大数据课程学习笔记

# HDFS简介及常用操作

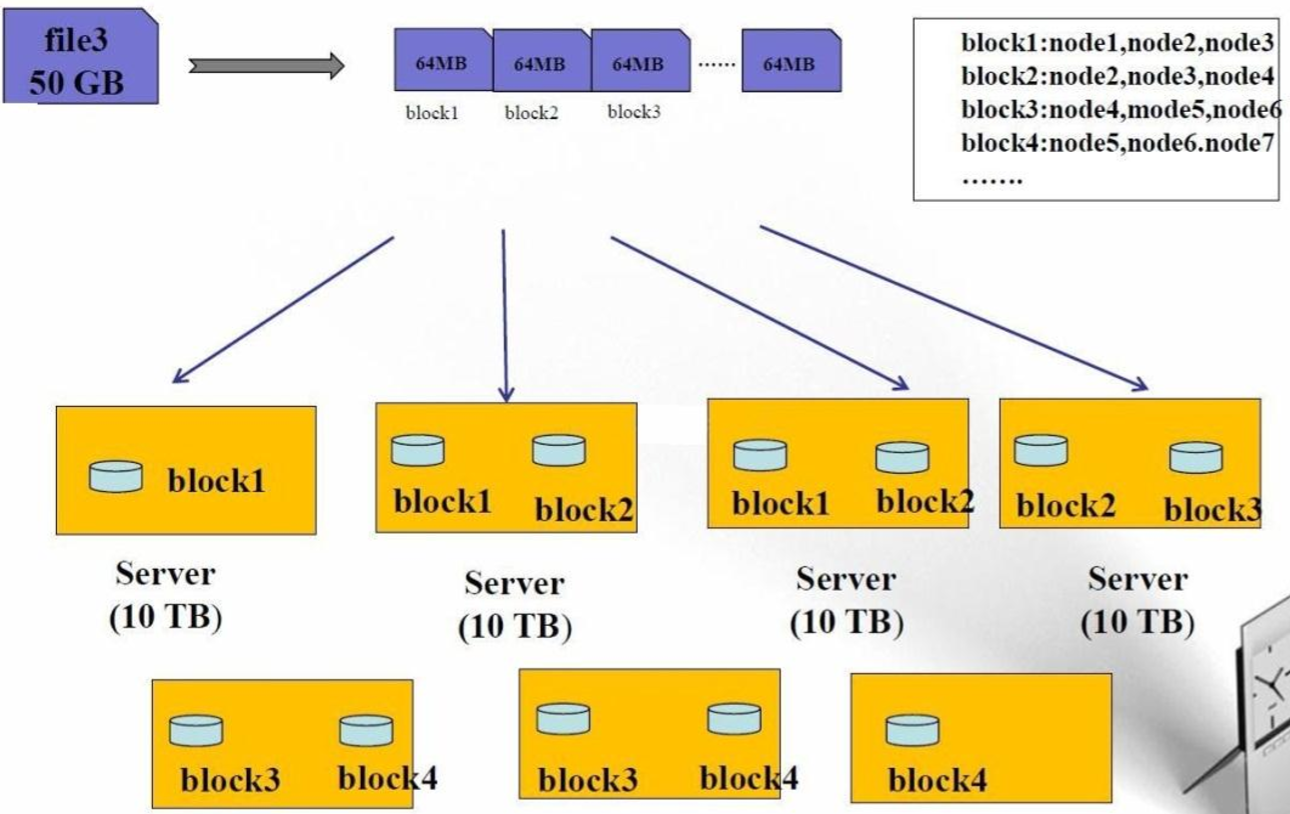
HDFS（Hadoop Distributed File System）， Hadoop 分布式文件系统，主要用来解决海量数据的存储问题，在大数据系统架构中为各类分布式运算框架（MapReduce，Spark，Tez，Flink，…）提供数据存储服务

* 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

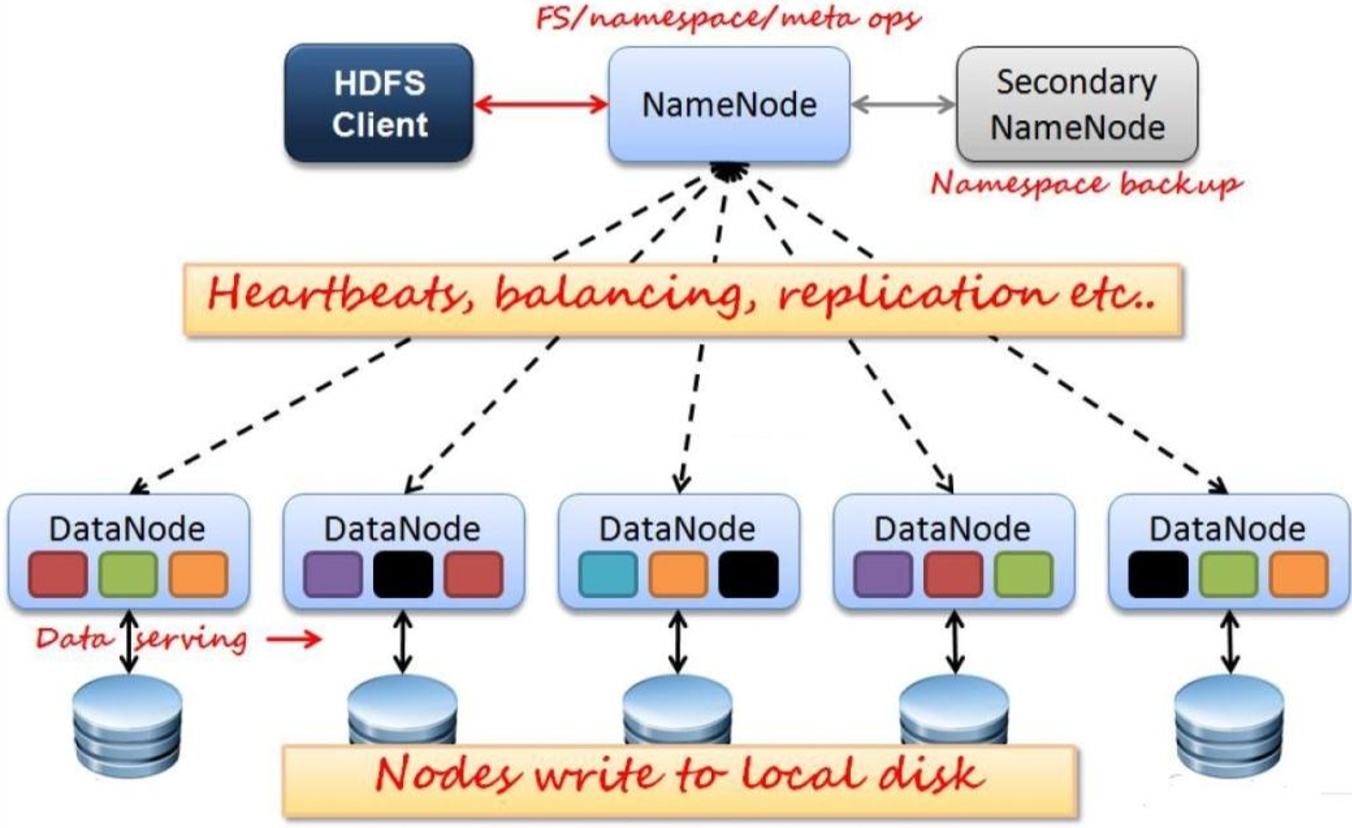
HDFS 被设计成用来使用低廉的服务器来进行海量数据的存储，那是怎么做到的呢？  
1、 大文件被切割成小文件，使用分而治之的思想让很多服务器对同一个文件进行联合管理  
2、 每个小文件做冗余备份，并且分散存到不同的服务器，做到高可靠不丢失



### 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

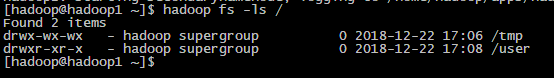
* HDFS 中的文件在物理上是分块存储（block），块的大小可以通过配置参数(dfs.blocksize) 来规定，默认大小在 hadoop2.x 版本中是 128M，1.x 是 64M ，3.x 是 256M
* HDFS 文件系统会给客户端提供一个统一的抽象目录树，客户端通过路径来访问文件，形如：hdfs://hadoop1:9000/soft/hadoop-2.7.5-centos-6.7.tar.gz
* 目录结构及文件分块位置信息（metadata,元数据）的管理由 namenode 节点承担。 namenode 是 HDFS 集群主节点，负责维护整个 hdfs 文件系统的目录树，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

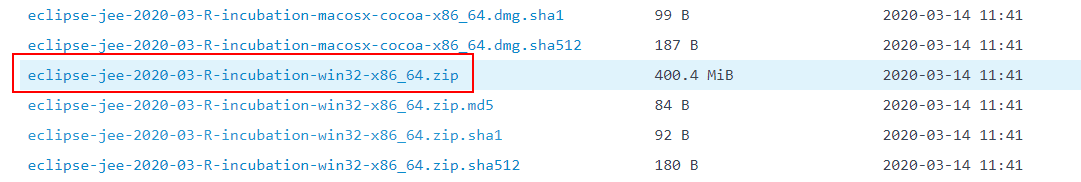
### 利用 Eclipse 可视化 HDFS 上的文件信息

下载eclipse

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

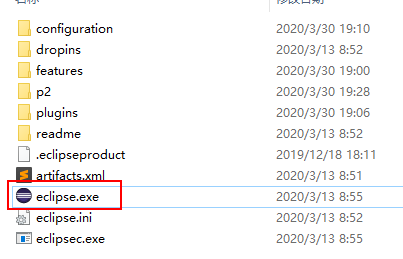
官网下载较慢，可以使用镜像下载：

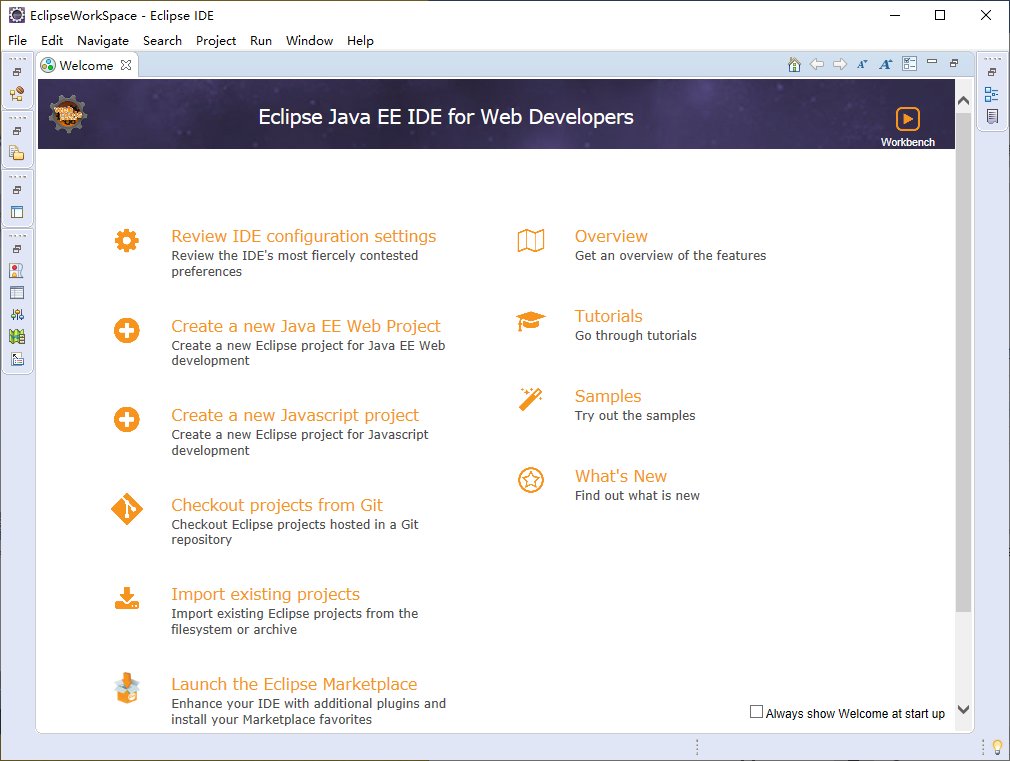
<https://mirrors.tuna.tsinghua.edu.cn/eclipse/technology/epp/downloads/release/>



选择eclipse-jee-xxx-win32-xxx版本

下载解压，然后运行exe



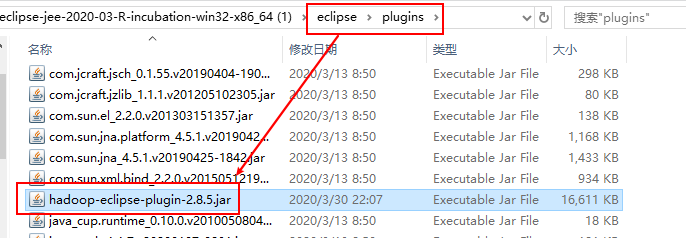


### 编译hadoop-eclipse插件

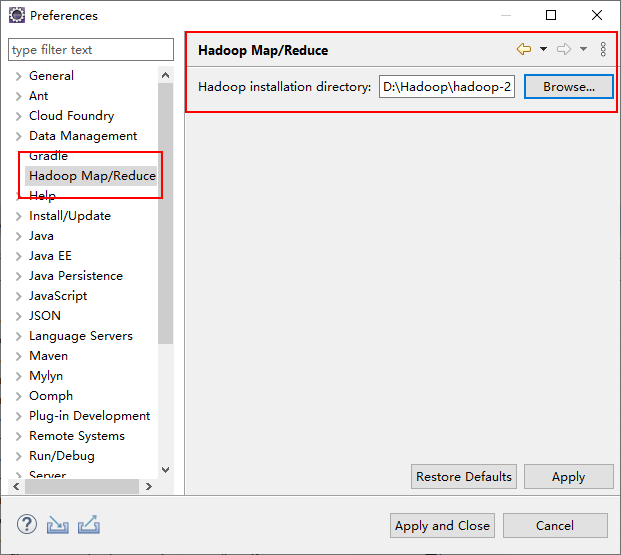
可以下载ant自行编译对应环境的插件，本次用的编译好的插件。

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

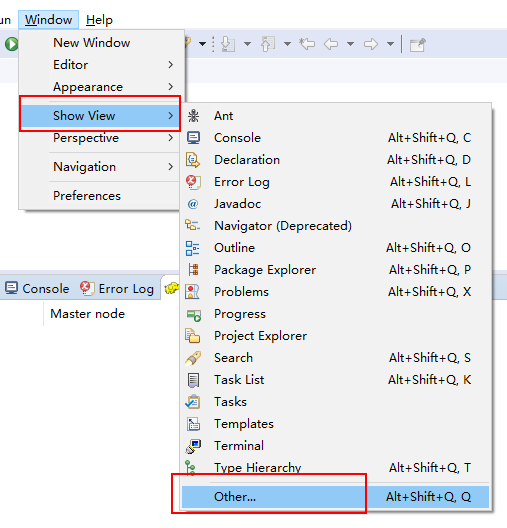


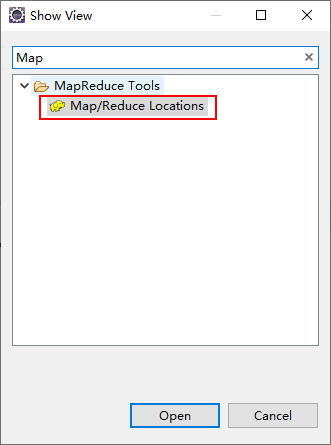
启动 eclipse ，点击 Window -> Prefrences ，在弹出的框选中Hadoop Map/Reduce ，然后在右边配置hadoop解压的路径

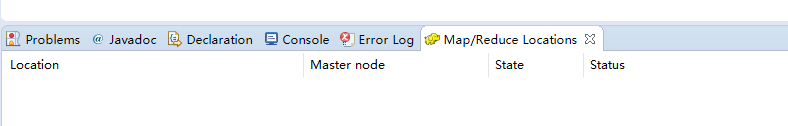


### 切换"Map/Reduce"工作目录

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



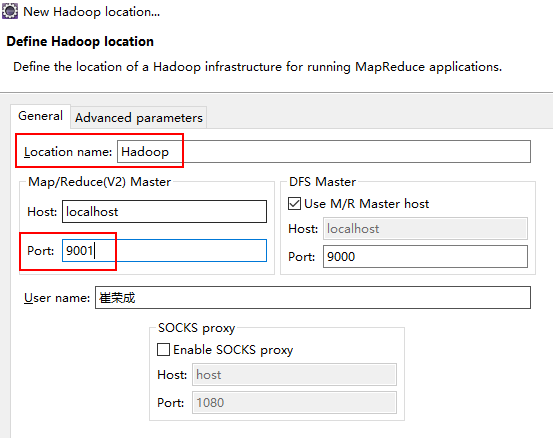




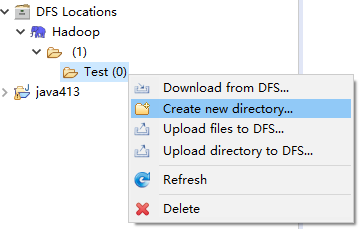
### 建立与Hadoop集群的连接

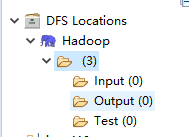
在Eclipse软件下面的"Map/Reduce Locations"进行右击，弹出一个选项，选择"New Hadoop Location"，然后弹出一个窗体。





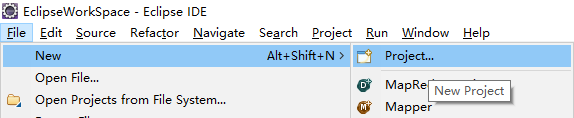
连接成功后可以看到HDFS结构，因为第一次连接所以没有东西，可以右击闯进文件夹。



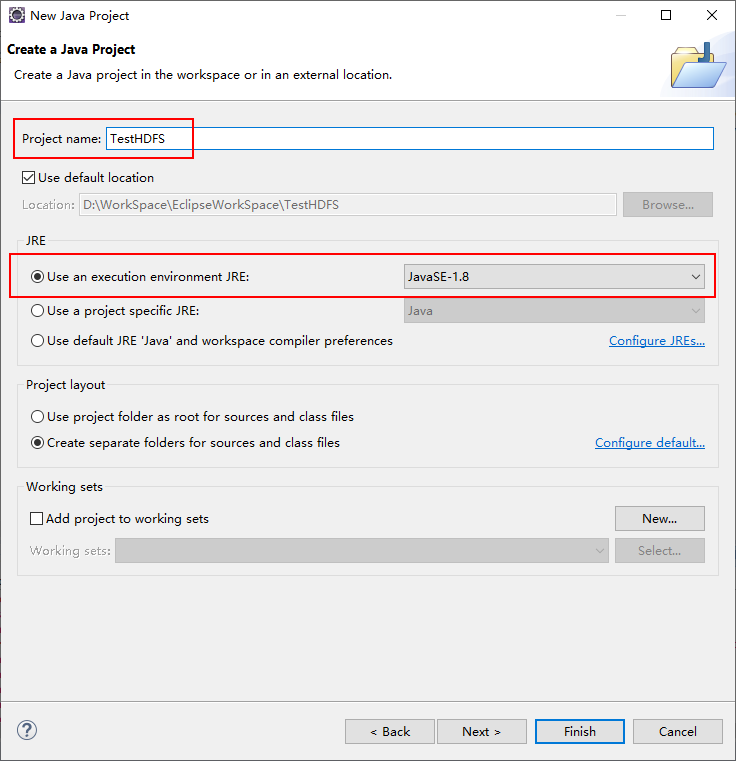


## Eclipse编程进行hdfs常用操作

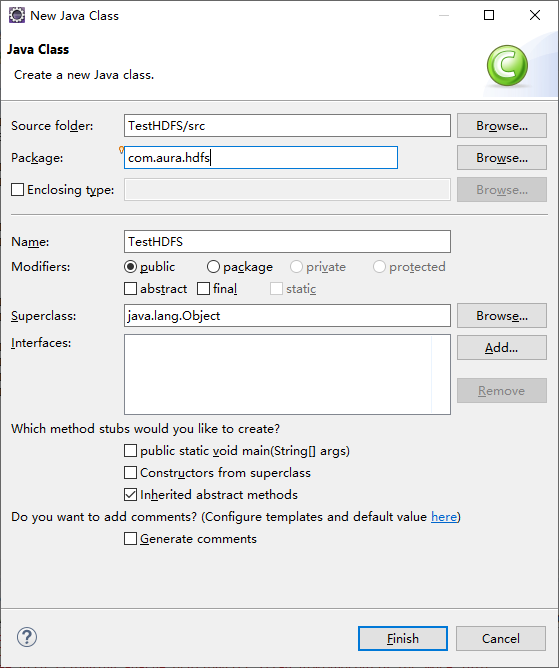
### 创建TestHDFS程序



选择Java程序

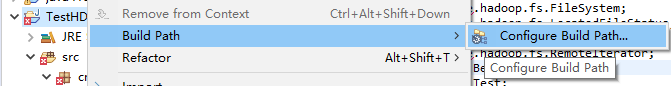


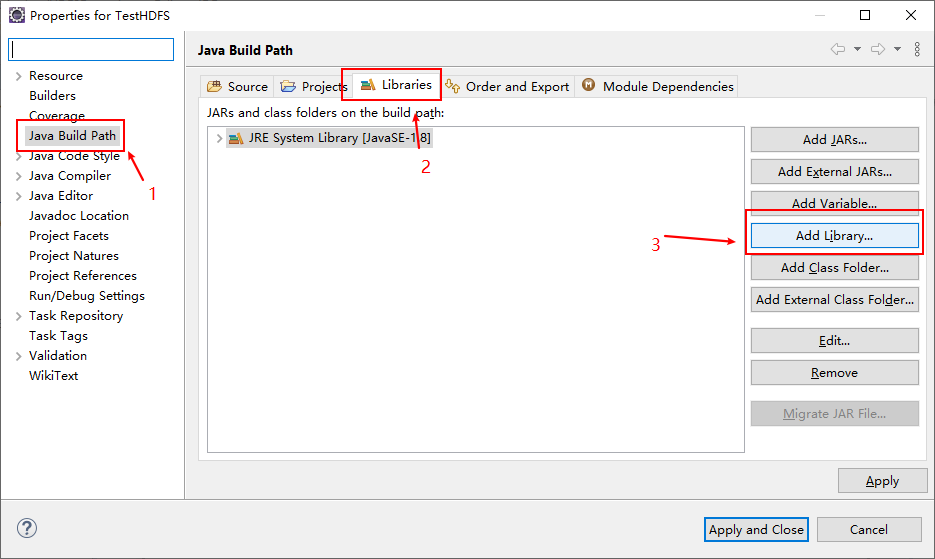
### 创建TestHDFS类

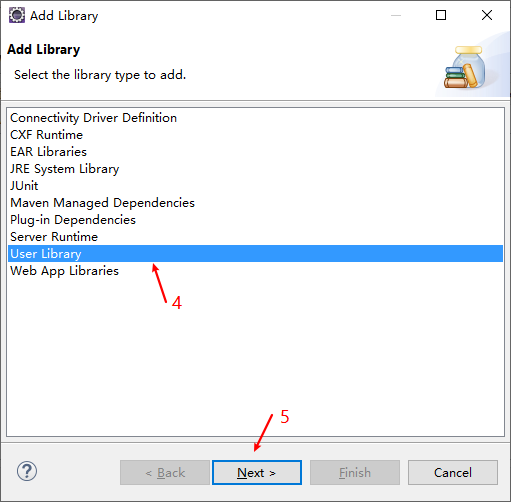


### 添加依赖库

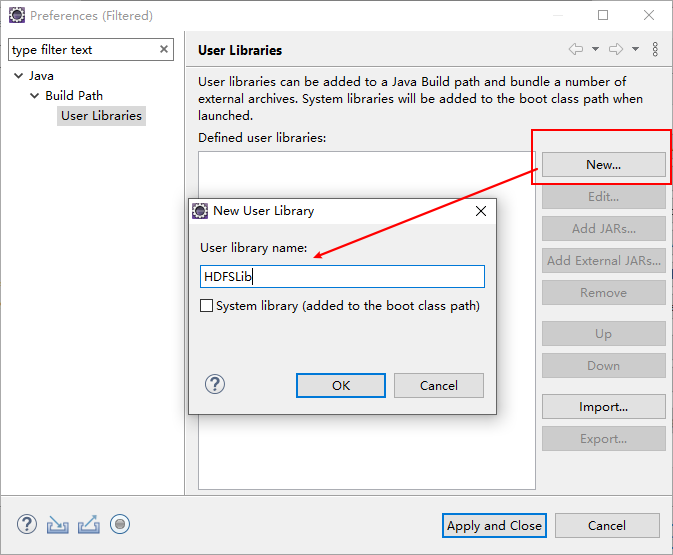
右击项目，选择build path，进行配置

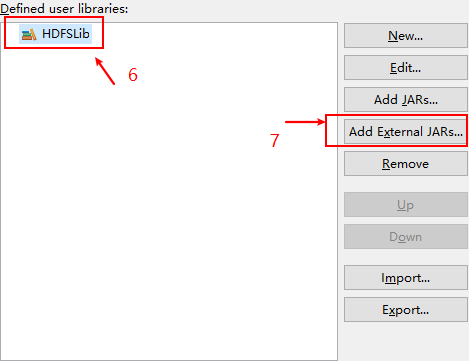




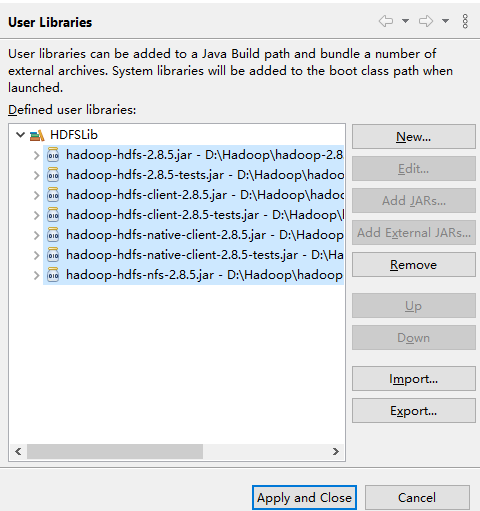


接下来选中User Libraries -> New ，输入HDFS库的名字

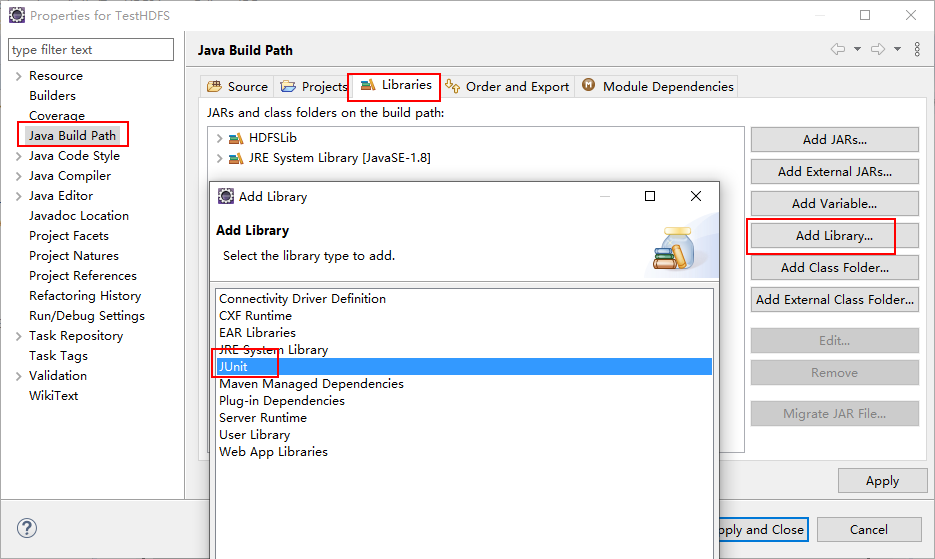


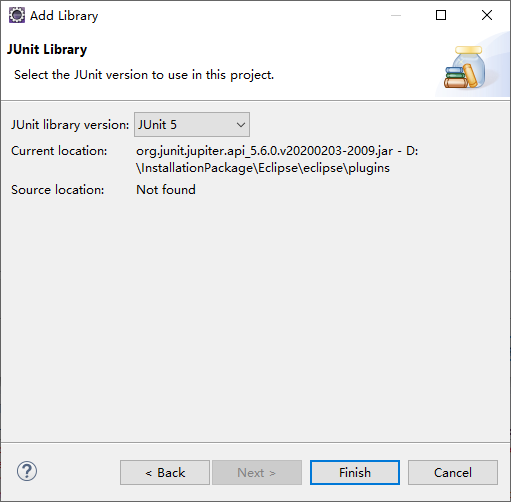


然后把hadoop目录下/share/hadoop/hdfs的所有jar包放进去



### 添加junit测试单元

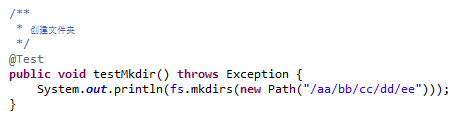


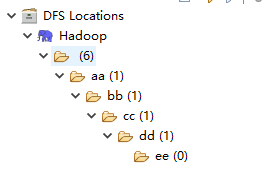


### 创建fs对象

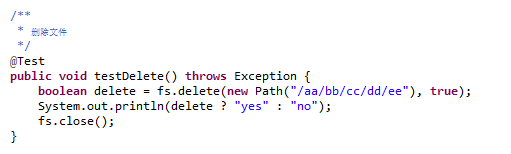


### 创建文件夹



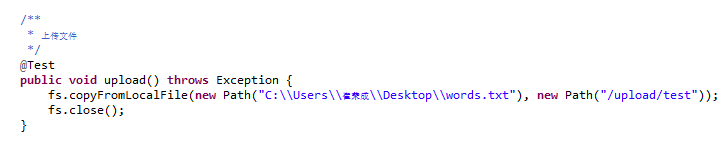


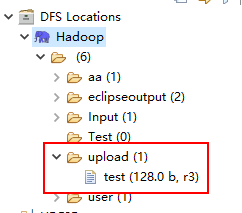
### 删除文件



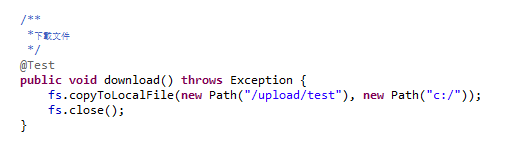


### 上传文件

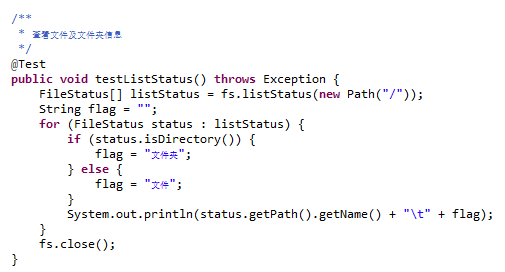


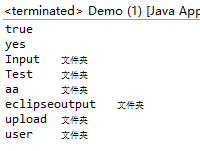


### 下载文件



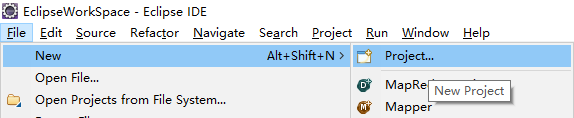
### 查看文件及文件夹信息

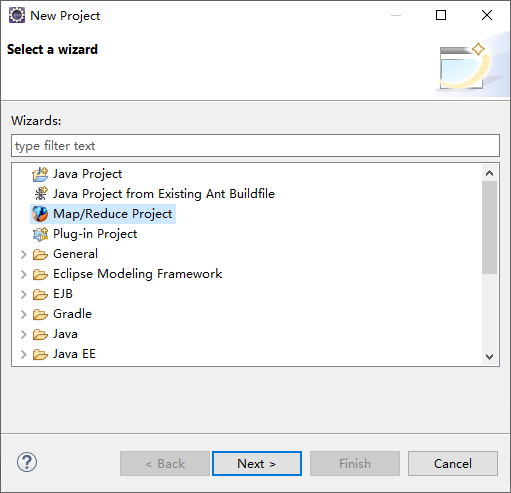


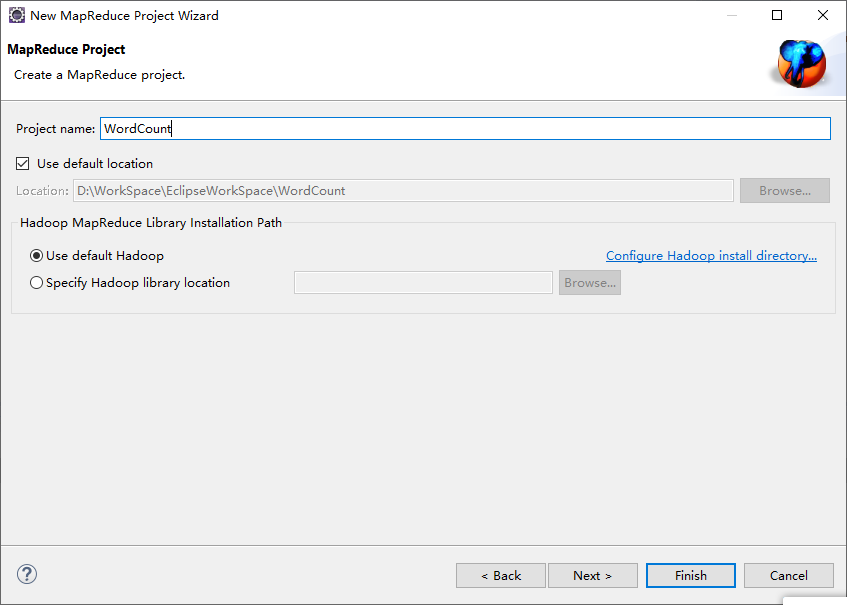


## 使用eclipse创建并运行wordcount程序

### 创建MapReduce程序



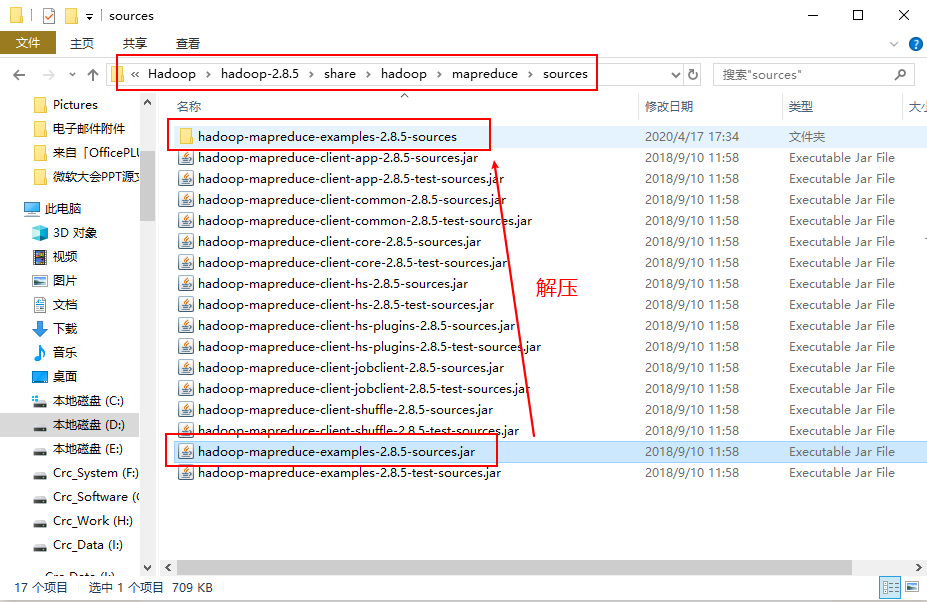


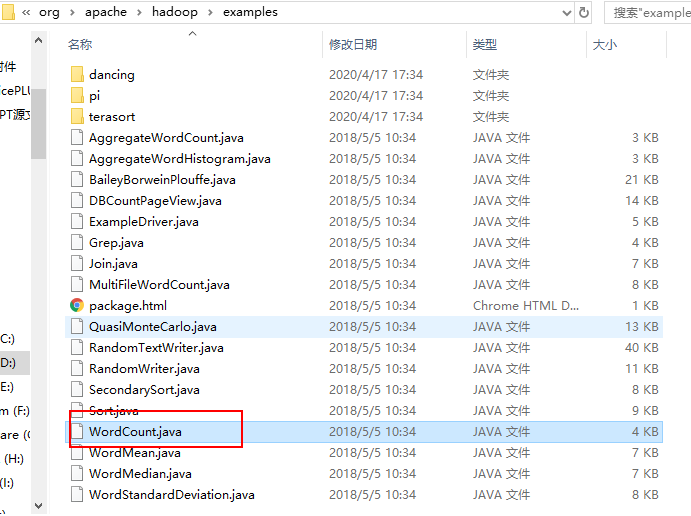


### 编写WordCount程序

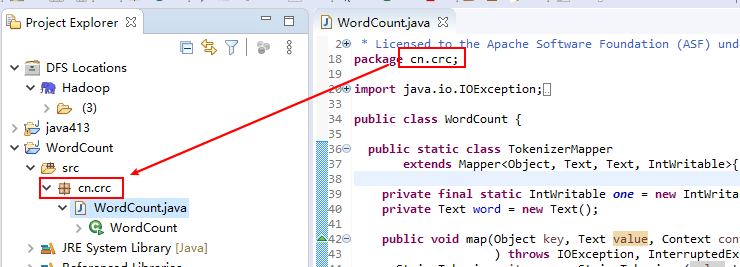
可以自己网上搜索代码

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

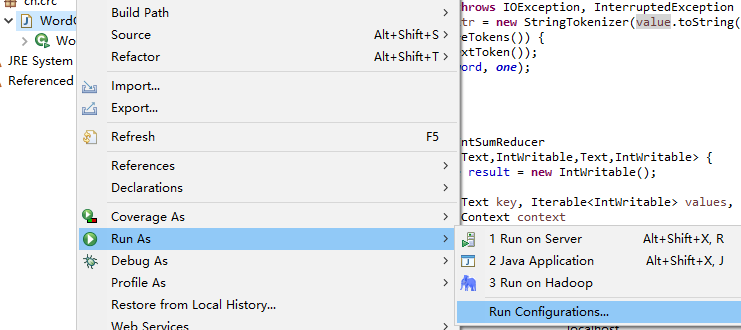


复制到当前项目下

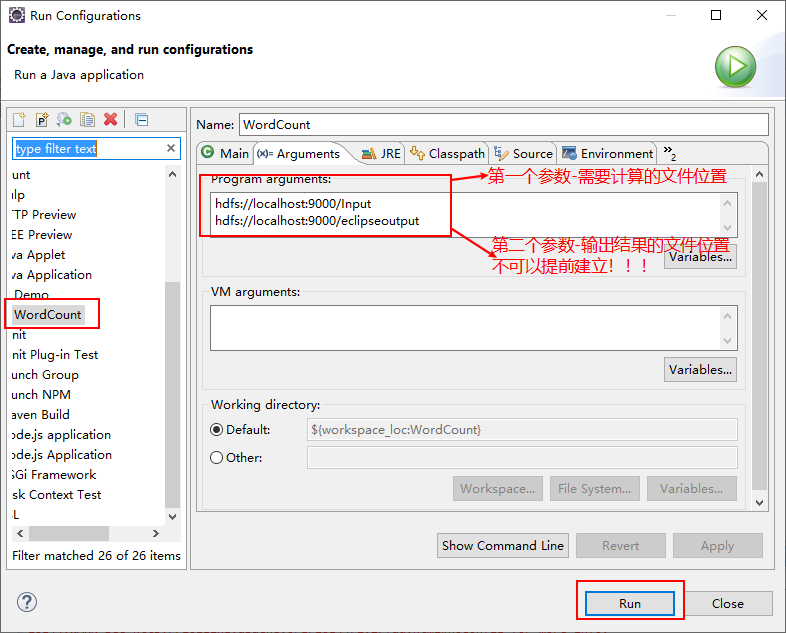


修改一下程序包名

右键点击Run As，选择Configurations



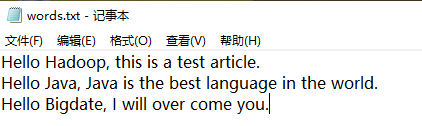
配置参数



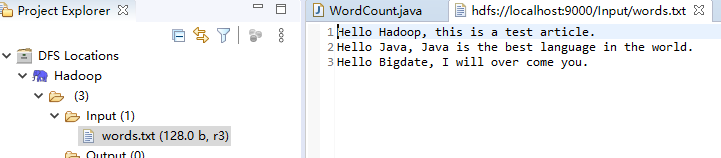
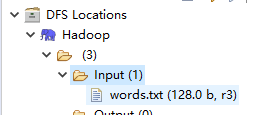
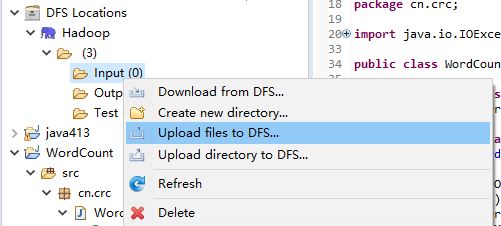


运行显示没有输入文件。

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。

