Hadoop分布式环境安装与配置

一、搭建虚拟机

本实验将建立三台虚拟机组成的集群，各虚拟机的节点分配如下

计算机名称	运行节点
hadoop01	NameNode、DataNode、ResourceManager、NodeManager
hadoop02	SecondaryNameNode、DataNode、NodeManager
hadoop03	DataNode、NodeManager

创建第一个名为hadoop01的虚拟机（略）

启动虚拟机后，输入命令安装需要的工具

```
sudo apt upgrade
sudo apt install openssh-server
sudo apt install net-tools
sudo apt install vim
```

二、安装hadoop环境包

1、准备好所有需要的安装包复制到之前设置的共享文件夹VMshare中，包含jdk、hadoop、spark等。建议使用与本实验版本一致的软件包，防止兼容问题。

用以下命令将共享文件夹所有的安装包复制到主目录

```
sudo cp -R /mnt/hgfs/VMShare/ /home/hadoop/
```

给此目录赋予操作权限

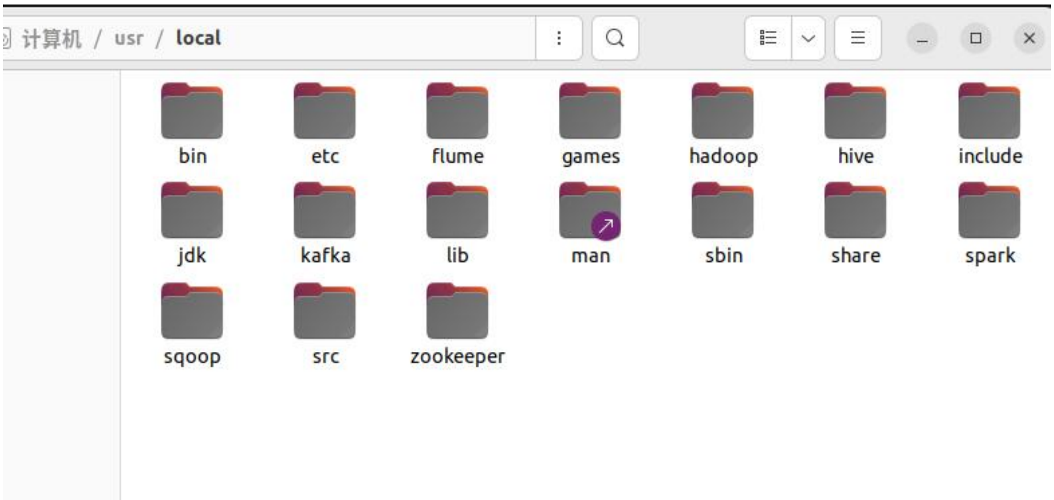
```
sudo chmod -R 777 /home/hadoop/VMShare
```

进入到主目录的VMShare，先安装jdk和hadoop，执行命令解压

```
tar -zxvf jdk-8u261-linux-x64.tar.gz
tar -zxvf hadoop-2.7.7.tar.gz
```

将解压后的文件夹移动到 /usr/local 目录下，同时改成简单的名字

```
sudo mv jdk1.8.0_261/ /usr/local/jdk
sudo mv hadoop-2.7.7/ /usr/local/hadoop
```



2、配置环境变量

使用以下命令打开配置文件

```
vi ~/.bashrc
```

在末尾插入以下代码（实际存放位置如不一致需要修改）

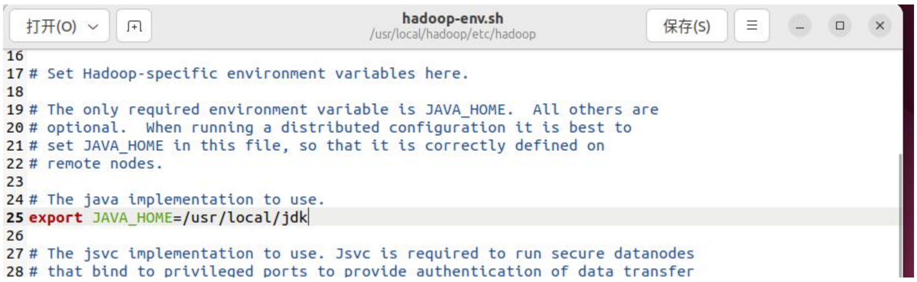
```
export JAVA_HOME=/usr/local/jdk
export PATH=$PATH:$JAVA_HOME/bin

export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

保存退出，执行以下代码使配置生效

```
source ~/.bashrc
```

3、配置hadoop

进入hadoop配置目录hadoop/etc/hadoop，打开hadoop-env.sh，设置JAVA_HOME的完整存放路径



```

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/local/jdk

# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
```

打开core-site.xml文件,在configuration标签中加入以下代码设置主节点名字，设置一个目录用于存放临时文件

```
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://hadoop01:9000</value>
    </property>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/usr/local/hadoop/data</value>
    </property>
```

打开hdfs-site.xml文件，往configuration加入设置，设置数据块的副本数量为3，设置hdfs的名字空间元数据、数据块存储位置，根据计划将SecondaryNameNode放在02节点

```
    <property>
        <name>dfs.replication</name>
        <value>3</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/usr/local/hadoop/data/dfs/name</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/usr/local/hadoop/data/dfs/data</value>
    </property>
    <property>
        <name>dfs.secondary.http.address</name>
        <value>hadoop02:50090</value>
    </property>
```

将mapred-site.xml.template复制一份并改名为mapred-site.xml，添加设置，mapreduce任务会提交到yarn上

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

打开yarn-site.xml文件，往configuration加入设置，使yarn在01节点运行，并启用mapreduce_shuffle"的辅助服务

```
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>hadoop01</value>
    </property>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
```

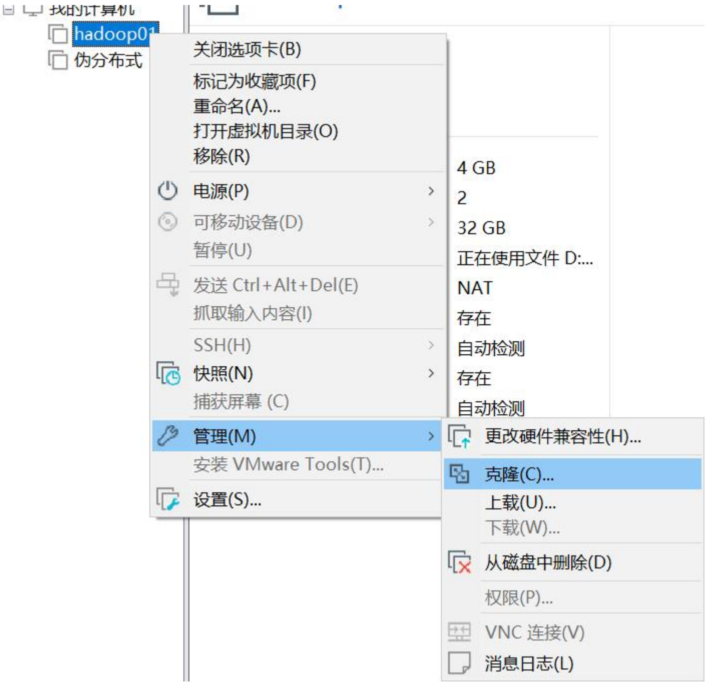
打开slaves文件，编写三台主机名

```
hadoop01
hadoop02
hadoop03
```

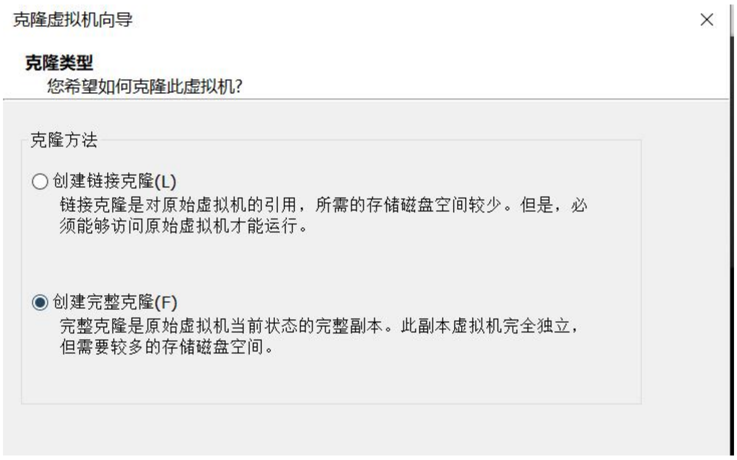
三、构建集群

1、克隆虚拟机

将节点01关机，右键——管理——克隆



创建完整克隆



克隆虚拟机向导

克隆类型

您希望如何克隆此虚拟机?

克隆方法

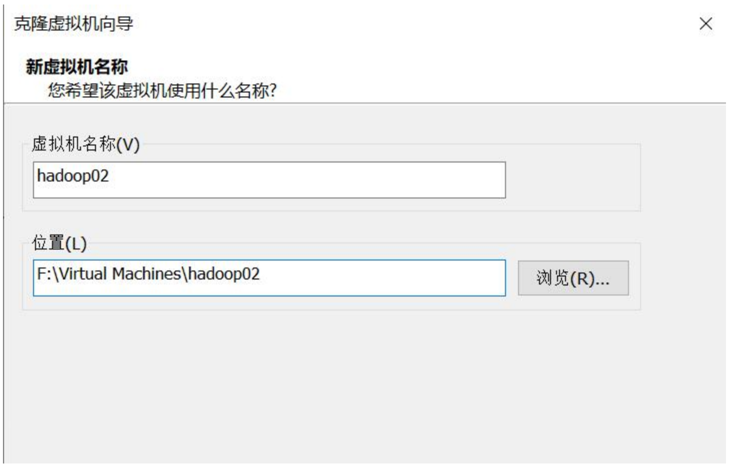
○ 创建链接克隆(L)

链接克隆是对原始虚拟机的引用，所需的存储磁盘空间较少。但是，必须能够访问原始虚拟机才能运行。

● 创建完整克隆(F)

完整克隆是原始虚拟机当前状态的完整副本。此副本虚拟机完全独立，但需要较多的存储磁盘空间。

设置节点名称和文件位置



克隆虚拟机向导

新虚拟机名称

您希望该虚拟机使用什么名称?

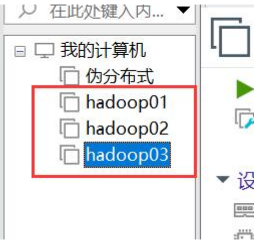
虚拟机名称(V)

hadoop02

位置(L)

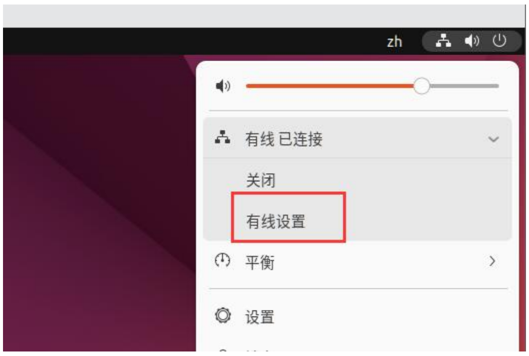
F:\Virtual Machines\hadoop02

用同样的方法，再将01节点克隆一份为03

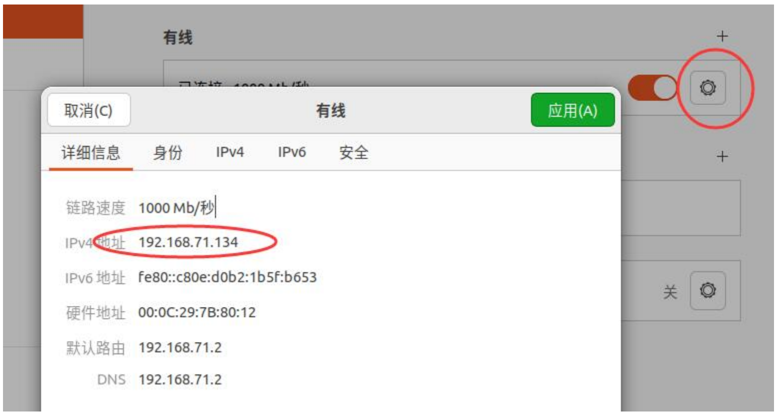


2、配置网络

将三个虚拟机全部打开，首先设置01节点的网络，打开有线设置

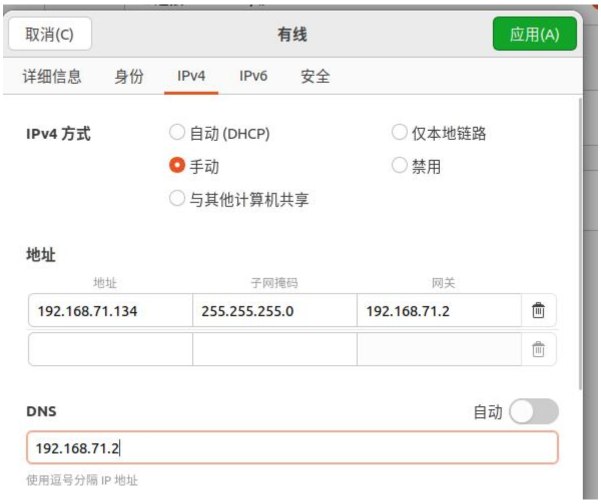


点开设置按钮，将ip地址复制



链路速度 1000 Mb/秒
IPv4 地址 192.168.71.134
IPv6 地址 fe80::c80e:d0b2:1b5f:b653
硬件地址 00:0C:29:7B:80:12
默认路由 192.168.71.2
DNS 192.168.71.2

切到IPv4页签，选择手动，地址、网关（默认路由）、DNS均与原本一致



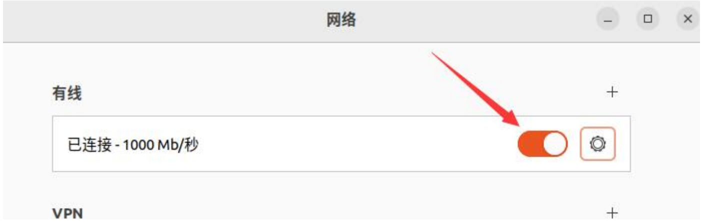
IPv4 方式
○ 自动 (DHCP) ○ 仅本地链路
● 手动 ○ 禁用
○ 与其他计算机共享

地址

地址	子网掩码	网关
192.168.71.134	255.255.255.0	192.168.71.2

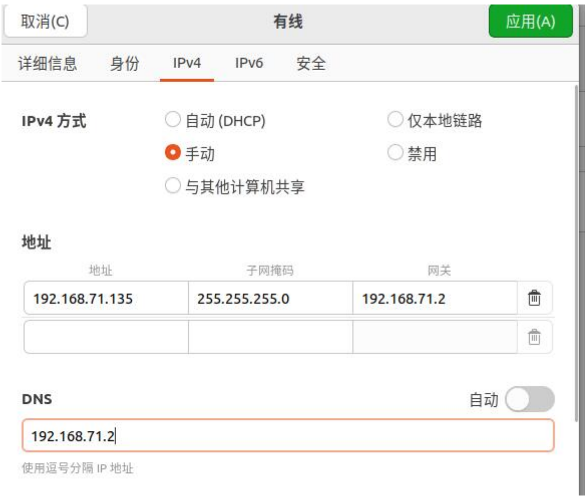
DNS 自动
192.168.71.2
使用逗号分隔 IP 地址

点击应用后，关闭再打开网络生效



已连接 - 1000 Mb/秒

其余两个节点重复同样的操作



IPv4 方式：○ 自动 (DHCP) ○ 仅本地链路 ● 手动 ○ 禁用 ○ 与其他计算机共享

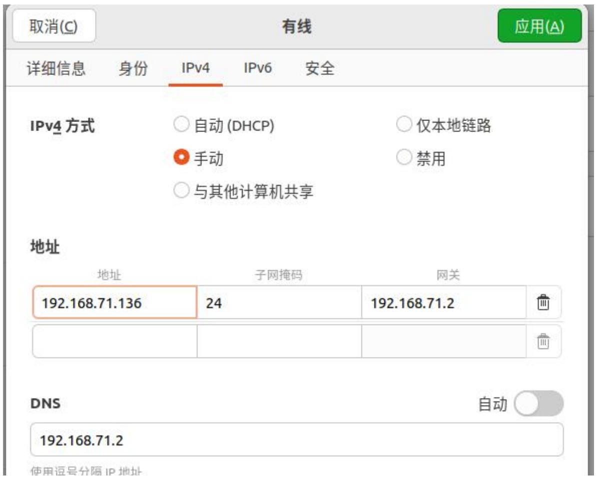
地址

地址	子网掩码	网关
192.168.71.135	255.255.255.0	192.168.71.2

DNS 自动 ○

192.168.71.2

使用逗号分隔 IP 地址



IPv4 方式：○ 自动 (DHCP) ○ 仅本地链路 ● 手动 ○ 禁用 ○ 与其他计算机共享

地址

地址	子网掩码	网关
192.168.71.136	24	192.168.71.2

DNS 自动 ○

192.168.71.2

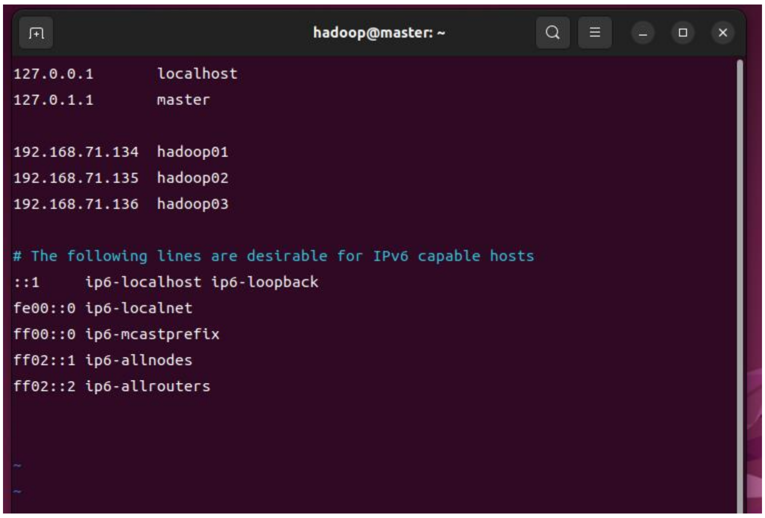
使用逗号分隔 IP 地址

3、配置主机IP

回到01节点，打开hosts文件

```
sudo vim /etc/hosts
```

添加以上3台虚拟机的实际ip地址和对应主机名（原本的内容可不删除）



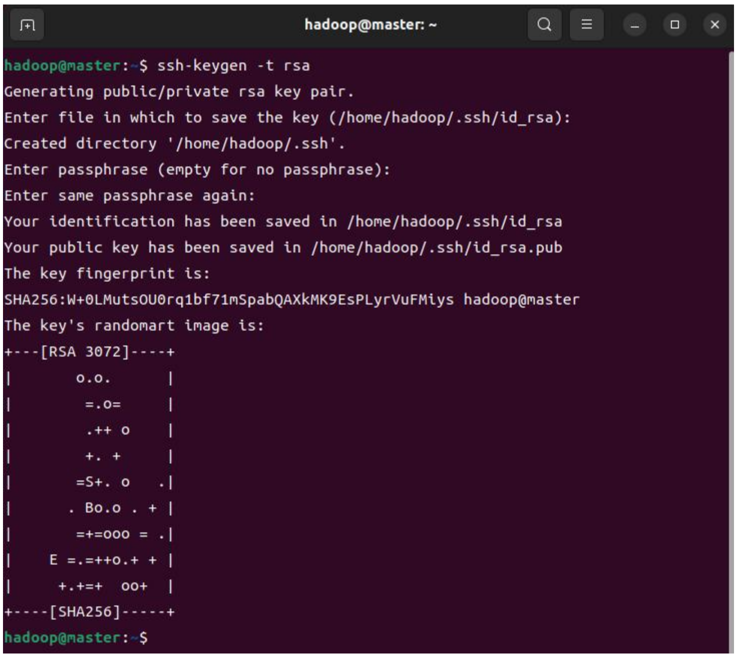
```
127.0.0.1       localhost
127.0.1.1       master

192.168.71.134  hadoop01
192.168.71.135  hadoop02
192.168.71.136  hadoop03

# The following lines are desirable for IPv6 capable hosts
::1     ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
~
~
```

同样的操作在02、03节点也重复一次

4、设置免密登录

回到01节点，打开终端输入命令，中途的对话全部留空直接按回车即可。

```
ssh-keygen -t rsa
```

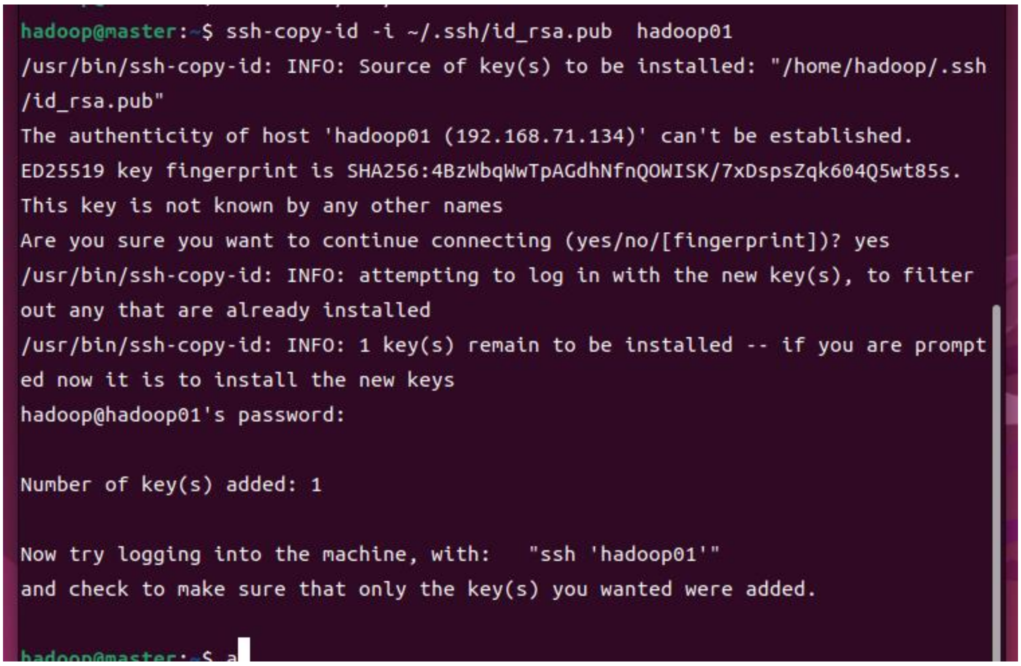


```
hadoop@master:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:W+0LMutsOU0rq1bf71mSpabQAXkMK9EsPLyrVuFMiys hadoop@master
The key's randomart image is:
+---[RSA 3072]----+
|       o.o.      |
|        =.o=     |
|        .++ o    |
|        +. +     |
|       =S+. o   .|
|      . Bo.o . + |
|       =+=ooo = .|
|    E =.=++o.+ + |
|     +.+=+  oo+  |
+----[SHA256]-----+
hadoop@master:~$
```

同样的操作在02、03节点也重复一次。

到01节点拷贝公钥，三个命令都需要输入yes和输密码。

```
ssh-copy-id -i ~/.ssh/id_rsa.pub  hadoop01
ssh-copy-id -i ~/.ssh/id_rsa.pub  hadoop02
ssh-copy-id -i ~/.ssh/id_rsa.pub  hadoop03
```

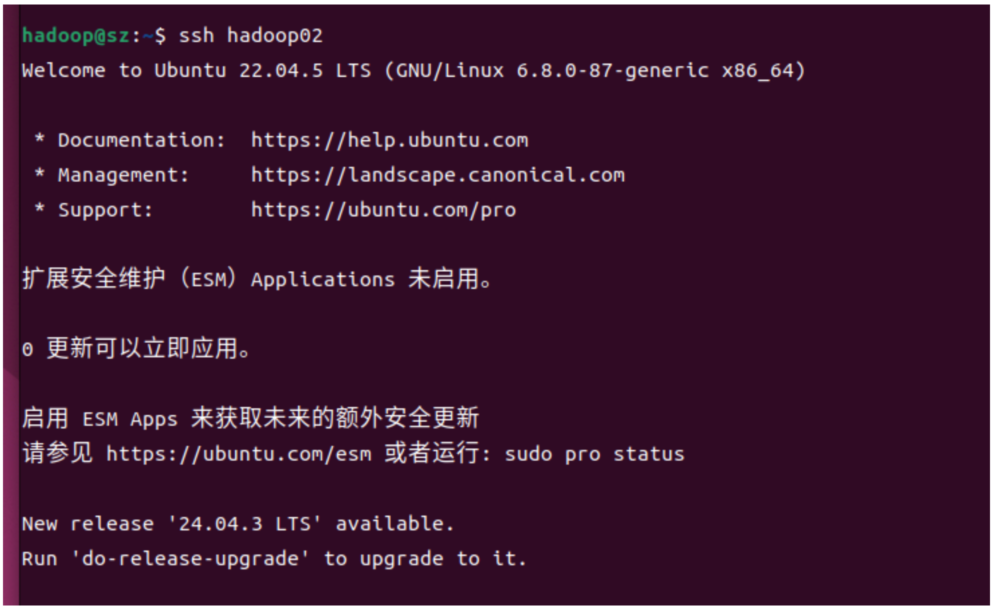


```
hadoop@master:~$ ssh-copy-id -i ~/.ssh/id_rsa.pub  hadoop01
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hadoop/.ssh
/id_rsa.pub"
The authenticity of host 'hadoop01 (192.168.71.134)' can't be established.
ED25519 key fingerprint is SHA256:4BzWbqWwTpAGdhNfnQOWISK/7xDspsZqk604Q5wt85s.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? 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 prompt
ed now it is to install the new keys
hadoop@hadoop01's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'hadoop01'"
and check to make sure that only the key(s) you wanted were added.

hadoop@master:~$ a
```

同样的操作在02、03节点也重复一次。(每个节点都要拷贝3个公钥)。

任意两个节点间使用ssh命令远程连接，均能直接出现以下结果，而不需要输密码，则设置成功

```
ssh 对方主机名
```

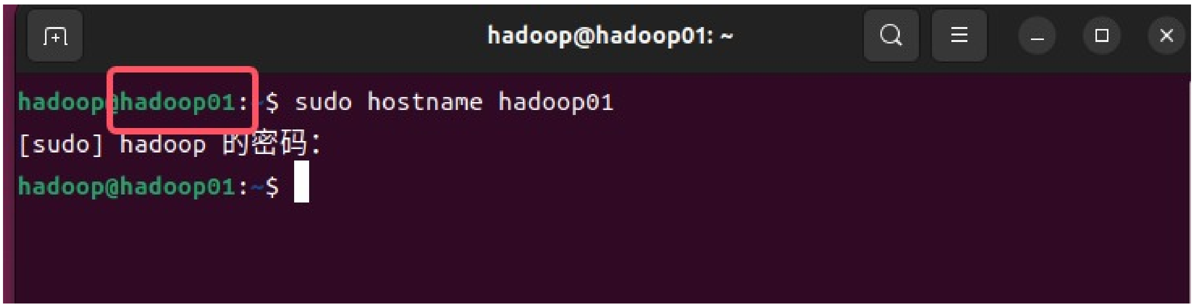


```
hadoop@sz:~$ ssh hadoop02
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-87-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

扩展安全维护（ESM）Applications 未启用。

0 更新可以立即应用。

启用 ESM Apps 来获取未来的额外安全更新
请参见 https://ubuntu.com/esm 或者运行: sudo pro status

New release '24.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
```

设置主机名，使用以下命令，将三个节点的主机名分别设为对应的hadoop01、hadoop02、hadoop03

```
sudo hostname 主机名
```

设置后，关闭终端重新打开，光标处应显示新的主机名



```
hadoop@hadoop01:~$ sudo hostname hadoop01
[sudo] hadoop 的密码:
hadoop@hadoop01:~$ 
```

四、启动测试

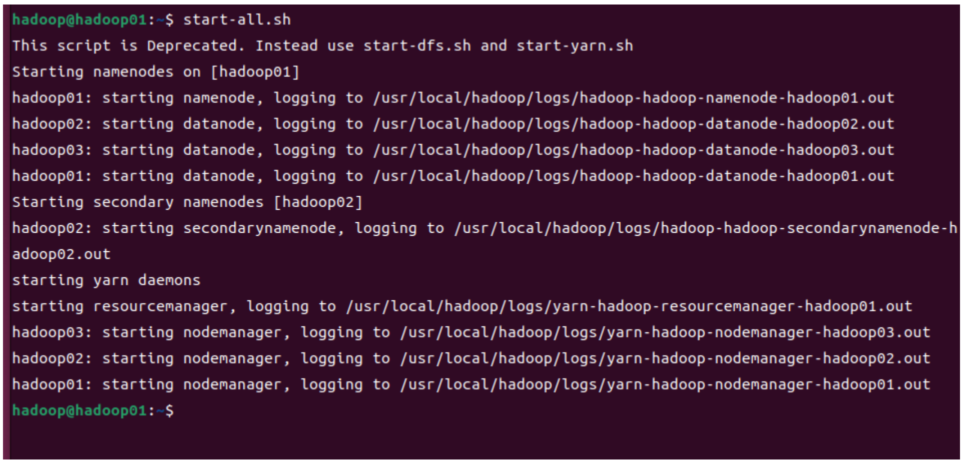
格式化、启动和停止都只需在01节点执行

1、格式化hdfs（首次使用）

```
hdfs namenode -format
```

2、启动hadoop，在输出日志中可以看到哪个节点启动了哪个进程

```
start-all.sh
```

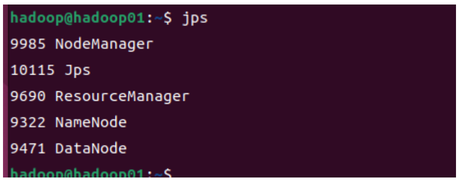


```
hadoop@hadoop01:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [hadoop01]
hadoop01: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-namenode-hadoop01.out
hadoop02: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-hadoop02.out
hadoop03: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-hadoop03.out
hadoop01: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-hadoop01.out
Starting secondary namenodes [hadoop02]
hadoop02: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-secondarynamenode-h
adoop02.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-resourcemanager-hadoop01.out
hadoop03: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-hadoop03.out
hadoop02: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-hadoop02.out
hadoop01: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-hadoop01.out
hadoop@hadoop01:~$
```

查分别查看3个节点的进程

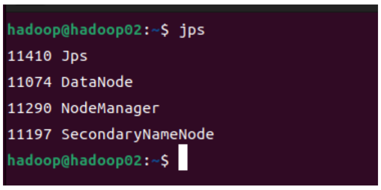
```
jps
```

01



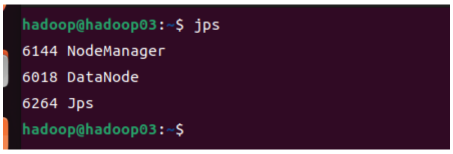
```
hadoop@hadoop01:~$ jps
9985 NodeManager
10115 Jps
9690 ResourceManager
9322 NameNode
9471 DataNode
hadoop@hadoop01:~$
```

02



```
hadoop@hadoop02:~$ jps
11410 Jps
11074 DataNode
11290 NodeManager
11197 SecondaryNameNode
hadoop@hadoop02:~$
```

03



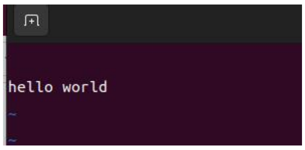
```
hadoop@hadoop03:~$ jps
6144 NodeManager
6018 DataNode
6264 Jps
hadoop@hadoop03:~$
```

与开头的规划表对应，部署成功。

4、测试上传文件

在本地创建一个文本文件，内容自定

```
vi test.txt
```



```
hello world
~
~
```

在hdfs上创建一个目录

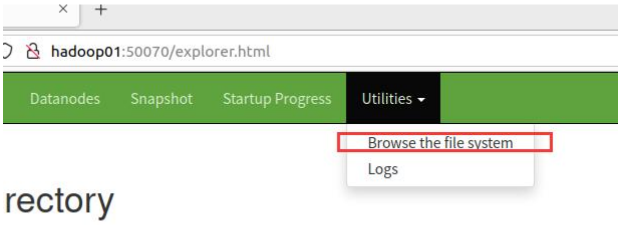
```
hdfs dfs -mkdir /home
```

上传文件到hdfs

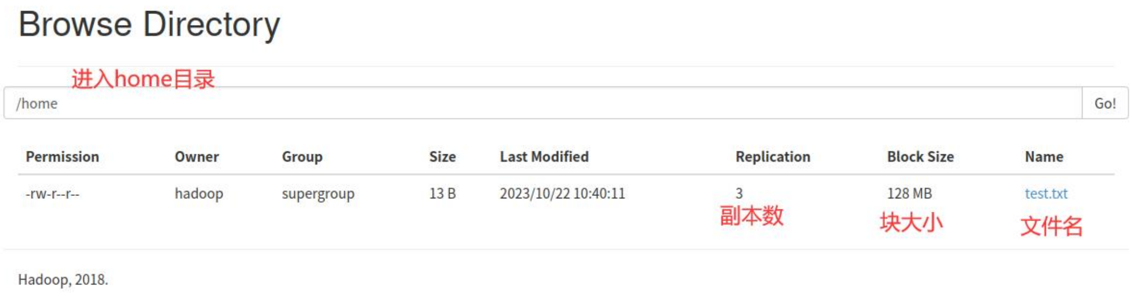
```
hdfs dfs -put test.txt /home
```

可通过浏览器访问hdfs的管理页面，地址：http://hadoop01:50070/

点击Utilities——Browse file system可以查看文件系统中的文件



查看刚才上传的文件信息



Browse Directory

进入home目录

/home

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	13 B	2023/10/22 10:40:11	3	128 MB	test.txt
					副本数	块大小	文件名

Hadoop, 2018.

要结束集群，在01节点执行

```
stop-all.sh
```