

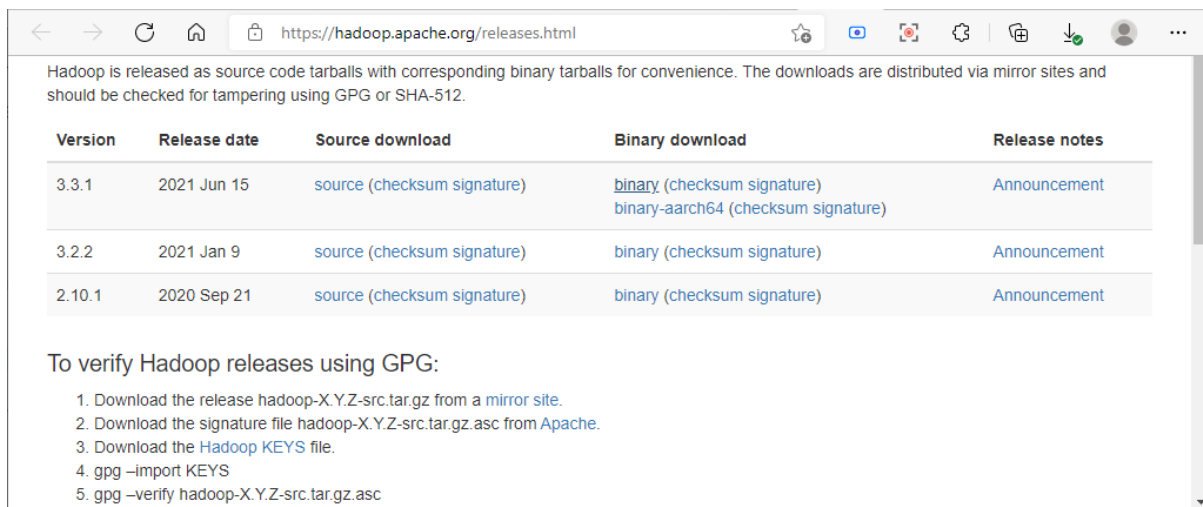
Bài 21: Cài đặt Hadoop 3.3

Bài này sẽ giúp bạn cài đặt Hadoop 3.3 trong Ubuntu để làm quen và trải nghiệm. Bạn cũng có thể áp dụng các lệnh tương tự cho môi trường Linux, MacOS hoặc Ubuntu trong Windows.

Có vài tình huống bạn muốn sử dụng Hadoop 2 thì tham khảo bài viết trong phần Phụ lục.

Tải phần mềm

Bạn vào trang web <https://hadoop.apache.org/releases.html> để xem các phiên bản hiện tại của Hadoop.



The screenshot shows the Hadoop releases page. It includes a table with columns: Version, Release date, Source download, Binary download, and Release notes. The table lists versions 3.3.1, 3.2.2, and 2.10.1. Below the table, there is a section titled 'To verify Hadoop releases using GPG:' with a list of 5 steps.

Version	Release date	Source download	Binary download	Release notes
3.3.1	2021 Jun 15	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
3.2.2	2021 Jan 9	source (checksum signature)	binary (checksum signature)	Announcement
2.10.1	2020 Sep 21	source (checksum signature)	binary (checksum signature)	Announcement

To verify Hadoop releases using GPG:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the signature file `hadoop-X.Y.Z-src.tar.gz.asc` from [Apache](#).
3. Download the [Hadoop KEYS](#) file.
4. `gpg --import KEYS`
5. `gpg --verify hadoop-X.Y.Z-src.tar.gz.asc`

Phần này sẽ giúp bạn cài nhanh Hadoop phiên bản 3.3 lên máy ảo Ubuntu.

Tạo thư mục soft trong thư mục home của user (dùng kí hiệu dấu ngã ~)

```
sudo mkdir ~/soft
```

Chuyển thư mục hiện hành vào thư mục soft mới tạo

```
cd ~/soft
```

Tải gói phần mềm hadoop phiên bản 3.3.1 về thư mục hiện hành bằng lệnh `wget` <url>:

```
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
```

Giải nén

Giải nén ra thư mục /opt

```
sudo tar -xvzf ./hadoop-3.3.1.tar.gz -C /opt
```

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

Tạo ánh xạ thư mục `/opt/hadoop-3.3.1` vào `/opt/hadoop`. Bước này giống như tạo shortcut trên Windows, thay vì truy cập vào đường dẫn dài `/opt/hadoop-3.3.1` thì tôi tạo một đường dẫn ngắn hơn gọi là alias hoặc shortcut `/opt/hadoop`. Ngoài ra khi cần thử nghiệm các phiên bản hadoop khác nhau thì chỉ cần ánh xạ lại khi cần. Sử dụng lệnh `ln` (`ln`)

```
sudo ln -nsf /opt/hadoop-3.3.1 /opt/hadoop
```

Kiểm tra lại nội dung thư mục bằng lệnh `ls` (`ls`):

```
ls /opt/hadoop
```

```
LICENSE-binary NOTICE-binary README.txt etc lib licenses-binary  
share LICENSE.txt NOTICE.txt bin include libexec sbin
```

Cấu hình các biến môi trường cho Hadoop

Sửa file `environment` bằng lệnh:

```
sudo nano /etc/environment
```

Thêm nội dung được bôi đậm:

```
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/  
bin:/usr/games:/usr/local/games:/opt/hadoop/bin"  
JAVA_HOME=/opt/jdk  
HADOOP_HOME=/opt/hadoop  
HADOOP_MAPRED_HOME=/opt/hadoop  
HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop/  
HDFS_NAMENODE_USER="root"  
HDFS_DATANODE_USER="root"  
HDFS_SECONDARYNAMENODE_USER="root"  
YARN_RESOURCEMANAGER_USER="root"  
YARN_NODEMANAGER_USER="root"  
# Hai biến bên dưới để dùng cho rhdfs trong R  
HADOOP_COMMON_LIB_NATIVE_DIR=/opt/hadoop/lib/native  
HADOOP_CMD=/opt/hadoop/bin/hadoop
```

Làm cho các thiết lập biến môi trường ở trên có tác dụng ngay luôn bằng lệnh:

```
source /etc/environment
```

Kiểm tra bằng cách xem giá trị của biến môi trường `HADOOP_HOME` bằng lệnh:

```
echo $HADOOP_HOME
```

Kết quả:

```
/opt/hadoop
```

Sửa file cấu hình của Hadoop

Sửa file `core-site.xml` bằng lệnh:

```
sudo nano /opt/hadoop/etc/hadoop/core-site.xml
```

Dán nội dung sau để thay thế cho nội dung 2 dòng `<configuration></configuration>` hiện tại:

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://192.168.146.128:9000</value>
  </property>
</configuration>
```

Cách làm như sau:

Bước 1: Copy đoạn cấu hình ở trên bằng phím `Ctrl + C`

Bước 2: Chạy lệnh `sudo nano...` ở trên trong máy ảo Ubuntu, bạn di chuyển con trỏ đến 2 dòng có thẻ `<configuration>` và `</configuration>` nhấn `Ctrl + K` để xóa.

Sau đó nhấn `Ctrl + Shift + V` để dán nội dung cấu hình vào file `core-site.xml`

Bước 3: Nhấn `Ctrl + O` để lưu

Bước 4: Nhấn `Ctrl + X` để thoát trình soạn thảo `nano`.

Thực hiện thay đổi file `hdfs-site.xml` với lệnh:

```
sudo nano /opt/hadoop/etc/hadoop/hdfs-site.xml
```

Với nội dung:

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

```
</property>
<property>
  <name>dfs.webhdfs.enabled</name>
  <value>true</value>
</property>
</configuration>
```

Tiếp tục thực hiện sửa file `mapred-site.xml` với lệnh:

```
sudo nano /opt/hadoop/etc/hadoop/mapred-site.xml
```

Với nội dung:

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

    <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

Tiếp tục sửa file `yarn-site.xml` bằng lệnh:

```
sudo nano /opt/hadoop/etc/hadoop/yarn-site.xml
```

Với nội dung:

```
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>

    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CON
```

```
F_DIR, CLASSPATH_PREPEND_DISTCACHE, HADOOP_YARN_HOME, HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

Thiết lập khóa cho lệnh ssh

Cần chuyển tài khoản sang root để thực hiện phần này bằng lệnh sau:

```
su -l
```

Tiếp theo thực hiện lệnh ssh để kết nối từ xa qua SSH:

```
ssh localhost
```

Nếu bạn chạy Ubuntu trong Windows thì có thể dịch vụ sshd chưa được cài. Thực hiện 2 lệnh sau rồi quay lại lệnh “ssh localhost” ở trên:

```
sudo apt install openssh-server
sudo service ssh start
```

Tiếp theo thực hiện 3 lệnh sau.

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```

Thực hiện lệnh `exit` hai lần để thoát ssh và trở về user bình thường.

Chuẩn bị dữ liệu cho Hadoop

```
sudo /opt/hadoop/bin/hdfs namenode -format
```

```
2021-10-11 19:47:49,659 INFO util.GSet: Computing capacity for map cache blocks
2021-10-11 19:47:49,659 INFO util.GSet: VM type = 64-bit
2021-10-11 19:47:49,660 INFO util.GSet: 0.25% max memory 15.6 GB = 40.1 MB
2021-10-11 19:47:49,660 INFO util.GSet: capacity = 2^22 = 4194304 entries
2021-10-11 19:47:49,667 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-10-11 19:47:49,667 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-10-11 19:47:49,668 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2021-10-11 19:47:49,670 INFO namenode.FSNamesystem: Retry cache on name node is enabled
2021-10-11 19:47:49,670 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2021-10-11 19:47:49,671 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-10-11 19:47:49,671 INFO util.GSet: VM type = 64-bit
```

```
2021-10-11 19:47:49,671 INFO util.GSet: 0.029999999329447746% max memory 15.6 GB = 4.8 MB
2021-10-11 19:47:49,671 INFO util.GSet: capacity = 2^19 = 524288 entries
2021-10-11 19:47:49,688 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1651829098-127.0.1.1-1633956469682
2021-10-11 19:47:49,704 INFO common.Storage: Storage directory /tmp/hadoop-root/dfs/name has been successfully formatted.
2021-10-11 19:47:49,746 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-root/dfs/name/current/fsimage.ckpt_00000000000000000000 using no compression
2021-10-11 19:47:49,831 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-root/dfs/name/current/fsimage.ckpt_00000000000000000000 of size 399 bytes saved in 0 seconds .
2021-10-11 19:47:49,851 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-10-11 19:47:49,896 INFO namenode.FSNamesystem: Stopping services started for active state
2021-10-11 19:47:49,896 INFO namenode.FSNamesystem: Stopping services started for standby state
2021-10-11 19:47:49,901 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-10-11 19:47:49,902 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at LAPTOP-I6R2E4C4/127.0.1.1
*****/
```

Khởi động hadoop

```
sudo /opt/hadoop/sbin/start-all.sh
```

Xem qua kết quả log của hadoop

Xem thư mục log bằng lệnh list:

```
ls /opt/hadoop/logs
```

Kết quả:

```
hadoop-root-datanode-ubuntu.log  userlogs
hadoop-root-datanode-ubuntu.out  yarn-root-nodemanager-ubuntu.log
hadoop-root-namenode-ubuntu.log  yarn-root-nodemanager-ubuntu.out
hadoop-root-namenode-ubuntu.out  yarn-root-resourcemanager-ubuntu.log
```

Thử xem file log “hadoop-root-datanode-ubuntu.log”:

```
sudo nano /opt/hadoop/logs/hadoop-root-datanode-ubuntu.log
```

Chữ “ubuntu” được bôi đậm là tên máy của bạn, hãy thay thế vào, hoặc lúc gõ lệnh dùng phím tab để hiển thị ra tên file cho đúng.

Không cần phải hiểu hết file này chứa cái gì, bạn chỉ cần đọc lướt qua để cảm nhận hadoop khi khởi động lên, nó ghi chú lại kết quả cho chúng ta biết nó khởi động như thế nào. Quá trình ghi chú của các phần mềm thì người ta gọi là **logging**, file được ghi chú ra gọi là **file log**.

Bạn sẽ thấy có dòng log có port 9864:

```
INFO org.apache.hadoop.hdfs.server.datanode.web.DatanodeHttpServer: Listening HTTP traffic on /0.0.0.0:9864
```

Bấm Ctrl + X để thoát lệnh nano.

Xem tiếp file log “hadoop-root-namenode-ubuntu.log”:

```
nano /opt/hadoop/logs/hadoop-root-namenode-ubuntu.log
```

Bạn sẽ thấy có dòng log có port 9000 như sau:

```
INFO org.apache.hadoop.ipc.Server: IPC Server Responder: starting
INFO org.apache.hadoop.ipc.Server: IPC Server listener on 9000: starting
INFO org.apache.hadoop.hdfs.server.namenode.NameNode: NameNode RPC up at: 192.168.146.128/192.168.146.128:9000
INFO org.apache.hadoop.hdfs.server.namenode.FSNamesystem: Starting services required for active state
INFO org.apache.hadoop.hdfs.server.namenode.FSDirectory: Initializing quota with 4 thread(s)
INFO org.apache.hadoop.hdfs.server.namenode.FSDirectory: Quota initialization completed in 58 milliseconds
```

Thử tạo thư mục, file trên Hadoop

Thử tạo thư mục mydata bên trong thư mục “/” của Hadoop. Sau đó copy file từ máy bạn (ví dụ file “myfile.txt” trong thư mục hiện tại) vào thư mục đã tạo:

```
hdfs dfs -mkdir /mydata
hdfs dfs -copyFromLocal ./myfile.txt /mydata/
```

Nếu bạn tò mò thì đặt câu hỏi là thư mục “mydata” và file “myfile.txt” trên Hadoop được lưu ở đâu?

Nếu bạn theo dõi file .log của Hadoop ở trên thì có thể khám phá thư mục mà Hadoop lưu trữ dữ liệu (theo cách cài mặc định như tài liệu ở trên) là:

```
/tmp/hadoop-root/dfs/data
```

Nếu bạn xem cấu trúc thư mục data này (dùng lệnh tree) thì kết quả như sau:

```
root@LAPTOP-I6R2E4C4:/tmp/hadoop-root/dfs/data# tree
.
├── current
│   ├── BP-867568824-127.0.1.1-1633956850322
│   │   ├── current
│   │   │   ├── VERSION
│   │   │   ├── finalized
│   │   │   │   └── subdir0
│   │   │   │       └── subdir0
│   │   │   │           ├── blk_1073741825
│   │   │   │           ├── blk_1073741825_1001.meta
│   │   │   │           ├── blk_1073741826
│   │   │   │           ├── blk_1073741826_1002.meta
│   │   │   │           ├── blk_1073741827
│   │   │   │           ├── blk_1073741827_1003.meta
│   │   │   │           ├── blk_1073741828
│   │   │   │           ├── blk_1073741828_1004.meta
│   │   │   │           ├── blk_1073741829
│   │   │   │           └── blk_1073741829_1005.meta
│   │   ├── rbw
│   │   ├── scanner.cursor
│   │   ├── tmp
│   │   └── VERSION
│   └── in_use.lock
```

Như vậy có thể suy đoán là Hadoop sẽ không lưu dữ liệu theo cách thông thường của hệ điều hành (có thể nhìn thấy tên thư mục và file bằng các lệnh dir, ls, hoặc các

phần mềm duyệt thư mục). Hadoop có thuật toán riêng để băm nhỏ dữ liệu (thư mục, file) để có thể lưu trữ lên nhiều máy (nodes) trên mạng. Vì vậy bạn không thể tìm thư mục, file bằng cách lệnh thông thường của hệ điều hành mà phải dùng các lệnh của Hadoop như:

```
hdfs dfs -ls /  
hdfs dfs -ls /mydata
```

Mở tường lửa để truy cập Hadoop từ xa

Câu hỏi đặt ra là bạn có thể truy cập vào Hadoop đã cài ở trên từ cái máy thật Windows được không? Câu trả lời là được nếu Ubuntu cho phép.

Chúng ta cho phép bằng cách mở port bằng các lệnh sau:

```
sudo firewall-cmd --permanent --add-port=9864/tcp  
sudo firewall-cmd --permanent --add-port=9000/tcp  
sudo firewall-cmd --permanent --add-port=50075/tcp  
sudo firewall-cmd --permanent --add-port=8088/tcp  
sudo firewall-cmd --reload
```

Kiểm tra lại các port đã mở trên máy

Cài đặt nmap

```
sudo apt install nmap
```

Quét port

```
sudo nmap -sT -O localhost
```

Truy cập Hadoop từ trình duyệt

Từ máy tính chạy Windows, bạn mở trình duyệt truy cập vào địa chỉ <http://192.168.146.128:9864/>

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

The screenshot shows the Hadoop DataNode web interface in a browser. The address bar indicates the URL is `192.168.204.128:9864/datanode.html`. The page has a green header with "Hadoop" and navigation links "Overview" and "Utilities". The main content area is titled "DataNode on LAPTOP-I6R2E4C4.localdomain:9866". Below this, there is a table with two rows: "Cluster ID" and "Version".

Cluster ID:	CID-ebbfa547-911f-4896-ab71-22ed5875808e
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2

Below the table is a section titled "Block Pools".

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
192.168.204.128:9000	BP-867568824-127.0.1.1-1633956850322	RUNNING	2s	4 minutes	0 B (128 MB)

Below the table is a section titled "Volume Information".

Truy cập <http://192.168.146.128:8088/>

The screenshot shows the Hadoop All Applications web interface in a browser. The address bar indicates the URL is `192.168.204.128:8088/cluster`. The page has a header with the Hadoop logo and the title "All Applications". On the left, there is a sidebar with a "Cluster" section containing links: "About", "Nodes", "Node Labels", "Applications", "NEW", "NEW SAVING", "SUBMITTED", "ACCEPTED", "RUNNING", "FINISHED", "FAILED", "KILLED", "Scheduler", and "Tools". The main content area displays "Cluster Metrics" and "Cluster Nodes Metrics".

Apps Submitted	Apps Pending	Apps Running	Apps Completed	
0	0	0	0	0

Below this is the "Cluster Nodes Metrics" section.

Active Nodes	Decommissioning Nodes
1	0

Below this is the "Scheduler Metrics" section.

Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[memory-mb (unit=Mi), vcores]

Below this is a table showing application entries. The table has columns: ID, User, Name, Application Type, Application Tags, Queue, Application Priority, StartTime, and Lau. The table is currently empty, showing "Showing 0 to 0 of 0 entries".

Như vậy đến đây, trong tay các bạn đã có một hệ thống Big Data với phần mềm Hadoop chạy trên máy ảo Ubuntu. Gọi là Big Data System nhưng chưa có data gì hết và chưa biết nếu dùng Python thì phân tích dữ liệu trên Hadoop này như thế nào?

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

Có nhiều câu hỏi cần phải trả lời. Nhưng thôi, hãy dừng lại và ăn mừng thành quả mà chúng ta đã học và làm được cái đã!

Khởi động lại Hadoop

Khi tắt và bật lại máy ảo Ubuntu thì bạn cần chạy lại Hadoop bằng các lệnh sau:

```
su -l
cd /opt/hadoop/sbin
rm -frd ../logs/*
./start-all.sh
tail -f ../logs/hadoop-root-datanode-ubuntu.log
```

Chú động dừng Hadoop

Tài liệu ở trên đã hướng dẫn bạn cài đặt, khởi động và trải nghiệm nhanh Hadoop. Khi cần dừng chạy Hadoop thì bạn thực hiện lệnh sau:

```
sudo /opt/hadoop/sbin/stop-all.sh
```

Trải nghiệm thêm với Hadoop

Cấu hình lại thư mục lưu trữ dữ liệu của Hadoop

Bạn có tham khảo tài liệu bên dưới để tìm hiểu thêm các tham số khác:

```
https://hadoop.apache.org/docs/r3.3.1/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml
```

Ví dụ khám phá tham số “dfs.datanode.data.dir” thử thay đổi thư mục chứa dữ liệu mặc định của Hadoop [file://\\$HADOOP_TMP_DIR/dfs/data](file://$HADOOP_TMP_DIR/dfs/data) sang thư mục riêng của bạn. Ví dụ tôi chạy Hadoop trong Ubuntu của Windows (gọi tắt là WSL2) thì tôi dùng đường dẫn trên Windows F:\Hadoop để chứa dữ liệu.

Thực hiện thay đổi file hdfs-site.xml với lệnh:

```
sudo nano /opt/hadoop/etc/hadoop/hdfs-site.xml
```

Bổ sung phần bôi vàng:

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

```
<name>dfs.webhdfs.enabled</name>
<value>>true</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>/mnt/f/Hadoop</value>
</property>
</configuration>
```

Chú ý các lệnh sau sẽ xóa dữ liệu đang có trên Hadoop của bạn. Hãy cẩn thận!

Các lệnh sau sẽ định dạng lại thư mục chứa dữ liệu, khởi động lại Hadoop:

```
sudo /opt/hadoop/bin/hdfs namenode -format
sudo /opt/hadoop/sbin/stop-all.sh
su -l
cd /opt/hadoop/sbin
rm -frd ../logs/*
./start-all.sh
```

Trải nghiệm lại các lệnh tạo thư mục, copy file trên máy bạn vào Hadoop và theo dõi thư mục data của bạn:

```
hdfs dfs -mkdir /mydata
hdfs dfs -copyFromLocal ./myfile.txt /mydata/
```