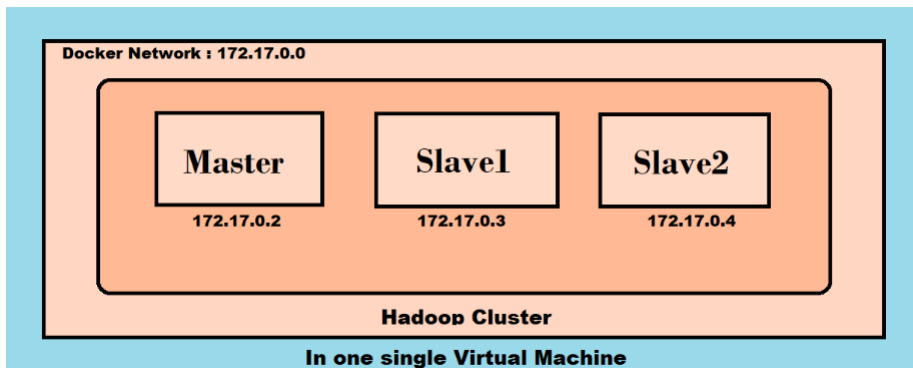


Hướng dẫn cài đặt Hadoop cluster

I. Mô tả:

- Chúng ta sẽ cài đặt Hadoop cluster gồm 1 master node và 2 slave node. Sử dụng Docker để mô phỏng lại 3 máy tính trong cụm chạy HDH ubuntu 20.04 và chúng kết nối với nhau bằng network hadoop.



- Địa chỉ IP của mỗi máy:
 - o Master 172.18.0.2
 - o Slave1 172.18.0.3
 - o Slave2 172.18.0.4
- Chạy thử chức năng HDFS và Mapreduce của hệ thống Hadoop.

II. Cài đặt:

Tạo mạng hadoop có tên là `hadoop_network` để các container có thể kết nối với nhau

```
PS C:\Users\phuoc> docker network create hadoop_network
```

Chạy một docker image ubuntu:20.04 và container được tạo ra sẽ là máy master

```
PS C:\Users\phuoc> docker run -it --name master -p 9870:9870 -p 8088:8088 -p 19888:19888 --hostname master --network hadoop_network ubuntu:20.04
```

Cập nhật và cài đặt các gói phần mềm

```
root@master:/# apt update
```

```
root@master:/# apt install -y wget tar ssh default-jdk
```

Tạo nhóm người dùng trên linux để tăng tính bảo mật khi chạy mỗi dịch vụ khác nhau của hadoop

```
root@master:/# groupadd hadoop
root@master:/# useradd -g hadoop -m -s /bin/bash hdfs
root@master:/# useradd -g hadoop -m -s /bin/bash yarn
root@master:/# useradd -g hadoop -m -s /bin/bash mapred
```

Tạo ssh-key cho mỗi user mới được tạo ra

```
root@master:/# su hdfs
hdfs@master:/$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hdfs/.ssh'.
Your identification has been saved in /home/hdfs/.ssh/id_rsa
Your public key has been saved in /home/hdfs/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:ZVZZlt1NSc0dm6lDoSiiKS0js5wa6tnTr4l/nCkUHII hdfs@master
The key's randomart image is:
+---[RSA 3072]-----+
| .          .++B0|
| E . .      ..0.0.@|
|  o... .+. . + |
|  oo. .+ . . |
|oo o  . S   o  |
|o.o  .      . |
|+   o . o      |
|o++..o.=       |
|*= oo+=.       |
+----[SHA256]-----+
hdfs@master:/$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hdfs@master:/$ chmod 0600 ~/.ssh/authorized_keys
hdfs@master:/$ exit
```

```

root@master:/# su yarn
yarn@master:/$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/yarn/.ssh'.
Your identification has been saved in /home/yarn/.ssh/id_rsa
Your public key has been saved in /home/yarn/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:dkgDktHmKKvfwUQjuwtHub0Bm63DUEGZb83cPe4vp+g yarn@master
The key's randomart image is:
+---[RSA 3072]-----+
| ..00+. |
| + ..0. |
| .00B .0. |
| 0=+. =..00 |
| .++. S... |
| .0.= . .. |
| =.* 0 . |
| oB * . .0 . |
| +=. . .E .=. |
+---[SHA256]-----+
yarn@master:/$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
yarn@master:/$ chmod 0600 ~/.ssh/authorized_keys
yarn@master:/$ exit

```

```

root@master:/# su mapred
mapred@master:/$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/mapred/.ssh'.
Your identification has been saved in /home/mapred/.ssh/id_rsa
Your public key has been saved in /home/mapred/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:rj0FWL3S6aUQ/s4PDAtqt36uevGkrt0S44/ksL+iqc mapred@master
The key's randomart image is:
+---[RSA 3072]-----+
| .0 . |
| 00 + . |
| ..0= 0 |
| 0+..0 |
| oS.+ . |
| .. ..= |
| . .+= = |
| .+ oX=00. 0 |
| E+o+=+@@Bo . |
+---[SHA256]-----+
mapred@master:/$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
mapred@master:/$ chmod 0600 ~/.ssh/authorized_keys
mapred@master:/$ exit

```

chạy dịch vụ ssh

```
root@master:/# service ssh start_
```

Tải Hadoop về máy

```
root@master:/# wget https://d1cdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
```

Giải nén

```
root@master:/# tar -xvzf hadoop-3.3.4.tar.gz
```

Di chuyển, thiết lập và cấp quyền cho hadoop cho phù hợp

```
root@master:/# mv hadoop-3.3.4 /lib/hadoop
root@master:/# mkdir /lib/hadoop/logs
root@master:/# chgrp hadoop -R /lib/hadoop
root@master:/# chmod g+w -R /lib/hadoop
```

Cấu hình biến môi trường /etc/bash.bashrc

```
export JAVA_HOME=/usr/lib/jvm/default-java
export HADOOP_HOME=/lib/hadoop
export PATH=$PATH:$HADOOP_HOME/bin

export HDFS_NAMENODE_USER="hdfs"
export HDFS_DATANODE_USER="hdfs"
export HDFS_SECONDARYNAMENODE_USER="hdfs"
export YARN_RESOURCEMANAGER_USER="yarn"
export YARN_NODEMANAGER_USER="yarn"

export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop_
# System-wide .bashrc file for interactive bash(1) .
```

Cập nhật biến môi trường \$HADOOP_HOME/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/default-java
#
```

Thiết lập cấu hình cho Hadoop bao gồm cấu hình lại 3 file:

- core-site.xml
- hdfs-site.xml
- yarn-site.xml

Cấu hình file: core-site.xml

Xem thêm tại: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/core-default.xml>

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/home/${user.name}/hadoop</value>
  </property>
</configuration>
```

Cấu hình file: hdfs-site.xml

Xem thêm tại: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml>

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.permissions.superusergroup</name>
    <value>hadoop</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir.perm</name>
    <value>774</value>
  </property>
</configuration>
```

Cấu hình file: yarn-site.xml

Xem thêm tại: <https://hadoop.apache.org/docs/current3/hadoop-yarn/hadoop-yarn-common/yarn-default.xml>

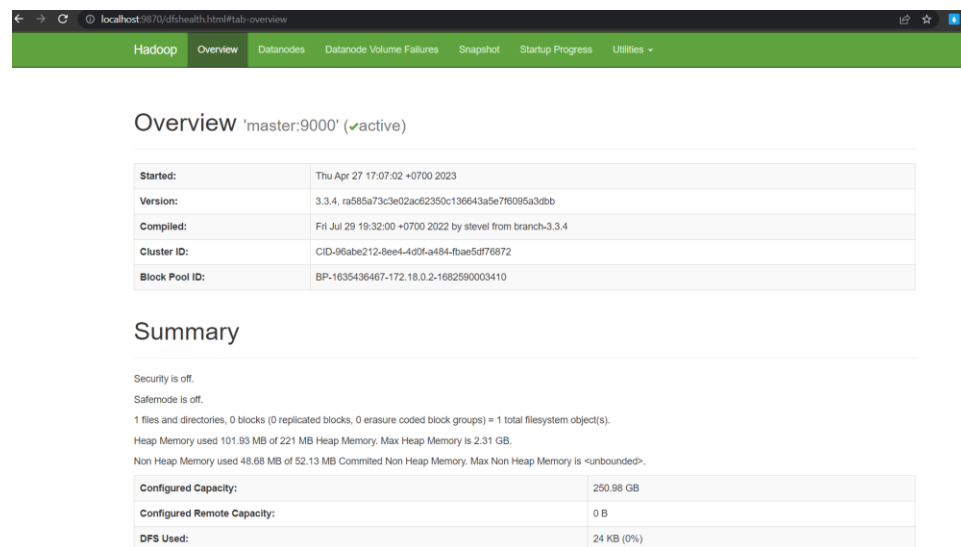
```
-->
<configuration>

<!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>master</value>
  </property>
</configuration>
```

Sau khi cấu hình xong 3 file của hadoop ta sẽ chạy các dịch vụ của hadoop

```
root@master:/# $HADOOP_HOME/sbin/start-all.sh
```

Kết quả: <http://localhost:9870/>



Overview 'master:9000' (✓active)

Started:	Thu Apr 27 17:07:02 +0700 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 19:32:00 +0700 2022 by stevel from branch-3.3.4
Cluster ID:	CID-96abe212-8ee4-4d0f-a484-fbae5df76872
Block Pool ID:	BP-1635436467-172.18.0.2-1682590003410

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 101.93 MB of 221 MB Heap Memory. Max Heap Memory is 2.31 GB.
Non Heap Memory used 48.68 MB of 52.13 MB Committed Non Heap Memory. Max Non Heap Memory is «unbounded».

Configured Capacity:	250.98 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)

Sau khi cài đặt và cấu hình xong master node thì ta sẽ tiếp tục cấu hình thêm 2 slave master node để hệ thống hoàn chỉnh.

Vì cấu hình của slave node khá giống với master node nên ta sẽ tận dụng image của master node (có tên là hadoop) để chạy thêm 2 slave node để tiết kiệm thời gian. Chúng ta cũng sẽ chạy trên hadoop_network.

```
PS C:\Users\phuc> docker run -it --name slave1 --hostname slave1 --network hadoop_network hadoop
```

Cập nhật IP và hostname cho slave1

```
GNU nano 4.8
172.18.0.3    slave1
172.18.0.2    master_
```

Cập nhật IP và hostname cho slave2

```
GNU nano 4.8
172.18.0.2    master
172.18.0.4    slave2
```

Cập nhật IP và hostname cho master

```
GNU nano 4.8
172.18.0.2    master
172.18.0.3    slave1
172.18.0.4    slave2_
```

Cập nhật file worker cho master để chúng có thể hoạt động với nhau.

```
master
slave1
slave2_
```

Cuối cùng chúng ta sẽ chạy tất cả các dịch vụ trên master node

Kết quả: <http://localhost:9870/dfshealth.html#tab-datanode>

In operation

DataNode State: All Show: 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✓ default-rack/slave1 9866 (172.18.0.3 9866)	http://slave1 9864	0s	4m	24 KB	8.28 GB	250.98 GB	0	24 KB (0%)	3.3.4
✓ default-rack/master 9866 (172.18.0.2 9866)	http://master 9864	0s	17m	28 KB	8.28 GB	250.98 GB	0	28 KB (0%)	3.3.4
✓ default-rack/slave2 9866 (172.18.0.4 9866)	http://slave2 9864	1s	1m	24 KB	8.28 GB	250.98 GB	0	24 KB (0%)	3.3.4

Showing 1 to 3 of 3 entries

Previous 1 Next

Thử nghiệm với dịch vụ HDFS:

Download Head the file (first 32K) Tail the file (last 32K)

Block information -- Block 4

Block ID: 1073741849

Block Pool ID: BP-1375821618-172.20.0.2-1666404487147

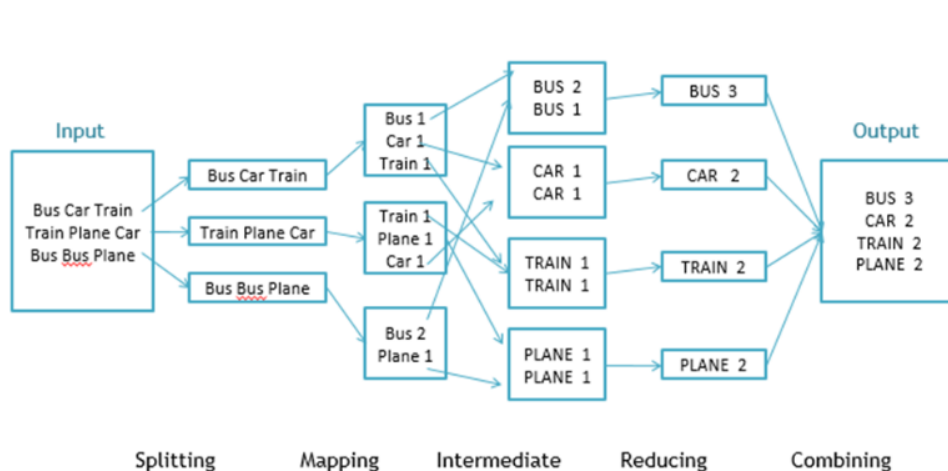
Generation Stamp: 1027

Size: 134217728

Availability:

- node02
- node03


Thử nghiệm với dịch vụ mapreduce




Kết quả:


```
Apache 1
Foundation 1
Software 1
The 1
This 1
by 1
developed 1
includes 1
product 1
software 1
```

Đưa Image của Master và Slave tạo ra lên Docker hub để có thể chia sẻ và tái sử dụng.

 txphuoc2001 / hadoop-docker

Description

This repository does not have a description 

 Last pushed: 2 hours ago

Docker commands





To push a new tag to this repository,

```
docker push txphuoc2001/hadoop-docker:tagname
```

[Public View](#)

Tags


This repository contains 2 tag(s).

Tag	OS	Type	Pulled	Pushed
 slave		Image	---	2 hours ago
 master		Image	---	2 hours ago

[See all](#) [Go to Advanced Image Management](#)

Automated Builds

Manually pushing images to Hub? Connect your account to GitHub or Bitbucket to automatically build and tag new images whenever your code is updated, so you can focus your time on creating.

Available with Pro, Team and Business subscriptions. [Read more about automated builds](#) .

[Upgrade](#)