

哈尔滨工业大学（深圳）

大数据实验指导书

实验一 Hadoop 环境配置与基本操作

1. 实验目的

- 1. 熟悉 Hadoop 分布式配置的方法。
- 2. 熟悉命令行运行 Mapreduce 作业的原理和操作。

2. 实验内容

- 1. 根据实验指导书，小组合作搭建 Hadoop 分布式环境并截取 WebUI 界面（HDFS 界面 DataNode 页面，该页面包含了 IP 地址，实验报告必须包含此截图）。
 - 2. 配置 Hadoop 客户端（自己电脑），实现文本文件（爬虫得到的文件）的上传（此处可以截图命令行）。
 - 3. 为 WordCount 项目添加集群运行配置，运行并截图（可以截取 HDFS WebUI, jobhistory webUI, yarn WebUI, IDEA 控制台, wordcount 输出文本文件等）。
- 2, 3 部分每个人根据自己作业二的内容进行完成。

由于机房环境重启会还原系统，如实验课上未完成集群的搭建，可以课后自己小组用个人电脑搭建；如实验课上已完成集群的搭建但是尚未完成 2, 3 部分，请向助教申请集群，或者自行搭建，不影响实验评分。

3. 实验环境

- 1. Ubuntu 16.04
- 2. Hadoop 3.2.2
- 3. Java 1.8

4. 实验步骤

真实分布式的 Hadoop 集群搭建过程较为复杂，请同学们仔细阅读该章节，注意突出显示的文本。

在本实验指导书中，Hadoop 集群计算机预设如表 4-1 所示，集群包含 3 台计算机，在此基础上实现分布式 Hadoop 集群的搭建。

Hadoop 集群的配置过程中有许多部分可以进行自定义，但是首次配置建议尽量与本指导书保持一致。

表 4-1 教程用 Hadoop 集群计算机资源概要

编号	计算机名	IP 地址
678-21	hitsz-master	10.248.5.21
678-22	hitsz-slave1	10.248.5.22

678-23	hitsz-slave2	10.248.5.23
--------	--------------	-------------

IP 地址每台机器都不同，请根据实际情况进行修改。

配置过程中涉及到多条 linux 命令行操作，示例如下：

主机名
需要执行的命令

```

root@hitsz-master:/home/Lenovo/# apt-get install vim

```

用户名
当前所在目录

4.1 预下载文件与准备工作

本次实验需要用到的文件有：

- (1) Hadoop3.22 压缩包
- (2) Jdk1.8.0_202 压缩包
- (3) 实验指导书
- (4) xsync 脚本
- (5) 用于 wordcount 的文本文件

(1) - (4) 会预先提供，请同学们将 (1) - (4) 文件拷贝到 U 盘上，实验时需拷贝到集群计算机上。点击左侧文件管理器的图标，即可进行拷贝，建议直接拷贝到主目录下的文件夹，方便查找。

注意：由于操作系统为 Ubuntu 16.04，请使用格式为 FAT32 的 U 盘，请勿使用其他格式的 U 盘或者默认为 exfat 格式的移动硬盘。否则系统将不能识别该存储设备。

此外，需要携带完成了作业二的计算机，存储有 (5) 用于 wordcount 的文本文件，用于完成 2, 3 部分。

4.2 登入网络

【以下操作三台计算机均要进行】

点击左上方的火狐浏览器，键入校园网网址 10.248.98.2，输入校园网用户名密码完成机器的联网。打开百度，确认联网成功。

4.3 基础配置

【以下操作三台计算机均要进行】

- (1) 切换到 root 用户，需要键入 root 用户的密码（密码保存在计算机桌面上）。

```
lenovo@687-21:~$ su root
```

- (2) apt 更新。

```
root@687-21:/home/Lenovo/# apt-get update
```

- (3) 查看防火墙状态。

```
root@687-21:/home/Lenovo/# ufw status
```

如果防火墙已经开启，关闭防火墙。防火墙会阻碍 Hadoop 集群内部的通信，在实际生产环境中，单个服务器的防火墙会关闭，通过设置集群整体对外防火墙来保证安全。

```
root@687-21:/home/Lenovo/# ufw disable
```

- (4) 修改主机名。此步并非必须，但是为了更好的区分各台机器，修改计算机主机名。将第一台机器的用户名修改为 **hitsz-master**；第二台机器修改为 **hitsz-slave1**；第三台机器修

改为 **hitsz-slave2**。

```
root@687-21:/home/Lenovo/# hostnamectl set-hostname hitsz-master
```

(5) 安装 vim。

```
root@hitsz-master:/home/Lenovo/# apt-get install vim
```

如果遇到依赖缺失问题，请按照提示键入：

```
root@hitsz-master:/home/Lenovo/# apt --fix-broken install
```

再重新安装 vim。

```
root@hitsz-master:/home/Lenovo/# apt-get install vim
```

4.4 创建用户

【以下操作三台计算机均要进行】

(1) 新增用户 hitsz，提示输入新的 UNIX 密码时需要键入 hitsz 的密码（如 1234），请记住自己设置的密码。其余项一律默认即可。

```
root@hitsz-master:/home/Lenovo/# adduser hitsz
```

```
正在添加用户"hitsz"...
正在添加新组"hitsz" (1001)...
正在添加新用户"hitsz" (1001) 到组"hitsz"...
创建主目录"/home/hitsz"...
正在从"/etc/skel"复制文件...
输入新的 UNIX 密码: 此处输入自定义密码如1234
重新输入新的 UNIX 密码: 此处重复输入自定义密码
passwd: 已成功更新密码
正在改变 hitsz 的用户信息
请输入新值，或直接敲回车键以使用默认值
  全名 []:
  房间号码 []:
  工作电话 []: 一律回车
  家庭电话 []:
  其它 []:
这些信息是否正确? [Y/n] Y 输入Y
```

(2) 添加用户到 sudo 用户组使其具有 sudo 权限。

```
root@hitsz-master:/home/Lenovo/# adduser hitsz sudo
```

注意：为保证部署的顺利进行，降低复杂度，在本教程中，三台机器的新用户名需保持一致，如均为 hitsz。

4.5 配置 Hosts

Hosts 是指域名和 IP 地址的映射，通过配置 Hosts，能够将复杂难记的 IP 地址转换成简单易记的域名，方便配置。在此教程中，将第一台机器 hitsz-master 的 IP 映射为 master，将第二台机器 hitsz-slave1 的 IP 映射为 slave1，第三台机器 hitsz-slave2 的 IP 映射为 slave2。

【以下操作三台计算机均要进行】

(1) 查看各台机器的 IP 地址

```
root@hitsz-master:/home/Lenovo/# ifconfig
```

```

root@ices-slave4:/home/ices# ifconfig
eno1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 10.249.181.15 netmask 255.255.248.0 broadcast 10.249.183.255
    inet6 fe80::4e86:8f01:379d:4e0f prefixlen 64 scopeid 0x20<link>
    ether 7c:d3:0a:c1:da:68 txqueuelen 1000 (以太网)
    RX packets 2237521 bytes 2801484068 (2.8 GB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 1074693 bytes 83200517 (83.2 MB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
    device memory 0xfb200000-fb27ffff

eno2: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500
    ether 7c:d3:0a:c1:da:69 txqueuelen 1000 (以太网)
    RX packets 0 bytes 0 (0.0 B)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 0 bytes 0 (0.0 B)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
    device memory 0xfb100000-fb17ffff

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536
    inet 127.0.0.1 netmask 255.0.0.0
    inet6 ::1 prefixlen 128 scopeid 0x10<host>
    loop txqueuelen 1000 (本地环回)
    RX packets 86627 bytes 8151415 (8.1 MB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 86627 bytes 8151415 (8.1 MB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

```

如果 ifconfig 命令不存在，需要先安装 net-tools。

```
root@hitsz-master:/home/Lenovo/# apt-get install net-tools
```

收集到三台计算机的 IP 地址如下：

计算机名	IP 地址	自定义域名
hitsz-master	10.248.5.21	master
hitsz-slave1	10.248.5.22	slave1
hitsz-slave2	10.248.5.23	Slave2

(2) 修改 host 文件，使用 vim 进入编辑。

注意：此操作至关重要，请仔细检查，确保 IP 地址键入的正确性。

```
root@hitsz-master:/home/Lenovo/# vim /etc/hosts
```

在 vim 界面按 i 进入编辑，在第一行 localhost 下键入以下内容，注意每行中间使用的是 Tab 键进行分隔：

```

10.248.5.21 master
10.248.5.22 slave1
10.248.5.23 slave2

```

IP 地址每台机器都不同，请根据实际情况进行修改。

最终效果如下：

```

127.0.0.1 localhost
10.249.182.54 master
10.249.183.85 slave1
10.249.178.243 slave2

# 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

```

按 Esc 键退出编辑，输入 :wq 命令保存退出。

(3) 重启网络使 hosts 文件生效

```
root@hitsz-master:/home/Lenovo/# /etc/init.d/networking restart
```

重启需稍等片刻，直至看到 ok。

4.6 安装 SSH 并配置 SSH 免密

【以下操作三台计算机均要进行】

(1) 切换至新用户 hitsz。在 su 命令后加“-”符号可以同时变更工作目录。

```
root@hitsz-master:/home/Lenovo/# su - hitsz
```

(2) 安装 ssh 服务器。此处需要键入 hitsz 的用户密码。

```
hitsz@hitsz-master:~$ sudo apt-get install openssh-server
```

(3) 修改 ssh 配置文件（注意是 sshd_config，不是 ssh_config）

```
hitsz@hitsz-master:~$ sudo vim /etc/ssh/sshd_config
```

找到 StrictModes 项，按 i 键进入编辑，删除#号取消注释，并将属性值由 yes 修改为 no。

找到 PermitRootLogin 项，取消注释，并将属性值由 prohibit-password 修改为 yes。

如果找不到，直接在任意位置添加以下内容：

```
PermitRootLogin yes
```

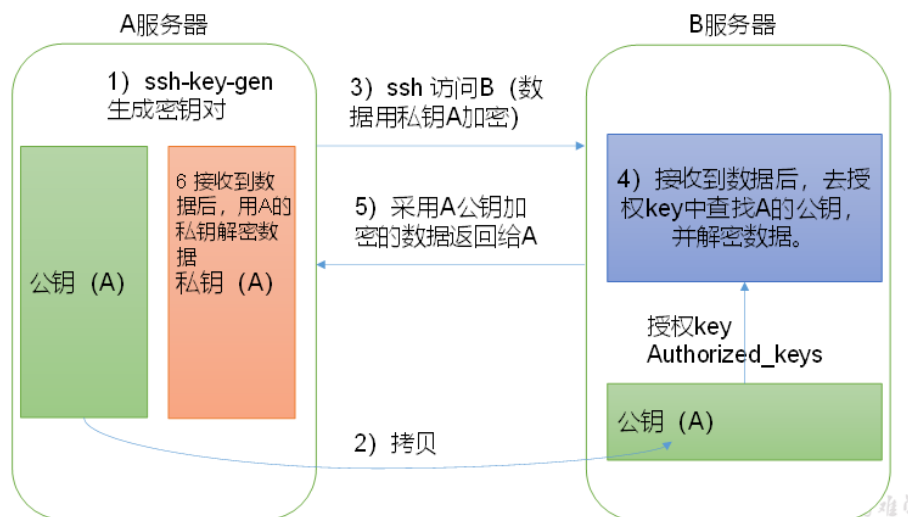
```
StrictModes no
```

按 Esc 退出编辑模式，输入 :wq 保存退出。

(4) 使 ssh 配置文件修改生效（每次修改 ssh 配置文件后都需要重新执行该命令）

```
hitsz@hitsz-master:~$ sudo service sshd reload
```

(5) 免密登录的原理



下面以 hitsz-master 为例演示 ssh 免密登录的操作过程。ssh 免密配置不是为了 hadoop 集群间的数据通信，而是为了传输指令。Hadoop 集群的主节点通过 ssh 协议控制从节点的启停。如果不配置 ssh 免密登录，则需要频繁地输入密码。配置免密登录也有利于后续的集群分发过程。

(6) 生成密钥对

```
hitsz@hitsz-master:~$ ssh-keygen -t rsa
```

然后敲（三个回车），就会生成两个文件 id_rsa（私钥）、id_rsa.pub（公钥），该密钥属于 hitsz-master 下的 hitsz 用户，存储在 /home/hitsz/.ssh 目录下。

可以使用 ssh 登录来验证配置是否成功。通过命令行@后接主机名可以快速判断是否登录成功。

```
root@hitsz-master:~# ssh slave1
```

注意：验证后一定要输入 exit 注销 ssh 登录。

假设当前已经 ssh 免密登录到 hitsz-slave1 上

```
root@hitsz-slave1:~# exit
```

4.7 创建目录

【以下操作三台计算机均要进行】

- (1) 切换到 hitsz 用户

```
root@hitsz-master:~# su - hitsz
```

- (2) 创建目录/opt/module 和/opt/software, 前者是软件的安装目录, 后者用来存放杂项文件。

```
hitsz@hitsz-master:~$ sudo mkdir /opt/module
```

```
hitsz@hitsz-master:~$ sudo mkdir /opt/software
```

- (3) 将新目录所属的用户和用户组均改为 hitsz

```
hitsz@hitsz-master:~$ sudo chown hitsz:hitsz /opt/module
```

```
hitsz@hitsz-master:~$ sudo chown hitsz:hitsz /opt/software
```

- (4) 可以移动到 opt 目录下使用 ll 命令查看是否修改成功

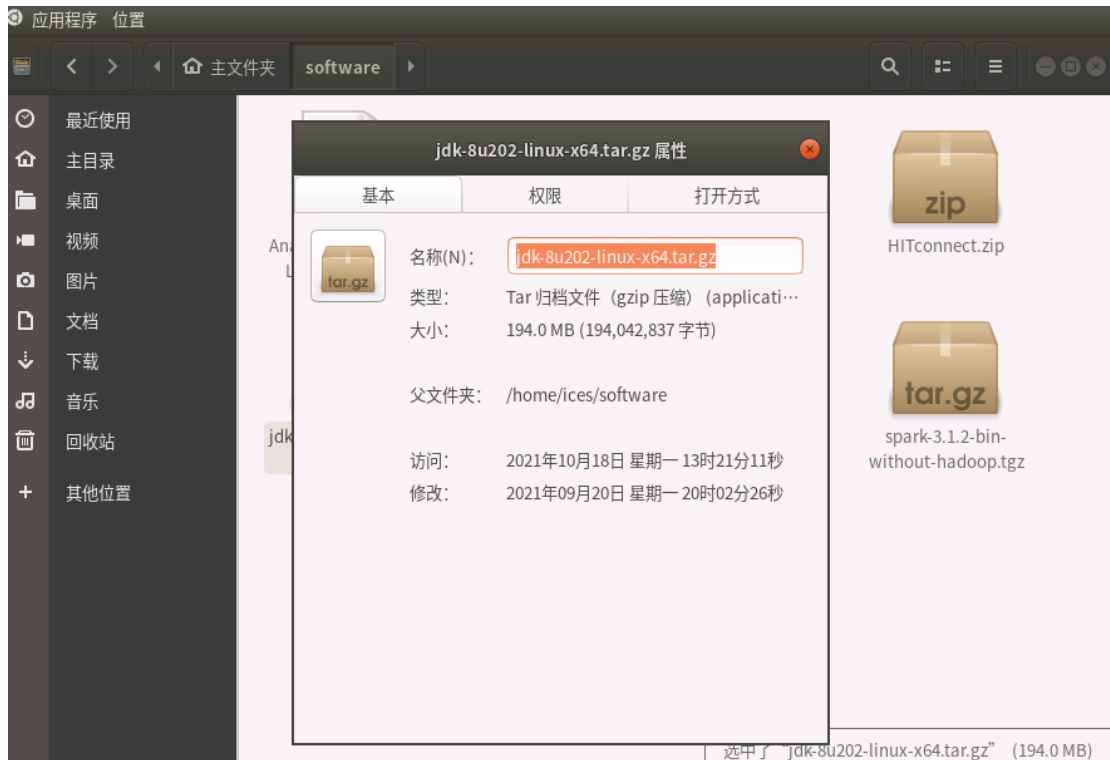
```
hitsz@hitsz-master:~$ cd /opt
```

```
hitsz@hitsz-master:/opt$ ll
```

4.8 安装 jdk

【以下操作在 hitsz-master 上进行】

- (1) 在文件管理器中找到拷贝到计算机上的 jdk 安装包 (最好不要在中文路径下, 因为机器没有安装输入法, 可在文件管理器中将安装包移动到主目录下), 右键, 点击属性即可查看该文件的实际位置。



- (2) 将该压缩包移动到 hitsz 用户的目录下，比如可以移动到/opt/software，jdk 安装包的名称复杂，可以使用 Tab 键进行自动补全。假设安装包完整位置为 /home/lenovo/software/

```
hitsz@hitsz-master:/opt$ sudo mv /home/lenovo/software/jdk-8u202-linux-x64.tar.gz /opt/software
```

- (3) 进入移动后的目录，解压 jdk 到/opt/module

```
hitsz@hitsz-master: /opt$ cd /opt/software
hitsz@hitsz-master: /opt/software $ tar -zxvf jdk-8u202-linux-x64.tar.gz -C /opt/module
```

- (4) 进入解压后的目录，重命名文件夹为 jdk18。该名字可以自定义，此举的目的主要是简化文件夹名称。

```
hitsz@hitsz-master: /opt/software$ cd /opt/module
hitsz@hitsz-master: /opt/module$ mv jdk1.8.202 jdk18
```

- (5) 编辑全局环境变量文件，添加 java 环境变量

```
hitsz@hitsz-master: /opt/module$ sudo vim /etc/profile.d/my_env.sh
```

按下 i 进入编辑模式，添加以下内容：

```
#JAVA_HOME
export JAVA_HOME=/opt/module/jdk18
export PATH=$PATH:$JAVA_HOME/bin
```

按下 Esc 后输入 :wq 保存退出。

- (6) 使环境变量修改生效

```
hitsz@hitsz-master: /opt/module$ source /etc/profile
```

- (7) 验证是否配置成功

```
hitsz@hitsz-master: /opt/module$ java -version
```

```
(base) ices@ices-master:~$ java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
```

出现版本号 1.8.0_202 即为正确。

4.9 安装 hadoop

【以下操作在 hitsz-master 上进行】

以下步骤与 4.8 非常相似。因此不过多赘述。

- (1) 在文件管理器中找到拷贝到计算机上的 hadoop 安装包，右键，点击属性即可查看该文件的实际位置。在该教程中，完整位置为/home/lenovo/software/
- (2) 将该压缩包移动到 hitsz 用户的目录下，比如可以移动到/opt/software。

```
hitsz@hitsz-master:/opt$ sudo mv /home/lenovo/software/j hadoop-3.2.2.tar.gz /opt/software
```

- (3) 进入移动后的目录，解压 hadoop 到/opt/module

```
hitsz@hitsz-master: /opt$ cd /opt/software
hitsz@hitsz-master: /opt/software $ tar -zxvf hadoop-3.2.2.tar.gz -C /opt/module
```

- (4) 进入解压后的目录，重命名文件夹为 hadoop。该名字可以自定义，此举的目的主要是简化文件夹名称。

```
hitsz@hitsz-master: /opt/software$ cd /opt/module
hitsz@hitsz-master: /opt/module$ mv hadoop-3.2.2 hadoop
```

- (5) 编辑全局环境变量文件，添加 hadoop 环境变量

```
hitsz@hitsz-master: /opt/module$ sudo vim /etc/profile.d/my_env.sh
```

按下 i 进入编辑模式，在 java 的环境变量下方添加以下内容：

```
#HADOOP_HOME
export HADOOP_HOME=/opt/module/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
```

按下 Esc 后输入 :wq 保存退出。

- (6) 使环境变量修改生效

```
hitsz@hitsz-master: /opt/module$ source /etc/profile
```

- (7) 验证是否配置成功，注意 version 前面没有“-”。

```
hitsz@hitsz-master: /opt/module$ hadoop version
```

```
(base) ices@ices-master:~$ hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /opt/module/hadoop/share/hadoop/common/hadoop-common-3.2.2.jar
```

出现以上内容即为安装成功。

4.10 hadoop 配置

【以下操作在 hitsz-master 上进行】

- (1) 集群规划

一个典型的集群规划如下：

	hitsz-master	hitsz-slave1	hitsz-slave2
HDFS	NameNode		SecondaryNameNode
	DataNode	DataNode	DataNode
YARN	NodeManager	ResourceManager	NodeManager
	NodeManager	NodeManager	NodeManager

表格中的值代表机器在集群中需要承担的角色。

- (2) 修改配置文件

进入到配置文件所在位置，通过环境变量可以快速的定位 hadoop 安装目录。

```
hitsz@hitsz-master: /opt/module$ cd $HADOOP_HOME/etc/hadoop
```

1. 修改核心配置文件

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim core-site.xml
```

文件内容如下，进入编辑模式删除原有内容在粘贴以下内容即可。Xml 文件无需对齐，但需要保证标签的完整性，请务必复制完整。由于排版原因，指导书中的一个配置文件可能出现分页的情况，请务必注意。

注意配置文件中的 master、slave1 等名称均是指 4.5 中配置的域名。

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <!-- 指定 NameNode 的地址 -->
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:8020</value>
  </property>

  <!-- 指定 hadoop 数据的存储目录 -->
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/module/hadoop/data</value>
  </property>

  <!-- 配置 HDFS 网页登录使用的静态用户为 hitsz -->
  <property>
    <name>hadoop.http.staticuser.user</name>
    <value>hitsz</value>
  </property>
</configuration>
```

2. 修改 HDFS 配置文件

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim hdfs-site.xml
```

文件内容如下：

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <!-- nn web 端访问地址-->
  <property>
    <name>dfs.namenode.http-address</name>
    <value>master:9870</value>
  </property>
  <!-- 2nn web 端访问地址-->
  <property>
    <name>dfs.namenode.secondary.http-address</name>
    <value>slave2:9868</value>
  </property>
</configuration>
```

3. 修改 yarn 配置文件

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim yarn-site.xml
```

文件内容如下:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
    <!-- 指定MR走 shuffle -->
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>

    <!-- 指定 ResourceManager 的地址-->
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>slave1</value>
    </property>

    <!-- 环境变量的继承 -->
    <property>
        <name>yarn.nodemanager.env-whitelist</name>
        <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_
DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HO
ME</value>
    </property>
    <!-- 开启日志聚集功能 -->
    <property>
        <name>yarn.log-aggregation-enable</name>
        <value>true</value>
    </property>
    <!-- 设置日志聚集服务器地址 -->
    <property>
        <name>yarn.log.server.url</name>
        <value>http://hadoop102:19888/jobhistory/logs</value>
    </property>
    <!-- 设置日志保留时间为 7 天 -->
    <property>
        <name>yarn.log-aggregation.retain-seconds</name>
        <value>604800</value>
    </property>
</configuration>
```

4. 修改 MapReduce 配置文件

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim mapred-site.xml
```

文件内容如下:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
    <!-- 指定 MapReduce 程序运行在 Yarn 上 -->
    <property>
        <name>mapreduce.framework.name</name>
```

```

        <value>yarn</value>
    </property>
    <!-- 历史服务器端地址 -->
    <property>
        <name>mapreduce.jobhistory.address</name>
        <value>master:10020</value>
    </property>

    <!-- 历史服务器 web 端地址 -->
    <property>
        <name>mapreduce.jobhistory.webapp.address</name>
        <value>master:19888</value>
    </property>
</configuration>

```

5. 修改 Worker 文件，配置 DataNode 节点

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim workers
```

删除原有的 localhost，加入以下内容：

注意：该文件中添加的内容结尾不允许有空格，文件中不允许有空行。

```

master
slave1
slave2

```

6. 修改 hadoop-env.sh

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ vim Hadoop-
env.sh
```

在任意位置加入以下内容，即再次显示声明 JAVA_HOME：

```
export JAVA_HOME=/opt/module/jdk18
```

4.11 hadoop 集群分发

【以下操作在 hitsz-master 上进行】

(1) 在文件管理器中找到拷贝到计算机上的 xsync 脚本，右键，点击属性即可查看该文件的实际位置。在该教程中，完整位置为/home/ices/software/。

(2) 将该脚本移动到/bin/目录下

```
hitsz@hitsz-master: /opt/module/Hadoop/etc/hadoop$ sudo mv
/home/ices/software/xsync /bin/
```

(3) **如果**你自定义的域名与本教程不同，则需要修改 xsync 文件（红色字体部分），将域名进行替换。如果与本教程一致，则无需修改。

Xsync 文件内容如下：

```

#!/bin/bash

#1. 判断参数个数
if [ $# -lt 1 ]
then
    echo Not Enough Argument!
    exit;
fi

#2. 遍历集群所有机器

```

```

for host in master slave1 slave2
do
    echo ===== $host =====
    #3. 遍历所有目录，挨个发送

    for file in $@
    do
        #4. 判断文件是否存在
        if [ -e $file ]
        then
            #5. 获取父目录
            pdir=$(cd -P $(dirname $file); pwd)

            #6. 获取当前文件的名称
            fname=$(basename $file)
            ssh $host "mkdir -p $pdir"
            rsync -av $pdir/$fname $host:$pdir
        else
            echo $file does not exists!
        fi
    done
done

```

(4) 赋予 xsync 执行权限

```

hitsz@hitsz-master: /bin$ cd /bin/
hitsz@hitsz-master: /bin$ sudo chmod +x xsync

```

(5) 分发 jdk 和 hadoop 安装文件

直接同步整个 /opt/module 文件夹

```

hitsz@hitsz-master: /bin$ xsync /opt/module

```

特别注意：虽然 xsync 可以自动对比文件的差异，只同步机器间不同的文件，但是启动集群后，/opt/module/hadoop 目录下会生成 data 和 logs 两个文件夹，这两个文件夹不同的机器会产生不同的文件，不可以使用 xsync 进行同步，后续如果修改了配置文件，同步配置文件或者配置文件所在文件夹即可。绝对不能直接同步整个 /opt/module/ 或者整个 hadoop 安装目录。如下所示，假设修改了 core-site.xml 文件。

```

hitsz@hitsz-master: /opt/module/hadoop/etc/hadoop$ xsync core-site.xml
或：
hitsz@hitsz-master: /opt/module/Hadoop $ xsync $HADOOP_HOME/etc/hadoop

```

(6) 同步环境变量

```

hitsz@hitsz-master: /bin$ sudo /bin/xsync /etc/profile.d/my_env.sh

```

注意，由于环境变量文件属于 root 用户，所以需要使用 sudo，而在使用 sudo 时，xsync 需要带上绝对路径。

(7) 让环境变量生效

【该操作需要在 hitsz-slave1, hitsz-slave2 上进行】

```

hitsz@hitsz-slave1:~ $ source /etc/profile
hitsz@hitsz-slave2:~ $ source /etc/profile

```

4.12 启动 Hadoop 集群

【以下操作在 hitsz-master 上进行】

(1) **如果集群是第一次启动**，需要在 hitsz-master 节点格式化 NameNode（注意：格式化 NameNode，会产生新的集群 id，导致 NameNode 和 DataNode 的集群 id 不一致，集群找不到已往数据。如果集群在运行过程中报错，需要重新格式化 NameNode 的话，

一定要先停止 namenode 和 datanode 进程,并且要删除所有机器的 data 和 logs 目录,然后再进行格式化。)

```
hitsz@hitsz-master: /bin$ cd $HADOOP_HOME
hitsz@hitsz-master: /opt/module/hadoop $ hdfs namenode -format
```

(2) 启动 hdfs

```
hitsz@hitsz-master: /opt/module/hadoop $ sbin/start-dfs.sh
```

(3) 在配置了 ResourceManager 的节点 (hitsz-slave1) 上启动 yarn

```
hitsz@hitsz-slave1: /bin$ cd $HADOOP_HOME
hitsz@hitsz-slave1: /opt/module/hadoop $ sbin/start-yarn.sh
```

(4) 启动历史记录服务器 (hitsz-master)

```
hitsz@hitsz-master: /opt/module/hadoop $ mapred --daemon start
historyserver
```

(5) 验证是否正常启动, 在每台机器使用 jps 命令查看

```
hitsz@hitsz-master: /opt/module/hadoop $ jps
```

每台机器的输出结果应该等同于 4.9 节集群规划中的内容。

查看 HDFS WebUI 界面, 点击 DataNode, 截图此界面 In Operation 部分。

<http://master:9870>

In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ ices-master:9866	http://ices-master:9864	1s	0m	1.79 TB	32	986.53 MB (0.05%)	3.2.2
✓ ices-slave1:9866	http://ices-slave1:9864	1s	0m	1.79 TB	32	986.53 MB (0.05%)	3.2.2
✓ ices-slave2:9866	http://ices-slave2:9864	1s	0m	1.79 TB	32	986.53 MB (0.05%)	3.2.2

Showing 1 to 3 of 3 entries

Previous 1 Next

4.13 配置 Hadoop 客户端

(1) 将服务端安装目录下 etc/hadoop 子目录下的四个配置 4 个配置文件替换客户端安装目录下 etc/hadoop 子目录下的文件。4 个文件分别为 core-site.xml, hdfs-site.xml, yarn-site.xml, mapred-site.xml。

(2) 配置 Host 文件

以 windows 客户端为例, macos 系统请自行百度, 在此不赘述。将 4.5 配置的 3 行 host 复制到 C:\Windows\System32\drivers\etc\hosts 文件内, 需要管理员权限, 建议先保存到桌面, 再从桌面复制到文件夹内进行覆盖。在 cmd 内输入

```
C:\> ipconfig /flushdns
```

4.14 上传文本文件

以 windows 客户端为例, 假设文本文件为 str.txt, 位于 D 盘下。

(1) 新建目录, 如果目录为多级, 需要加 -p 参数

```
D:\> hadoop fs -mkdir -p /user/hitsz/input
```

(3) 上传文件

```
D:\> hadoop fs -put D:\str.txt /user/hitsz/input
```

4.15 执行 WordCount 程序

此部分可以查看视频教程。

- (1) 将 core-site.xml, hdfs-site.xml, yarn-site.xml, mapred-site.xml 四个配置文件添加到 resource 目录下。
- (2) 在 WordCountDriver 文件中添加以下配置代码,

```
//设置允许跨平台运行
conf.set("mapreduce.app-submission.cross-platform", "true");
//设置用户名
System.setProperty("HADOOP_USER_NAME", "hitsz");
Job job = Job.getInstance(conf);
//关联 jar 包, 此处应该填写自己的 jar 包位置。
job.setJar("E:\\workplace\\idea\\test\\hadoopdemo\\out\\artifacts\\wc_jar\\hadoopdemo.jar");
```

- (3) 重新 build 项目生成 jar 包。
- (4) 修改运行配置, Edit Configuration -> Program arguments, 填写输入输出路径, 两个路径之间用空格分割。Output 路径必须不存在。

```
hdfs://master:8020/user/hitsz/input
hdfs://master:8020/user/hitsz/output01
```

- (5) 点击运行
- (6) 下载查看运行结果

```
D:\> hadoop fs -get /user/hitsz/output01/part-r-00000 D:\
```

```
极写 1
极冷 16
极准 7
极出 1
极刑 26
极则必反 1
极则辱 1
极制 1
极前 1
极力 709
极力争取 3
极力回避 5
极力推荐 165
极加 1
极劣 2
极劲 1
极劲 1
极化 10
极化处理 1
极北 6
极华春莹 1
极卡 1
极压 1
极庆 1
极反转 1
极叔 1
极受 4
极右 87
极右份子 2
极右派 5
极右翼 25
极舍 3
极后 1
极味 1
极命 1
极品 1304
极品飞车 81
```

626296,1 62%

4.16 Hadoop 集群关闭

注意: 该部分了解即可

如果存在问题需要关闭集群，按照启动的相反顺序依次进行关闭。

(1) 关闭历史记录服务器 (hitsz-master)

```
hitsz@hitsz-master: /opt/module/hadoop $ mapred --daemon stop  
historyserver
```

(2) 在配置了 ResourceManager 的节点 (hitsz-slave1) 上关闭 yarn

```
hitsz@hitsz-slave1: /bin$ cd $HADOOP_HOME  
hitsz@hitsz-slave1: /opt/module/hadoop $ sbin/stop-yarn.sh
```

(3) 关闭 hdfs (hitsz-master)

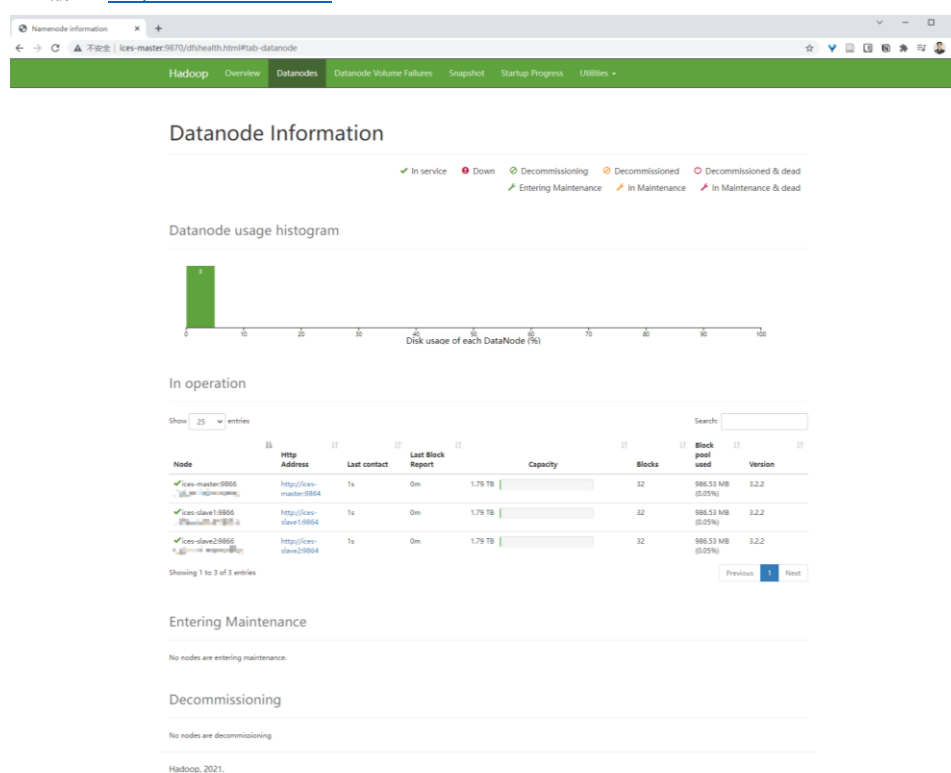
```
hitsz@hitsz-master: /opt/module/hadoop $ sbin/stop-dfs.sh
```

4.17 Hadoop 集群 WebUI 界面

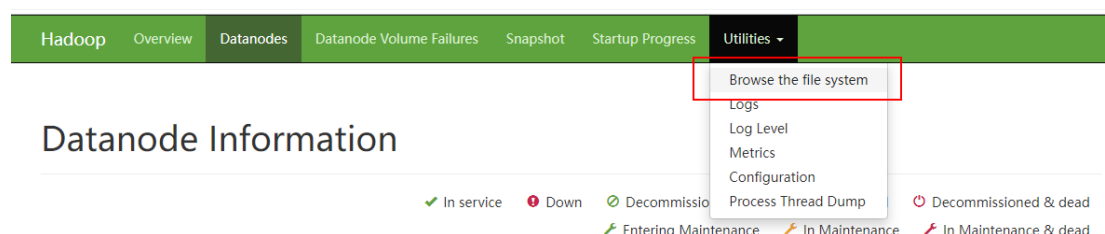
注意：该部分了解即可

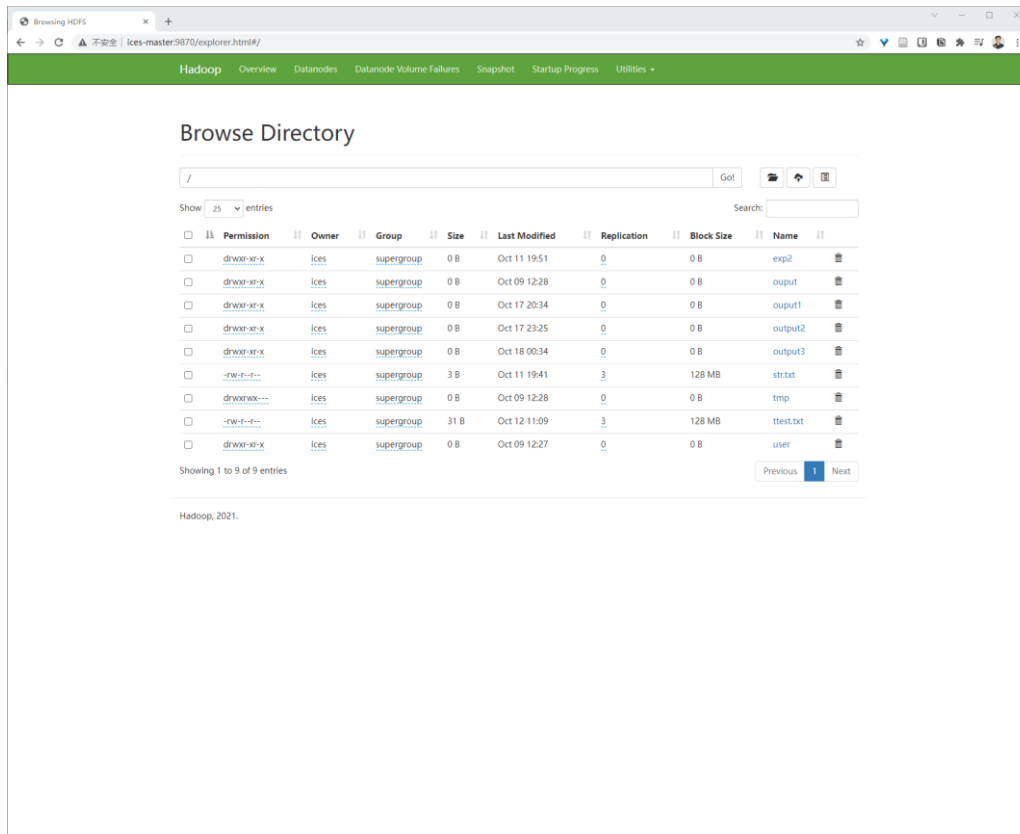
(1) Web 端查看 HDFS 的 NameNode

浏览器中输入 <http://master:9870>

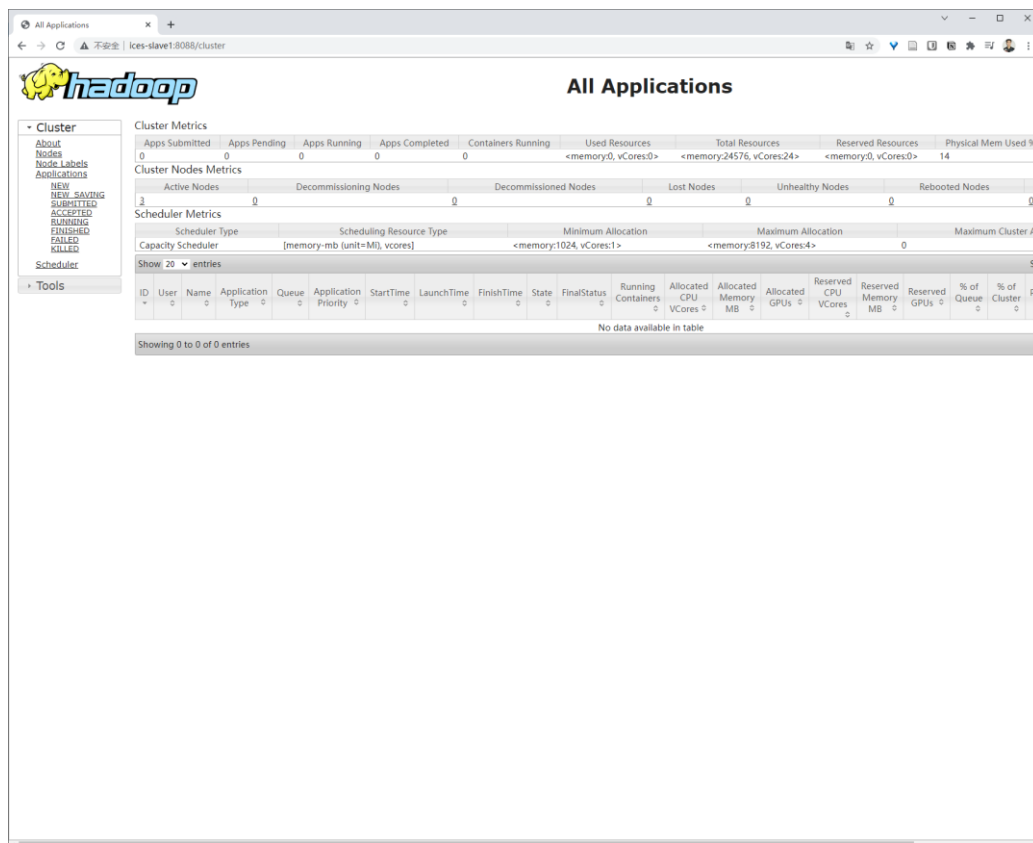


点击 Utilities 下的 Browse the file system 可以直接访问集群的 HDFS 系统。

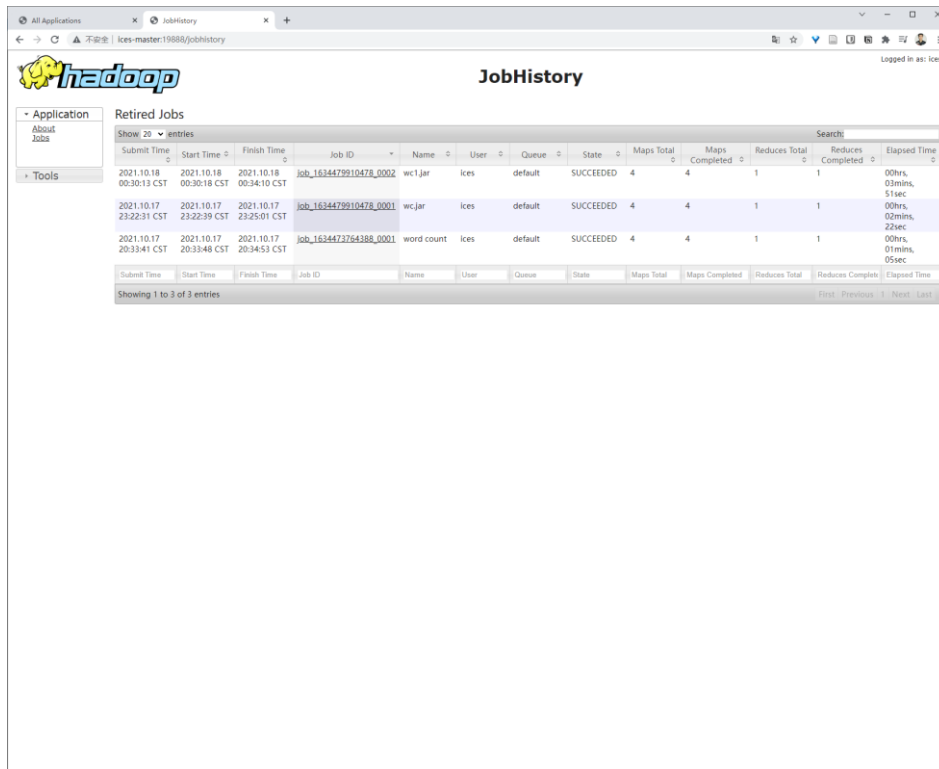




(2) Web 端查看 YARN 的 ResourceManager
浏览器中输入 <http://slave1:8088>



(3) Web 端查看 Mapreduce 运行历史记录
浏览器中输入 <http://master:19888/jobhistory>



The screenshot shows the Hadoop JobHistory web interface. The title is "JobHistory" and it says "Logged in as: ices". On the left, there is a sidebar with "Application" (About, Jobs) and "Tools". The main content area is titled "Retired Jobs" and shows a table of jobs. The table has columns: Submit Time, Start Time, Finish Time, Job ID, Name, User, Queue, State, Maps Total, Maps Completed, Reduces Total, Reduces Completed, and Elapsed Time. There are three rows of job data, all with a state of "SUCCEEDED".

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
2021.10.18 00:30:13 CST	2021.10.18 00:30:18 CST	2021.10.18 00:34:10 CST	Job_1634479910478_0002	wc1.jar	ices	default	SUCCEEDED	4	4	1	1	00hrs, 01min, 51sec
2021.10.17 23:22:31 CST	2021.10.17 23:22:39 CST	2021.10.17 23:25:01 CST	Job_1634479910478_0001	wc.jar	ices	default	SUCCEEDED	4	4	1	1	00hrs, 02min, 22sec
2021.10.17 20:33:41 CST	2021.10.17 20:33:49 CST	2021.10.17 20:34:53 CST	Job_1634473764388_0001	word count	ices	default	SUCCEEDED	4	4	1	1	00hrs, 01min, 05sec

Showing 1 to 3 of 3 entries

5. 常见错误和解决方案

(1) 集群部分正常，部分无法启动
此错误主要有两个原因，配置文件出错以及 Hosts 文件出错。请仔细检查配置文件和 hosts 文件。
可以查看对应的日志来查找问题的所在。比如 secondlynamenode 没有启动，在 hitsz-slave2 上查看日志

```
hitsz@hitsz-master: ~ $ cd $HADOOP_HOME/logs
hitsz@hitsz-master: ~ $ vim hadoop-ices-secondarynamenode-ices-slave2.log
```

(2) hitsz-master 可以正常启动，从属节点无法启动
考虑 SSH 配置和 Hosts 文件配置的问题。

(3) 端口号被占用导致的无法启动
正常情况下本教程所使用的端口在机房环境下均不会出现端口占用的问题。在这种情况下，尝试关闭集群，重启机器，再启动集群，必要时重新格式化 namenode。如果问题依旧存在，基本为以下两个原因。

1. 配置文件出错，端口重复使用了。
2. Hosts 文件出错，将两个域名映射到了同一个 IP 地址。

(4) Java.net.UnknownHostException: master:master
重新配置 Hosts 文件。同时主机名称和域名不要使用 hadoop hadoop000 等特殊名称。

(5) 命令不存在的问题

检查环境变量的设置，以及修改环境变量后是否使用 source 命令使之生效。

(6) vim 环境下输入命令无效

检查输入法，vim 相关命令均需要在英文环境下输入。

(7) vim 误删

按下 Esc, 按 u 进行回退。

或者按下 Esc, 输入 :q! 退出，该操作会完全撤销这次的 vim 编辑。

(8) jps 不生效

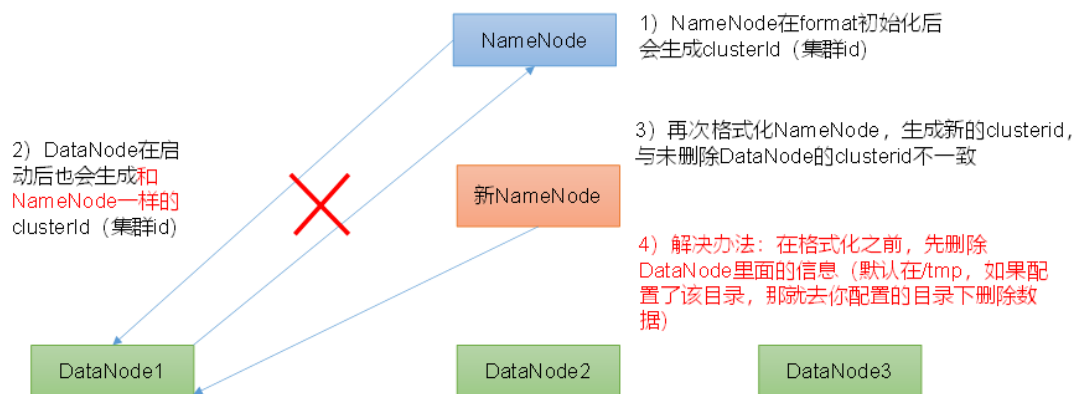
全局变量 hadoop java 没有生效。解决办法：检查 java 环境变量是否设置正确，同时 source /etc/profile 文件。

(9) jps 发现进程已经没有，但是重新启动集群，提示进程已经开启。

原因是在 Linux 的根目录下/tmp 目录中存在启动的进程临时文件，将集群相关进程删除掉，再重新启动集群。

(10) DataNode 和 NameNode 进程同时只能工作一个

DataNode和NameNode进程同时只能有一个工作问题分析



如果无法判断问题的原因，一直无法解决问题，可以考虑删除 /opt/module 目录，删除环境变量 my_env.sh，删除 hosts 文件，删除所有.ssh 文件夹，重启机器再进行重新配置。