大数据课程学习笔记

# Hadoop整体介绍及集群搭建

## 大数据简介

### 数据

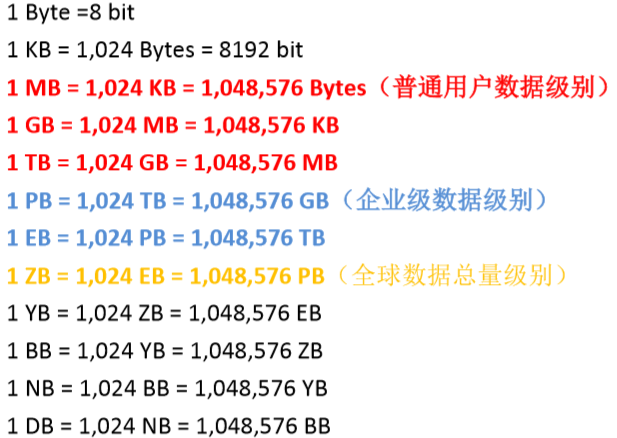
数据就是数值，也就是我们通过观察、实验或计算得出的结果。数据有很多种，最简单的就是数字。数据也可以是文字、图像、声音等。在计算机系统中，数据以二进制信息单元 0,1 的形式表示。

### 大数据

大数据（big data），指的是在一定时间范围内不能以常规软件工具处理（存储和计算）的大而复杂的数据集。

**数据衡量单位**

* 从小到大排列：bit、Byte、KB、MB、GB、TB、PB、EB、 ZB、YB、BB、NB、DB



据国际数据公司(IDC)统计，全球数据总量预计 2020 年达到 44ZB，中国数据量将达到 8060EB， 占全球数据总量的 18%

**大数据的特点**

* 容量（Volume）：数据的大小决定所考虑的数据的价值和潜在的信息
  + 新浪微博，3 亿用户，每天上亿条微博
  + 朋友圈，8 亿用户，每天亿级别朋友圈
* 种类（Variety）：数据类型的多样性，包括文本，图片，视频，音频
  + 结构化数据：可以用二维数据库表来抽象，抽取数据规律
  + 半结构化数据：介于结构化和非结构化之间，主要指 XML，HTML 等，也可称非结构化
  + 非结构化数据：不可用二维表抽象，比如图片，图像，音频，视频等
* 速度（Velocity）：指获得数据的速度以及处理数据的速度
  + 数据的产生呈指数式爆炸式增长
  + 处理数据要求的延时越来越低
* 价值（Value）：合理运用大数据，以低成本创造高价值
  + 单条数据记录无价值，无用数据多
  + 综合价值大，隐含价值大

一句话：容量大，种类多，速度快，价值高

## Hadoop整体概念

### 产生背景

* Hadoop 是Apache Lucene 创始人 Doug Cutting 创建的，最早起源于 Apache Nutch项目。Nutch 的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，遇到了严重的可扩展性问题 ——如何解决数十亿网页的存储和索引问题
* 2003 年、2004 年谷歌发表的三篇论文为该问题提供了可行的解决方案
  + 分布式文件系统 GFS，可用于处理海量网页的存储
  + 分布式计算框架 MapReduce，可用于处理海量网页的索引计算问题
  + 分布式数据库 BigTable，每一张表可以存储上 billions 行和 millions 列
* Nutch 的开发人员完成了相应的开源实现 HDFS 和 MapReduce，并从 Nutch 中剥离成为独立项目 Hadoop，到 2008 年 1 月，Hadoop 成为 Apache 顶级项目，迎来了它的快速发展期

### 什么是Hadoop

Hadoop官网：<https://hadoop.apache.org/>

Hadoop 是 Apache 旗下的一套开源分布式软件平台，用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用集群的威力进行高速运算和存储。你可以把 Hadoop 理解为一个分布式的操作系统。

什么是**分布式程序**？

* 该软件系统会划分成多个子系统或模块，各自运行在不同的机器上，子系统或模块之间通过网络通信进行协作，实现最终的整体功能，这多个机器就构成了**集群**。通俗的讲，分布式系统就是利用集群的多个节点共同协作完成一项或多项具体业务功能的系统。

Hadoop 的**核心组件**

* Common（基础功能组件：工具包，RPC 框架）
* HDFS（Hadoop Distributed File System 分布式文件系统）
* MapReduce（分布式运算编程框架）
* YARN（Yet Another Resources Negotiator 运算资源调度系统）

广义的Hadoop指一个更广泛的概念——**Hadoop 生态圈** ：

* HDFS：Hadoop 的分布式文件存储系统
* MapReduce：Hadoop 的分布式程序运算框架，也可以叫做一种编程模型
* YARN：Hadoop 的资源调度系统
* Hive：基于 Hadoop 的类 SQL 数据仓库工具
* HBase：基于 Hadoop 的列式分布式 NoSQL 数据库
* ZooKeeper：分布式协调服务组件
* Oozie/Azkaban：工作流调度引擎
* Sqoop：数据迁入迁出工具
* Flume：日志采集工具
* ……

## Windows下安装Hadoop

### 安装jdk1.8.0并配置环境

**下载地址：**

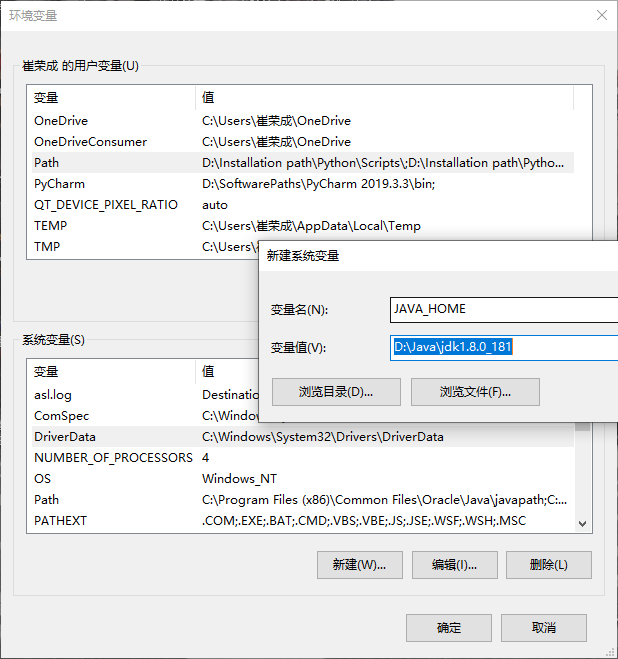
<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

下载对应的jdk

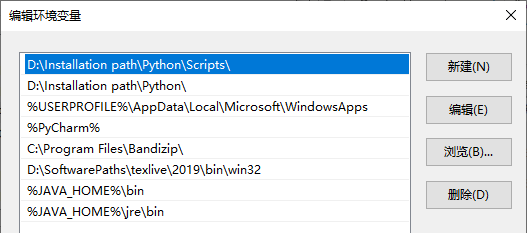


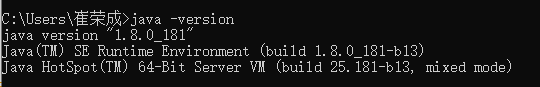
**配置环境变量：**

新建JAVA\_HOME



找到Path，编辑，在最后面加上;%JAVA\_HOME%\bin;%JAVA\_HOME%\jre\bin;





Cmd命令测试，安装成功。

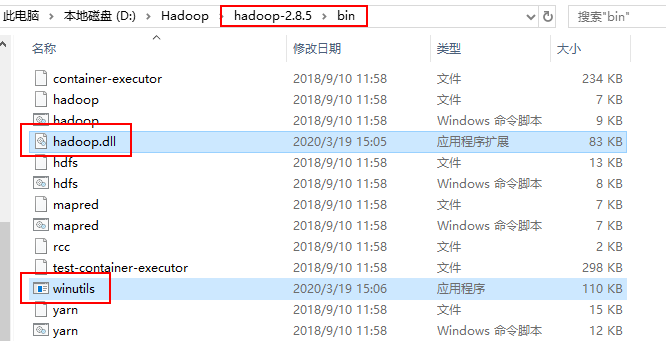
### 安装Hadoop2.8.5，并配置环境

**1.下载Hadoop2.8.5并解压到本地**

下载链接<https://mirrors.tuna.tsinghua.edu.cn/apache/hadoop/common/>

**2.下载对应版本的配置文件**

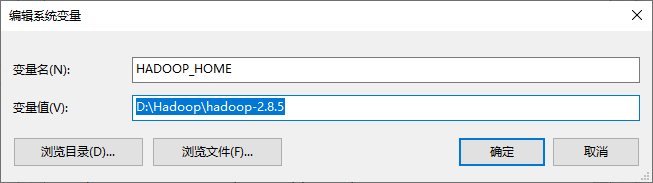
下载链接<https://github.com/cdarlint/winutils>



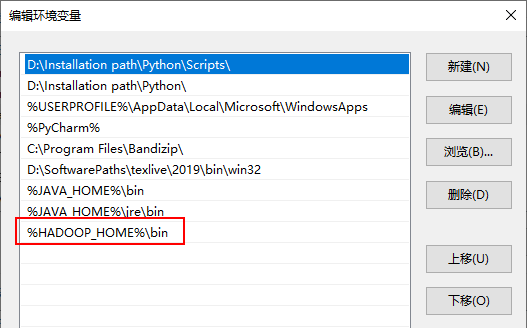
关键是将hadoop.dll和winutils.exe两个文件放入到/bin目录下

**3.配置hadoop环境变量**

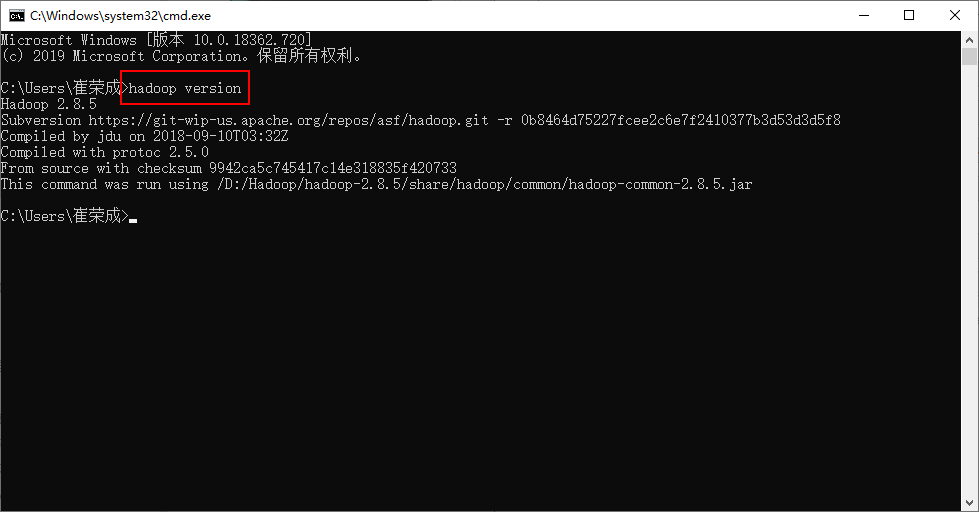
新建HADOOP\_HOME



添加path属性,将;%HADOOP\_HOME%\bin;添加到path环境变量中



验证hadoop环境,在windows控制台输入: hadoop version 结果如下图:



**4.修改Hadoop配置文件**

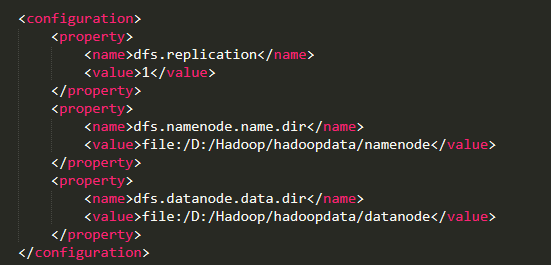
打开Hadoop文件夹下的\etc\hadoop目录对配置文件进行修改，

**hadoop-env.cmd**

找到如下代码（26行），将路径修改为你自己的JAVA\_HOME路径



**hdfs-site.xml**



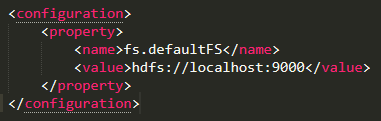
namenode的文件路径：

D:/Hadoop/hadoopdata/namenode

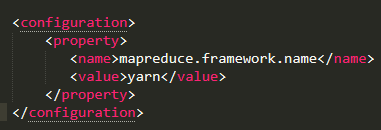
datanode的文件路径：

D:/Hadoop/hadoopdata/datanode

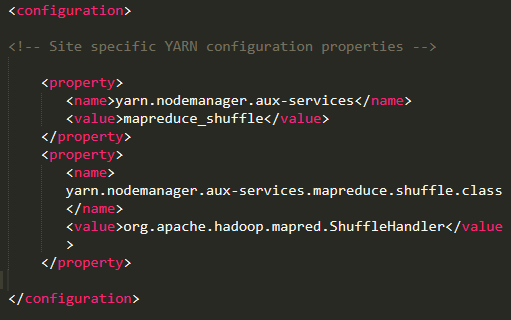
**core-site.xml**



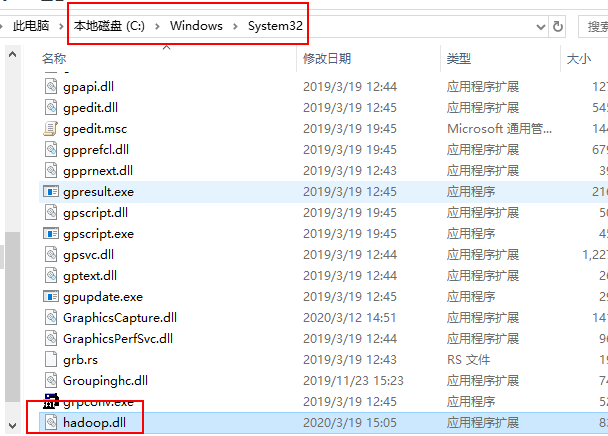
**mapred-site.xml**



**yarn-site.xml**



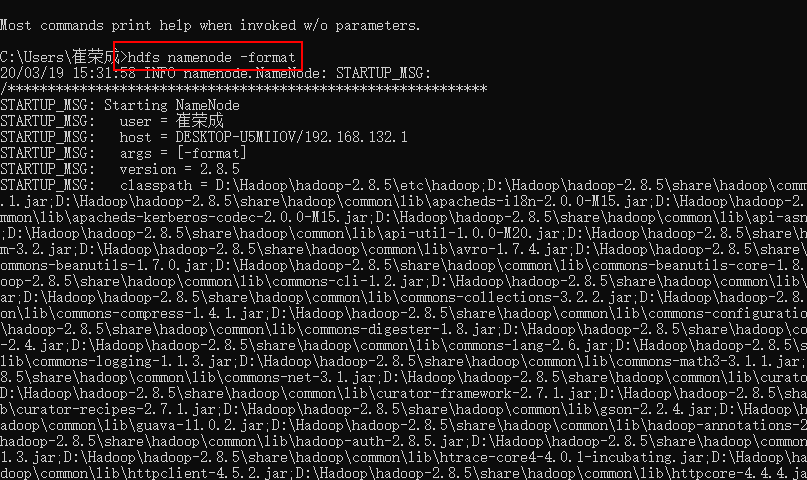
**5.** **将Hadoop文件夹\bin下的hadoop.dll复制到C:\Windows\System32**



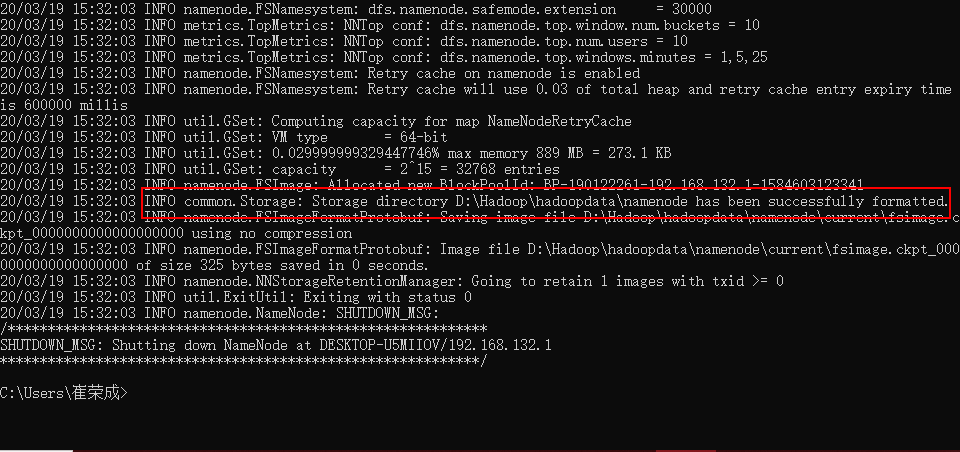
**6.初始化HDFS**

运行如下命令：

hdfs namenode -format



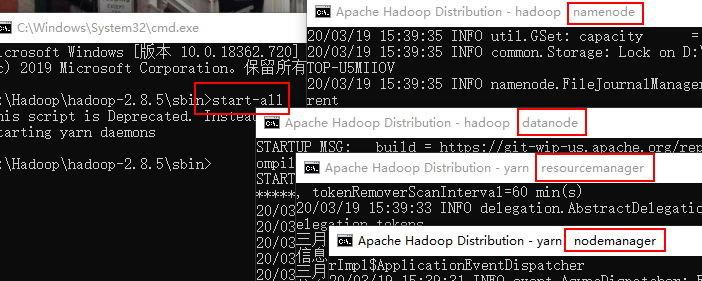
初始化成功



**7.启动Hadoop**

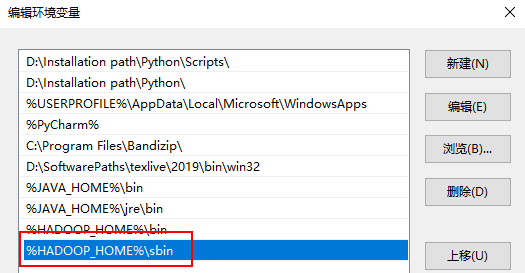
进入hadoop目录下sbin，执行启动命令：

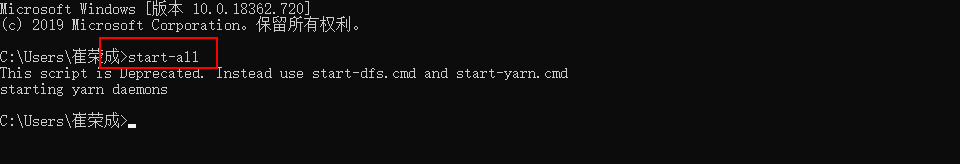
start -all



弹出其他四个控制台输入框说明启动成功。

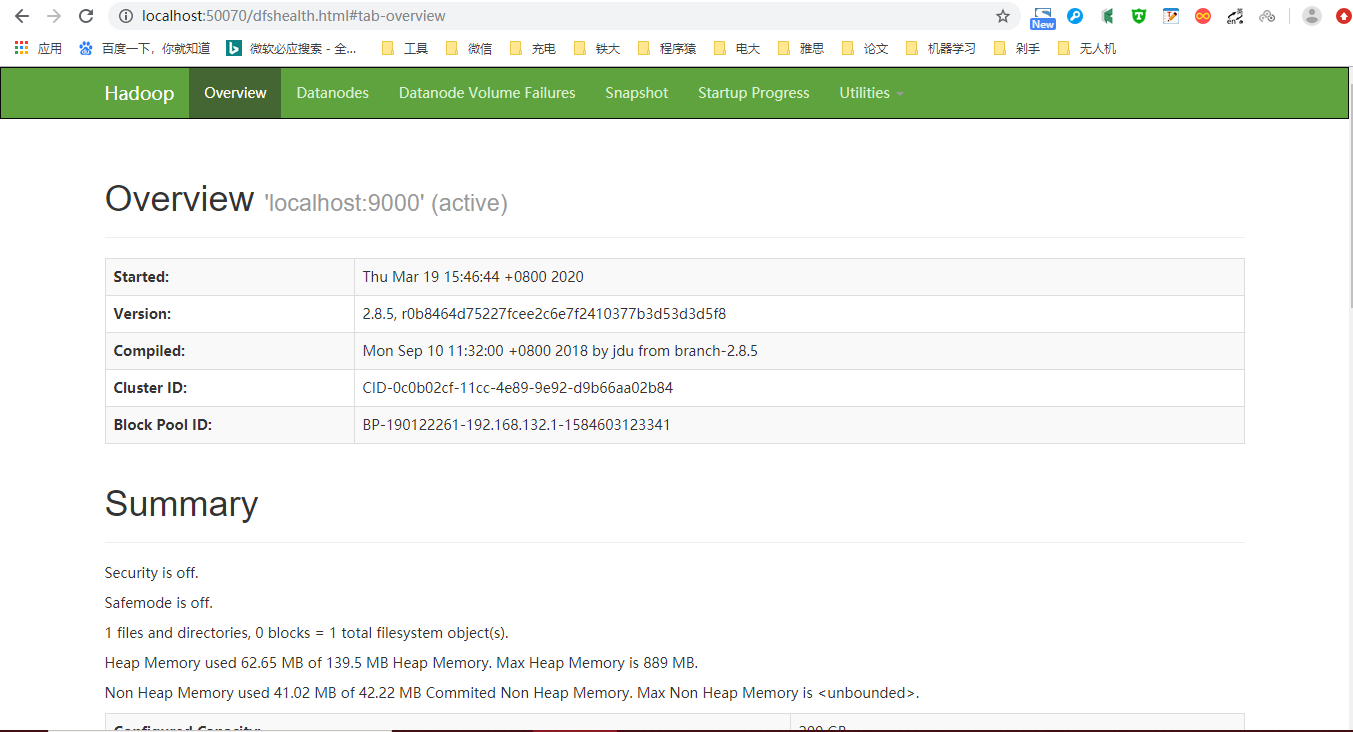
如果把hadoop的sbin也放到系统Path变量，可以直接从系统cmd打开



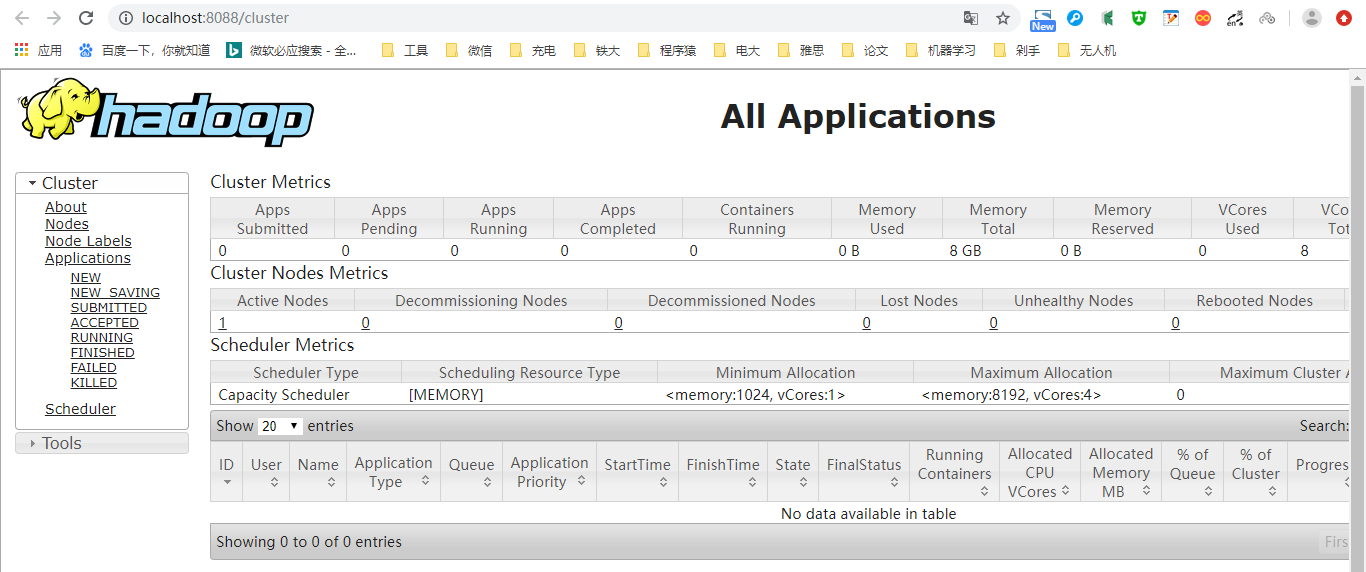


进入网页查看：

[localhost:50070](http://localhost:50070/)

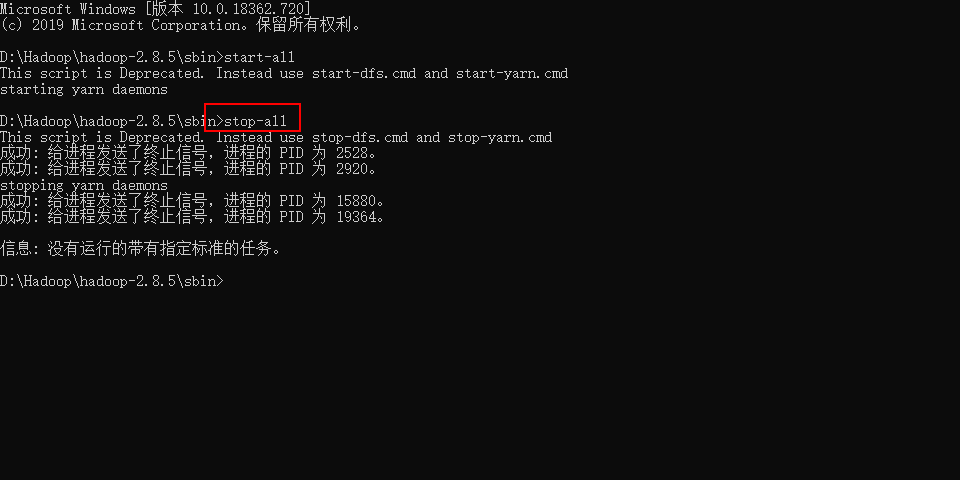


[localhost:8088](http://localhost:8088/)



**8.关闭Hadoop**

关闭命令：stop-all



## linux下安装Hadoop

### 安装JDK环境

1.9节中已经安装

### 安装Hadoop

**1.准备工作**

集群规划

|  | **hadoop1** | **hadoop2** | **hadoop3** |
| --- | --- | --- | --- |
| NameNode | √ | × | × |
| SecondryNameNode | × | √ | × |
| DataNode | √ | √ | √ |
| ResourceManager | × | × | √ |
| NodeManager | √ | √ | √ |

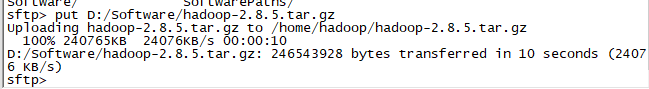
两种思路：  
**1）** 先准备好3台空机器，再把所有需要的软件挨个安装一次,需要修改的配置挨个修改一次  
**2）** 先准备好1台空机器，把所有需要的软件安装一次，把通用的配置改好，再来克隆它，这样我们就可以少做很多重复性工作。  
这里，我们采用第二种思路:

2.上传hadoop安装包，解压到规划的安装目录

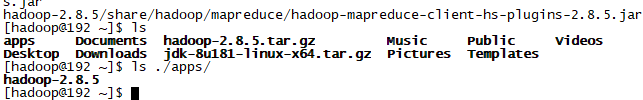
切换到 hadpoop 用户，并打开sftp

pwd 查看工作目录，不满意当前目录就使用 cd 切换

put windos上hadoop安装包路径

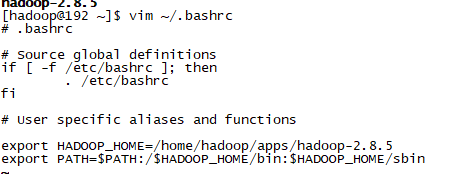


tar -zxvf hadoop-2.8.5.tar.gz -C ~/apps/（-C 的意思是指定解压路径，目标路径不存在则需要先创建：mkdir ~/apps）



3. 配置环境变量

vim ~/.bashrc进入文件



在最后面加入以下语句：

export HADOOP\_HOME=/home/hadoop/apps/hadoop-2.8.5  
export PATH=$PATH:/$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin  
然后保存并退出vi编辑器

source ~/.bashrc（使用该指令重新加载一次环境变量，使新配置的环境变量生效）

4. 配置 hadoop 用户 sudoer 权限（之前已配置好，可忽略）

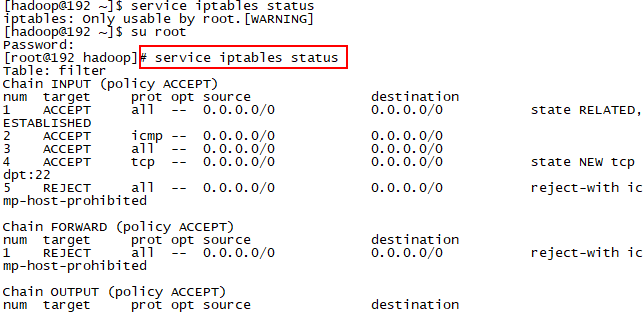
在 root 用户下，命令行输入：vi /etc/sudoers

找到 root ALL=(ALL) ALL 这一行，然后在它下面添加一行：  
hadoop ALL=(ALL) ALL

wq!（保存，退出）

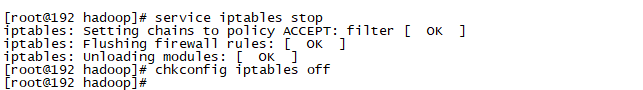
5. 关闭防火墙

service iptables status （查看）



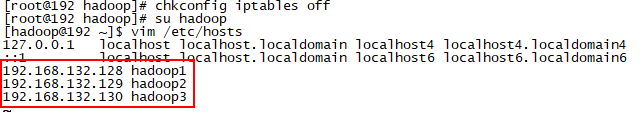
service iptables stop （关闭）

chkconfig iptables off （关闭开机自启）



6.配置主机映射

vim /etc/hosts

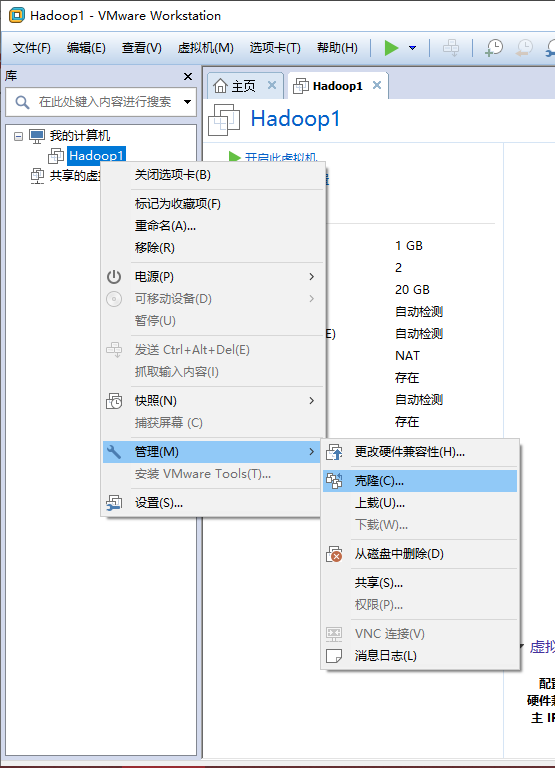


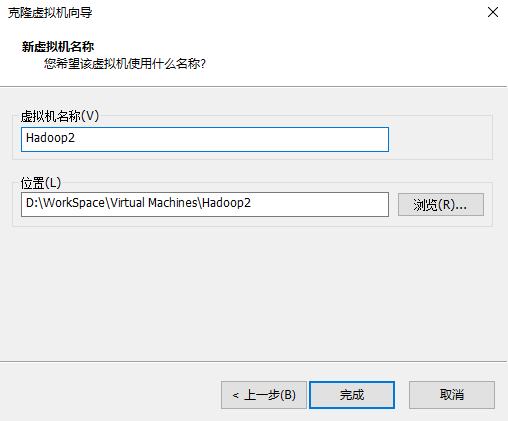
加入对应的映射

保存并退出vi编辑器

7. 接下来，关机 ，克隆虚拟机

在 VMware Workstation 库面板里选中要克隆的虚拟机，右键→管理→克隆，下一步下一步，选中“创建克隆链接”，然后继续下一步，指定好保存的位置及名字，点击完成。重复这个过程，克隆到你规划的数量为止。







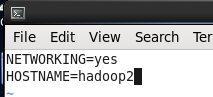
虚拟机克隆完毕之后，我们原始的虚拟机就可以不使用了（最好不要使用，避免被污染）。

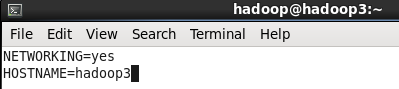
8.克隆虚拟机配置

对于克隆好的虚拟机，有一些麻烦必须要解决：主机名及ip地址

修改主机名：

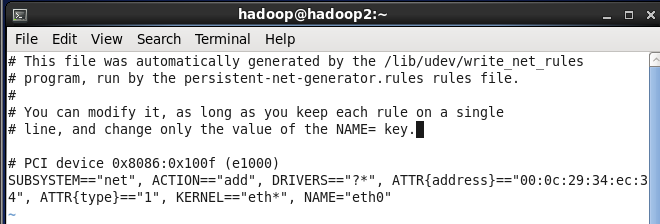
sudo vi /etc/sysconfig/network ，编辑hostname的值





ip问题：

vi /etc/udev/rules.d/70-peresitent-net.rules，将 eth0 删除，将 eth1 修改为eth0

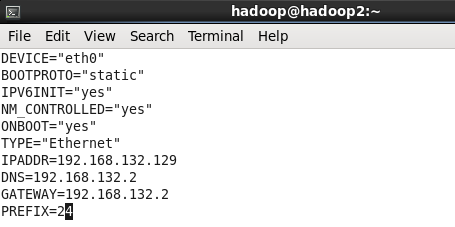


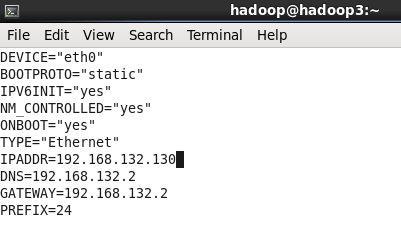
sudo vi /etc/sysconfig/network-scripts/ifcfg-eth0

删除里面的 hwaddr 和 uuid

修改 bootproto 为 static

添加 IPADDR=192.168.132.129/130，DNS1=192.168.25.2，GATEWAY=192.168.25.2，PREFIX=24

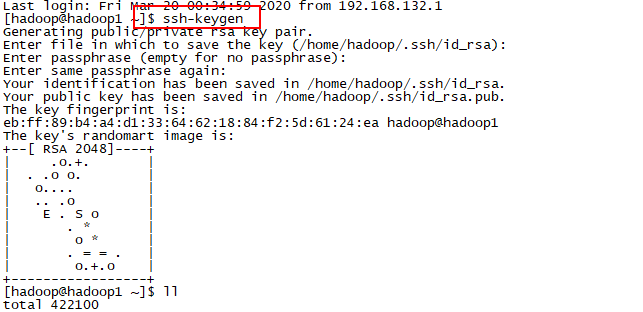


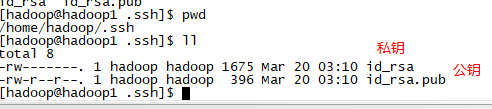


ip 和 hostname 要修改成与主机映射文件一致

9.配置免密登录

在 hadoop 用户下，输入命令 ssh-keygen ，连按 3 次回车，之后你会发现，在/home/hadoop/.ssh 目录下生成了一对密钥





在 hadoop1 的 hadoop 用户家目录使用 ssh-copy-id hadoop1 ，这条指令的意思是复制 ssh公钥到 hadoop1 的 .ssh/authorized\_keys 文件里，这个文件就是免密登录的关键文件