1. **Hướng dẫn cài đặt Hadoop**

Trong hướng dẫn này, mình sẽ mô phỏng cài Hadoop (single node) trên 1 máy ảo Ubuntu 22.04 sử dụng VMware.

1. **Cài đặt máy ảo**

**Bước 1:** Cài đặt VMware

**Bước 2:** Tải image Ubuntu 22.04 tại đây: [Link](https://www.linuxvmimages.com/images/ubuntu-2204/).

**Bước 3:** Load image vào VMware

* Giải nén Image Ubuntu 22.04 vừa tải về.
* Mở file Ubuntu\_22.04\_VM\_LinuxVMImages.COM.vmx bằng VMware.

**Bước 4:** Đăng nhập vào Ubuntu: Mật khẩu mặt định là ubuntu.

A screenshot of a computer

Description automatically generated

1. Cài đặt Hadoop

Bước 1: Cài đặt Java 8 để hỗ trợ runtime cho Hadoop

**sudo apt install openjdk-8-jdk -y**

Kiểm tra lại xem java đã cài có lỗi gì không, nếu thấy hiện version là không bị lỗi.

**java -version; javac -version**

A screen shot of a computer

Description automatically generated

Bước 2: Tải Hadoop

Truy cập vào dường link sau và tải xuống phiên bản Hadoop-3.3.6 binary: [Link](https://hadoop.apache.org/releases.html).

A screenshot of a computer

Description automatically generated

Giải nén vào thư mục Home

A screenshot of a computer

Description automatically generated

Bước 3: Cài đặt open SSH

**sudo apt install openssh-server openssh-client -y**

Tạo và cài đặt SSH Certificates

**ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa**

Lưu lại thông tin key vừa tạo vào authorized\_keys

**cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys**

Cấp quyền cho người dùng ubuntu hiện tại có thể tương tác với authorized\_keys

**chmod 0600 ~/.ssh/authorized\_keys**

Tiến hành thử truy cập ssh xem có kết nối được không

**ssh localhost**

Nếu hiện như bên dưới là đã kết nối thành công ( ghi nhớ thông tin địa chỉ localhost là 127.0.0.1).  
A computer screen shot of a computer program

Description automatically generated

Bước 4: Cấu hình bashrc

Tìm đường dẫn javac vừa cài bằng lệnh

**which javac**

Sau đó dùng lệnh này để lấy toàn bộ dường dẫn

**readlink -f /usr/bin/javac**

A screenshot of a computer program

Description automatically generated

Lưu lại thông tin đường dẫn để setup bashrc.

Mở bashrc bằng lệnh

**gedit ~/.bashrc**

Cuộn xuống cuối file và bắt đầu cấu hình tại đây

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export HADOOP\_HOME=/home/ubuntu/hadoop-3.3.6

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

Lưu lại file và chạy file để cập nhật thay đổi

**source ~/.bashrc**

Bước 5: Cấu hình hadoop-env.sh

Quay lại đường dẫn home và cd vào thư mục hadoop đã giải nén

**cd ~**

**cd hadoop-3.3.6/etc/hadoop/**

Mở file hadoop-env.sh

**gedit hadoop-env.sh**

Xóa dòng thứ 37 và thay thế bằng đường dẫn java đã cài và lưu lại file.

A computer screen shot of a computer code

Description automatically generated

Bước 6: Cấu hình file core-site.xml

Tạo thư mục tmp trong đường dẫn hadoop hiện tại

**mkdir -p /home/ubuntu/hadoop-3.3.6/tmp**

Mở file core-site.xml để cấu hình

**gedit core-site.xml**

Thêm đoạn cấu hình sau vào thẻ <configuration> và lưu lại.

<property>

<name>hadoop.tmp.dir</name>

<value>/home/ubuntu/hadoop-3.3.6/tmp</value>

</property>

<property>

<name>fs.default.name</name >

<value>hdfs://localhost:9000</value>

</property>

A screenshot of a computer program

Description automatically generated

Bước 7: Cấu hình mapred-site.xml

Mở file mapred-site.xml

**gedit mapred-site.xml**

Thêm đoạn configuration sau và thẻ

<property>

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

<value>yarn</value>

</property>

A computer screen shot of a computer screen

Description automatically generated

Bước 8: Cấu hình file hdfs-site.xml

Tạo 2 thư mục namenode và datanode

mkdir -p /home/ubuntu/hadoop-3.3.6/data/namenode

mkdir -p /home/ubuntu/hadoop-3.3.6/data/datanode

Mở file hdfs-site.xml và thêm các dòng lệnh vào thẻ configuration

**gedit hdfs-site.xml**

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>/home/ubuntu/hadoop-3.3.6/data/namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>/home/ ubuntu/hadoop-3.3.6/data/datanode</value>

</property>

A screenshot of a computer program

Description automatically generated

Bước 9: Cấu hình file yarn-site.xml

Mở file yarn-site.xml và thêm đoạn cấu hình sau vào thẻ configuration.

**gedit yarn-site.xml**

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

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

<value>127.0.0.1</value>

</property>

<property>

<name>yarn.acl.enable</name>

<value>0</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PERPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

A computer screen shot of a computer code

Description automatically generated

Bước 10: Chạy Hadoop

Tiến hành format lại các file hệ thống vừa cấu hình bằng lệnh

**hdfs namenode -format**

Khởi chạy namenode, datanode, secondary namenode bằng lệnh

**start-dfs.sh**

A screenshot of a computer screen

Description automatically generated

Khởi chạy yarn

**start-yarn.sh**

**A screenshot of a computer program

Description automatically generated**

Kiểm tra các chương trình đang chạy, nếu thấy 6 chương trình đang chạy như bên dưới là đã thành công cài đặt Hadoop và có thể sử dụng.

**jps**

**A screenshot of a computer

Description automatically generated**

Kiểm tra web UI

Namenode : <http://localhost:9870>

A screenshot of a computer

Description automatically generated

Yarn resource manager: <http://localhost:8088>

A screenshot of a computer

Description automatically generated