12653363]  
  
phu@DESKTOP-04NCH03:~$
```

Cấp quyền

```
phu@DESKTOP-04NCH03:~$ sudo mv apache-zookeeper-3.6.4-bin /opt/zookeeper  
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/zookeeper  
phu@DESKTOP-04NCH03:~$
```

Tạo thư mục để tiến hành lưu dữ liệu

```
phu@DESKTOP-04NCH03:~$ mkdir -p ~/zookeeper_data  
phu@DESKTOP-04NCH03:~$ cd /opt/zookeeper/conf  
phu@DESKTOP-04NCH03:/opt/zookeeper/conf$ cp zoo_sample.cfg zoo.cfg  
phu@DESKTOP-04NCH03:/opt/zookeeper/conf$ nano zoo.cfg  
phu@DESKTOP-04NCH03:/opt/zookeeper/conf$
```

Code cấu hình file đó, chỉnh lại dataDir

```
phu@DESKTOP-04NCH03: /opt/zookeeper/conf  
GNU nano 7.2 zoo.cfg *  
# The number of milliseconds of each tick  
tickTime=2000  
# The number of ticks that the initial  
# synchronization phase can take  
initLimit=10  
# The number of ticks that can pass between  
# sending a request and getting an acknowledgement  
syncLimit=5  
# the directory where the snapshot is stored.  
# do not use /tmp for storage, /tmp here is just  
# example sakes.  
  
dataDir=/home/phu/zookeeper_data  
clientPort=2181  
# the maximum number of client connections.  
# increase this if you need to handle more clients  
#maxClientCnxns=60  
#  
# Be sure to read the maintenance section of the  
# administrator guide before turning on autopurge.  
#  
# http://zookeeper.apache.org/doc/current/zookeeperAdmin.html#sc_maintenance  
#  
# The number of snapshots to retain in dataDir  
#autopurge.snapRetainCount=3  
# Purge task interval in hours  
# Set to "0" to disable auto purge feature  
#autopurge.purgeInterval=1  
  
## Metrics Providers  
#  
# https://prometheus.io Metrics Exporter  
#metricsProvider.className=org.apache.zookeeper.metrics.prometheus.PrometheusMetricsProvider  
#metricsProvider.httpPort=7000  
#metricsProvider.exportJvmInfo=true  
  
^C Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location ^U Undo ^A Set Mark
```

Áp dụng các cấu hình bashrc

```
# APACHE ZOOKEEPER CONFIGURATION
export ZK_HOME=/opt/zookeeper
export PATH=$PATH:$ZK_HOME/bin|
<
Ln 149, Col 31    100%    Unix (LF)    UTF-8
phu@DESKTOP-04NCH03:~$ mkdir -p ~/zookeeper_data
```

Khởi động dịch vụ zookeeper

```
phu@DESKTOP-04NCH03:/opt/zookeeper/conf$ zkServer.sh start
ZooKeeper JMX enabled by default
Using config: /opt/zookeeper/bin/../conf/zoo.cfg
Starting zookeeper ... STARTED
phu@DESKTOP-04NCH03:/opt/zookeeper/conf$
```

Kiểm tra kết quả : /opt/drill/bin/drill-embedded

```
phu@DESKTOP-04NCH03: /opt/drill
phu@DESKTOP-04NCH03:/opt/drill$ /opt/drill/bin/drill-embedded
Apache Drill 1.19.0
"Good friends, good books and Drill cluster: this is the ideal life."
apache drill>
apache drill>
apache drill>
apache drill>
```

PHÂN TÀI VÀ CÀI ĐẶT VỀ HBASE

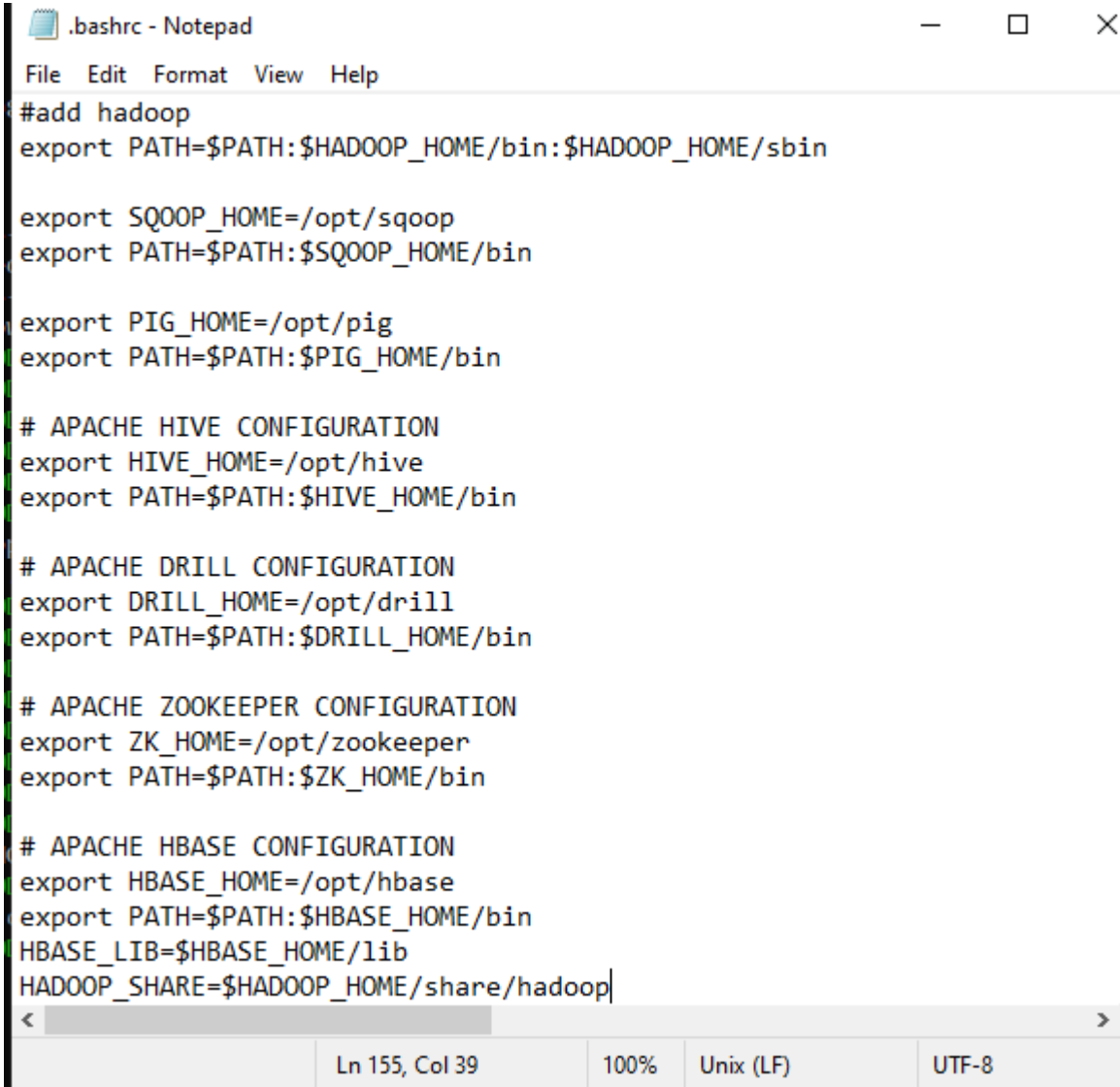
1. Tải và cài đặt

```
phu@DESKTOP-O4NCH03: ~  
phu@DESKTOP-O4NCH03:~$ wget https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz  
--2025-10-13 20:42:01-- https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz  
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2  
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 283496242 (270M) [application/x-gzip]  
Saving to: 'hbase-2.4.9-bin.tar.gz'  
  
hbase-2.4.9-bin.tar.gz      0%[                               ] 117.02K  70.4KB/s  
  
phu@DESKTOP-O4NCH03:~$ wget https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz  
--2025-10-13 20:42:01-- https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz  
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2  
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 283496242 (270M) [application/x-gzip]  
Saving to: 'hbase-2.4.9-bin.tar.gz'  
  
hbase-2.4.9-bin.tar.gz      100%[=====] 270.36M  111KB/s  in 75m 36s  
2025-10-13 21:57:38 (61.0 KB/s) - 'hbase-2.4.9-bin.tar.gz' saved [283496242/283496242]  
  
phu@DESKTOP-O4NCH03:~$ wget https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz  
--2025-10-13 22:28:40-- https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz  
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644  
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.  
HTTP request sent, awaiting response... 404 Not Found  
2025-10-13 22:28:40 ERROR 404: Not Found.  
  
phu@DESKTOP-O4NCH03:~$ wget https://dlcdn.apache.org/zeppelin/zeppelin-0.11.0/zeppelin-0.11.0-bin-all.tgz  
--2025-10-13 22:28:48-- https://dlcdn.apache.org/zeppelin/zeppelin-0.11.0/zeppelin-0.11.0-bin-all.tgz  
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644  
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 884230650 (843M) [application/x-gzip]  
Saving to: 'zeppelin-0.11.0-bin-all.tgz'  
  
zeppelin-0.11.0-bin-all.tgz 100%[=====] 843.27M  6.50MB/s  in 2m 3s  
2025-10-13 22:32:50 (6.84 MB/s) - 'zeppelin-0.11.0-bin-all.tgz' saved [884230650/884230650]  
  
phu@DESKTOP-O4NCH03:~$ wget https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz  
--2025-10-13 22:34:51-- https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz  
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2  
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 400446614 (382M) [application/x-gzip]  
Saving to: 'spark-3.5.1-bin-hadoop3.tgz'  
  
spark-3.5.1-bin-hadoop3.tgz 82%[=====] 314.59M  74.9KB/s  eta 14m 2s  
  
phu@DESKTOP-O4NCH03:~$ wget https://archive.apache.org/dist/phoenix/phoenix-5.1.3/phoenix-hbase-2.4-5.1.3-bin.tar.gz  
--2025-10-13 23:52:41-- https://archive.apache.org/dist/phoenix/phoenix-5.1.3/phoenix-hbase-2.4-5.1.3-bin.tar.gz  
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2  
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 341278569 (325M) [application/x-gzip]  
Saving to: 'phoenix-hbase-2.4-5.1.3-bin.tar.gz'  
  
phoenix-hbase-2.4-5.1.3-bin.ta 100%[=====] 325.47M  192KB/s  in 48m 54s  
2025-10-14 00:41:37 (114 KB/s) - 'phoenix-hbase-2.4-5.1.3-bin.tar.gz' saved [341278569/341278569]
```

Thực hiện giải nén và đổi tên dễ nhớ

```
phu@DESKTOP-04NCH03:~$ sudo mv hbase-2.4.9 /opt/hbase
[sudo] password for phu:
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/hbase
phu@DESKTOP-04NCH03:~$ source ~/.bashrc
phu@DESKTOP-04NCH03:~$
```

Cấu hình file bashrc



```
.bashrc - Notepad
File Edit Format View Help
#add hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

export SQOOP_HOME=/opt/sqoop
export PATH=$PATH:$SQOOP_HOME/bin

export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin

# APACHE HIVE CONFIGURATION
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin

# APACHE DRILL CONFIGURATION
export DRILL_HOME=/opt/drill
export PATH=$PATH:$DRILL_HOME/bin

# APACHE ZOOKEEPER CONFIGURATION
export ZK_HOME=/opt/zookeeper
export PATH=$PATH:$ZK_HOME/bin

# APACHE HBASE CONFIGURATION
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin
HBASE_LIB=$HBASE_HOME/lib
HADOOP_SHARE=$HADOOP_HOME/share/hadoop
```

Cấu hình hbase

Chỉnh sửa hbase-env

phu@DESK IOP-U4NCHU3: /opt/hbase/conf

```
GNU nano 7.2 /opt/hbase/conf/hbase-env.sh *
#!/usr/bin/env bash
#
#/**
# * Licensed to the Apache Software Foundation (ASF) under one
# * or more contributor license agreements. See the NOTICE file
# * distributed with this work for additional information
# * regarding copyright ownership. The ASF licenses this file
# * to you 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.
# */
#
# Set environment variables here.
#
# This script sets variables multiple times over the course of starting an hbase process,
# so try to keep things idempotent unless you want to take an even deeper look
# into the startup scripts (bin/hbase, etc.)
#
# The java implementation to use. Java 1.8+ required.
# export JAVA_HOME=/usr/java/jdk1.8.0/
export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"
#
# Ngăn HBase tự khởi động ZooKeeper của riêng nó (dùng ZooKeeper đã cài đặt)
export HBASE_MANAGES_ZK=false
#
# Extra Java CLASSPATH elements. Optional.
# export HBASE_CLASSPATH=
#
# The maximum amount of heap to use. Default is left to JVM default.
# export HBASE_HEAPSIZE=1G
#
# Uncomment below if you intend to use off heap cache. For example, to allocate 8G of
```

Chỉnh hbase-site

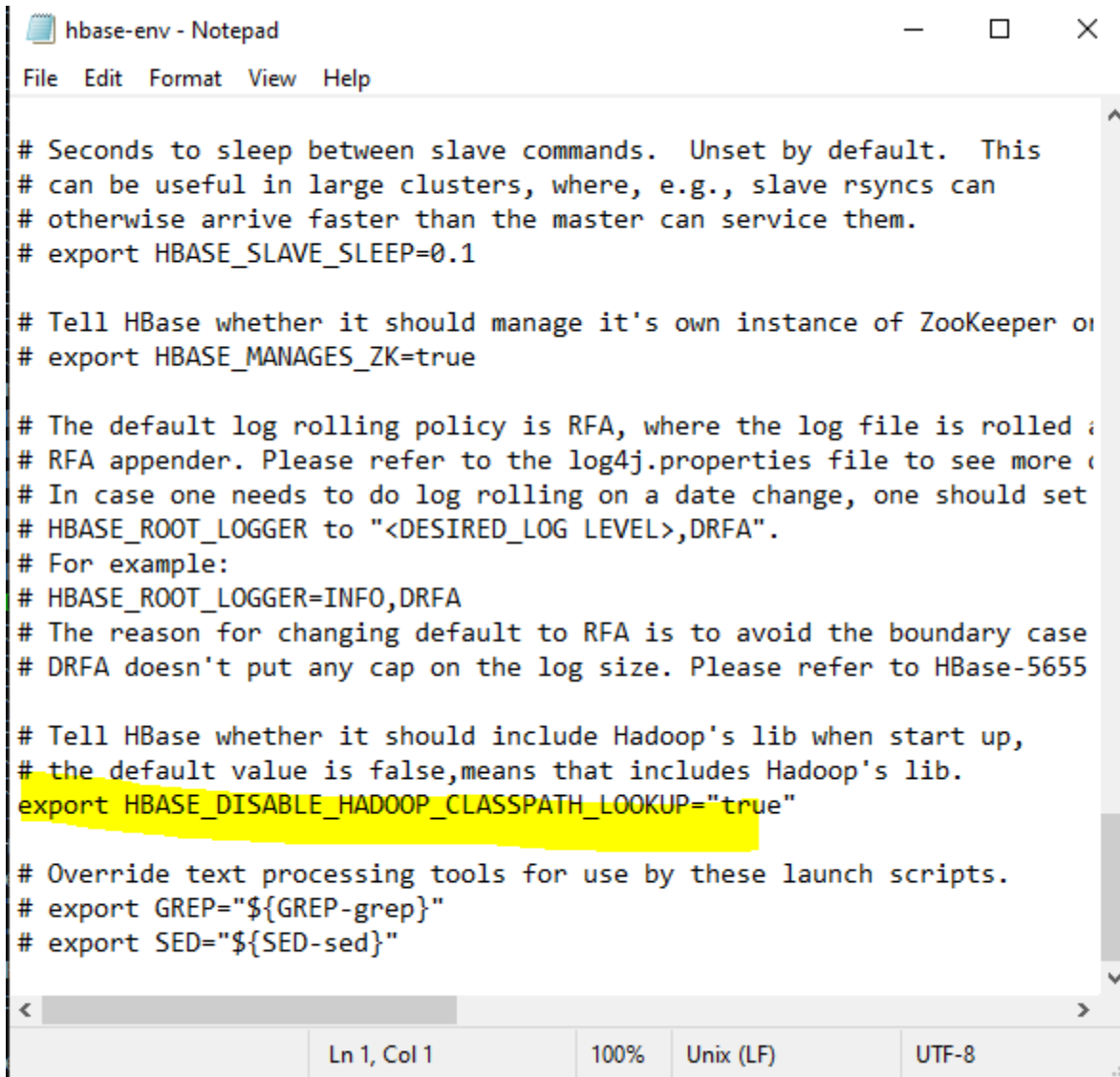
phu@DESKTOP-O4NCH03: /opt/hbase/conf

```
GNU nano 7.2 /opt/hbase/conf/hbase-site.xml *
<configuration>
  <!--
    The following properties are set for running HBase as a single process on a
    developer workstation. With this configuration, HBase is running in
    "stand-alone" mode and without a distributed file system. In this mode, and
    without further configuration, HBase and ZooKeeper data are stored on the
    local filesystem, in a path under the value configured for `hbase.tmp.dir`.
    This value is overridden from its default value of `/tmp` because many
    systems clean `/tmp` on a regular basis. Instead, it points to a path within
    this HBase installation directory.

    Running against the `LocalFileSystem`, as opposed to a distributed
    filesystem, runs the risk of data integrity issues and data loss. Normally
    HBase will refuse to run in such an environment. Setting
    `hbase.unsafe.stream.capability.enforce` to `false` overrides this behavior,
    permitting operation. This configuration is for the developer workstation
    only and __should not be used in production!__

    See also https://hbase.apache.org/book.html#standalone\_dist
  -->
  <property>
    <name>hbase.cluster.distributed</name>
    <value>true</value>
  </property>
  <property>
    <name>hbase.zookeeper.quorum</name>
    <value>localhost</value>
    <description>Địa chỉ ZooKeeper</description>
  </property>
  <property>
    <name>hbase.tmp.dir</name>
    <value>./tmp</value>
  </property>
  <property>
    <name>hbase.unsafe.stream.capability.enforce</name>
    <value>>false</value>
  </property>
</configuration>
```

Bỏ bình luận dòng này để tránh trùng classpath



```
# Seconds to sleep between slave commands. Unset by default. This
# can be useful in large clusters, where, e.g., slave rsyncs can
# otherwise arrive faster than the master can service them.
# export HBASE_SLAVE_SLEEP=0.1

# Tell HBase whether it should manage it's own instance of ZooKeeper or
# export HBASE_MANAGES_ZK=true

# The default log rolling policy is RFA, where the log file is rolled a
# RFA appender. Please refer to the log4j.properties file to see more
# In case one needs to do log rolling on a date change, one should set
# HBASE_ROOT_LOGGER to "<DESIRED_LOG_LEVEL>,DRFA".
# For example:
# HBASE_ROOT_LOGGER=INFO,DRFA
# The reason for changing default to RFA is to avoid the boundary case
# DRFA doesn't put any cap on the log size. Please refer to HBase-5655

# Tell HBase whether it should include Hadoop's lib when start up,
# the default value is false, means that includes Hadoop's lib.
export HBASE_DISABLE_HADOOP_CLASSPATH_LOOKUP="true"

# Override text processing tools for use by these launch scripts.
# export GREP="${GREP-grep}"
# export SED="${SED-sed}"
```

Khởi chạy dịch vụ và kiểm tra jps

```
phu@DESKTOP-04NCH03:/opt/hbase/conf$ start-hbase.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
running master, logging to /opt/hbase/logs/hbase-phu-master-DESKTOP-04NCH03.out
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
: running regionserver, logging to /opt/hbase/logs/hbase-phu-regionserver-DESKTOP-04NCH03.out
: SLF4J: Class path contains multiple SLF4J bindings.
: SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
: SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
: SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
phu@DESKTOP-04NCH03:/opt/hbase/conf$ jps
331973 Jps
331813 HRegionServer
331639 HMaster
325127 NodeManager
324817 SecondaryNameNode
324472 NameNode
324632 DataNode
329579 QuorumPeerMain
325003 ResourceManager
phu@DESKTOP-04NCH03:/opt/hbase/conf$
```

Kiểm tra hbase shell

```
phu@DESKTOP-04NCH03:/opt/hbase$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hbase/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
2025-10-14 01:07:26,877 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.9, rc49f7f63fca144765bf7c2da41791769286dfccc, Fri Dec 17 19:02:09 PST 2021
Took 0.0019 seconds
hbase:001:0> list
TABLE
0 row(s)
Took 0.5388 seconds
=> []
hbase:002:0>
```

PHẦN VỀ CẤU HÌNH SPARK

1. Lệnh cài đặt

```
phu@DESKTOP-04NCH03:~$ wget https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz
--2025-10-13 20:42:01-- https://archive.apache.org/dist/hbase/2.4.9/hbase-2.4.9-bin.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 283496242 (270M) [application/x-gzip]
Saving to: 'hbase-2.4.9-bin.tar.gz'

hbase-2.4.9-bin.tar.gz      100%[=====] 270.36M  111KB/s   in 75m 36s
2025-10-13 21:57:38 (61.0 KB/s) - 'hbase-2.4.9-bin.tar.gz' saved [283496242/283496242]

phu@DESKTOP-04NCH03:~$ wget https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
--2025-10-13 22:28:40-- https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-10-13 22:28:40 ERROR 404: Not Found.

phu@DESKTOP-04NCH03:~$ wget https://dlcdn.apache.org/zeppelin/zeppelin-0.11.0/zeppelin-0.11.0-bin-all.tgz
--2025-10-13 22:28:48-- https://dlcdn.apache.org/zeppelin/zeppelin-0.11.0/zeppelin-0.11.0-bin-all.tgz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 884230650 (843M) [application/x-gzip]
Saving to: 'zeppelin-0.11.0-bin-all.tgz'

zeppelin-0.11.0-bin-all.tgz  100%[=====] 843.27M  6.50MB/s   in 2m 3s
2025-10-13 22:32:50 (6.84 MB/s) - 'zeppelin-0.11.0-bin-all.tgz' saved [884230650/884230650]

phu@DESKTOP-04NCH03:~$ wget https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
--2025-10-13 22:34:51-- https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400446614 (382M) [application/x-gzip]
Saving to: 'spark-3.5.1-bin-hadoop3.tgz'

spark-3.5.1-bin-hadoop3.tgz  82%[=====] 314.59M  74.9KB/s   eta 14m 2s
```

2. Giải nén đối tên và di chuyển vào thư mục

phu@DESKTOP-04NCH03: ~

```

spark-3.5.1-bin-hadoop3/R/lib/SparkR/Meta/features.rds
spark-3.5.1-bin-hadoop3/R/lib/SparkR/doc/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/doc/index.html
spark-3.5.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.5.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.5.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdx
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/paths.rds
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdb
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/AnIndex
spark-3.5.1-bin-hadoop3/R/lib/SparkR/help/aliases.rds
spark-3.5.1-bin-hadoop3/R/lib/SparkR/NAMESPACE
spark-3.5.1-bin-hadoop3/R/lib/SparkR/tests/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/tests/testthat/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/tests/testthat/test_basic.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.5.1-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.5.1-bin-hadoop3/R/lib/SparkR/profile/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/profile/general.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/profile/shell.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/worker/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/worker/daemon.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/worker/worker.R
spark-3.5.1-bin-hadoop3/R/lib/SparkR/html/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/html/00Index.html
spark-3.5.1-bin-hadoop3/R/lib/SparkR/html/R.css
spark-3.5.1-bin-hadoop3/R/lib/SparkR/R/
spark-3.5.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx
spark-3.5.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb
spark-3.5.1-bin-hadoop3/R/lib/SparkR/R/SparkR
spark-3.5.1-bin-hadoop3/R/lib/sparkr.zip
phu@DESKTOP-04NCH03:~$ ls
CentOS-7-x86_64-DVD-2009.iso          drill                                phoenix-hbase-2.4-5.1.3-bin.tar.gz.1
apache-drill-1.19.0.tar.gz             hadoop                              spark-3.5.1-bin-hadoop3
apache-drill-1.19.0.tar.gz.1           hadoop-3.3.6.tar.gz.1              spark-3.5.1-bin-hadoop3.tgz
apache-drill-1.21.2.tar.gz             hbase-2.4.9-bin.tar.gz             sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
apache-zookeeper-3.6.4-bin.tar.gz      mysql-connector-j-8.0.33            zeppelin-0.11.0-bin-all.tgz
commons-lang-2.6.jar                   mysql-connector-j-8.0.33.tar.gz      zookeeper_data
crawl                                  phoenix-hbase-2.4-5.1.3-bin.tar.gz
phu@DESKTOP-04NCH03:~$ sudo mv spark-3.5.1-bin-hadoop3 /opt/spark
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/spark
phu@DESKTOP-04NCH03:~$

```

Chỉnh lại cấu hình bashrc

```
.bashrc - Notepad
File Edit Format View Help
export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin

# APACHE HIVE CONFIGURATION
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin

# APACHE DRILL CONFIGURATION
export DRILL_HOME=/opt/drill
export PATH=$PATH:$DRILL_HOME/bin

# APACHE ZOOKEEPER CONFIGURATION
export ZK_HOME=/opt/zookeeper
export PATH=$PATH:$ZK_HOME/bin

# APACHE HBASE CONFIGURATION
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin
HBASE_LIB=$HBASE_HOME/lib
HADOOP_SHARE=$HADOOP_HOME/share/hadoop

# APACHE SPARK CONFIGURATION
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
# Yêu cầu Spark sử dụng cấu hình Hadoop đã có
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop|
< | Ln 161, Col 47 100% Unix (LF) UTF-8
```

Sao chép các JAR của hadoop vào lib của SPARK

```
phu@DESKTOP-O4NCH03:~$ cp $HADOOP_SHARE/common/lib/*.jar $SPARK_HOME/jars/
phu@DESKTOP-O4NCH03:~$ cp $HADOOP_SHARE/hdfs/lib/*.jar $SPARK_HOME/jars/
phu@DESKTOP-O4NCH03:~$
```

Sửa spark.env

```
phu@DESKTOP-O4NCH03: /opt/spark/conf
GNU nano 7.2 spark-env.sh *
# Xác định HADOOP_CONF_DIR cho các ứng dụng Spark chạy trên YARN
export HADOOP_CONF_DIR=/home/phu/hadoop/etc/hadoop
```

Kiểm tra cấu hình cuối cùng


```
phu@DESKTOP-04NCH03: ~  
phu@DESKTOP-04NCH03:~$ source ~/.bashrc  
phu@DESKTOP-04NCH03:~$ cp $HADOOP_SHARE/common/lib/*.jar $SPARK_HOME/jars/  
phu@DESKTOP-04NCH03:~$ cp $HADOOP_SHARE/hdfs/lib/*.jar $SPARK_HOME/jars/  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$ cd $SPARK_HOME/conf  
phu@DESKTOP-04NCH03:/opt/spark/conf$ nano spark-env.sh  
phu@DESKTOP-04NCH03:/opt/spark/conf$ ls  
fairscheduler.xml.template  metrics.properties.template  spark-env.sh.template  
log4j2.properties.template  spark-defaults.conf.template  workers.template  
phu@DESKTOP-04NCH03:/opt/spark/conf$ nano spark-env.sh  
phu@DESKTOP-04NCH03:/opt/spark/conf$ ls  
fairscheduler.xml.template  metrics.properties.template  spark-env.sh  workers.template  
log4j2.properties.template  spark-defaults.conf.template  spark-env.sh.template  
phu@DESKTOP-04NCH03:/opt/spark/conf$ cd  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$ spark-shell  
25/10/14 01:14:38 WARN Utils: Your hostname, DESKTOP-04NCH03 resolves to a loopback address: 127.0.1.1; using 172.30.212.45  
instead (on interface eth0)  
25/10/14 01:14:38 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/10/14 01:14:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe  
s where applicable  
Spark context Web UI available at http://172.30.212.45:4040  
Spark context available as 'sc' (master = local[*], app id = local-1760379297614).  
Spark session available as 'spark'.  
Welcome to  
  
      _ _ _ _ _  
     / _ _ _ _ \  
    / _ _ _ _ \  
   / _ _ _ _ \  
  / _ _ _ _ \  
 / _ _ _ _ \  
/_ _ _ _ _ \  
version 3.5.1  
  
Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 1.8.0_462)  
Type in expressions to have them evaluated.  
Type :help for more information.  
  
scala>
```

PHẦN VỀ CẤU HÌNH PHOENIX

1. Các lệnh cài đặt

```
phu@DESKTOP-04NCH03:~$ wget https://archive.apache.org/dist/phoenix/phoenix-5.1.3/phoenix-hbase-2.4-5.1.3-bin.tar.gz
--2025-10-13 23:52:41-- https://archive.apache.org/dist/phoenix/phoenix-5.1.3/phoenix-hbase-2.4-5.1.3-bin.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 341278569 (325M) [application/x-gzip]
Saving to: 'phoenix-hbase-2.4-5.1.3-bin.tar.gz'

phoenix-hbase-2.4-5.1.3-bin.ta 100%[=====] 325.47M 192KB/s in 48m 54s

2025-10-14 00:41:37 (114 KB/s) - 'phoenix-hbase-2.4-5.1.3-bin.tar.gz' saved [341278569/341278569]
```

2. Giải nén thư mục và tiến hành đổi tên, di chuyển

```
phoenix-hbase-2.4-5.1.3-bin/examples/wto_start_queries.sql
phoenix-hbase-2.4-5.1.3-bin/lib/slf4j-reload4j-1.7.36.jar
phoenix-hbase-2.4-5.1.3-bin/lib/reload4j-1.2.24.jar
phoenix-hbase-2.4-5.1.3-bin/lib/sqlline-1.9.0-jar-with-dependencies.jar
phu@DESKTOP-04NCH03:~$ ls
CentOS-7-x86_64-DVD-2009.iso          drill                                phoenix-hbase-2.4-5.1.3-bin.tar.gz
apache-drill-1.19.0.tar.gz            hadoop                              phoenix-hbase-2.4-5.1.3-bin.tar.gz.1
apache-drill-1.19.0.tar.gz.1         hadoop-3.3.6.tar.gz.1             spark-3.5.1-bin-hadoop3.tgz
apache-drill-1.21.2.tar.gz           hbase-2.4.9-bin.tar.gz           sqoop-1.4.7-bin_hadoop-2.6.0.tar.gz
apache-zookeeper-3.6.4-bin.tar.gz    mysql-connector-j-8.0.33          zeppelin-0.11.0-bin-all.tgz
commons-lang-2.6.jar                mysql-connector-j-8.0.33.tar.gz   zookeeper_data
crawl                                phoenix-hbase-2.4-5.1.3-bin
phu@DESKTOP-04NCH03:~$ sudo mv phoenix-hbase-2.4-5.1.3-bin /opt/phoenix
phu@DESKTOP-04NCH03:~$ sudo chown -R $USER:$USER /opt/phoenix
phu@DESKTOP-04NCH03:~$
```

3. Tích hợp Phoenix Server JAR vào HBASE

```
phu@DESKTOP-04NCH03:~$ jps
334838 HRegionServer
325127 NodeManager
324817 SecondaryNameNode
324472 NameNode
324632 DataNode
334713 HMaster
329579 QuorumPeerMain
335738 Jps
325003 ResourceManager
phu@DESKTOP-04NCH03:~$ kill -9 334838 334713
-bash: kill: (334838) - No such process
-bash: kill: (334713) - No such process
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$ jps
325127 NodeManager
324817 SecondaryNameNode
324472 NameNode
324632 DataNode
335771 Jps
329579 QuorumPeerMain
325003 ResourceManager
phu@DESKTOP-04NCH03:~$ PHOENIX_SERVER_JAR="/opt/phoenix/phoenix-server-hbase-2.4-5.1.3.jar"
phu@DESKTOP-04NCH03:~$ HBASE_LIB_DIR="/opt/hbase/lib"
phu@DESKTOP-04NCH03:~$ sudo cp $PHOENIX_SERVER_JAR $HBASE_LIB_DIR
phu@DESKTOP-04NCH03:~$
```

4. Cấu hình biến môi trường bashrc

```
.bashrc - Notepad
File Edit Format View Help
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin

# APACHE DRILL CONFIGURATION
export DRILL_HOME=/opt/drill
export PATH=$PATH:$DRILL_HOME/bin

# APACHE ZOOKEEPER CONFIGURATION
export ZK_HOME=/opt/zookeeper
export PATH=$PATH:$ZK_HOME/bin

# APACHE HBASE CONFIGURATION
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin
HBASE_LIB=$HBASE_HOME/lib
HADOOP_SHARE=$HADOOP_HOME/share/hadoop

# APACHE SPARK CONFIGURATION
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
# Yêu cầu Spark sử dụng cấu hình Hadoop đã có
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop

# APACHE PHOENIX CONFIGURATION
export PHOENIX_HOME=/opt/phoenix
export PATH=$PATH:$PHOENIX_HOME/bin|
< >
Ln 165, Col 36 100% Unix (LF) UTF-8
```

Khởi động lại hbase và kiểm tra hoạt động của phoenix

```
phu@DESKTOP-04NCH03:~$ start-hbase.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hbase/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
running master, logging to /opt/hbase/logs/hbase-phu-master-DESKTOP-04NCH03.out
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hbase/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
: running regionserver, logging to /opt/hbase/logs/hbase-phu-regionserver-DESKTOP-04NCH03.out
: SLF4J: Class path contains multiple SLF4J bindings.
: SLF4J: Found binding in [jar:file:/opt/hbase/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
: SLF4J: Found binding in [jar:file:/opt/hbase/lib/client-facing-thirdparty/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
: SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
: SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
phu@DESKTOP-04NCH03:~$ /opt/phoenix/bin/sqlline.py localhost:2181
/usr/bin/env: 'python': No such file or directory
phu@DESKTOP-04NCH03:~$ /opt/phoenix/bin/sqlline.py localhost:2181
/usr/bin/env: 'python': No such file or directory
phu@DESKTOP-04NCH03:~$ sudo ln -s /usr/bin/python3 /usr/bin/python
phu@DESKTOP-04NCH03:~$ /opt/phoenix/bin/sqlline.py localhost:2181
Setting property: [incremental, false]
Setting property: [isolation, TRANSACTION_READ_COMMITTED]
issuing: lconnect -p driver org.apache.phoenix.jdbc.PhoenixDriver -p user "none" -p password "none" "jdbc:phoenix:localhost:2181"
Connecting to jdbc:phoenix:localhost:2181
25/10/14 01:27:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

THỰC HIỆN PHẦN CÀI ZEPPELIN

1. Cài đặt và giải nén

```
phu@DESKTOP-04NCH03: /opt/zeppelin/conf
hadoop
hadoop-3.3.6.tar.gz.1
hbase-2.4.9-bin.tar.gz
hbase_conflicts_bak
hs_err_pid8626.log
mysql-connector-j-8.0.33
mysql-connector-j-8.0.33.tar.gz
phoenix-hbase-2.4-5.1.3-bin.tar.gz
phoenix-hbase-2.4-5.1.3-bin.tar.gz.1
spark-3.5.1-bin-hadoop3.tgz
sqoop-1.4.7-bin__hadoop-2.6.0.tar.gz
tmp
zeppelin-0.11.0-bin-all
zeppelin-0.11.0-bin-all.tgz
zookeeper_data
phu@DESKTOP-04NCH03:~$ sudo mv zeppelin-0.11.0-bin-all /opt/zeppelin
[sudo] password for phu:
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$
phu@DESKTOP-04NCH03:~$ nano /opt/zeppelin/conf/zeppelin-env.sh
phu@DESKTOP-04NCH03:~$ cd /opt/zeppelin/conf
phu@DESKTOP-04NCH03:/opt/zeppelin/conf$ ls
configuration.xsl  log4j_yarn_cluster.properties
interpreter-list   shiro.ini.template
log4j.properties  zeppelin-env.cmd.template
log4j.properties2 zeppelin-env.sh.template
log4j2.properties zeppelin-site.xml.template
phu@DESKTOP-04NCH03:/opt/zeppelin/conf$ cp zeppelin-env.sh.template zeppelin-env.sh
phu@DESKTOP-04NCH03:/opt/zeppelin/conf$
```

2. Chỉnh sửa lại file zeppelin-env

```
zeppelin-env - Notepad
File Edit Format View Help
#!/bin/bash
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You 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.
#
# Đường dẫn bắt buộc cho Zeppelin
export HADOOP_HOME="/opt/hadoop"
export HBASE_HOME="/opt/hbase"

# Thêm các JAR của HBase vào Classpath của Zeppelin (Giúp tải Driver Phoenix)
# LƯU Ý: Đảm bảo /opt/hbase/lib/* tồn tại nếu bạn không dùng HBASE_HOME
export ZEPPELIN_CLASSPATH="$ZEPPELIN_CLASSPATH:/opt/hbase/conf:/opt/hbase/lib/*"

# export JAVA_HOME=
```

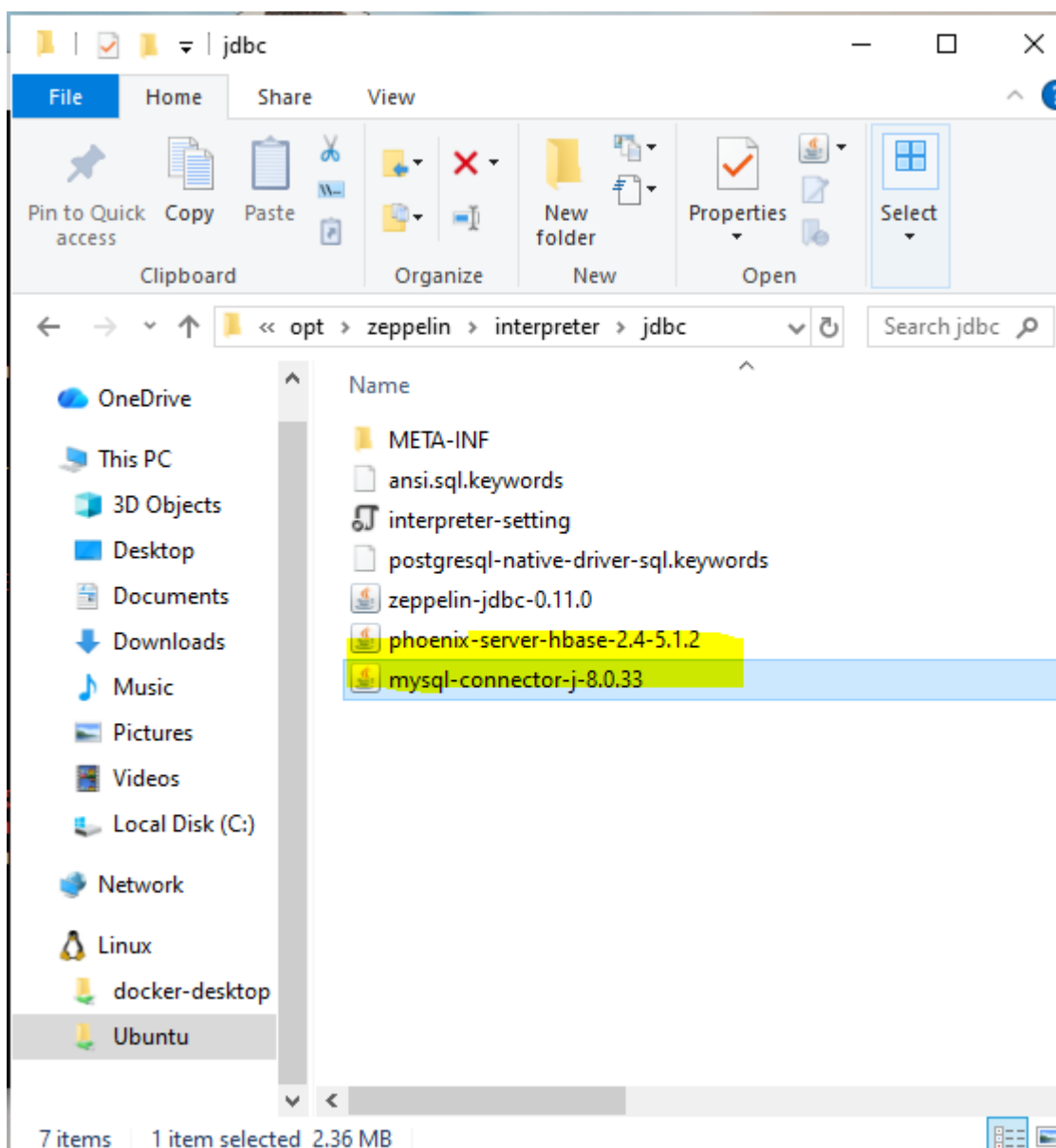
3. Thêm các cấu hình bashrc

```
phu@DESKTOP-O4NCH03: ~  
GNU nano 7.2 /opt/zeppelin/conf/zeppelin-env.sh *  
#!/bin/bash  
#  
# Licensed to the Apache Software Foundation (ASF) under one or more  
# contributor license agreements. See the NOTICE file distributed with  
# this work for additional information regarding copyright ownership.  
# The ASF licenses this file to You 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.  
#  
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64  
export HADOOP_HOME=/home/phu/hadoop  
export HADOOP_CONF_DIR=/home/phu/hadoop/etc/hadoop  
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin  
# Cho phép truy cập từ mọi IP (không chỉ localhost)  
export ZEPPELIN_ADDR=0.0.0.0  
export ZEPPELIN_PORT=8080  
# XÓA HOÀN TOÀN DÒNG ZEPPELIN_CLASSPATH THỦ CÔNG:  
# (Để Zeppelin tự xử lý phần này)  
^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
```

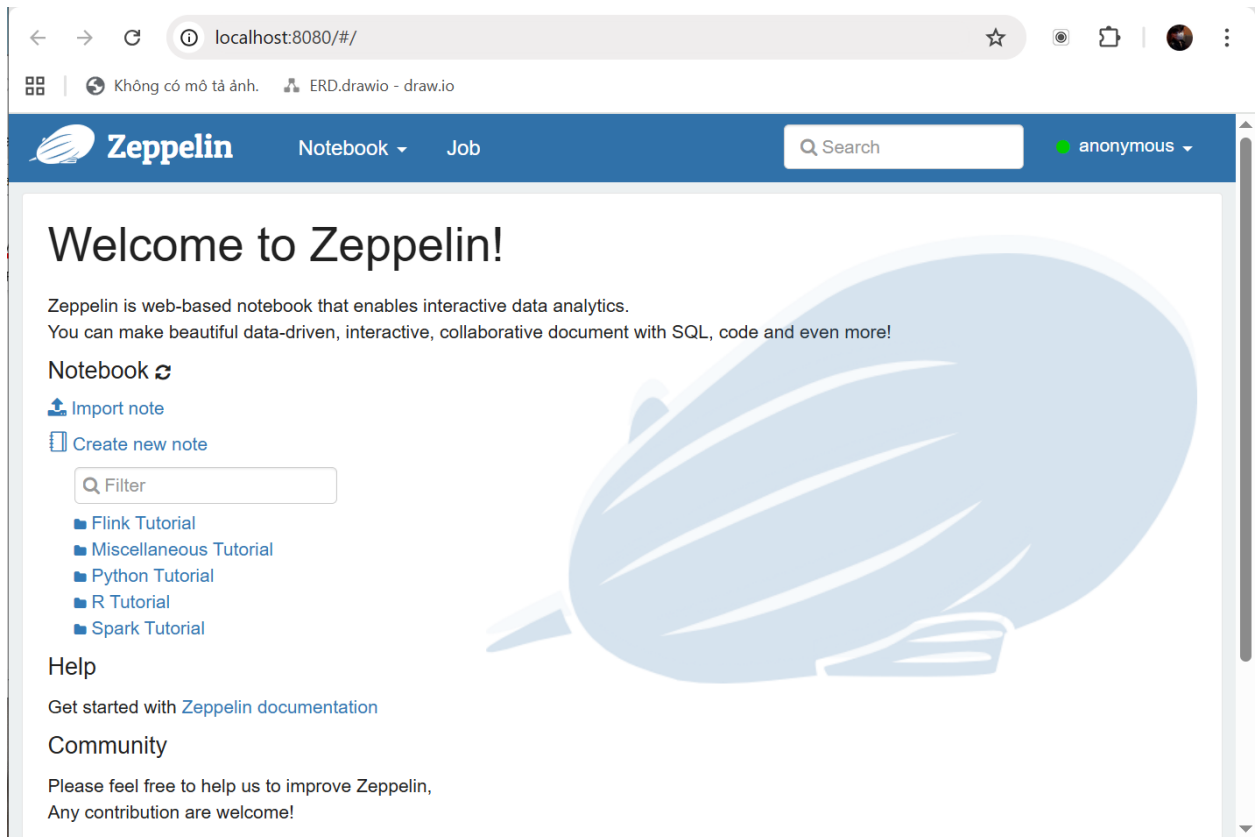
4. Khởi động thử các dịch vụ


```
phu@DESKTOP-04NCH03: ~  
cLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/home/phu/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
Zeppelin is restarting  
ZEPPELIN_CLASSPATH: /opt/zeppelin/lib/*:/opt/zeppelin/*:/opt/hbase/conf:/opt/hbase/lib/*:/opt/zeppelin/conf:/home/phu/hadoop/etc/hadoop:/home/phu/hadoop/etc/hadoop  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/opt/zeppelin/lib/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/opt/hbase/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
phu@DESKTOP-04NCH03:~$ sudo nano /opt/zeppelin/conf/zeppelin-env.sh  
phu@DESKTOP-04NCH03:~$ sudo /opt/zeppelin/bin/zeppelin-daemon.sh start [ OK ]  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$ sudo /opt/zeppelin/bin/zeppelin-daemon.sh start  
phu@DESKTOP-04NCH03:~$ sudo /opt/zeppelin/bin/zeppelin-daemon.sh start  
Zeppelin is already running  
phu@DESKTOP-04NCH03:~$ sudo lsof -i :8080  
COMMAND PID USER  FD  TYPE DEVICE SIZE/OFF NODE NAME  
java    7154 root  621u  IPv6  53586      0t0  TCP *:http-alt (LISTEN)  
phu@DESKTOP-04NCH03:~$
```

5. Thêm các JAR của JDBC vào



Tiến hành kiểm tra trên webUI



Sau khi cài xong tất cả tôi sẽ có các dịch vụ chạy webUI ở các port sau :

HDFS: 9870

YARN:8088

HBASE:16010

ZEPPELIN : 8080

HIVE: 10002

DRILL: 8047

Các lệnh để khởi chạy

1. HDFS và YARN : start-all.sh
 2. HBASE VÀ Phoenix: /opt/hbase/bin/start-hbase.sh
 3. DRILL : cd /opt/drill và ./bin/drill-embedded
 4. HIVE : cd /opt/hive và ./bin/hive --service metastore & và ./bin/hive --service hiveserver2 &
- Truy cập hive : ./bin/hive
- 5: PIG : cd /opt/pig và ./bin/pig ten_script_cua_ban.pig và mở sell : ./bin/pig

THỰC HIỆN PHẦN VỀ CÀO DỮ LIỆU

THỰC HIỆN VIỆC ĐƯA DỮ LIỆU LÊN HDFS

1. Tạo 1 thư mục gốc trên hdfs

1.1 Tạo thư mục cha là raw_data

```
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir -p /raw_data
phu@DESKTOP-04NCH03:~$
```

1.2 Tạo thư mục chứa các sản phẩm

```
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir /raw_data/products_csv
phu@DESKTOP-04NCH03:~$
```

1.3 Tạo thư mục chứa các khuyến mãi

```
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir /raw_data/promotions_json
phu@DESKTOP-04NCH03:~$
```

2. Đưa các file vào các thư mục đã có trên hdfs

2.1 Put dữ liệu của cellphoneS lên trước

```
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir -p /raw_data
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir /raw_data/products_csv
phu@DESKTOP-04NCH03:~$ hdfs dfs -mkdir /raw_data/promotions_json
phu@DESKTOP-04NCH03:~$ ls
CentOS-7-x86_64-DVD-2009.iso  hadoop-3.3.6.tar.gz.1
commons-lang-2.6.jar          mysql-connector-j-8.0.33
crawl                         mysql-connector-j-8.0.33.tar.gz
hadoop                       sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
phu@DESKTOP-04NCH03:~$ cd crawl
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -put cellphones_raw_data.csv /raw_data/products_csv/
phu@DESKTOP-04NCH03:~/crawl$
```

2.2 Put dữ liệu của thegioididong

```
phu@DESKTOP-04NCH03:~/crawl$ ls
cellphones_promotions_nosql.json  kich_hoat_venv.txt
cellphones_raw_data.csv          laptops_enriched_data.csv
crawl_cellphones_promotions.py   tgdd_promotions_nosql.json
crawl_promotions.py             venv
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -put "laptops_enriched_data.csv" /raw_data/products_csv/
phu@DESKTOP-04NCH03:~/crawl$
```

3. Đưa các file JSON lên hdfs

3.1 Dữ liệu khuyến mãi của CellPhoneS

```
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -put cellphones_promotions_nosql.json /raw_data/promotions_json/
phu@DESKTOP-04NCH03:~/crawl$
```

3.2 Dữ liệu khuyến mãi của TheGioiDiDong

```
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -put tgdd_promotions_nosql.json /raw_data/promotions_json/
phu@DESKTOP-04NCH03:~/crawl$
```

4. Kiểm tra xác nhận lại

```
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -ls /raw_data/products_csv/
Found 2 items
-rw-r--r-- 1 phu supergroup 146502 2025-10-13 17:39 /raw_data/products_csv/cellphones_raw_data.csv
-rw-r--r-- 1 phu supergroup 109995 2025-10-13 17:41 /raw_data/products_csv/laptops_enriched_data.csv
phu@DESKTOP-04NCH03:~/crawl$ hdfs dfs -ls /raw_data/promotions_json/
Found 2 items
-rw-r--r-- 1 phu supergroup 619041 2025-10-13 17:42 /raw_data/promotions_json/cellphones_promotions_nosql.json
-rw-r--r-- 1 phu supergroup 477619 2025-10-13 17:42 /raw_data/promotions_json/tgdd_promotions_nosql.json
phu@DESKTOP-04NCH03:~/crawl$
```

Kiểm tra xác nhận trên web

← → ↻ localhost:9870/explorer.html#/raw_data/products_csv ☆ ⚙️ 📄 🗑️

⏏️ Không có mô tả ảnh. ERD.drawio - draw.io

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/raw_data/products_csv Go! 📁 ⬆️ 📄 🗑️

Show 25 ▾ entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	phu	supergroup	143.07 KB	Oct 13 17:39	1	128 MB	cellphones_raw_data.csv	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	phu	supergroup	107.42 KB	Oct 13 17:41	1	128 MB	laptops_enriched_data.csv	<input type="checkbox"/>

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2023.

Và 2 file JSON

⏏️ Không có mô tả ảnh. ERD.drawio - draw.io

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/raw_data/promotions_json Go! 📁 ⬆️ 📄 🗑️

Show 25 ▾ entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	phu	supergroup	604.53 KB	Oct 13 17:42	1	128 MB	cellphones_promotions_nosql.json	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	phu	supergroup	466.42 KB	Oct 13 17:42	1	128 MB	tgdd_promotions_nosql.json	<input type="checkbox"/>

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2023.

PHẦN VỀ DÙNG APACHE PIG ĐỂ TIẾN HÀNH XỬ LÝ DỮ LIỆU TRƯỚC KHI DÙNG SQOOP ĐẨY VÀO MYSQL

1. Chuẩn bị Môi trường và File

1.1 Tạo file pig_script để tạo môi trường làm việc

```
phu@DESKTOP-04NCH03: ~  
phu@DESKTOP-04NCH03:/opt/hive$ cd /opt/pig  
phu@DESKTOP-04NCH03:/opt/pig$ cd ..  
phu@DESKTOP-04NCH03:/opt$ cd  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$  
phu@DESKTOP-04NCH03:~$ mkdir -p ~/pig_scripts  
phu@DESKTOP-04NCH03:~$ ls  
CentOS-7-x86_64-DVD-2009.iso      hbase_conflicts_bak  
Downloads                        hs_err_pid8626.log  
apache-drill-1.19.0.tar.gz        mysql-connector-j-8.0.33  
apache-drill-1.19.0.tar.gz.1      mysql-connector-j-8.0.33.tar.gz  
apache-drill-1.21.2.tar.gz        phoenix-hbase-2.4-5.1.3-bin.tar.gz  
apache-zookeeper-3.6.4-bin.tar.gz phoenix-hbase-2.4-5.1.3-bin.tar.gz.1  
commons-lang-2.6.jar             pig_scripts  
crawl                             spark-3.5.1-bin-hadoop3.tgz  
drill                             sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz  
hadoop                             tmp  
hadoop-3.3.6.tar.gz.1             zeppelin-0.11.0-bin-all.tgz  
hbase-2.4.9-bin.tar.gz           zookeeper_data  
phu@DESKTOP-04NCH03:~$
```

1.2 Tạo file mysql_brand_etl.pig


```

phu@DESKTOP-O4NCH03: ~
GNU nano 7.2 /home/phu/pig_scripts/mysql_brand_etl.pig *
-- 2. Xử lý Dữ liệu TGDD (Thêm source_brand)
-- =====

TGDD_DATA = LOAD '/raw_data/products_csv/laptops_enriched_data.csv'
            USING PigStorage(',')
            AS (id:chararray, product_name:chararray, current_price:long, list_price:long, brand:chararray);
-- Định nghĩa toàn bộ schema để đảm bảo đọc đúng

-- Lọc dòng Header
TGDD_FILTERED = FILTER TGDD_DATA BY NOT (id == 'id' AND product_name == 'product_name');

-- Chuẩn hóa và thêm trường 'source_brand'
TGDD_CLEANED = FOREACH TGDD_FILTERED GENERATE
               id,
               product_name,
               current_price,
               list_price,
               'thegioididong' AS source_brand:chararray;

-- =====
-- 3. Hợp nhất Dữ liệu và Lưu Tạm ra HDFS
-- =====

-- Hợp nhất hai bộ dữ liệu (UNION)
ALL_DATA_FOR_MYSQL = UNION CELLPHONES_CLEANED, TGDD_CLEANED;

-- Lưu kết quả đã xử lý (5 cột) vào thư mục tạm thời trên HDFS
-- Thứ tự cột: id, product_name, current_price, list_price, source_brand
STORE ALL_DATA_FOR_MYSQL INTO '/processed_data/mysql_export_temp' USING PigStorage(',');

^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

```

Code của toàn bộ phần này : 'output/mysql_export_temp_fixed'

```

-- Cấu hình: Thiết lập số lượng job song song
SET default_parallel 5;

-- Thiết lập encoding là UTF-8 để xử lý ký tự Tiếng Việt
SET default_charset utf-8;

-- BẮT BUỘC: Đăng ký và định nghĩa Piggybank UDFs
REGISTER /opt/pig/lib/piggybank.jar;
-- Hàm này sẽ được dùng để tách chuỗi số hợp lệ từ chuỗi giá Cellphones
DEFINE REGEX_EXTRACT org.apache.pig.piggybank.evaluation.string.RegexExtract;

-----
-- 1. Xử lý Dữ liệu CellphoneS (LÀM SẠCH GIÁ VÀ CHUYỂN ĐỔI KIỂU DỮ LIỆU)
-----

-- LOAD từ Local File System (Đường dẫn cần tồn tại trên máy local)
CELLPHONES_DATA = LOAD
'file:///home/phu/raw_data/products_csv/cellphones_raw_data.csv'
  USING org.apache.pig.piggybank.storage.CSVExcelStorage()
  AS (id:chararray, product_name:chararray, current_price_raw:chararray,
      list_price_raw:chararray, raw_specs:chararray, url:chararray);

-- Lọc dòng Header

```

```
CELLPHONES_FILTERED = FILTER CELLPHONES_DATA BY NOT (id == 'id' AND
product_name == 'product_name');

-- LÀM SẠCH VÀ CHUYỂN ĐỔI KIỂU DỮ LIỆU
CELLPHONES_CLEANED = FOREACH CELLPHONES_FILTERED {
  -- 1. Tách chuỗi số (bao gồm dấu chấm) khỏi đơn vị tiền tệ và chữ cái (ví dụ: '22.990.000đ' ->
'22.990.000')
  -- Biểu thức chính quy: '([0-9\\.]+)' lấy ra chuỗi số và dấu chấm.
  price_cur_str = REGEX_EXTRACT(current_price_raw, '([0-9\\.]+)', 1);
  price_list_str = REGEX_EXTRACT(list_price_raw, '([0-9\\.]+)', 1);

  -- 2. Loại bỏ dấu chấm phân cách hàng nghìn (ví dụ: '22.990.000' -> '22990000')
  price_cur_clean = REPLACE(price_cur_str, '\\.', '');
  price_list_clean = REPLACE(price_list_str, '\\.', '');

  GENERATE
    (long)id AS id,
    product_name,
    -- Chuyển đổi sang double. Nếu chuỗi không hợp lệ, Pig sẽ đặt là NULL.
    (double)price_cur_clean AS current_price,
    (double)price_list_clean AS list_price,
    'cellphones' AS source_brand;
};

-----
-- 2. Xử lý Dữ liệu TGDD (LOAD từ LOCAL)
-----

-- LOAD từ Local File System (Giá đã là số hợp lệ)
TGDD_DATA = LOAD 'file:///home/phu/raw_data/products_csv/laptops_enriched_data.csv'
  USING org.apache.pig.piggybank.storage.CSVExcelStorage()
  AS (id:chararray, product_name:chararray, current_price:double, list_price:double,
brand:chararray, category:chararray, cpu:chararray, ram:chararray, storage:chararray,
screen_size:chararray, screen_resolution:chararray, os:chararray, software:chararray,
average_rating:float, product_url:chararray);

-- Lọc dòng Header
TGDD_FILTERED = FILTER TGDD_DATA BY NOT (id == 'id' AND product_name ==
'product_name');

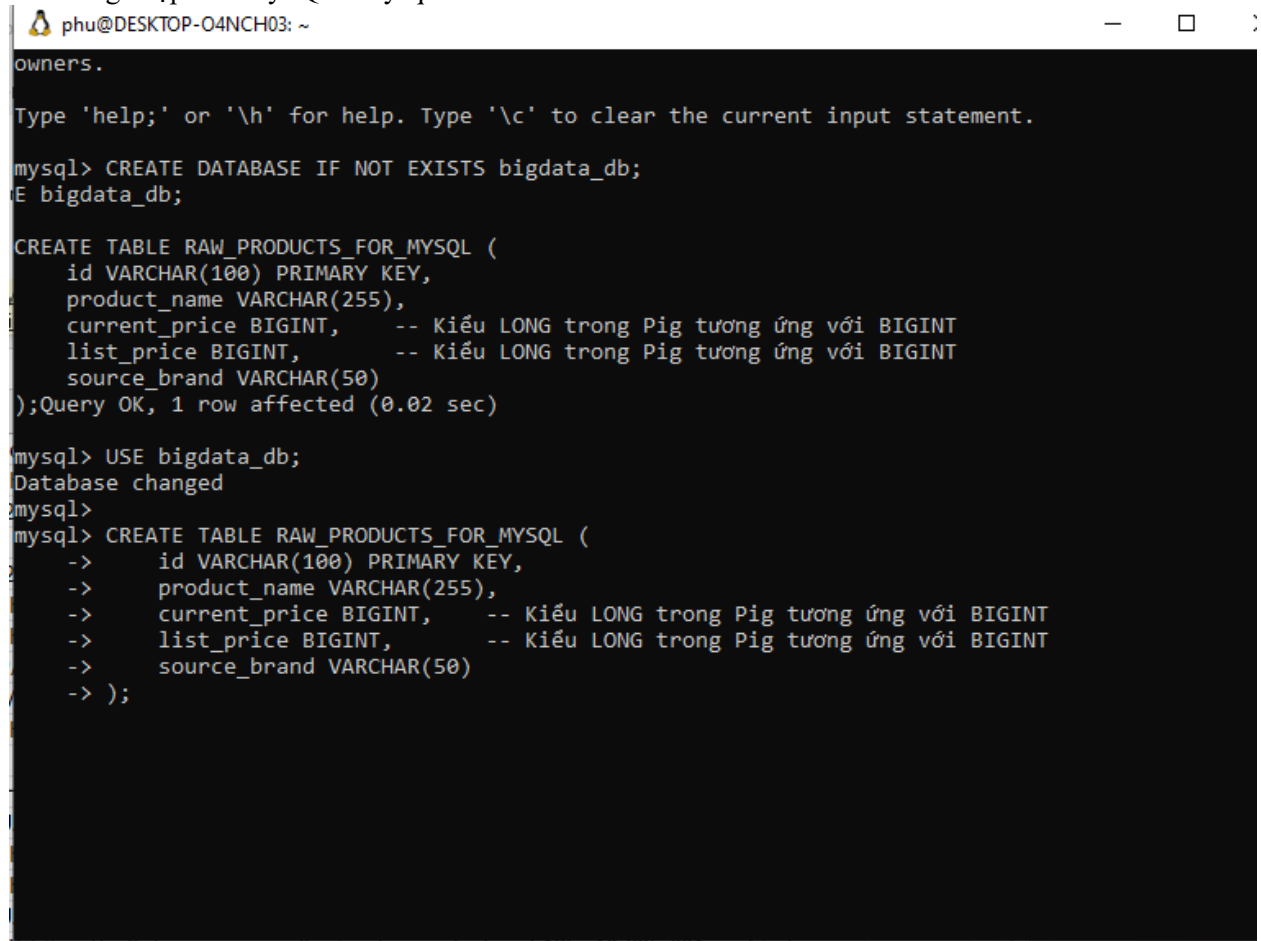
-- Chuẩn hóa: giữ nguyên các cột cần thiết (Giá đã là double)
TGDD_CLEANED = FOREACH TGDD_FILTERED GENERATE
  (long)id AS id,
  product_name,
  current_price,
  list_price,
  'thegioididong' AS source_brand;

-----
-- 3. Hợp nhất Dữ liệu và Lưu Tạm
```

```
-----  
-- Hợp nhất hai bộ dữ liệu (UNION) với Schema thống nhất: (long, chararray, double, double,  
chararray)  
ALL_DATA_FOR_MYSQL = UNION CELLPHONES_CLEANED, TGDD_CLEANED;  
  
-- Lưu dữ liệu. NullStorage('\N') sẽ xử lý các giá trị NULL thành '\N', phù hợp cho MySQL.  
STORE ALL_DATA_FOR_MYSQL  
  INTO 'output/mysql_export_temp_fixed'  
  USING PigStorage(',', 'NullStorage(\\\\"N\\")');
```

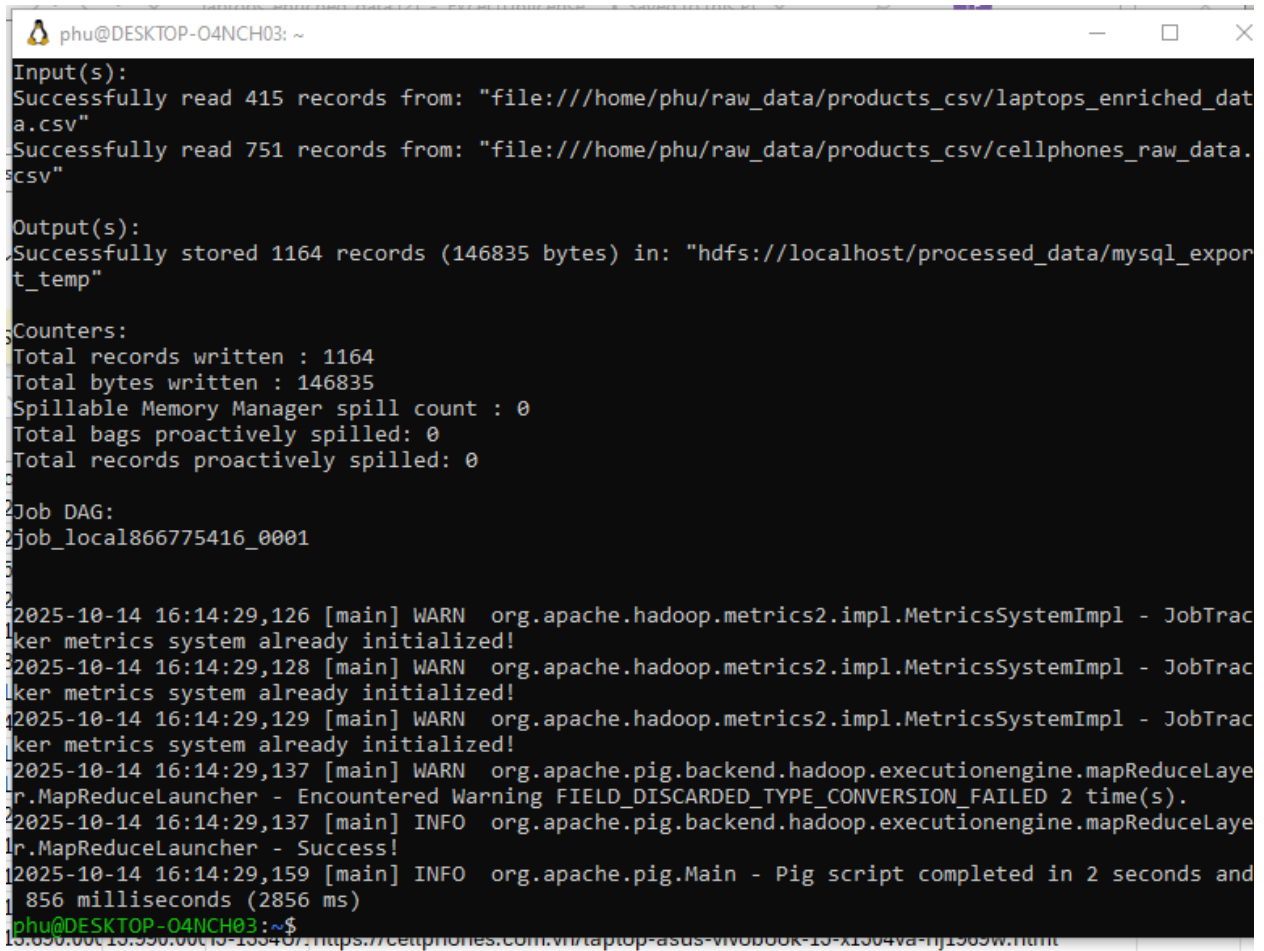
2. Chuẩn bị bảng đích trên MYSQL

2.1 Đăng nhập vào MySQL : mysql -u root



```
phu@DESKTOP-04NCH03: ~  
owners.  
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.  
mysql> CREATE DATABASE IF NOT EXISTS bigdata_db;  
CREATE bigdata_db;  
  
mysql> CREATE TABLE RAW_PRODUCTS_FOR_MYSQL (  
  id VARCHAR(100) PRIMARY KEY,  
  product_name VARCHAR(255),  
  current_price BIGINT,      -- Kiểu LONG trong Pig tương ứng với BIGINT  
  list_price BIGINT,        -- Kiểu LONG trong Pig tương ứng với BIGINT  
  source_brand VARCHAR(50)  
); Query OK, 1 row affected (0.02 sec)  
  
mysql> USE bigdata_db;  
Database changed  
mysql>  
mysql> CREATE TABLE RAW_PRODUCTS_FOR_MYSQL (  
  -> id VARCHAR(100) PRIMARY KEY,  
  -> product_name VARCHAR(255),  
  -> current_price BIGINT,      -- Kiểu LONG trong Pig tương ứng với BIGINT  
  -> list_price BIGINT,        -- Kiểu LONG trong Pig tương ứng với BIGINT  
  -> source_brand VARCHAR(50)  
  -> );
```

3. Thực thi ETL : Pig (HDFS → HDFS tạm) để làm sạch dữ liệu trước



Kiểm tra sơ qua các file

```
phuong@DESKTOP-04NCH03: ~$ hdfs dfs -cat /processed_data/mysql_export_temp/part-m-00000 | head -n 5
1,Laptop Gigabyte G6 MF-H2PH853KH,,,cellphones
2,Laptop Lenovo IdeaPad Slim 5 14IRH10R 83J0006CVN,,,cellphones
3,MacBook Pro 14 M4 Pro 14CPU 20GPU 24GB 512GB Nano | Chính hãng Apple Việt Nam,,,cellphones
4,Laptop ASUS Gaming V16 K3607VJ-IR106W,,,cellphones
5,Laptop Lenovo IdeaPad Slim 3 15IRH10 83K1000GVN,,,cellphones
cat: Unable to write to output stream.
phuong@DESKTOP-04NCH03:~$
```

Tiến hành sqoop dữ liệu lên mysql