****

**大数据技术原理及应用**

实验报告



学 院： 计算机科学与技术学院

专 业： 电力信息技术

报告名称： Hadoop MapReduce编程实践

学生姓名： 崔荣成 学号： 19108002

指导老师： 周 平

时 间： 2020年4月17日

**目 录**

[一、 实验目的 3](#_Toc38126277)

[二、 实验内容 3](#_Toc38126278)

[三、 实验步骤 3](#_Toc38126279)

[1. Hadoop整体介绍及集群搭建 3](#_Toc38126280)

[1.1. 大数据简介 3](#_Toc38126281)

[1.2. Hadoop整体概念 4](#_Toc38126282)

[1.3. Windows下安装Hadoop 5](#_Toc38126283)

[2. HDFS简介及常用操作 15](#_Toc38126284)

[2.1. HDFS 相关概念和特性 15](#_Toc38126285)

[2.2. HDFS Shell 操作（命令行操作） 17](#_Toc38126286)

[2.3. HDFS 的 Java API 操作 18](#_Toc38126287)

[3. Eclipse编程进行hdfs常用操作 24](#_Toc38126288)

[3.1. 创建TestHDFS程序 24](#_Toc38126289)

[3.2. 创建TestHDFS类 25](#_Toc38126290)

[3.3. 添加依赖库 25](#_Toc38126291)

[3.4. 添加junit测试单元 28](#_Toc38126292)

[3.5. 创建fs对象 29](#_Toc38126293)

[3.6. 创建文件夹 30](#_Toc38126294)

[3.7. 删除文件 30](#_Toc38126295)

[3.8. 上传文件 30](#_Toc38126296)

[3.9. 下载文件 31](#_Toc38126297)

[3.10. 查看文件及文件夹信息 31](#_Toc38126298)

[4. 使用eclipse创建并运行wordcount程序 32](#_Toc38126299)

[4.1. 创建MapReduce程序 32](#_Toc38126300)

[4.2. 编写WordCount程序 34](#_Toc38126301)

[4.3. 配置参数 36](#_Toc38126302)

[4.4. 运行结果 37](#_Toc38126303)

# 实验目的

掌握Hadoop平台的安装配置方法，认识Hadoop平台的构成；熟悉Hadoop平台基本命令操作；

掌握HDFS文件系统结构，熟悉HDFS的基本命令操作和编程操作。

# 实验内容

1、安装配置Hadoop平台；

2、熟悉Hadoop常用指令；

3、安装配置Eclipse-Hadoop环境；

4、运行HDFS基本示例；

# 实验步骤

## Hadoop整体介绍及集群搭建

### 大数据简介

#### 数据

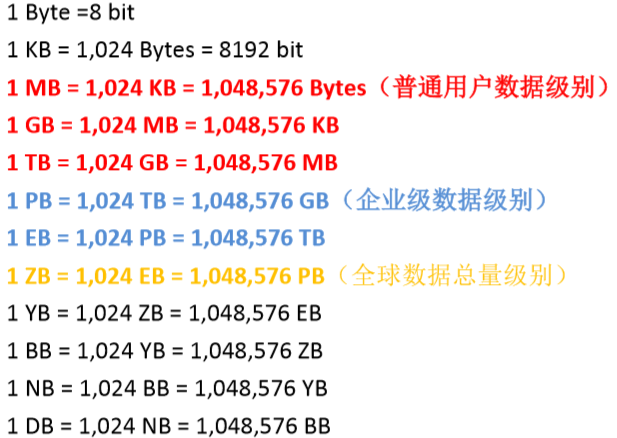
数据就是数值，也就是我们通过观察、实验或计算得出的结果。数据有很多种，最简单的就是数字。数据也可以是文字、图像、声音等。在计算机系统中，数据以二进制信息单元 0,1 的形式表示。

#### 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：**

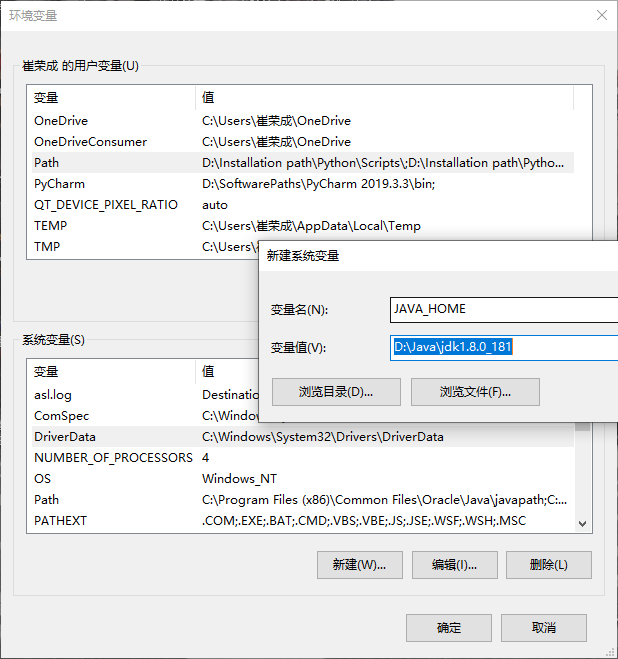
<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

下载对应的jdk

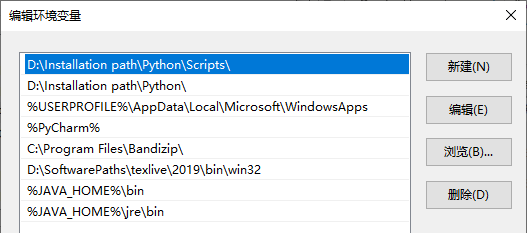


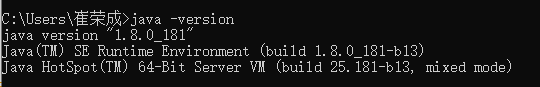
**配置环境变量：**

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。

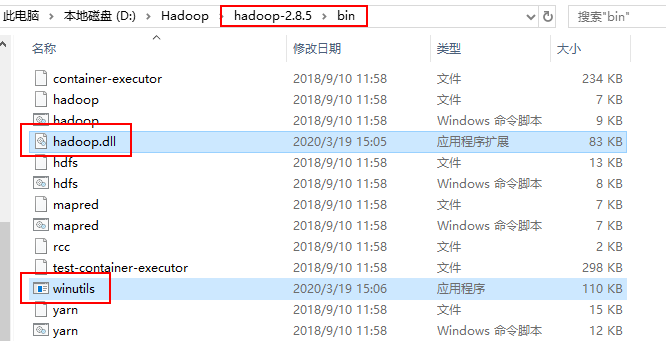
#### 安装Hadoop2.8.5并配置环境

**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**

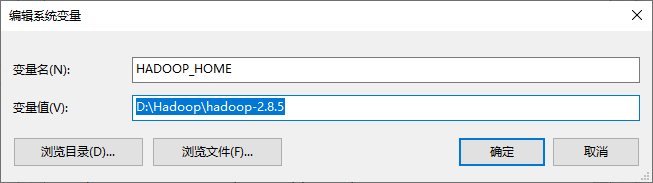
下载链接<https://github.com/cdarlint/winutils>



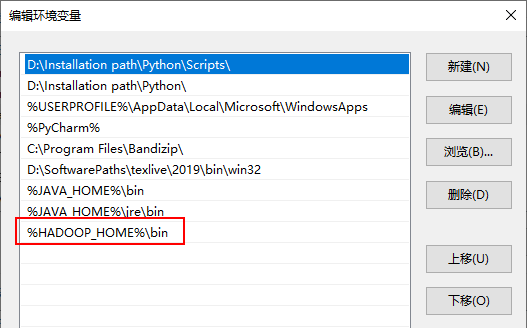
关键是将hadoop.dll和winutils.exe两个文件放入到/bin目录下

**3.配置hadoop环境变量**

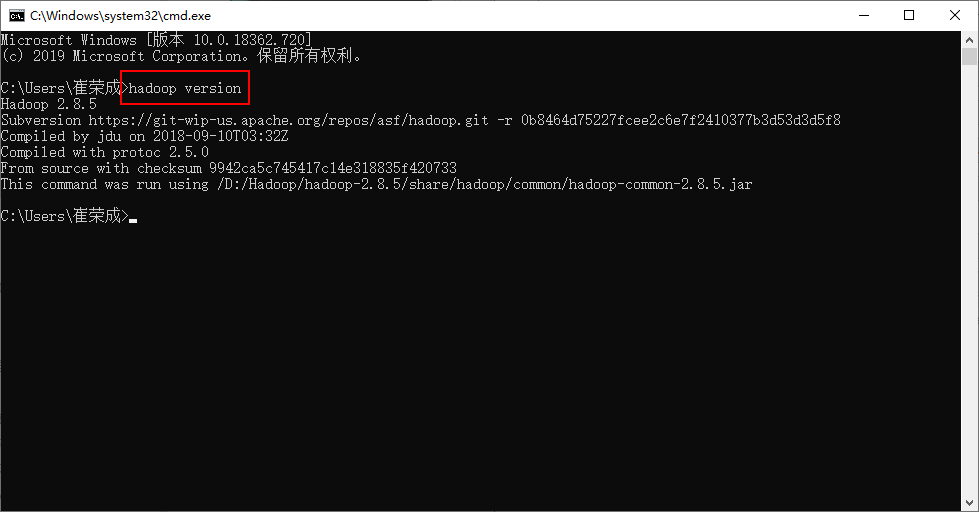
新建HADOOP\_HOME



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**

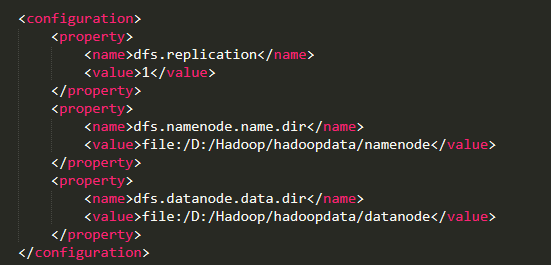
打开Hadoop文件夹下的\etc\hadoop目录对配置文件进行修改，

**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**



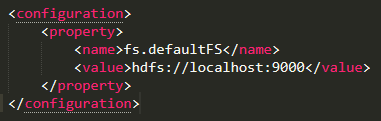
namenode的文件路径：

D:/Hadoop/hadoopdata/namenode

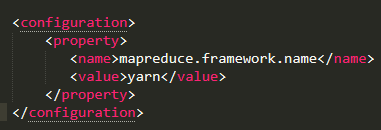
datanode的文件路径：

D:/Hadoop/hadoopdata/datanode

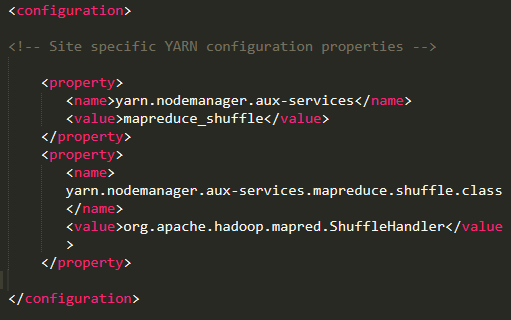
**core-site.xml**



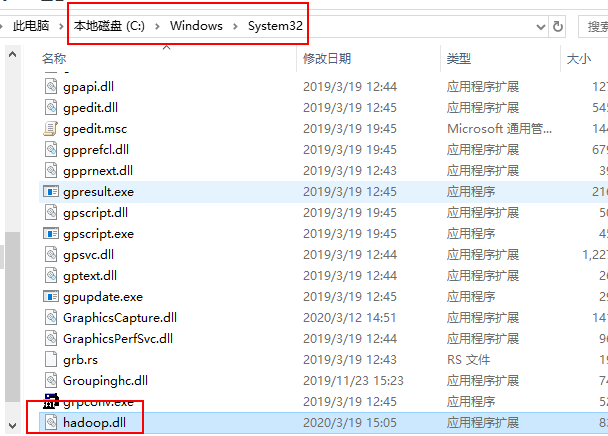
**mapred-site.xml**



**yarn-site.xml**



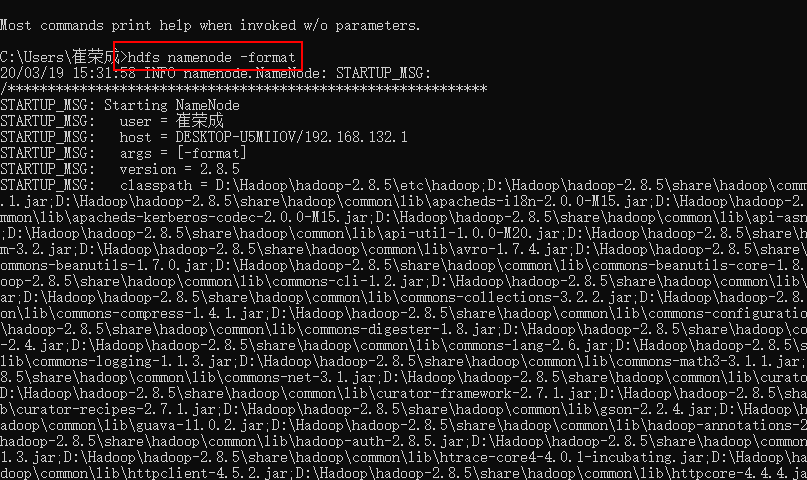
**5.** **将Hadoop文件夹\bin下的hadoop.dll复制到C:\Windows\System32**



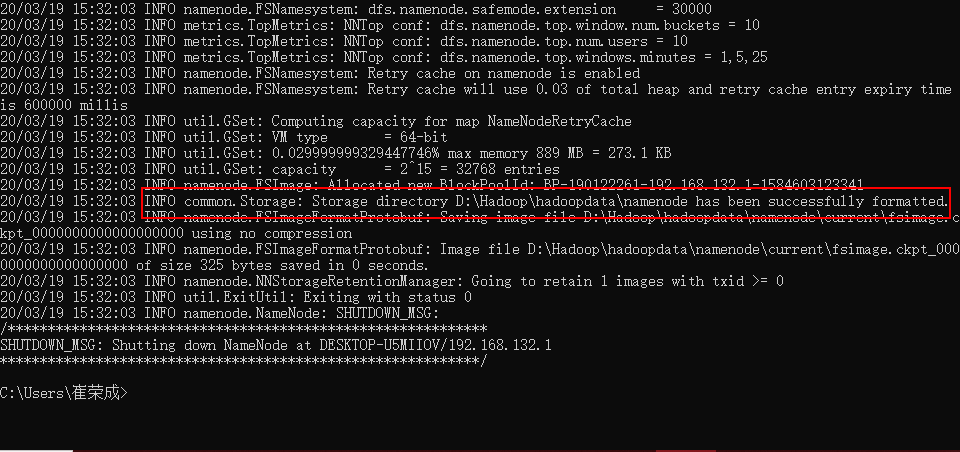
**6.初始化HDFS**

运行如下命令：

hdfs namenode -format



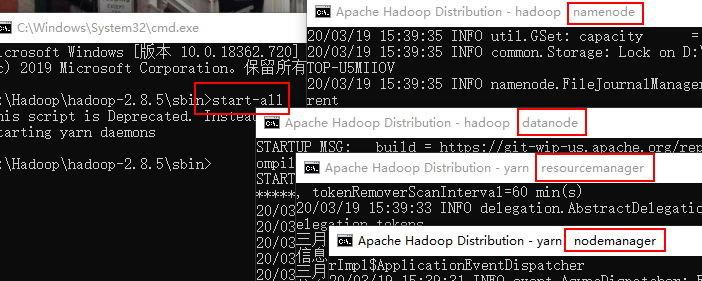
初始化成功



**7.启动Hadoop**

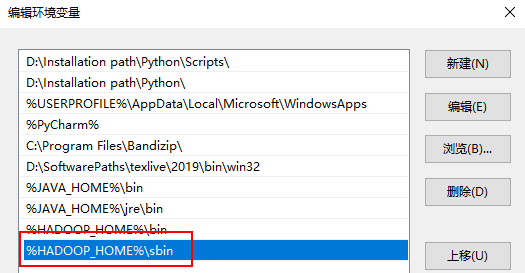
进入hadoop目录下sbin，执行启动命令：

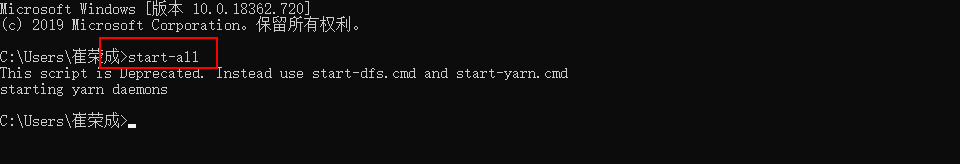
start -all



弹出其他四个控制台输入框说明启动成功。

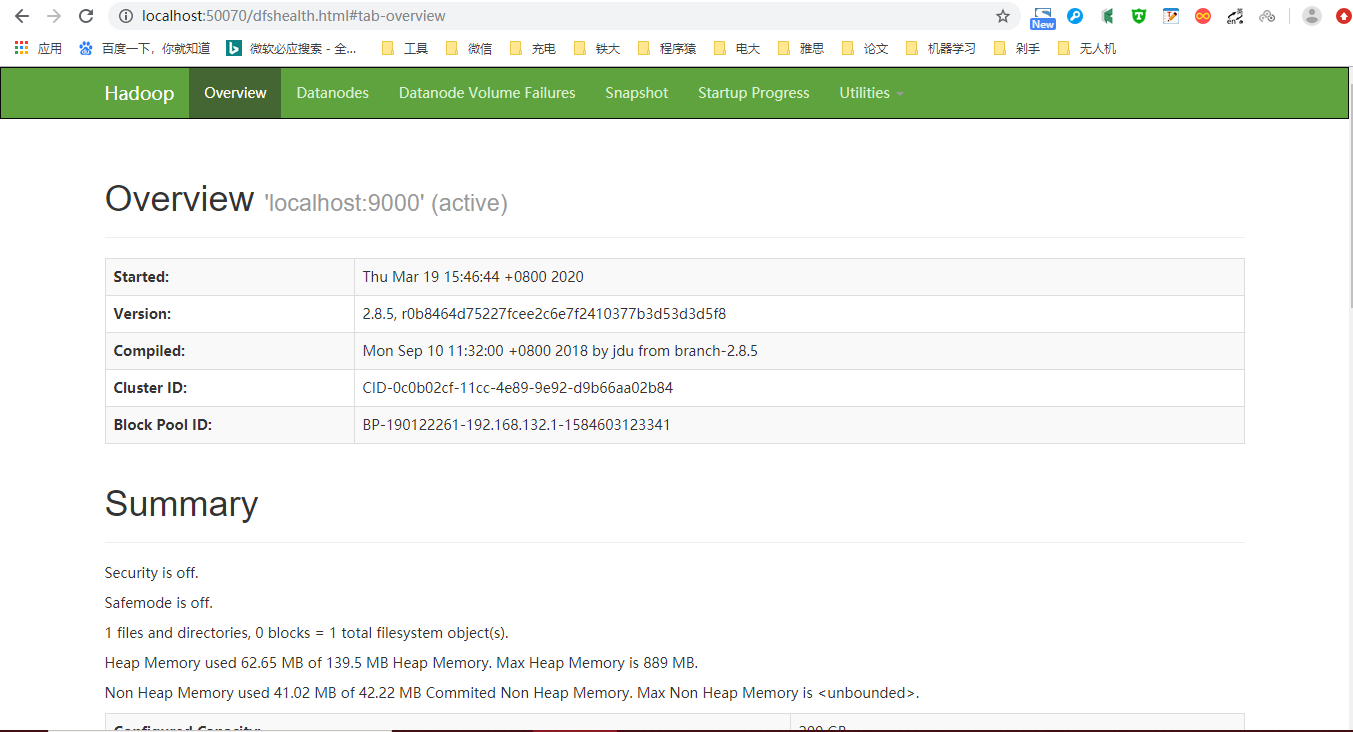
如果把hadoop的sbin也放到系统Path变量，可以直接从系统cmd打开



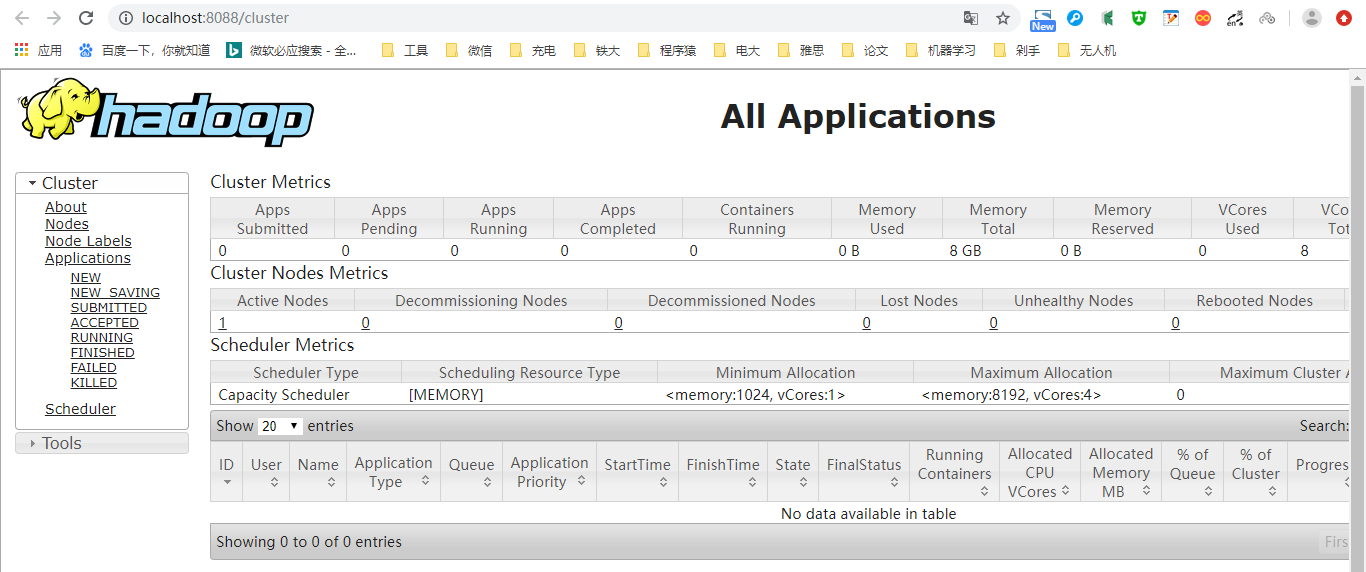


进入网页查看：

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

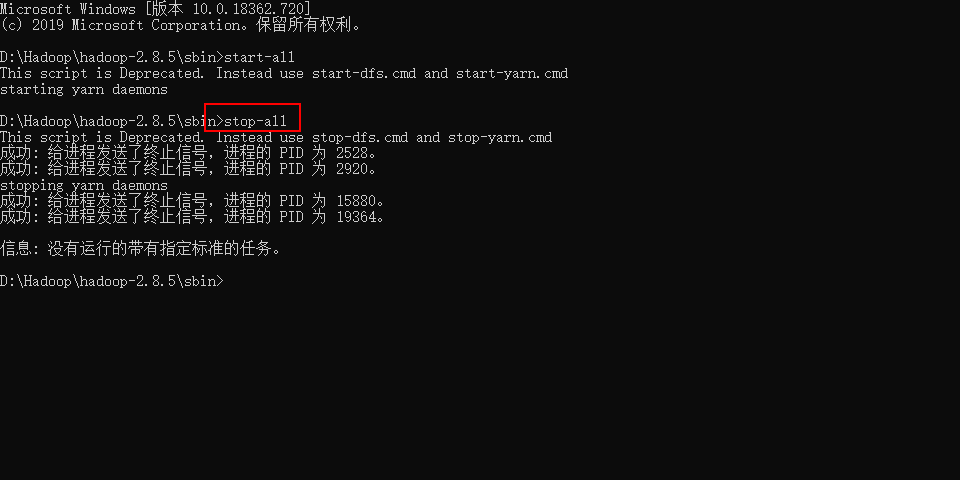


[localhost:8088](http://localhost:8088/)



**8.关闭Hadoop**

关闭命令：stop-all



## HDFS简介及常用操作

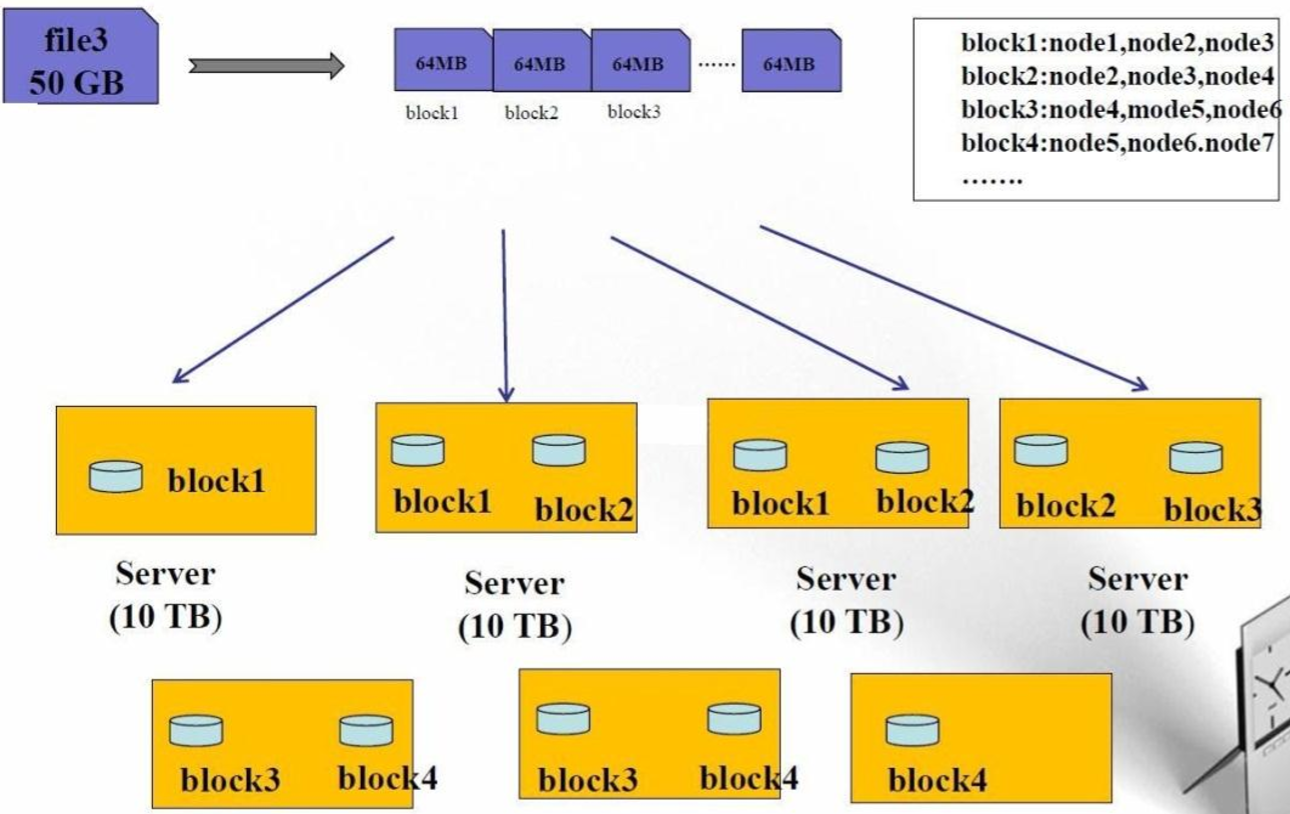
HDFS（Hadoop Distributed File System）， Hadoop 分布式文件系统，主要用来解决海量数据的存储问题，在大数据系统架构中为各类分布式运算框架（MapReduce，Spark，Tez，Flink，…）提供数据存储服务

* 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

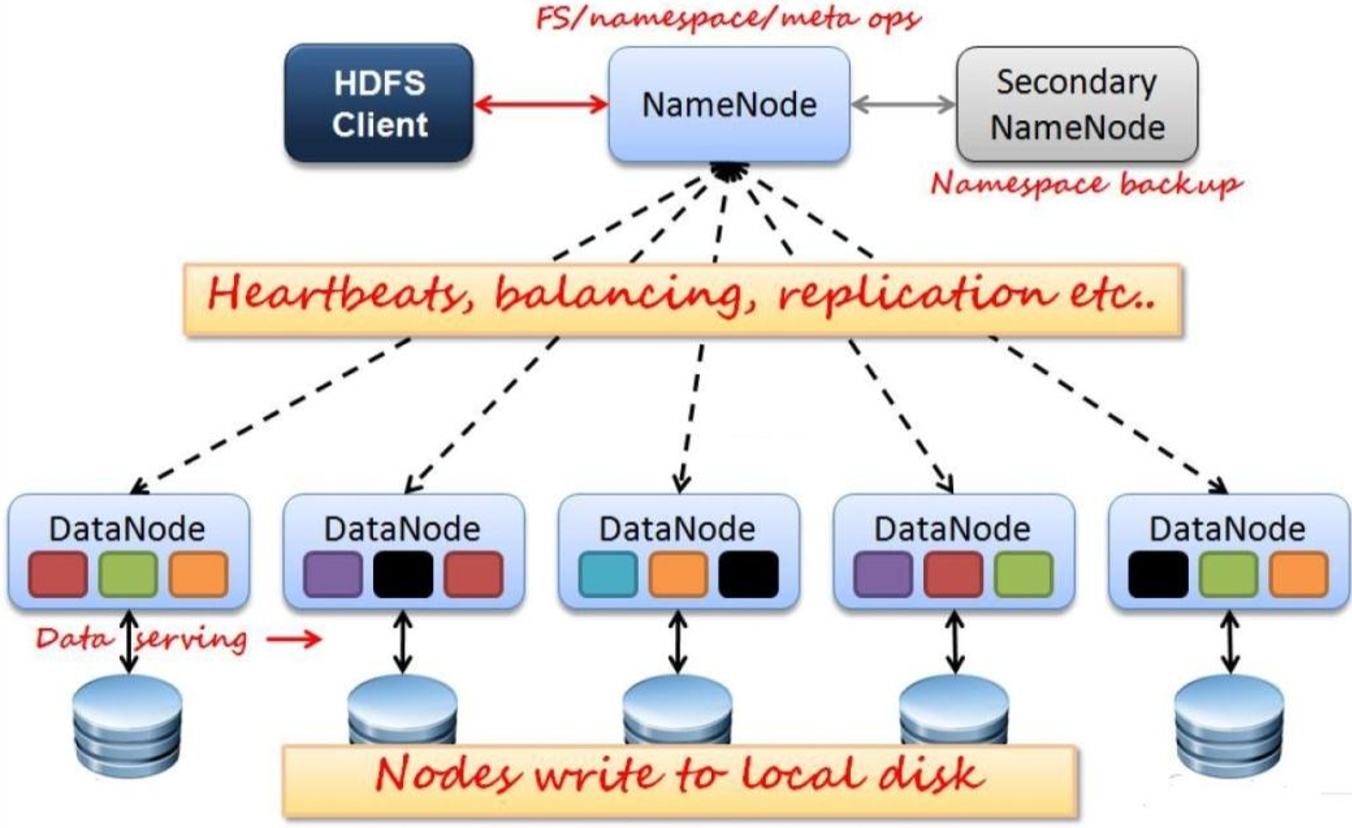
HDFS 被设计成用来使用低廉的服务器来进行海量数据的存储，那是怎么做到的呢？  
1、 大文件被切割成小文件，使用分而治之的思想让很多服务器对同一个文件进行联合管理  
2、 每个小文件做冗余备份，并且分散存到不同的服务器，做到高可靠不丢失



#### HDFS 架构

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 特性

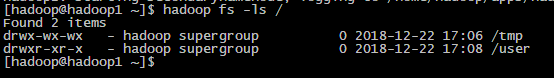
* HDFS 中的文件在物理上是分块存储（block），块的大小可以通过配置参数(dfs.blocksize) 来规定，默认大小在 hadoop2.x 版本中是 128M，1.x 是 64M ，3.x 是 256M
* HDFS 文件系统会给客户端提供一个统一的抽象目录树，客户端通过路径来访问文件，形如：hdfs://hadoop1:9000/soft/hadoop-2.7.5-centos-6.7.tar.gz
* 目录结构及文件分块位置信息（metadata,元数据）的管理由 namenode 节点承担。 namenode 是 HDFS 集群主节点，负责维护整个 hdfs 文件系统的目录树，以及每一个路径（文件）所对应的 block 块信息（block 的 id，及所在的 datanode 服务器）
* 文件的各个 block 的存储管理由 datanode 节点承担。 datanode 是 HDFS 集群从节点，每一个 block 都可以在多个 datanode 上存储多个副本（副本数量也可以通过参数设置： dfs.replication，默认是 3）
* HDFS 是设计成适应一次写入，多次读出的场景，且不支持文件的随机修改

#### HDFS 优缺点

* 优点
  + 可构建在廉价机器上, 通过多副本提高可靠性，提供了容错和恢复机制
  + 高容错性，数据自动保存多个副本，副本丢失后，自动恢复
  + 适合大数据处理， GB、TB、甚至 PB 级数据，百万规模以上的文件数量，10K+节点规模
  + 流式文件访问， 一次性写入，多次读取，保证数据一致性
* 缺点
  + 延迟高（涉及到多个节点之间相互通信）
  + 不适合存储小文件（因为元信息存储在 NameNode 内存中 ，一个 block 元信息消耗大约 150 byte ,存储 1 亿个 block，大约需要 15GB 内存来存储对应的元数据，浪费资源）

### 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：



常用命令参数介绍：

* -help 功能：输出这个命令参数手册
  + hadoop -help
  + hadoop fs -help
  + hadoop fs -help ls
* -ls 功能：显示目录信息
  + hadoop fs -ls hdfs://hadoop1:9000/
  + hadoop fs -ls / （二者等效）
* -mkdir 功能：在 hdfs 上创建目录
  + hadoop fs -mkdir -p /aa/bb/cc/dd
* -put 功能：等同于 -copyFromLocal，进行文件上传
  + hadoop fs -put /aaa/jdk.tar.gz /bbb/jdk.tar.gz
* -get 功能：等同于 -copyToLocal，就是从 hdfs 下载文件到本地
  + hadoop fs -get /aaa/jdk.tar.gz
* -getmerge 功能：合并下载多个文件
  + hadoop fs -getmerge /aaa/log.\* ./log.sum （ 假如 hdfs 的目录 /aaa/下有多个文件:log.1, log.2,log.3,... ）
* -cp 功能：从 hdfs 的一个路径拷贝 hdfs 的另一个路径
  + hadoop fs -cp /aaa/jdk.tar.gz /bbb/jdk.tar.gz.2
* -mv 功能：在 hdfs 目录中移动文件
  + hadoop fs -mv /aaa/jdk.tar.gz /
* -rm 功能：删除文件或文件夹
  + ：hadoop fs -rm -r /aaa/bbb/
* -appendToFile 功能：追加一个文件到已经存在的文件末尾
  + hadoop fs -appendToFile ./hello.txt /hello.txt
* -cat 功能：显示文件内容
  + hadoop fs -cat /hello.txt
* -tail 功能：显示一个文件的末尾
  + ：hadoop fs -tail /weblog/access\_log.1
* -text 功能：以字符形式打印一个文件的内容
  + hadoop fs -text /weblog/access\_log.1
* hdfs dfsadmin -report 查看 dfs 集群工作状态的命令

### HDFS 的 Java API 操作

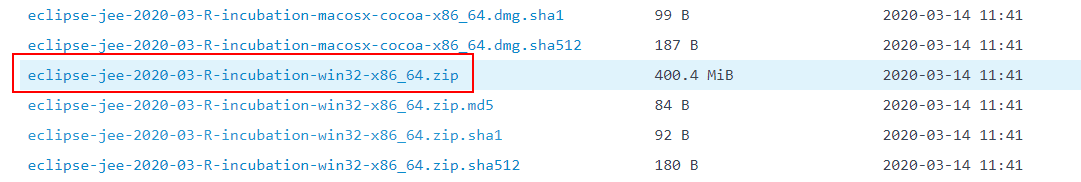
#### 利用 Eclipse 可视化 HDFS 上的文件信息

下载eclipse

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

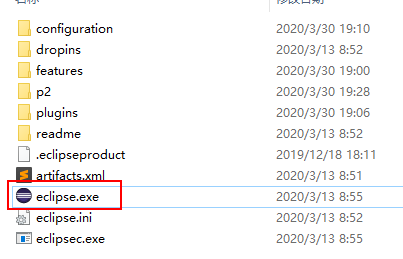
官网下载较慢，可以使用镜像下载：

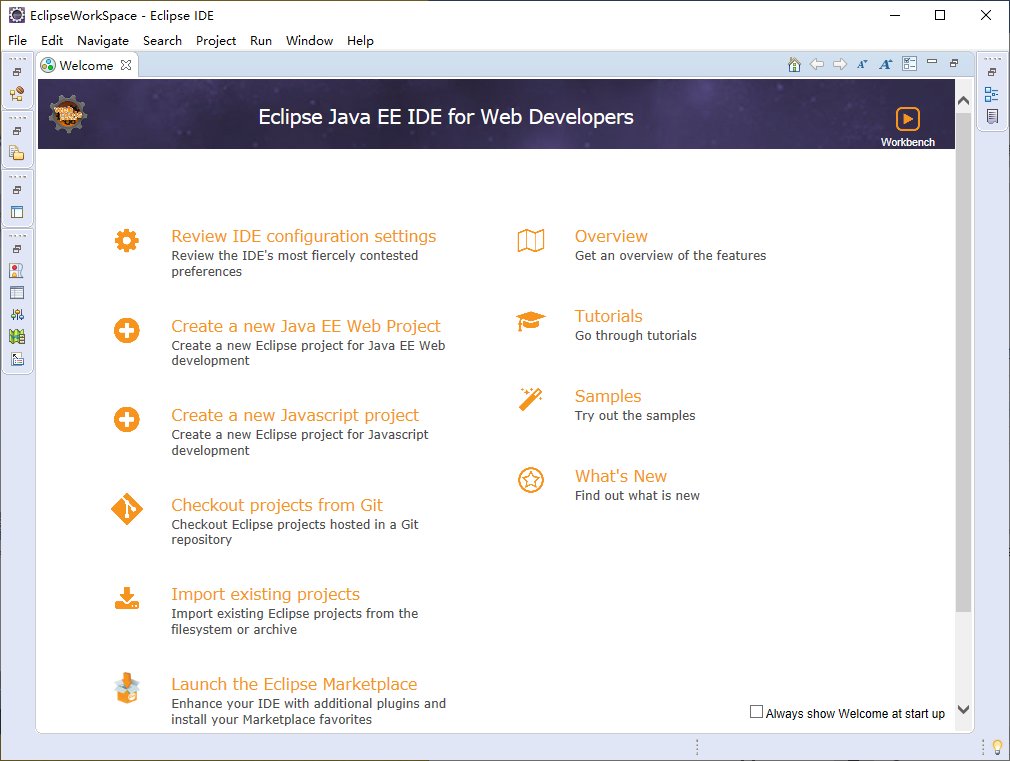
<https://mirrors.tuna.tsinghua.edu.cn/eclipse/technology/epp/downloads/release/>



选择eclipse-jee-xxx-win32-xxx版本

下载解压，然后运行exe



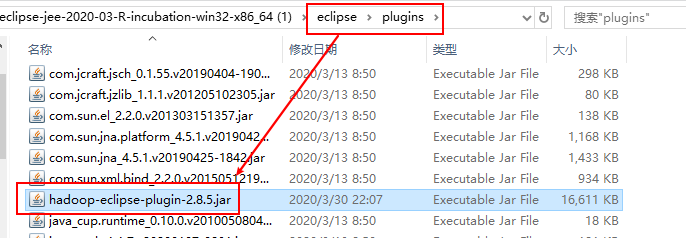


#### 编译hadoop-eclipse插件

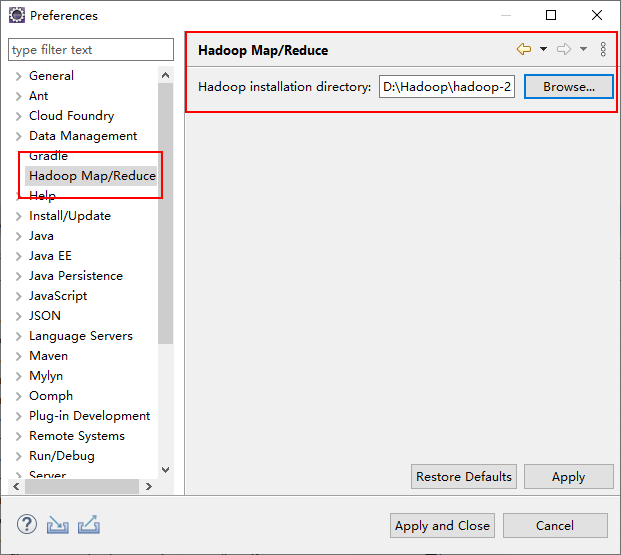
可以下载ant自行编译对应环境的插件，本次用的编译好的插件。

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

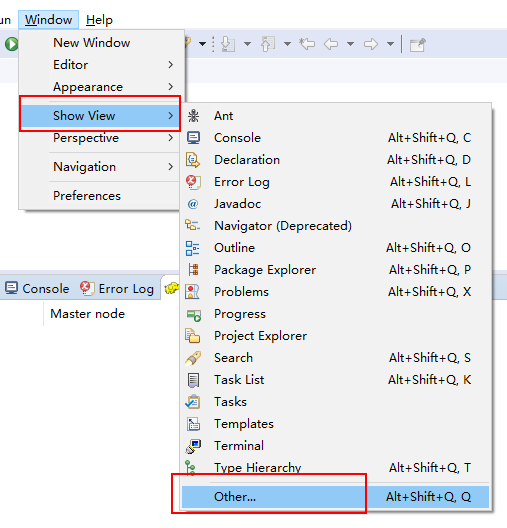


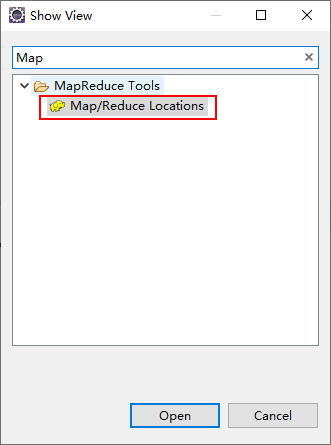
启动 eclipse ，点击 Window -> Prefrences ，在弹出的框选中Hadoop Map/Reduce ，然后在右边配置hadoop解压的路径

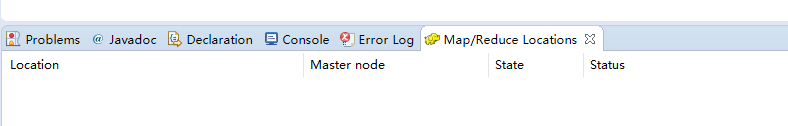


#### 切换"Map/Reduce"工作目录

选择"Window"菜单下选择"Show view"，选择others，弹出一个窗体，搜索Map，从中选择"Map/Reduce"选项即可进行切换



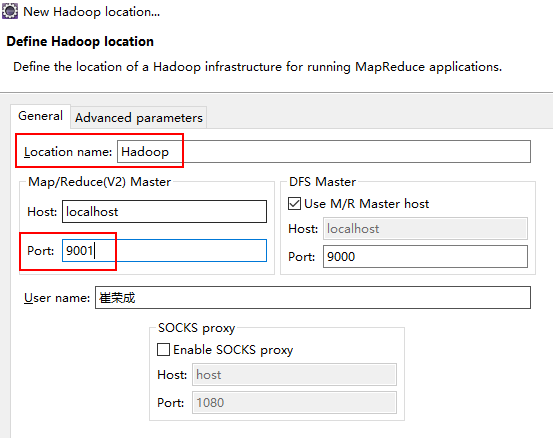




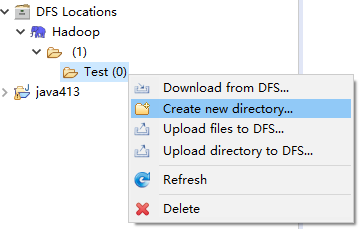
#### 建立与Hadoop集群的连接

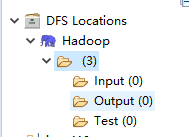
在Eclipse软件下面的"Map/Reduce Locations"进行右击，弹出一个选项，选择"New Hadoop Location"，然后弹出一个窗体。





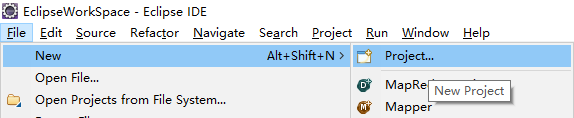
连接成功后可以看到HDFS结构，因为第一次连接所以没有东西，可以右击闯进文件夹。



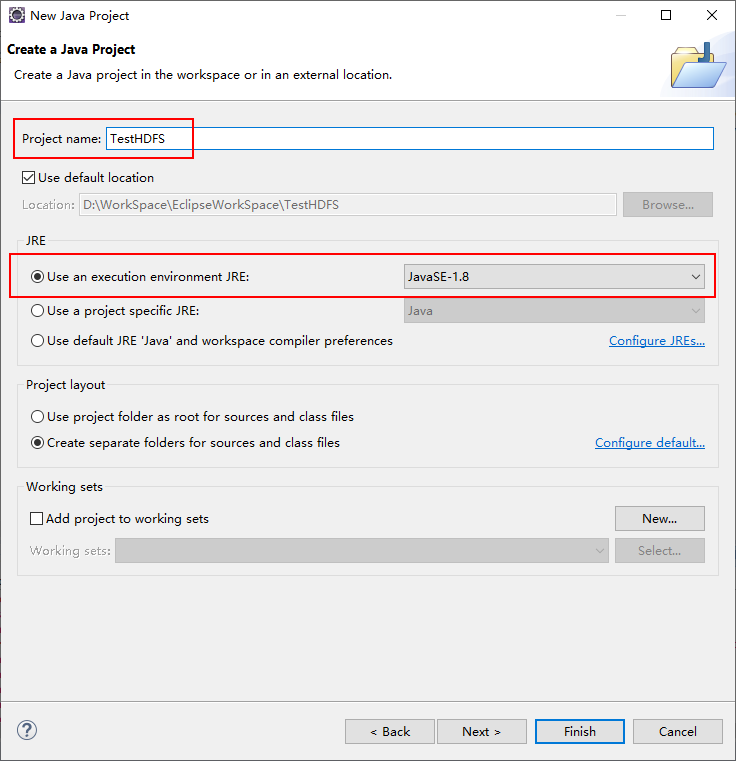


## Eclipse编程进行hdfs常用操作

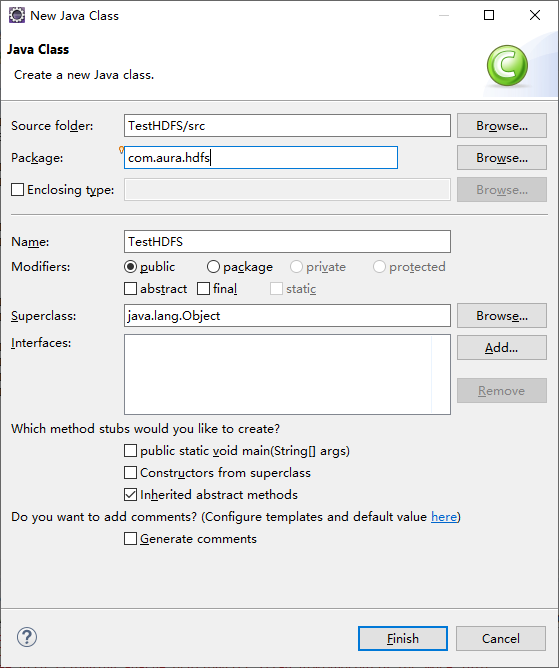
### 创建TestHDFS程序



选择Java程序

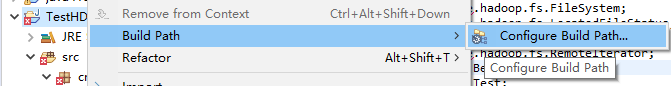


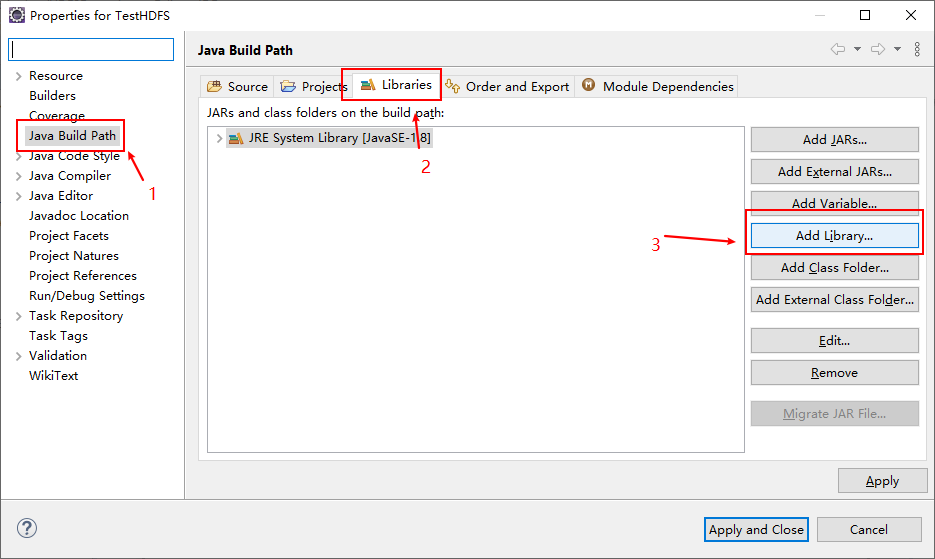
### 创建TestHDFS类

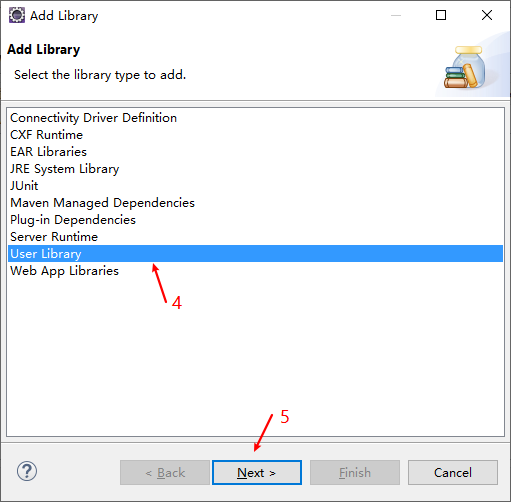


### 添加依赖库

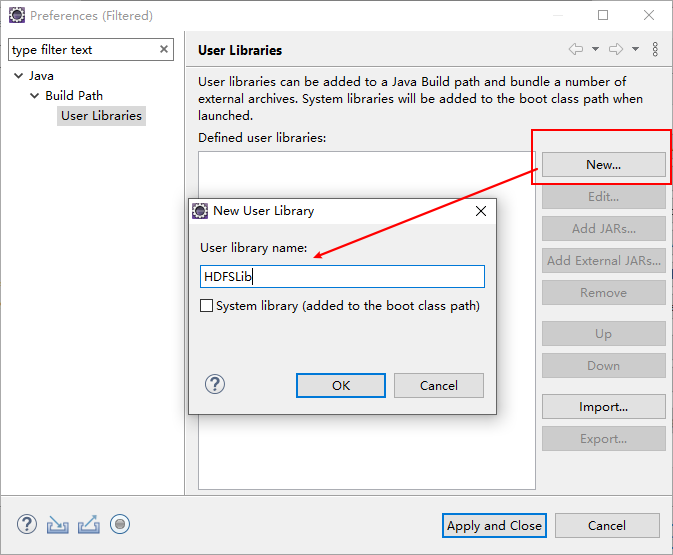
右击项目，选择build path，进行配置

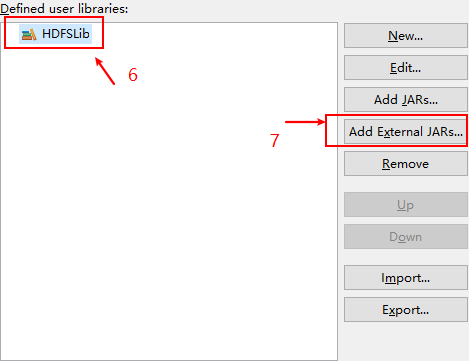




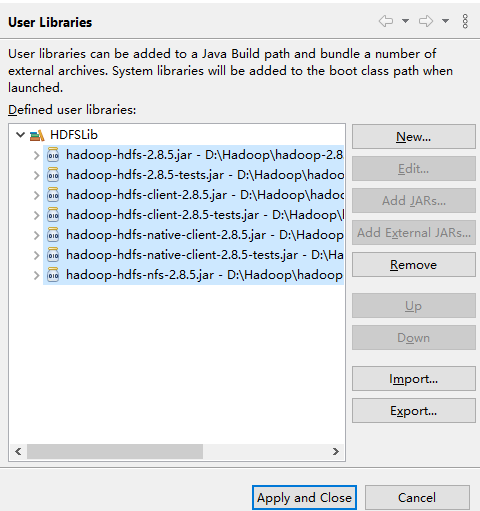


接下来选中User Libraries -> New ，输入HDFS库的名字

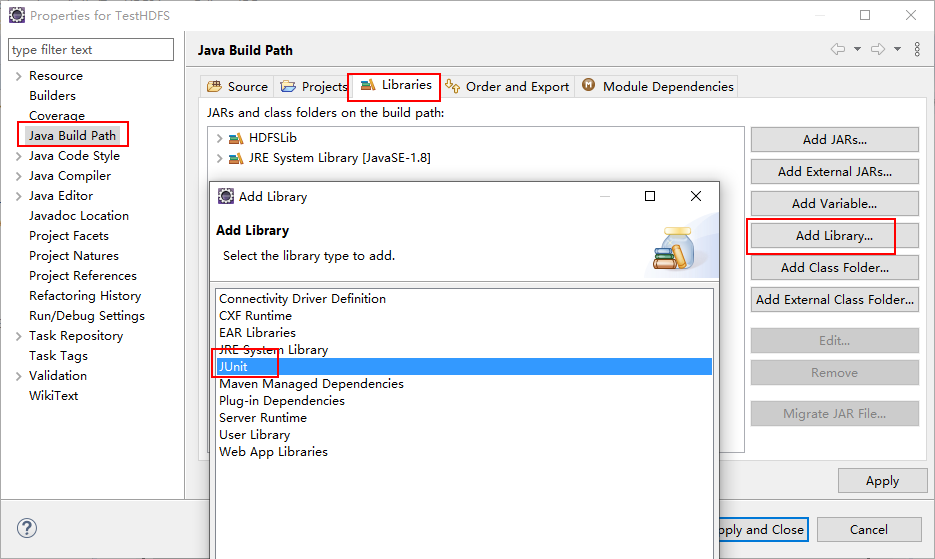


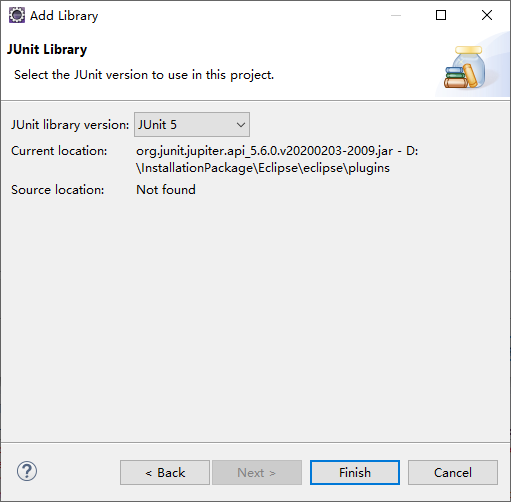


然后把hadoop目录下/share/hadoop/hdfs的所有jar包放进去



### 添加junit测试单元

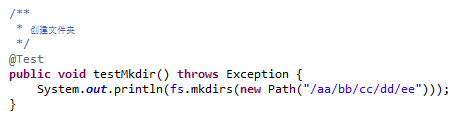


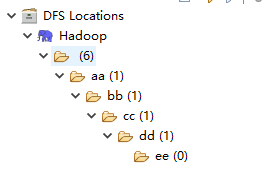


### 创建fs对象

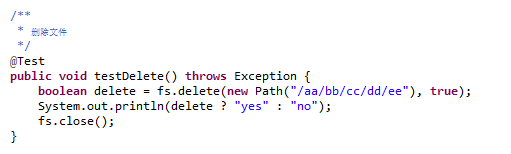


### 创建文件夹



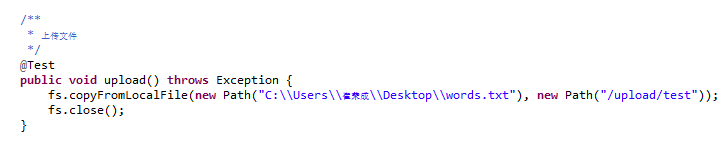


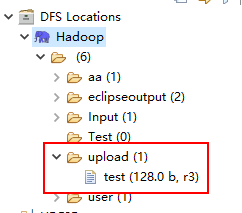
### 删除文件



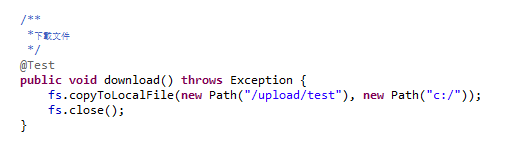


### 上传文件

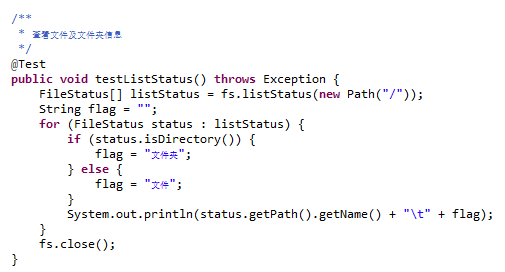


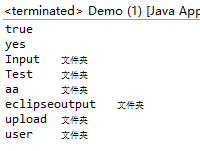


### 下载文件



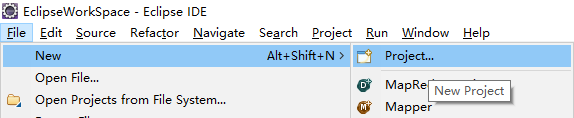
### 查看文件及文件夹信息

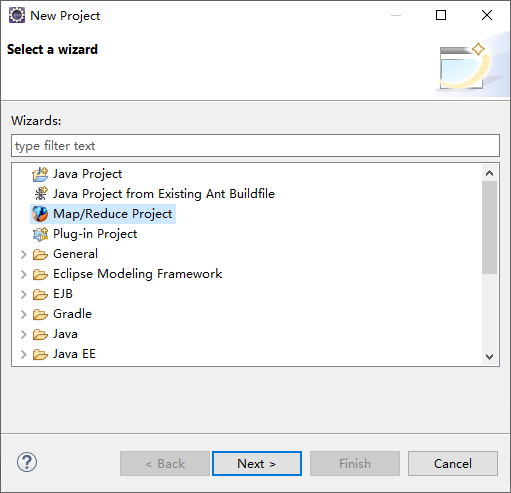


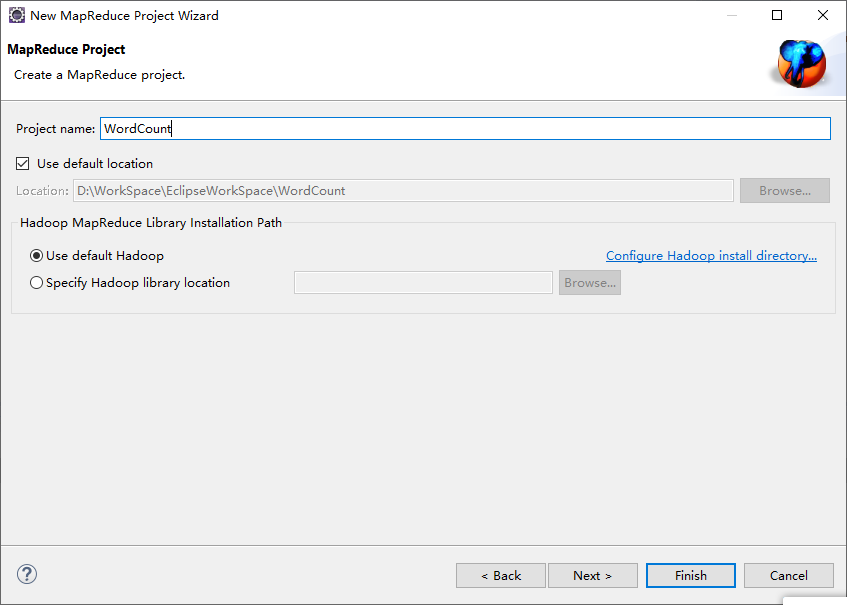


## 使用eclipse创建并运行wordcount程序

### 创建MapReduce程序



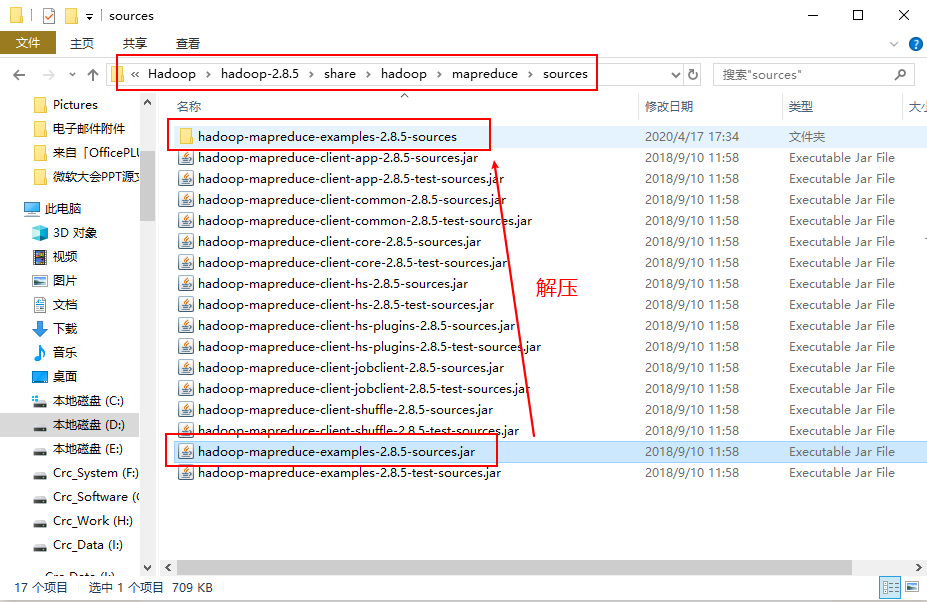


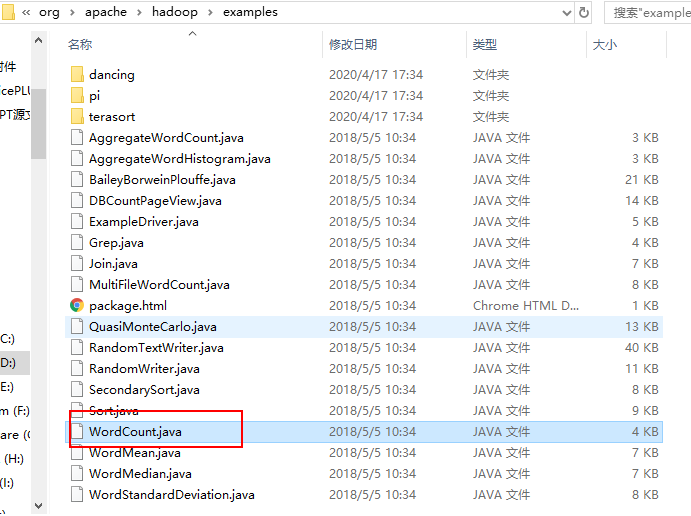


### 编写WordCount程序

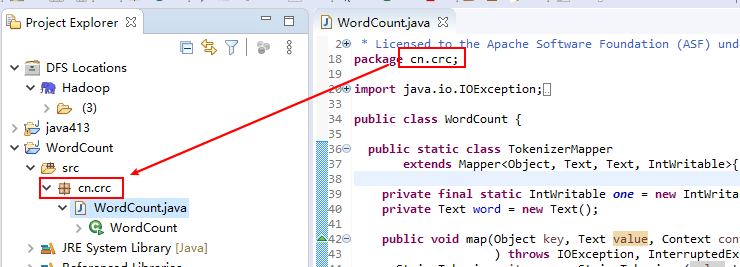
可以自己网上搜索代码

或者到Hadoop路径下share/Hadoop/mapreduce/sources下解压对应版本的examples

  
在解压文件的org\apache\hadoop\examples这个文件夹下面找到wordcount程序

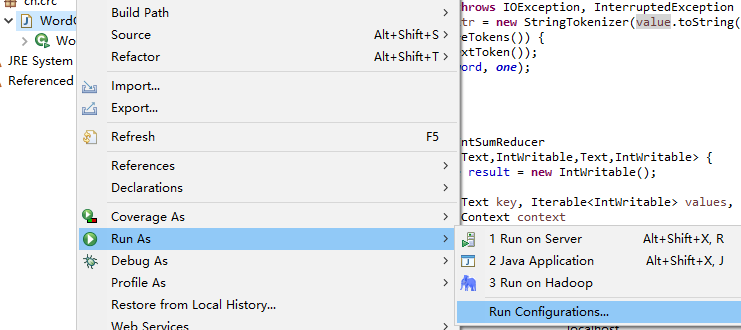


复制到当前项目下

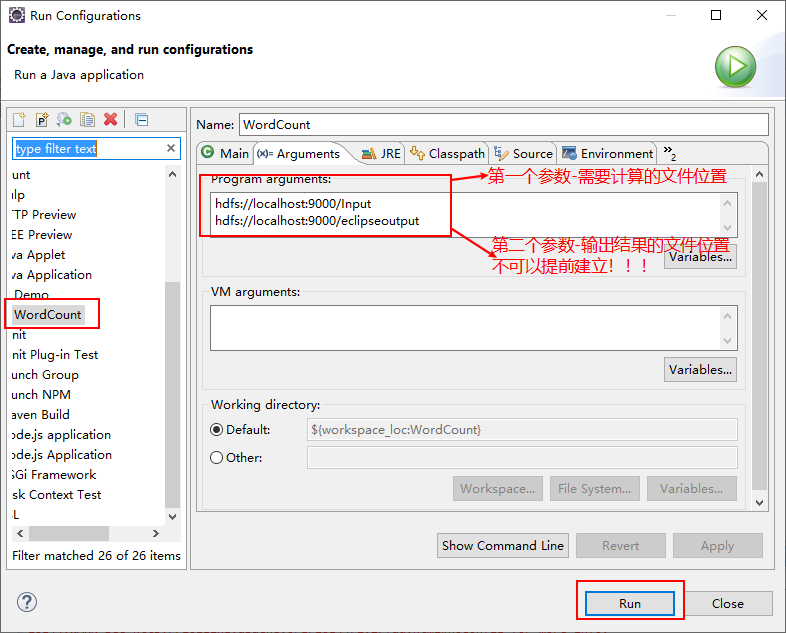


修改一下程序包名

右键点击Run As，选择Configurations



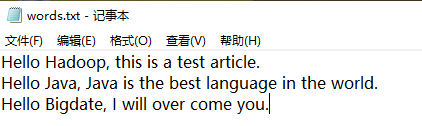
### 配置参数



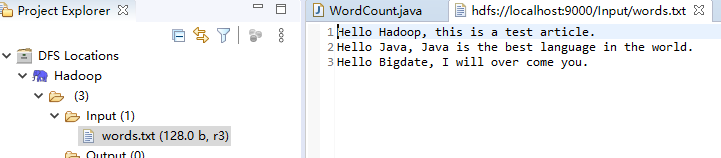
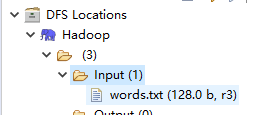
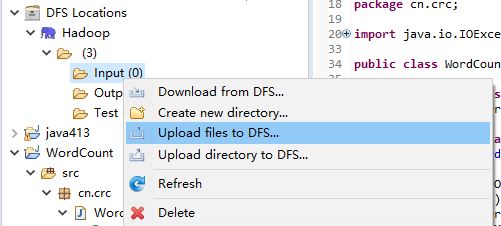


运行显示没有输入文件。

创建一个待统计的文件



上传到配置的输入路径



### 运行结果

运行之后可以看到在HDFS文件下创建了结果文件夹，同时出现了统计结果。

