Hadoop Global Distributed Environment Construction Manual

Node planning

CPU name	IP address	size	operating system
master	192.168.20.10	2 核 2G 50G	Centos7.5
node1	192.168.20.11	2 核 2G 50G	Centos7.5
node2	192.168.20.12	2 核 2G 50G	Centos7.5

一、Prepare three virtual machines

The three host names are set to master node1 node2 respectively

1. turn off firewall selinux

```
    [root@localhost ~]# systemctl stop firewalld
    [root@localhost ~]# systemctl disable firewalld
    Removed symlink /etc/systemd/system/multi-
user.target.wants/firewalld.service
                              .
    Removed symlink /etc/systemd/system/dbus-
org.fedoraproject.FirewallD1.service
                                .
    [root@localhost ~]# vi /etc/selinux/config
    [root@localhost ~]# setenforce 0
    [root@localhost ~]# getenforce
    Permissive
```

2. set hosts file

```
    (1). Modify hostname
    [root@localhost ~]# hostnamectl set-hostname master
```

```
[root@localhost ~]# bash
[root@master ~]# vi /etc/hosts
192.168.20.10 master
192.168.20.11 node1
192.168.20.12 node2
(2). Send configuration to other hosts
[root@master ~]# scp -r /etc/hosts root@node1:/etc/
[root@master ~]# scp -r /etc/hosts root@node2:/etc/

[root@master ~]# scp -r /etc/selinux/config root@node1:/etc/
[root@master ~]# scp -r /etc/selinux/config root@node2:/etc/
(3). test
ping node1
ping node2
[root@master ~]# ping node1 -c 3
PING node1 (192.168.20.11) 56(84) bytes of data.
64 bytes from node1 (192.168.20.11): icmp_seq=1 ttl=64 time=0.531 ms
64 bytes from node1 (192.168.20.11): icmp_seq=2 ttl=64 time=0.595 ms
64 bytes from node1 (192.168.20.11): icmp_seq=3 ttl=64 time=0.661 ms
--- node1 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2006ms
rtt min/avg/max/mdev = 0.531/0.595/0.661/0.060 ms
[root@master ~]# ping node2 -c 3
PING node2 (192.168.20.12) 56(84) bytes of data.
64 bytes from node2 (192.168.20.12): icmp_seq=1 ttl=64 time=0.662 ms
64 bytes from node2 (192.168.20.12): icmp_seq=2 ttl=64 time=0.556 ms
64 bytes from node2 (192.168.20.12): icmp_seq=3 ttl=64 time=0.654 ms
--- node2 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2004ms
rtt min/avg/max/mdev = 0.556/0.624/0.662/0.048 ms
```

```
[root@master ~]#
```

二、Install and configure jdk

1. Copy the jdk compressed file to the virtual machine

```
[root@master ~]# ll
total 586276
-rw-------. 1 root root      1260 May  6 14:33 anaconda-ks.cfg
-rw-r--r--. 1 root root 408587111 May  7 09:19 hadoop-2.10.1.tar.gz
-rw-r--r--. 1 root root 191753373 May  7 09:19 jdk-8u191-linux-x64.tar.gz
```

2. Unzip the compressed file to the /opt/ directory

```
[root@master ~]# tar -zxvf jdk-8u191-linux-x64.tar.gz -C /opt/
[root@master ~]# ll /opt/
total 0
drwxr-xr-x. 7 10 143 245 Oct  6  2018 jdk1.8.0_191
```

3. Configure jdk environment variables

```
[root@master ~]# vi /etc/profile
export JAVA_HOME=/opt/jdk1.8.0_191
export PATH=$PATH:$JAVA_HOME/bin


# 立即生效环境变量
[root@master ~]# source /etc/profile
```

4. copy jdk file to other host

```
[root@master ~]# scp -r /opt/jdk1.8.0_191 root@node1:/opt/
```

```
[root@master ~]# scp -r /opt/jdk1.8.0_191 root@node2:/opt/
```

5. Copy environment variable configuration files to other hosts

```
[root@master ~]# scp -r /etc/profile root@node1:/etc/

[root@master ~]# scp -r /etc/profile root@node2:/etc/
```

Execute separately and take effect immediately

```
[root@node1 ~]# source /etc/profile

[root@node2 ~]# source /etc/profile
```

Test if jdk is installed successfully

```
[root@master ~]# java -version
```

Displaying the java version information indicates that the jdk installation is successful

```
[root@master ~]# java -version

java version "1.8.0_191"

Java(TM) SE Runtime Environment (build 1.8.0_191-b12)

Java HotSpot(TM) 64-Bit Server VM (build 25.191-b12, mixed mode)


[root@node1 ~]# java -version

java version "1.8.0_191"

Java(TM) SE Runtime Environment (build 1.8.0_191-b12)

Java HotSpot(TM) 64-Bit Server VM (build 25.191-b12, mixed mode)


[root@node2 ~]# java -version

java version "1.8.0_191"

Java(TM) SE Runtime Environment (build 1.8.0_191-b12)

Java HotSpot(TM) 64-Bit Server VM (build 25.191-b12, mixed mode)
```

三、Install hadoop cluster

1. Download hadoop zip file

2. Upload the compressed file to the virtual machine

```
[root@master ~]# ll
total 586276
-rw-------. 1 root root      1260 May  6 14:33 anaconda-ks.cfg
-rw-r--r--. 1 root root 408587111 May  7 09:19 hadoop-2.10.1.tar.gz
-rw-r--r--. 1 root root 191753373 May  7 09:19 jdk-8u191-linux-x64.tar.gz
```

3. Unzip to /opt/ directory

```
[root@master ~]# tar -zxvf hadoop-2.10.1.tar.gz -C /opt/
[root@master ~]# ll /opt/
total 0
drwxr-xr-x. 9 1000 1000 149 Sep 14  2020 hadoop-2.10.1
drwxr-xr-x. 7   10  143 245 Oct  6  2018 jdk1.8.0_191
```

4. Configure environment variables

```
    [root@master ~]# vi /etc/profile
    export PATH=$PATH:$JAVA_HOME/bin:/opt/hadoop-2.10.1/sbin:/opt/hadoop-2.10.1/bin
    # 立即生效
    [root@master ~]# source /etc/profile
    # 查看版本
    [root@master ~]# hadoop version
    Hadoop 2.10.1
    Subversion https://github.com/apache/hadoop -r
1827467c9a56f133025f28557bfc2c562d78e816
    Compiled by centos on 2020-09-14T13:17Z
    Compiled with protoc 2.5.0
    From source with checksum 3114edef868f1f3824e7d0f68be03650
    This command was run using /opt/hadoop-2.10.1/share/hadoop/common/hadoop-common-
2.10.1.jar
```

5. Configure hadoop cluster

```
xml file corresponding to each component

common component------>core-site.xml

HDFS component------>hdfs-site.xml

MapReduce component------>mapred-site.xml

YARN  component------>yam-site.xml [root@master ~]# cd /opt/hadoop-2.10.1/etc/hadoop/
```

5.1 Configure hadoop-env.sh file

```
[root@master hadoop]# vi hadoop-env.sh

export JAVA_HOME=/opt/jdk1.8.0_191     # JDK 的安装路径
```

5.2 Configure the core-site.xml file

```
[root@master hadoop]# vi core-site.xml

<configuration>

    <!-- Specify the address of the namenode in hdfs -->

    <property>

        <name>fs.defaultFS</name>

        <value>hdfs://master:9000</value>

    </property>


    <!-- Specifies the storage directory for files generated when hadoop is running -->

    <property>

        <name>dfs.tmp.dir</name>

        <value>file:///opt/hadoop-data/</value>

    </property>

</configuration>
```

5.3 Configure hdfs-site.xml file

```
[root@master hadoop]# vi hdfs-site.xml

<configuration>

  <!-- Set the number of dfs replicas, the default is 3 if not set -->

  <property>

    <name>dfs.replication</name>

    <value>1</value>

  </property>


  <!-- Set the port for secoundname -->

  <property>

    <name>dfs.namenode.secondary.http-address</name>

    <value>node1:50090</value>

  </property>

</configuration>
```

5.4 Configure maperd-env.sh file

```
[root@master hadoop]# vi mapred-env.sh

# Find export JAVA_HOME=Add  java environment variable  afterexport

JAVA_HOME=/opt/jdk1.8.0_191
```

5.5 Configure the mapred-site.xml file

```
If there is no such file, copy the mapred-site.xml.template file  to mapred-site.xml

[root@master hadoop]# cp mapred-site.xml.template mapred-site.xml

[root@master hadoop]# vi mapred-site.xml

<configuration>

<!--Specify mapreduce to run on yarn -->

   <property>      <name>mapreduce.framework.name</name>
```

```
        <value>yarn</value>
    </property>
</configuration>
```

5.6 configure yarn-env.sh

```
[root@master hadoop]# vi yarn-env.sh
# Find export JAVA_HOME=Add  java environment variable behind
export JAVA_HOME=/opt/jdk1.8.0_191
export JAVA_HOME=${JAVA_HOME}
```

5.7 Configure yarn-site.xml file

```
[root@master hadoop]# vi yarn-site.xml
<configuration>
    <!-- Specify the address of the ResourceManager-->
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>master</value>
    </property>
    <!-- Specify how the reducer gets data-->
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>
```

5.8 Configure slaves file

```
[root@master hadoop]# vi slaves
```

```
master
node1
node2
```

6. Distribute the installed hadoop to other host nodes

```
[root@master ~]# scp -r /opt/hadoop-2.10.1 root@node1:/opt/
[root@master ~]# scp -r /opt/hadoop-2.10.1 root@node2:/opt/
```

7. Copy the environment variable configuration file to other hosts

```
[root@master ~]# scp -r /etc/profile root@node1:/etc/
[root@master ~]# scp -r /etc/profile root@node2:/etc/
# Execute separately and take effect immediately
[root@node1 ~]# source /etc/profile
[root@node2 ~]# source /etc/profile
```

View version

```
[root@node1 ~]# hadoop version
Hadoop 2.10.1
Subversion https://github.com/apache/hadoop -r 1827467c9a56f133025f28557bfc2c562d78e816
Compiled by centos on 2020-09-14T13:17Z
Compiled with protoc 2.5.0
From source with checksum 3114edef868f1f3824e7d0f68be03650
This command was run using /opt/hadoop-2.10.1/share/hadoop/common/hadoop-common-2.10.1.jar
[root@node2 ~]# hadoop version
Hadoop 2.10.1
Subversion https://github.com/apache/hadoop -r
```

```
1827467c9a56f133025f28557bfc2c562d78e816
Compiled by centos on 2020-09-14T13:17Z
Compiled with protoc 2.5.0
From source with checksum 3114edef868f1f3824e7d0f68be03650
This command was run using /opt/hadoop-2.10.1/share/hadoop/common/hadoop-common-2.10.1.jar
```

8、Set up SSH password-free access

Require password-free access between any two hosts

```
Execute the following commands between the three hosts
Take the master node as an example

[root@master ~]# ssh-keygen always press Enter when executing this command
[root@master ~]# ssh-copy-id master
[root@master ~]# ssh-copy-id node1
[root@master ~]# ssh-copy-id node2

[root@master ~]# ssh-copy-id master
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/root/.ssh/id_rsa.pub"
The authenticity of host 'master (192.168.20.10)' can't be established.
ECDSA key fingerprint is SHA256:Os3CLxJnNK5r6yjp351a2ITXWb3zXDfPnyZKq8tDmHk.
ECDSA key fingerprint is MD5:dd:cd:01:92:ee:c3:d9:ee:a7:4b:5d:f3:36:f0:e2:bb.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys
root@master's password: node password
```

```
Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'master'"
and check to make sure that only the key(s) you wanted were added.

[root@master ~]# ssh-copy-id node1
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/root/.ssh/id_rsa.pub"
The authenticity of host 'node1 (192.168.20.11)' can't be established.
ECDSA key fingerprint is SHA256:Os3CLxJnNK5r6yjp351a2ITXWb3zXDfPnyZKq8tDmHk.
ECDSA key fingerprint is MD5:dd:cd:01:92:ee:c3:d9:ee:a7:4b:5d:f3:36:f0:e2:bb.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys
root@node1's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'node1'"
and check to make sure that only the key(s) you wanted were added.

[root@master ~]# ssh-copy-id node2
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/root/.ssh/id_rsa.pub"
The authenticity of host 'node2 (192.168.20.12)' can't be established.
ECDSA key fingerprint is SHA256:Os3CLxJnNK5r6yjp351a2ITXWb3zXDfPnyZKq8tDmHk.
ECDSA key fingerprint is MD5:dd:cd:01:92:ee:c3:d9:ee:a7:4b:5d:f3:36:f0:e2:bb.
```

```
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys
root@node2's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'node2'"
and check to make sure that only the key(s) you wanted were added.

[root@master ~]#
```

test

```
[root@master ~]# ssh master
Last login: Sat May  7 09:10:49 2022 from 192.168.20.1
[root@master ~]# exit
logout
Connection to master closed.
[root@master ~]# ssh node1
Last login: Sat May  7 09:10:54 2022 from 192.168.20.1
[root@node1 ~]# exit
logout
Connection to node1 closed.
[root@master ~]# ssh node2
Last login: Sat May  7 09:10:52 2022 from 192.168.20.1
[root@node2 ~]# exit
```

```
logout
Connection to node2 closed.
```

9、Start the cluster is only executed on the master node

We configured the runtime file storage location hadoop-data in core-site.xml

In fact, there is no directory in the generated directory. You need to create it yourself. If you do not configure this directory, it will be automatically stored in the tmp directory under the root directory. [root@master ~]# cd /opt/

```
[root@master opt]# mkdir hadoop-data
```

9.1 The first time you start the cluster, you need to format the namenode

```
[root@master ~]# hdfs namenode -format
[root@master ~]# sh start-dfs.sh
```

9.2 start yarn

On the host master, because we configured to start on the master in yarn-site.xml, it needs to be started separately

```
[root@master ~]#sh start-yarn.sh
```

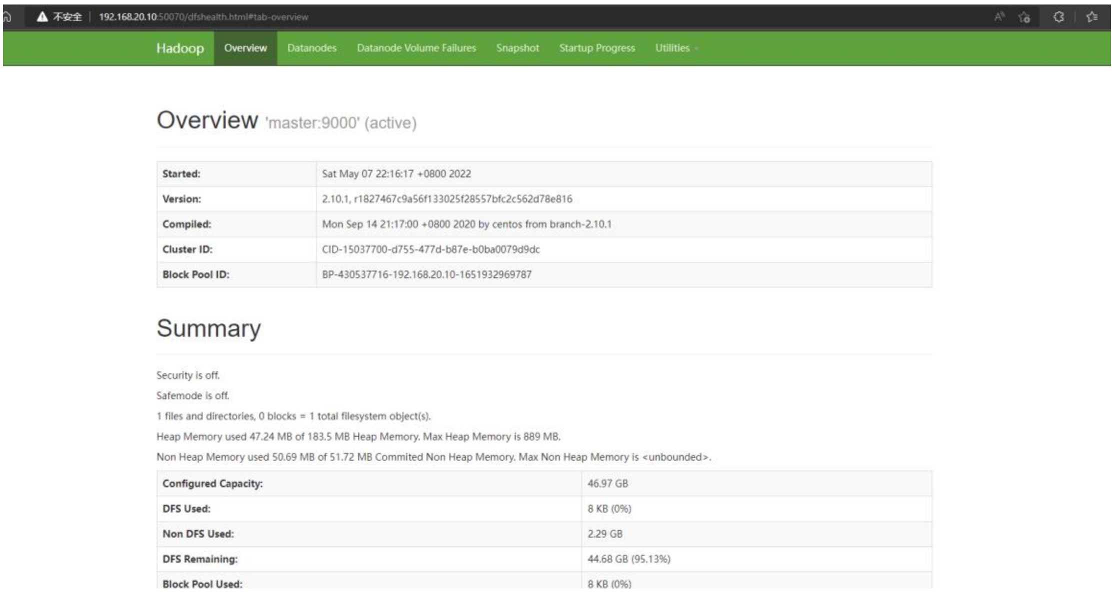
9.3 jps view process

```
[root@master ~]# jps
11602 NameNode
12717 NodeManager
13069 Jps
12607 ResourceManager
```

```
[root@node1 ~]# jps
3187 SecondaryNameNode
3299 NodeManager
3415 Jps

[root@node2 ~]# jps
3556 NodeManager
3672 Jps
3321 DataNode
```

9.4 View on the web

Enter master in the address bar of the web page: 50070



Overview 'master:9000' (active)

Started:	Sat May 07 22:16:17 +0800 2022
Version:	2.10.1, r1827467c9a56f133025f28557bfc2c562d78e816
Compiled:	Mon Sep 14 21:17:00 +0800 2020 by centos from branch-2.10.1
Cluster ID:	CID-15037700-d755-477d-b87e-b0ba0079d9dc
Block Pool ID:	BP-430537716-192.168.20.10-1651932969787

Summary

Security is off.

Safemode is off.

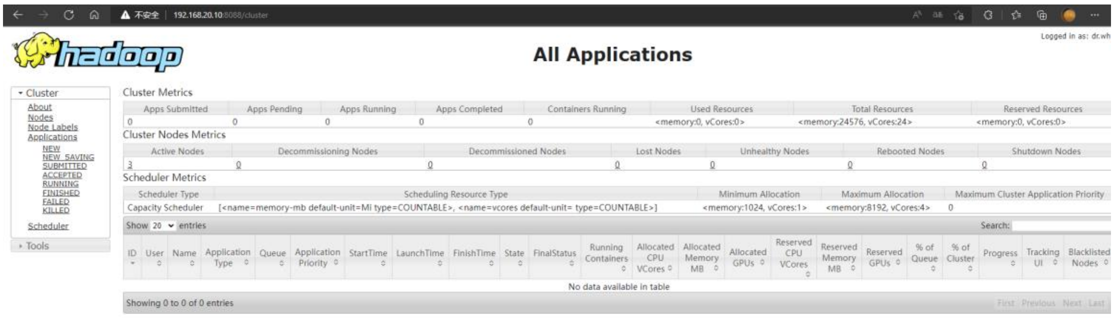
1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 47.24 MB of 183.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 50.69 MB of 51.72 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	46.97 GB
DFS Used:	8 KB (0%)
Non DFS Used:	2.29 GB
DFS Remaining:	44.68 GB (95.13%)
Block Pool Used:	8 KB (0%)

Explorer view master:8088



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources
0	0	0	0	0	<memory:0, vCores:0>	<memory:24576, vCores:24>	<memory:0, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
3	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

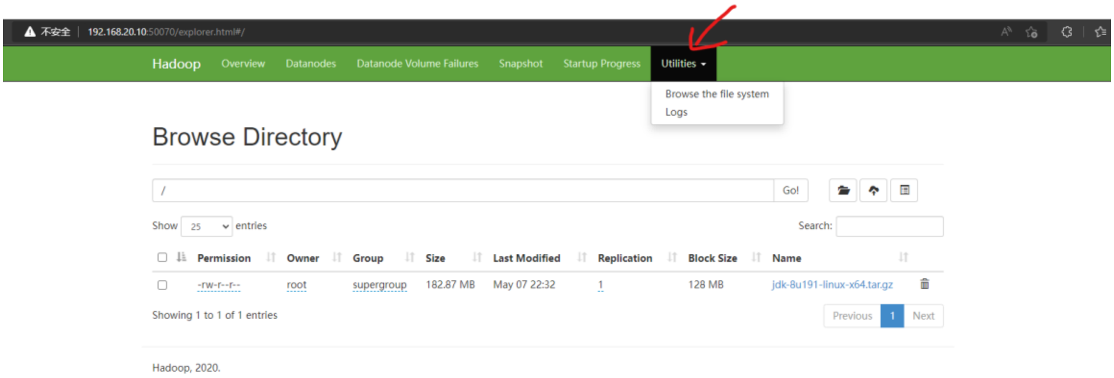
ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																						

Showing 0 to 0 of 0 entries

9.5 test upload download

```
This test uses the Hadoop API test

[root@master ~]# hadoop fs -put /root/jdk-8u191-linux-x64.tar.gz /

[root@master ~]# hadoop fs -ls /

Found 1 items

-rw-r--r--   1 root supergroup   191753373 2022-05-07 10:32 /jdk-8u191-linux-x64.tar.gz
```

web view:



Browse Directory

/

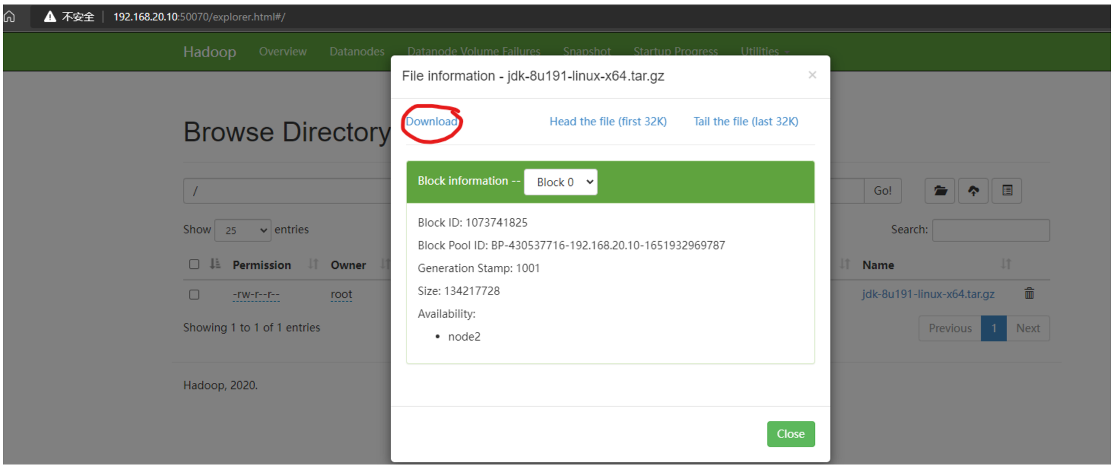
Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	root	supergroup	182.87 MB	May 07 22:32	1	128 MB	jdk-8u191-linux-x64.tar.gz

Showing 1 to 1 of 1 entries

Hadoop, 2020.

download



File information - jdk-8u191-linux-x64.tar.gz

Block information -- Block 0

Block ID: 1073741825

Block Pool ID: BP-430537716-192.168.20.10-1651932969787

Generation Stamp: 1001

Size: 134217728

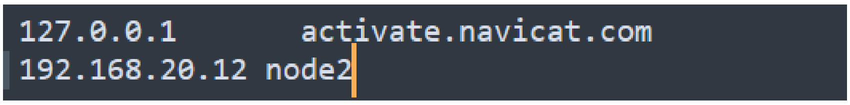
Availability:

- node2

The download needs to modify the hosts file in the win environment to add node2 domain name resolution

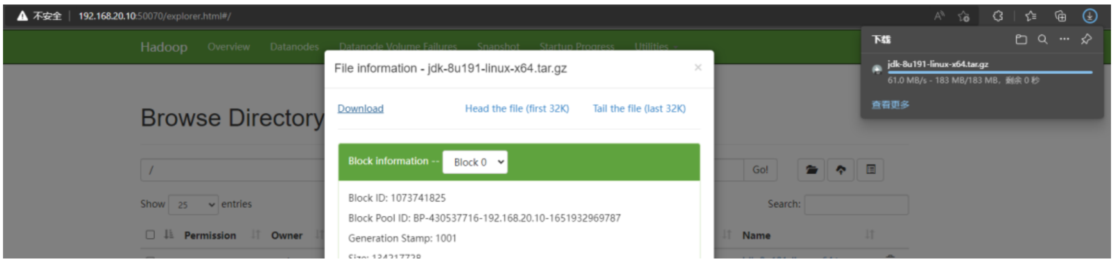
There is a hosts file under C:\Windows\System32\drivers\etc

Copy it out and add the domain name resolution of the node2 node, and move it to the source file location



```
127.0.0.1        activate.navicat.com
192.168.20.12 node2
```

click to download



File information - jdk-8u191-linux-x64.tar.gz

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

Block information -- Block 0

Block ID: 1073741825

Block Pool ID: BP-430537716-192.168.20.10-1651932969787

Generation Stamp: 1001

Size: 134217728

jdk-8u191-linux-x64.tar.gz

61.0 MB/s - 183 MB/183 MB，剩余 0 秒