大数据课程学习笔记

# Linux安装及基本操作

## Linux背景简介

Linux是一种得到广泛应用的计算机操作系统。由芬兰人Linus Torvalds最早开发。第一次正式公布时间是1991年10月5日。

通常提到“Linux”一词时，并不是单纯的值操作系统，而是泛指已下三个部分的总体概念。

1.一种类UNIX，名为“Linux”的计算机操作系统

2.Linux操作系统环境下的开发工具和开发环境

3.Linux操作系统下的个中国应用软件和工具软件

Linux与GNU计划密不可分。GNU是Richard Stallman在1984年创立，其目标是发展一个完全免费的自由软件——一种类UNIX计算机操作系统以及运行在其上的软件开发工具和各种应用程序。实际发展中，GNU自己的操作系统内核并没用真正的开发出来并应用，但是大量GNU软件（包括开发工具，调试工具和各种个应用软件等）与Linux操作系统完美融合在一起。

Linux从最开始就是一个开放的系统，并且始终遵循源代码开放的原则，是一个成熟而稳定的网络操作系统。优势有：

1.开放性

2.多用户

3.多任务

4.出色的稳定性和速度性能

5.丰富的网络性能

6.可靠的系统安全

7.可移植性：适用于大多数计算机平台

Linux的发展一方面得益于开放源码的措施。众多计算机爱好者和开发者的不懈努力。另一方面得益于众多像Red Hat这样的商业软件公司，加快其商品化的步伐。

常见的Linux发行版：Debian   ，  Ubuntu   ，  Fedora    ，Red Hat  ,CentOS, SUSE  国内的 红旗，新华

## 安装虚拟机（VMware14）（大小限制，此处省略很多截图）

下载VMwareWorkstation软件

下载链接：链接：https://pan.baidu.com/s/1DhEqfV36tJAesN1zUaJ6eg

提取码：c2zb



下载之后双击安装

此处安装路径为：D:\SoftwarePaths\VMware\

输入许可证：ZC5XK-A6E0M-080XQ-04ZZG-YF08D进行激活，完成安装。



安装好之后打开虚拟机

添加CentOS镜像文件

设置用户名和密码

输入之前设置的密码即可进入对应Linux

## Linux文件系统

文件系统格式

Linux：ext4、ext3、xfs（查看分区文件系统：parted -l df -T）

Windows：NTFS、fat32

常见Linux结构目录

1、/bin：二进制(binary) 可执行的文件命令存放其中

2、/boot：Linux启动的时候使用的核心文件

3、/dev：device 设备文件目录

4、/etc：系统和应用程序的配置信息

5、/home：用户信息文件目录，子目录的个数通常对应用户的个数

6、/lib：library库文件目录

7、/media：可移动设备挂载点 U盘

8、/mnt：外部设备挂载点，光盘和软盘

9、/opt：默认安装软件的位置

10、/root：根用户的文件信息

11、/sbin：超级用户的可执行命令

12、/tmp：临时文件

13、/usr：静态的用户级别程序/usr/include

## Linux基本指令

命令行含义：

[用户名@主机名 当前目录]#

#root用户，$普通用户

常用shell命令：

Su（su 用户名，切换用户）

Pwd（显示当前工作目录）

Cd（改变工作目录）

Cd path（进入指定目录）

Cd ..（返回上级）

Cd ~（返回用户主目录）

Cd -（返回上次操作目录）

Ls（ls [参数]路径，显示目录内容列表）

-l（详细信息显示，可用ll）

-a（显示所有子目录和文件信息，包括隐藏）

-A（类似-a，不现实“.”和“..”目录的信息）

-R（递归显示内容）

Touch（touch路径，创建空文件或者更新文件的时间标记）

Mkdir（mkdir[参数]路径，创建文件夹）

-p（级联的方式创建文件夹，上级目录不存在自动创建）

File（file路径，查看文件类型）

Cp（cp[参数] 源文件或目录...目录文件或者目录）

-r（递归复制整个目录）

Mv（mv[参数]源文件或目录...目录文件或目录）

源和目标 目标一样时就是重命名的效果

Cat（cat[参数]文件路径，查看目录文件和内容）

-n（显示行号）

Rm（rm[选项]文件或目录，删除指定文件或目录）

-f

-i

-rf（从删库到跑路）

## Vi编辑器（非常重要）

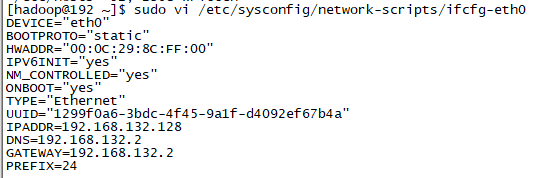
vi编辑器是Linux和Unix上的标准文本编辑器，相当于windows上的记事本，功能更强大。vim是增强版的vi。

vi编辑器的三种模式

* 命令模式
  + 使用vi/vim 文件名 进入编辑器后就是命令模式
  + 可以进行复制、剪切、粘贴、删除等操作
  + 命令行模式下保存并退出：输入 ZZ
  + 命令模式下按一下冒号进入底行模式（shift+:）
* 插入模式（左下方出现 INSERT）
  + 在命令模式下按字母 i 进入 Insert mode，插入模式下可以输入文字
  + 插入模式按esc返回命令模式
* 底行模式（最下方出现冒号）
  + w 文件名（以指定的文件名保存）
  + wq （存盘并退出vi）
  + q! (不存盘强制退出vi）

## 一些常用操作

* 查看网络接口状态
  + ifconfig（对应windows里 ipconfig）
  + 配置静态IP
    - sudo vi /etc/sysconfig/network-scripts/ifcfg-eth0
    - 修改 bootproto 为 static
    - 添加 IPADDR=192.168.132.128，DNS1=192.168.132.2，GATEWAY=192.168.132.2，PREFIX=24



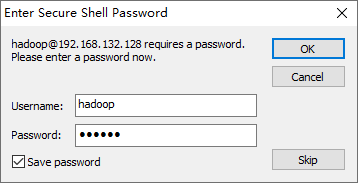
* 测试网络
  + ping 地址（自动无限循环，windows要加 -t）
* 主机名
  + 查看：hostname
  + 修改（临时生效）：hostname 名字
  + 修改（重启永久生效，centos6）：vi /etc/sysconfig/network
* 对防火墙的操作
  + 防火墙是位于内部网和外部网之间的屏障，它按照系统管理员预先定义好的规则来控制数据包的进出
  + service iptables status：查看防火墙状态
  + service iptables start：开启防火墙
  + service iptables stop：关闭防火墙
  + chkconfig iptables on：开机自启
  + chkconfig iptables off：关闭开机自启
* 系统启动级别
  + 查看级别：runlevel
  + 修改启动级别：vi /etc/inittab（centos6，0-6，改为3）
* 切换用户
  + su 用户（可带参数“-”，表示切换用户及用户对应的环境变量）
* 文件权限
  + **权限修改**
    - chmod u=rwx，g=rx，o=rx abc：同上u=用户权限，g=组权限，o=不同组其他用户权限
    - chmod u-x，g+w abc：给abc去除用户执行的权限，增加组写的权限
    - chmod a+r abc：给所有用户添加读的权限
    - chmod 755 abc：赋予abc权限rwxr-xr-x（用数字表示为：r=4，w=2，x=1）

## 远程连接Linux（部分截图已删除）

SecureCRT、XShell等软件都可以，这里使用SecureCRT

Ifconfig查看Linux地址

SecureCRT远程连接，输入对应地址和用户名



输入用户密码进行连接

绿色箭头表示连接成功。

## SFTP文件传输

sftp是Secure File Transfer Protocol的缩写，安全文件传送协议。可以为传输文件提供一种安全的加密方法。sftp 是一个交互式文件传输程式。它类似于 ftp, 但它进行加密传输，比FTP有更高的安全性。下边就简单介绍一下如何远程连接主机，进行文件的上传和下载，以及一些相关操作。

**常用命令**

远端路径：cd、ls、pwd

本地路径：lcd、lls、lpwd

cd 路径 更改远程目录到“路径”

lcd 路径 更改本地目录到“路径”

ls [选项] [路径] 显示远程目录列表

lls [选项] [路径] 显示本地目录列表

put 本地路径 上传文件

get 远程路径 下载文件

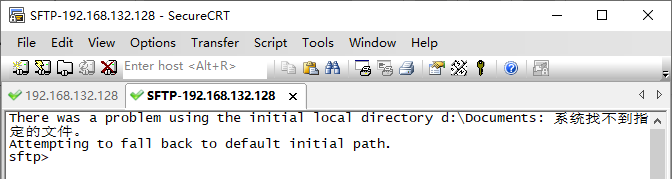
**上传文件**

put /path/filename(本地主机) /path/filename(远端主机)

**下载文件**

get /path/filename(远端主机) /path/filename(本地主机)

本地和虚拟机的文件传输



在SecureCRT中按Alt+P，可以打开sftp

## Linux上安装JDK（下面有windows安装教程，受文档大小限制此处截图删除）

### 从windows上传压缩包到Linux

put jdk安装包路径（路径不要带中文）

### 解压到指定目录

tar -zxvf 压缩包路径 -C 目标路径（目标路径不存在则需要先创建：mkdir xxxxx）

用上面的解压命令解压到/home/hadoop下

配置好hadoop权限之后在/usr/下创建/java文件夹

Jdk文件移到java文件下

### 配置环境变量

vi /etc/profile（打开保存环境变量的文件）

在最后面加入以下语句：

export JAVA\_HOME=/usr/java/jdk1.8.0\_131

export JRE\_HOME=${JAVA\_HOME}/jre

export CLASSPATH=.:${JAVA\_HOME}/lib:${JRE\_HOME}/lib:$CLASSPATH

export JAVA\_PATH=${JAVA\_HOME}/bin:${JRE\_HOME}/bin

export PATH=$PATH:${JAVA\_PATH}  
然后保存并退出vi编辑器

source /etc/profile（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）

输入 java -version 检验jdk是否安装成功

看到以下内容表示安装成功

# Hadoop整体介绍及集群搭建

## 大数据简介

### 数据

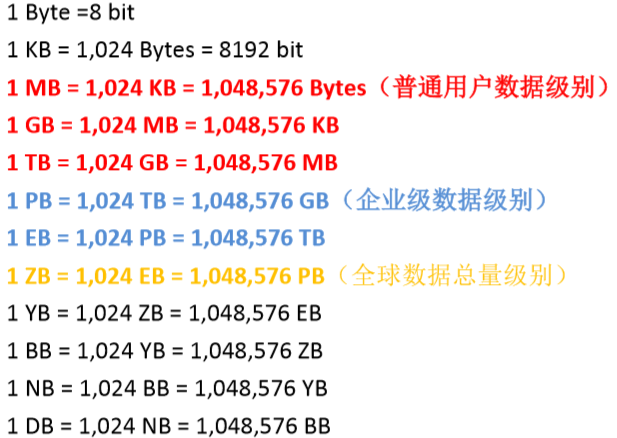
数据就是数值，也就是我们通过观察、实验或计算得出的结果。数据有很多种，最简单的就是数字。数据也可以是文字、图像、声音等。在计算机系统中，数据以二进制信息单元 0,1 的形式表示。

### 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

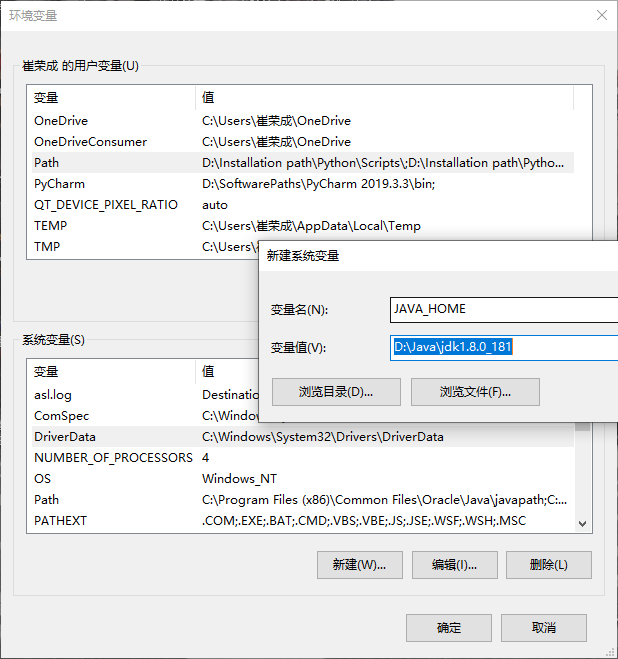
**下载地址：**

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

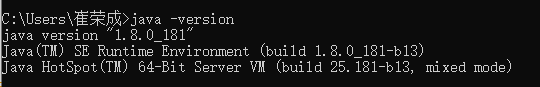
下载对应的jdk

**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。

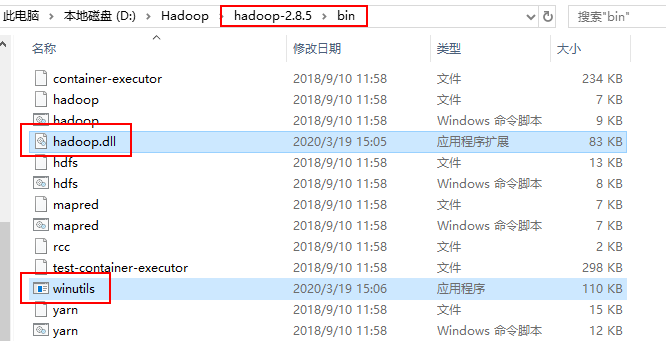
### 安装Hadoop2.8.5，并配置环境

**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**

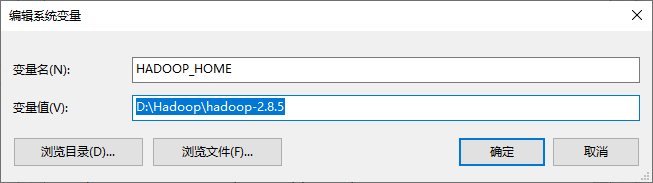
下载链接<https://github.com/cdarlint/winutils>



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

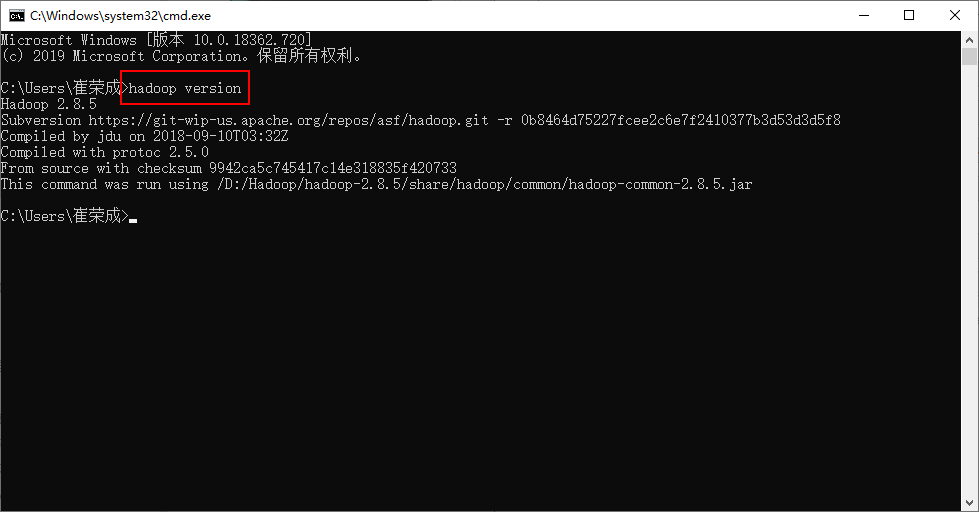
**3.配置hadoop环境变量**

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**

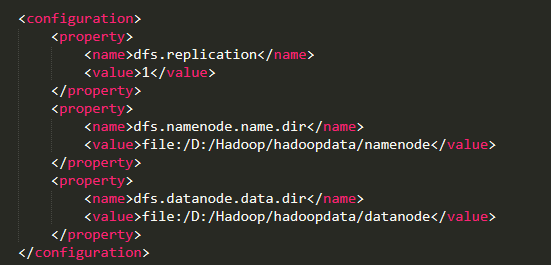
打开Hadoop文件夹下的\etc\hadoop目录对配置文件进行修改，

**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**



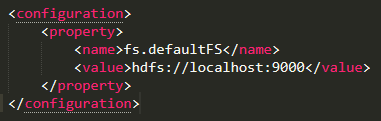
namenode的文件路径：

D:/Hadoop/hadoopdata/namenode

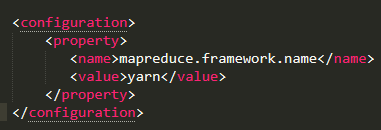
datanode的文件路径：

D:/Hadoop/hadoopdata/datanode

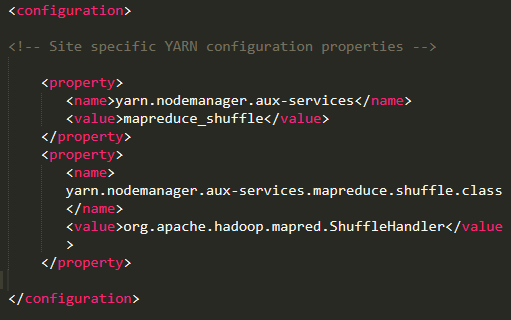
**core-site.xml**



**mapred-site.xml**



**yarn-site.xml**



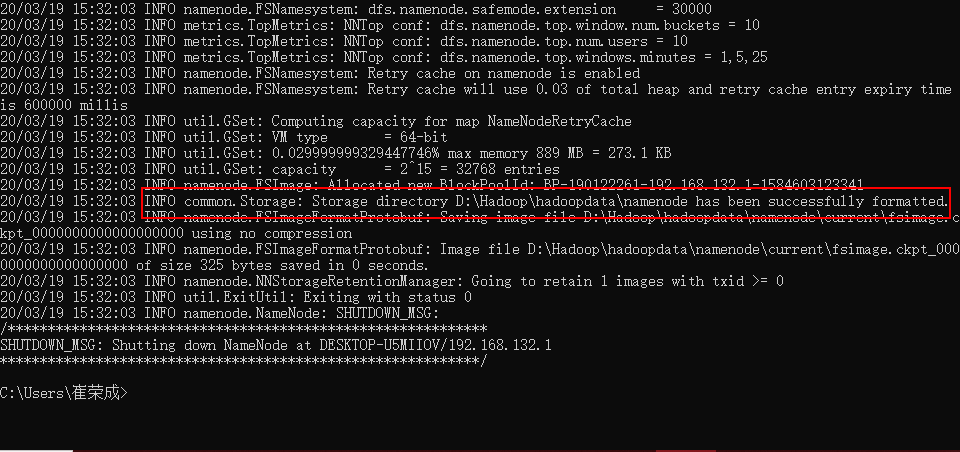
**5.** **将Hadoop文件夹\bin下的hadoop.dll复制到C:\Windows\System32**

**6.初始化HDFS**

运行如下命令：

hdfs namenode -format

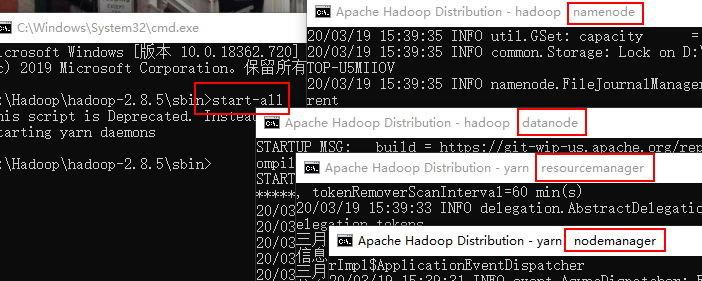
初始化成功



**7.启动Hadoop**

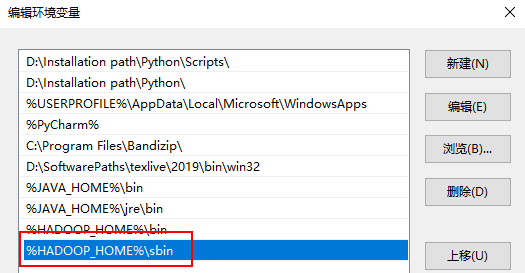
进入hadoop目录下sbin，执行启动命令：

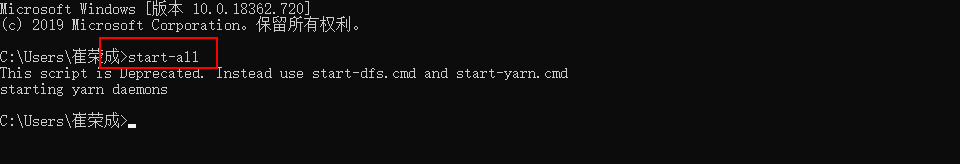
start -all



弹出其他四个控制台输入框说明启动成功。

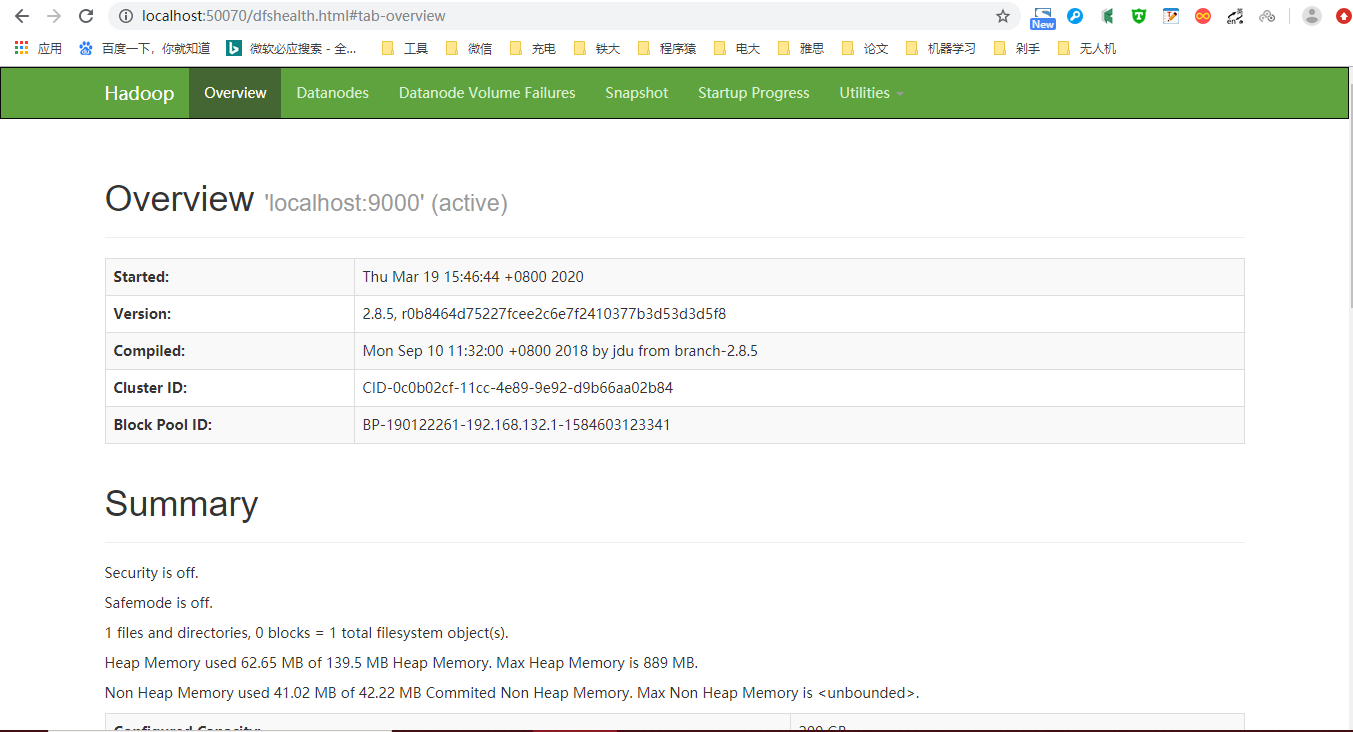
如果把hadoop的sbin也放到系统Path变量，可以直接从系统cmd打开



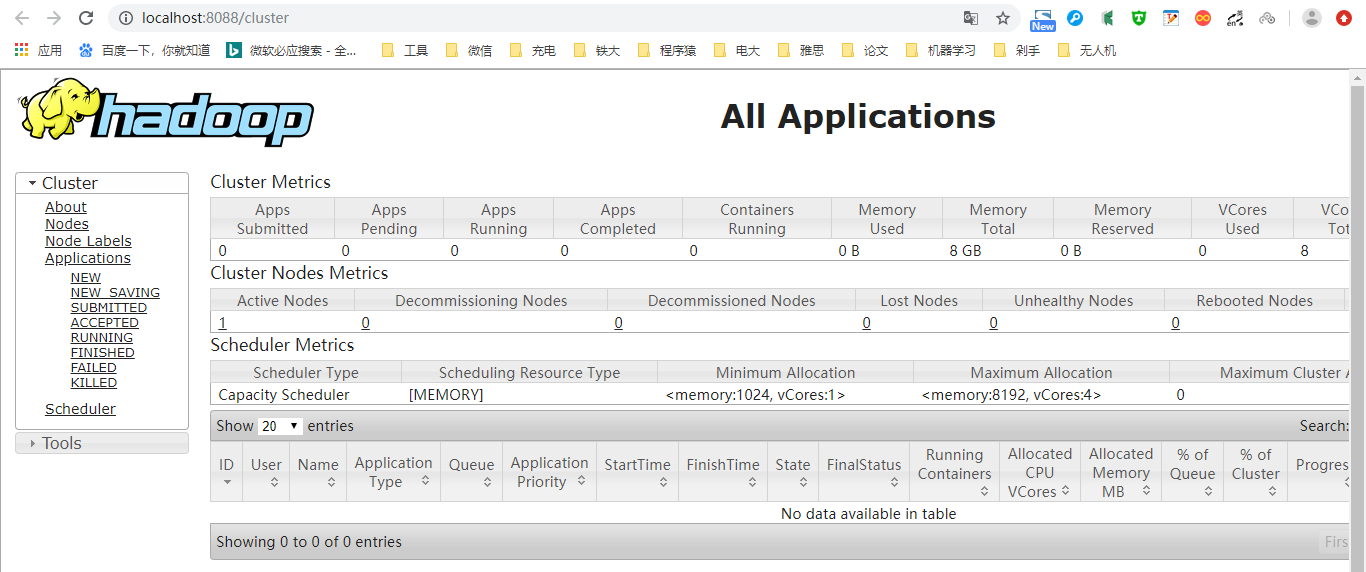


进入网页查看：

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

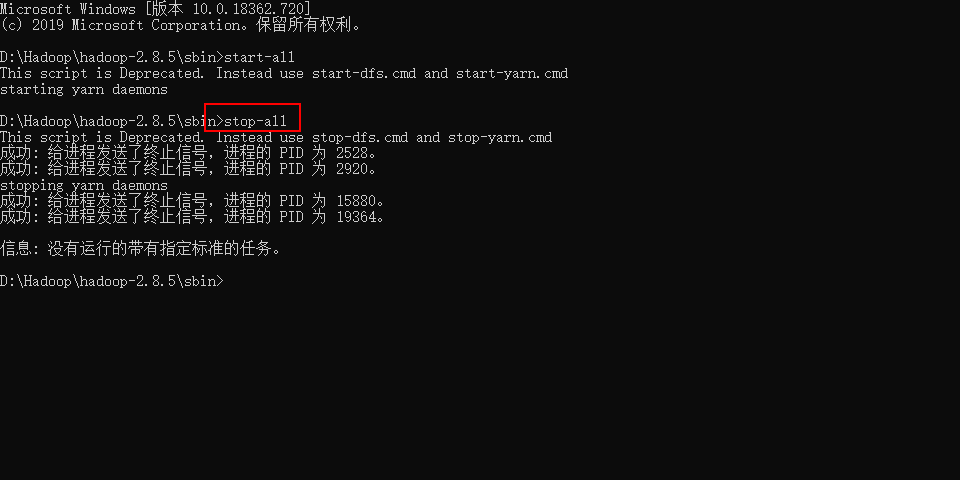


[localhost:8088](http://localhost:8088/)



**8.关闭Hadoop**

关闭命令：stop-all



## linux下安装Hadoop（因为后续以Windows为主，本步骤内容删除，有需要可单独查看学习笔记二）

# HDFS简介及常用操作（未完整）

HDFS（Hadoop Distributed File System）， Hadoop 分布式文件系统，主要用来解决海量数据的存储问题，在大数据系统架构中为各类分布式运算框架（MapReduce，Spark，Tez，Flink，…）提供数据存储服务

* 设计思想
  + 分而治之：将大文件，大批量文件，分布式的存放于大量服务器上。以便于采取分而治 之的方式对海量数据进行运算分析
* 重要概念
  + 数据块/副本、负载均衡、心跳机制、副本存放策略、元数据/元数据管理、 安全模式……

## HDFS 相关概念和特性

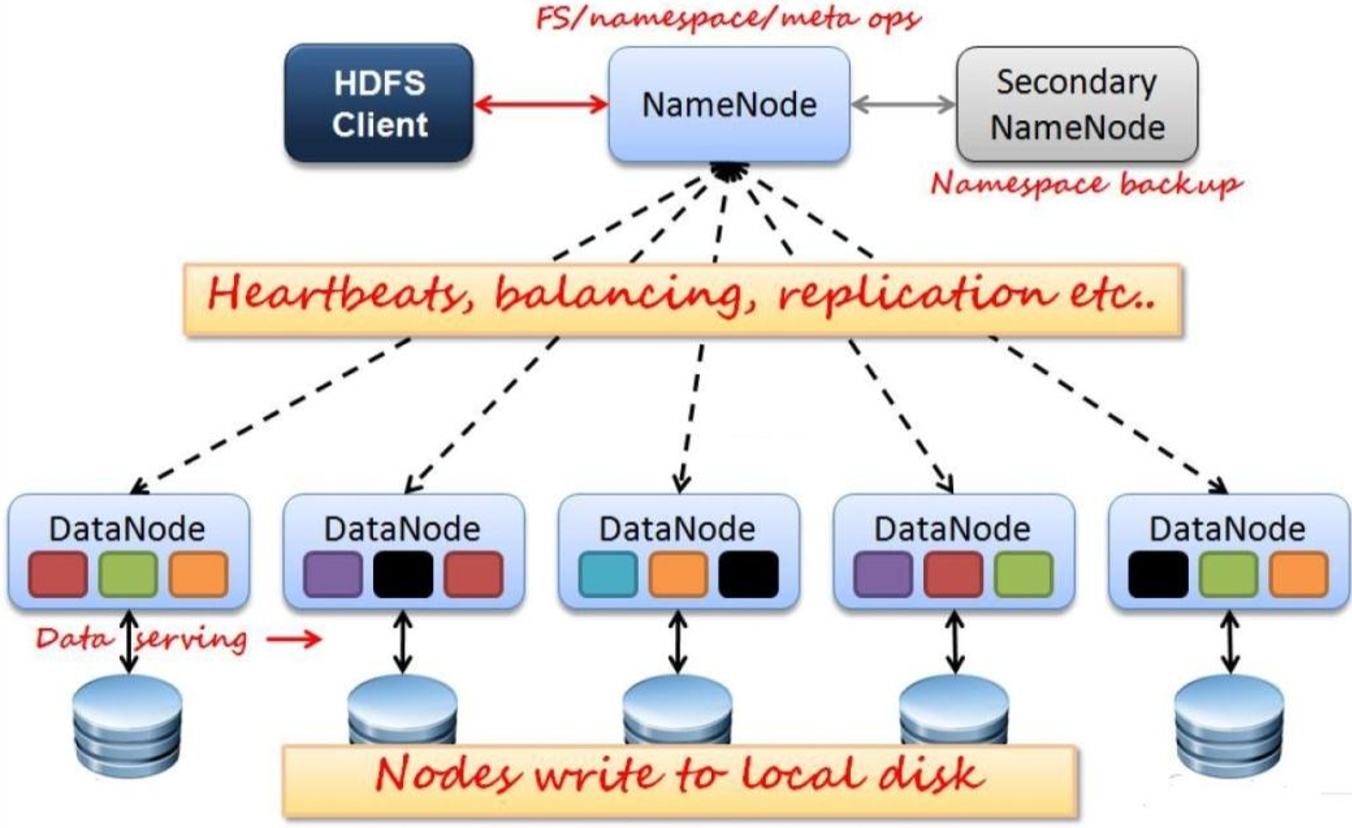
### HDFS 设计思路[¶](file:///H:\%E5%A4%A7%E6%95%B0%E6%8D%AE%E5%AE%9E%E4%B9%A0%E5%AE%9E%E8%AE%AD\%E5%AE%9E%E8%AE%AD%E6%96%87%E6%A1%A3\day3%20HDFS%E7%AE%80%E4%BB%8B%E5%8F%8A%E5%B8%B8%E7%94%A8%E6%93%8D%E4%BD%9C.html#HDFS-设计思路)

HDFS 被设计成用来使用低廉的服务器来进行海量数据的存储，那是怎么做到的呢？  
1、 大文件被切割成小文件，使用分而治之的思想让很多服务器对同一个文件进行联合管理  
2、 每个小文件做冗余备份，并且分散存到不同的服务器，做到高可靠不丢失

### HDFS 架构[¶](file:///H:\%E5%A4%A7%E6%95%B0%E6%8D%AE%E5%AE%9E%E4%B9%A0%E5%AE%9E%E8%AE%AD\%E5%AE%9E%E8%AE%AD%E6%96%87%E6%A1%A3\day3%20HDFS%E7%AE%80%E4%BB%8B%E5%8F%8A%E5%B8%B8%E7%94%A8%E6%93%8D%E4%BD%9C.html#HDFS-架构)

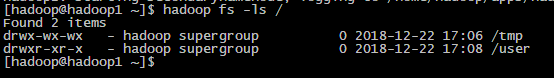
主从架构设计：

* 主节点：NameNode（集群老大，管理整个文件系统的元数据，处理客户端读且请求负责 ）
  + 普通分布式集群还有一个 Secondary NameNode （用来帮 NameNode 分担压力）
* 从节点：DataNode（存储整个集群所有数据块，真正处理数据读写 ）



## HDFS Shell 操作（命令行操作）

HDFS 提供 shell 命令行客户端，使用方法如下：



## HDFS 的 Java API 操作[¶](file:///H:\\%E5%A4%A7%E6%95%B0%E6%8D%AE%E5%AE%9E%E4%B9%A0%E5%AE%9E%E8%AE%AD\\%E5%AE%9E%E8%AE%AD%E6%96%87%E6%A1%A3\\day3%20HDFS%E7%AE%80%E4%BB%8B%E5%8F%8A%E5%B8%B8%E7%94%A8%E6%93%8D%E4%BD%9C.html" \l "HDFS-的-Java-API-操作)

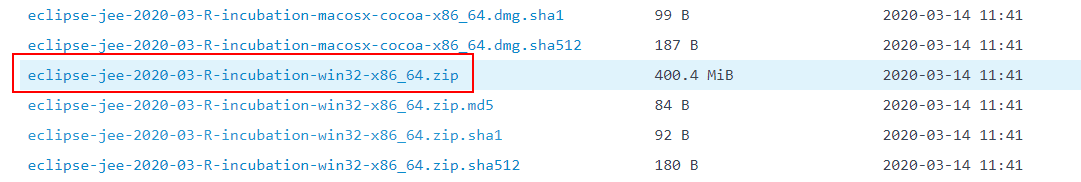
### 利用 Eclipse 可视化 HDFS 上的文件信息

下载eclipse

官网下载链接：<https://www.eclipse.org/downloads/>

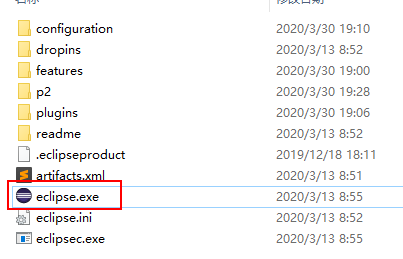
官网下载较慢，可以使用镜像下载：

<https://mirrors.tuna.tsinghua.edu.cn/eclipse/technology/epp/downloads/release/>



选择eclipse-jee-xxx-win32-xxx版本

下载解压，然后运行exe

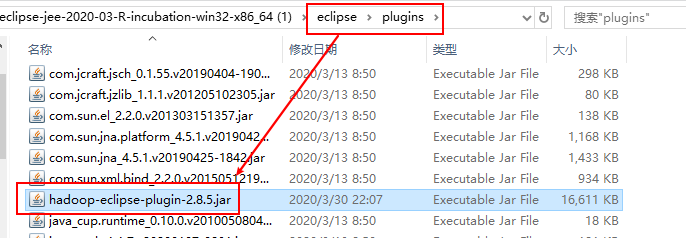


### 编译hadoop-eclipse插件

可以下载ant自行编译对应环境的插件，本次用的编译好的插件。

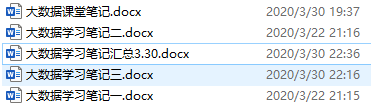
下载链接：<https://github.com/DoubleBirdsU/Hadoop-eclipse-plugin>

将 hadoop-eclipse-plugin-2.8.5.jar 插件包放到 eclipse/plugins 文件夹下



启动 eclipse ，点击 Window -> Prefrences ，在弹出的框选中Hadoop Map/Reduce ，然后在右边配置hadoop解压的路径

P.S.



之前的笔记是按照周次记录的，因为文档大小限制，本文档是前三次笔记的简单汇总，且只保留了关键步骤的截图，后续笔记会继续完善。