细细品味Hadoop

——Hadoop 集群（第 7 期）

精

华

集

锦

csAxp

虾皮工作室

http://www.cnblogs.com/xia520pi/

2012 年 5 月 20 日

Hadoop 集群（第 7 期）

——Eclipse 开发环境设置

1、Hadoop开发环境简介

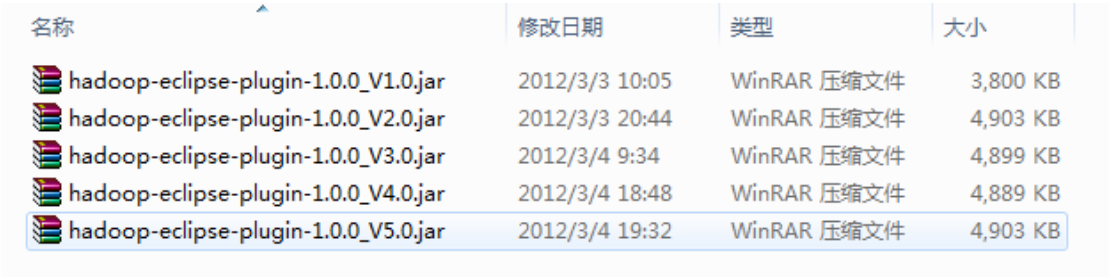
1.1 Hadoop集群简介

Java 版本：jdk-6u31-linux-i586.bin
Linux 系统：CentOS6.0
Hadoop 版本：hadoop-1.0.0.tar.gz

1.2 Windows开发简介

Java 版本：jdk-6u31-windows-i586.exe
Win 系统：Windows 7 旗舰版
Eclipse 软件：eclipse-jee-**indigo**-SR1-win32.zip | eclipse-jee-**helios**-SR2-win32.zip
Hadoop 软件：hadoop-1.0.0.tar.gz
Hadoop Eclipse 插件：hadoop-eclipse-plugin-1.0.0.jar
下载地址：http://download.csdn.net/detail/xia520pi/4113746

备注：下面是网上收集的收集的“hadoop-eclipse-plugin-1.0.0.jar”，除“版本 2.0”是根据“V1.0”按照“常见问题 FAQ_**1**”改的之外，剩余的“V3.0”、“V4.0”和“V5.0”和“V2.0”一样是别人已经弄好的，而且我已经都测试过，没有任何问题，可以放心使用。我们这里选择第“V5.0”使用。记得在使用时**重新命名**为“**hadoop-eclipse-plugin-1.0.0.jar**”。



名称	修改日期	类型	大小
hadoop-eclipse-plugin-1.0.0_V1.0.jar	2012/3/3 10:05	WinRAR 压缩文件	3,800 KB
hadoop-eclipse-plugin-1.0.0_V2.0.jar	2012/3/3 20:44	WinRAR 压缩文件	4,903 KB
hadoop-eclipse-plugin-1.0.0_V3.0.jar	2012/3/4 9:34	WinRAR 压缩文件	4,899 KB
hadoop-eclipse-plugin-1.0.0_V4.0.jar	2012/3/4 18:48	WinRAR 压缩文件	4,889 KB
hadoop-eclipse-plugin-1.0.0_V5.0.jar	2012/3/4 19:32	WinRAR 压缩文件	4,903 KB

2、Hadoop Eclipse简介和使用

2.1 Eclipse插件介绍

Hadoop 是一个强大的并行框架，它允许任务在其分布式集群上并行处理。但是编写、调试 Hadoop 程序都有很大难度。正因为如此，Hadoop 的开发者开发出了 Hadoop Eclipse 插件，它在 Hadoop 的开发环境中嵌入了 Eclipse，从而实现了开发环境的图形化，降低了编

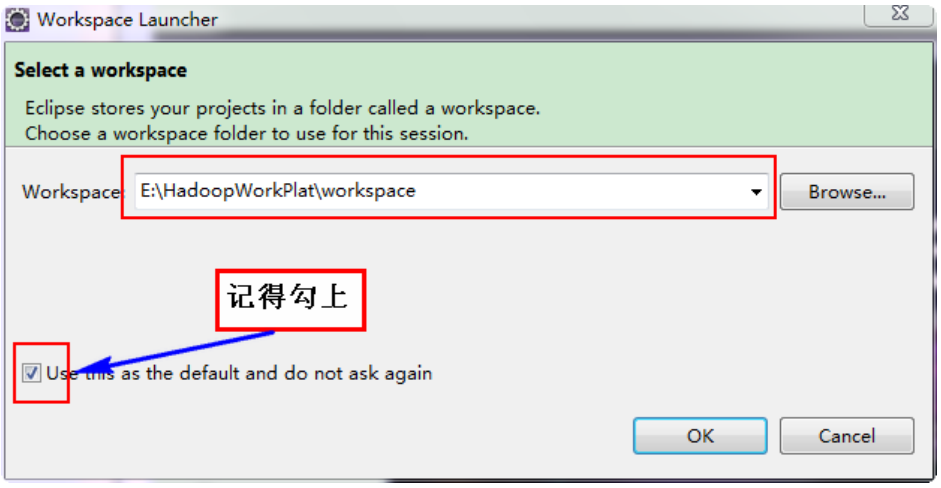
程难度。在安装插件，配置 Hadoop 的相关信息之后，如果用户创建 Hadoop 程序，插件会自动导入 Hadoop 编程接口的 JAR 文件，这样用户就可以在 Eclipse 的图形化界面中编写、调试、运行 Hadoop 程序（包括单机程序和分布式程序），也可以在其中查看自己程序的实时状态、错误信息和运行结果，还可以查看、管理 HDFS 以及文件。总地来说，Hadoop Eclipse 插件安装简单，使用方便，功能强大，尤其是在 Hadoop 编程方面，是 Hadoop 入门和 Hadoop 编程必不可少的工具。

2.2 Hadoop工作目录简介

为了以后方便开发，我们按照下面把开发中用到的软件安装在此目录中，JDK 安装除外，我这里把 JDK 安装在 C 盘的默认安装路径下，下面是我的工作目录：

```
系统磁盘（E:）
    |---HadoopWorkPlat
        |--- eclipse
        |--- hadoop-1.0.0
        |--- workplace
        |---……
```

按照上面目录把 Eclipse 和 Hadoop 解压到“E:\HadoopWorkPlat”下面，并创建“workplace”作为 Eclipse 的工作空间。



备注：大家可以按照自己的情况，不一定按照我的结构来设计。

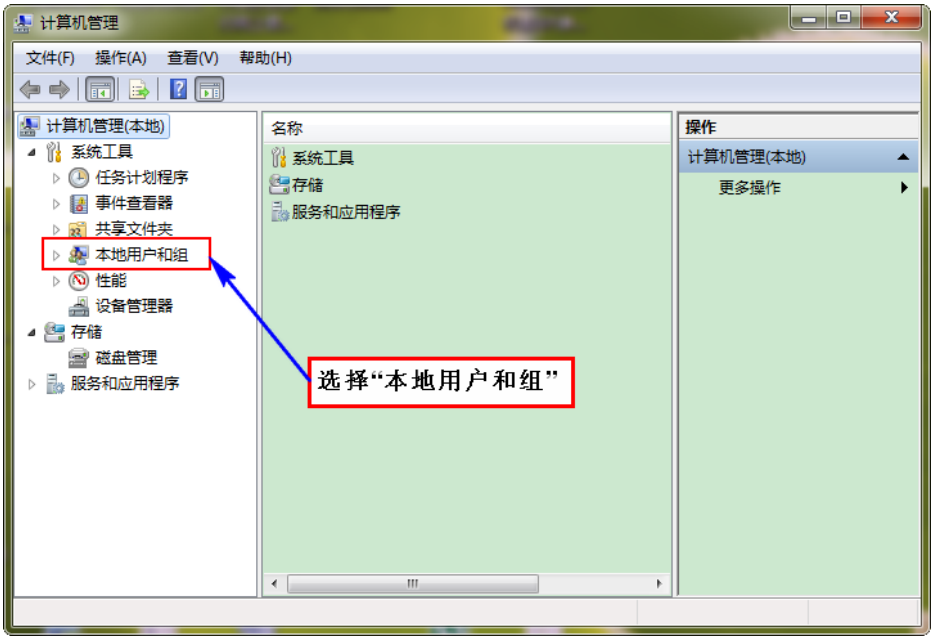
2.3 修改系统管理员名字

经过两天多次探索，为了使 Eclipse 能正常对 Hadoop 集群的 HDFS 上的文件能进行修改和删除，所以修改你工作时所用的 Win7 系统管理员名字，默认一般为“**Administrator**”，把它修改为“**hadoop**”，此用户名与 Hadoop 集群普通用户一致，大家应该记得我们 Hadoop 集群中所有的机器都有一个普通用户——hadoop，而且 Hadoop 运行也是用这个用户进行的。

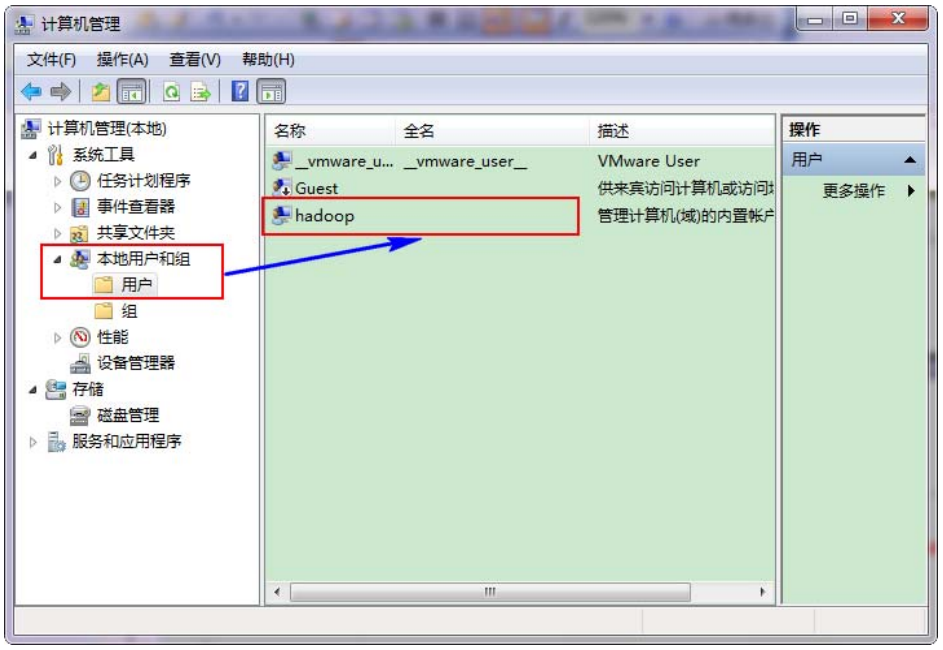
为了不至于为权限苦恼，我们可以修改 Win7 上系统管理员的姓名，这样就避免出现该用户在 Hadoop 集群上没有权限等都疼问题，会导致在 Eclipse 中对 Hadoop 集群的 HDFS 创建和删除文件受影响。

你可以做一下实验，查看 Master.Hadoop 机器上“/usr/hadoop/logs”下面的日志。发现权限不够，不能进行“Write”操作，网上有几种解决方案，但是对 Hadoop1.0 不起作用，详情见“常见问题 FAQ_2”。下面我们进行修改管理员名字。

首先“右击”桌面上图标“我的电脑”，选择“管理”，弹出界面如下：



接着选择“本地用户和组”，展开“用户”，找到系统管理员“Administrator”，修改其为“hadoop”，操作结果如下图：



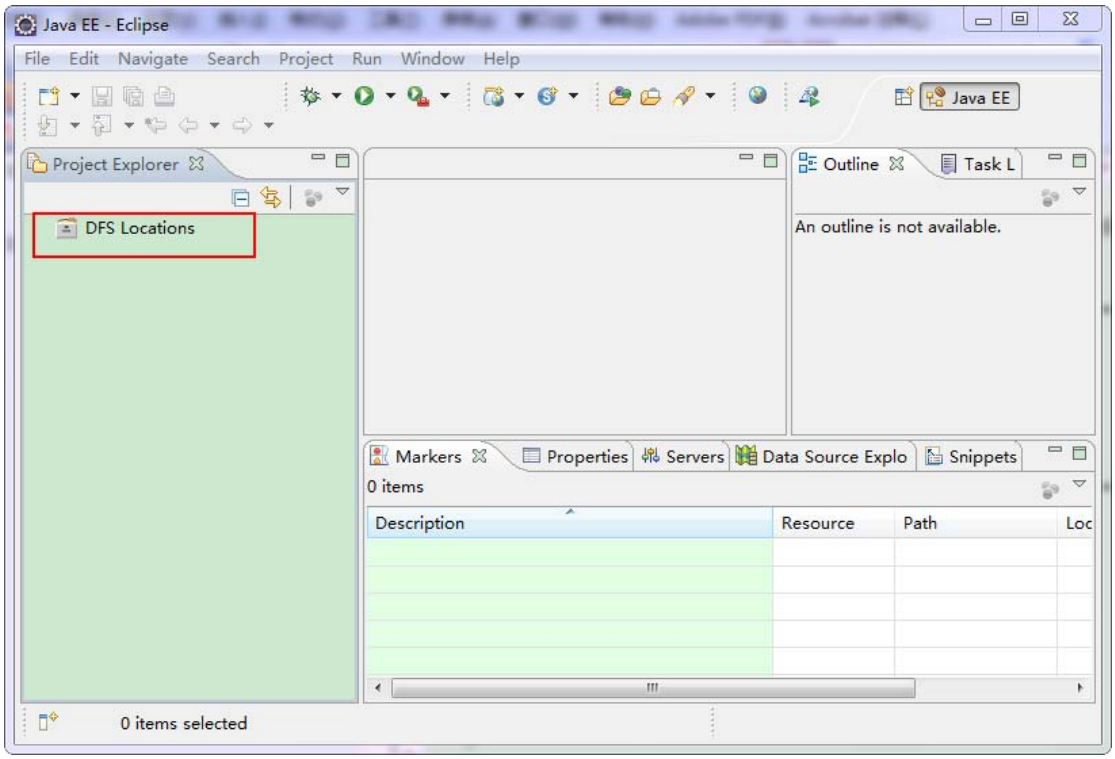
最后，把电脑进行“注销”或者“重启电脑”，这样才能使管理员才能用这个名字。

2.4 Eclipse插件开发配置

第一步：把我们的“hadoop-eclipse-plugin-1.0.0.jar”放到 Eclipse 的目录的“plugins”中，然后重新 Eclipse 即可生效。

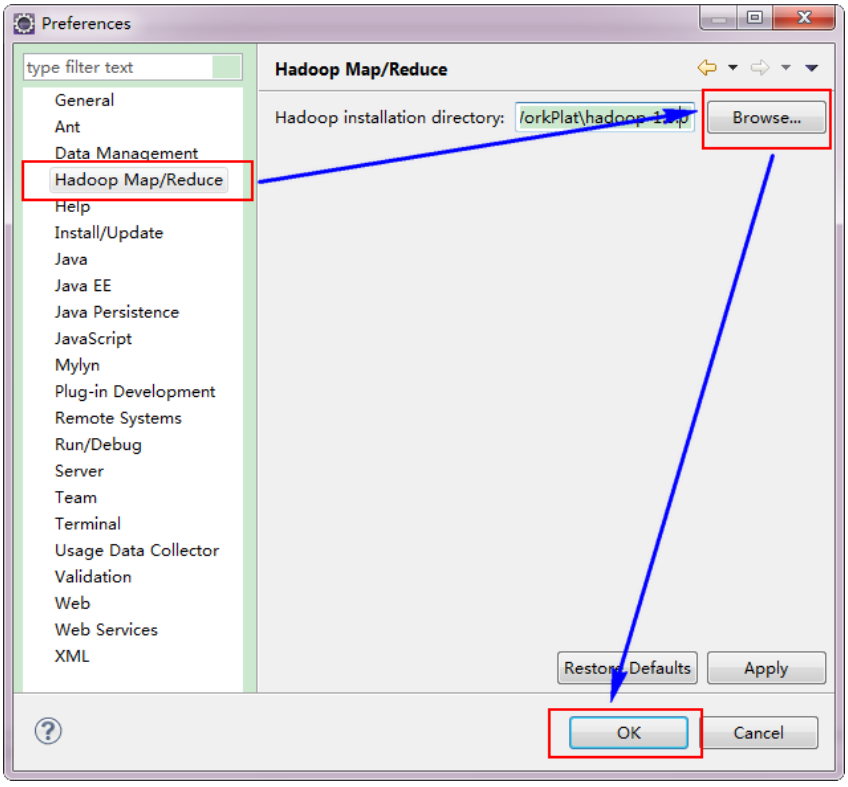
```
系统磁盘（E:）
    |---HadoopWorkPlat
        |--- eclipse
            |--- plugins
                |--- hadoop-eclipse-plugin-1.0.0.jar
```

上面是我的“hadoop-eclipse-plugin”插件放置的地方。重启 Eclipse 如下图：



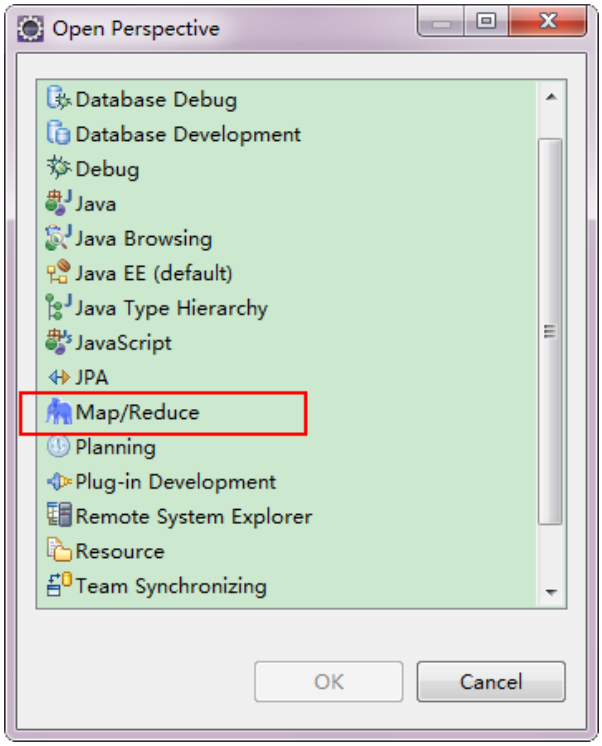
细心的你从上图中左侧“Project Explorer”下面发现“DFS Locations”，说明 Eclipse 已经识别刚才放入的 Hadoop Eclipse 插件了。

第二步：选择“Window”菜单下的“Preference”，然后弹出一个窗体，在窗体的左侧，有一列选项，里面会多出“Hadoop Map/Reduce”选项，点击此选项，选择 Hadoop 的安装目录（如我的 Hadoop 目录：E:\HadoopWorkPlat\hadoop-1.0.0）。结果如下图：



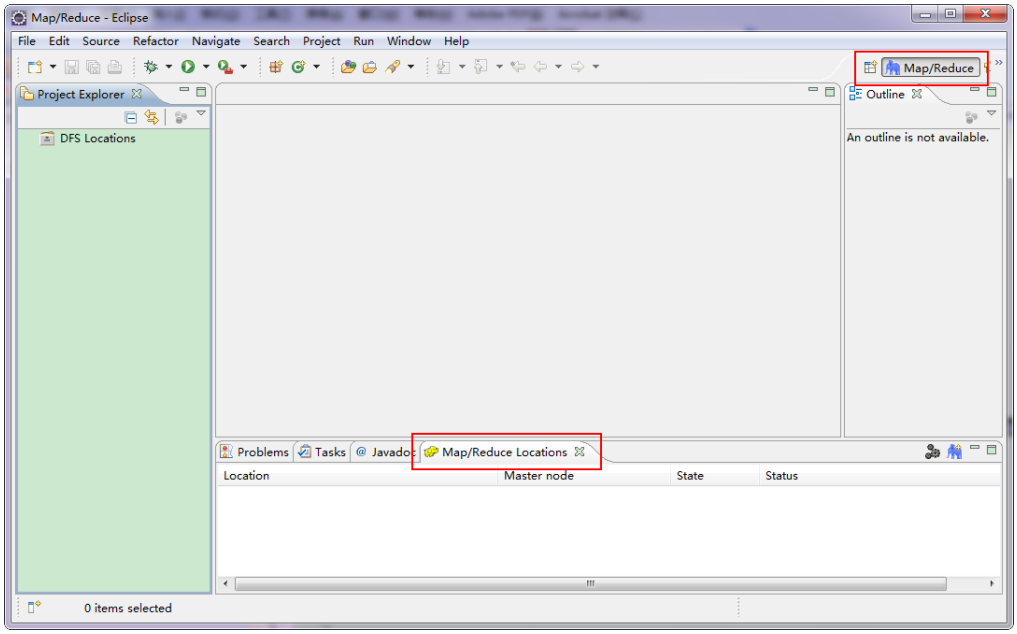
第三步：切换“Map/Reduce”工作目录，有两种方法：

1)选择“Window”菜单下选择“Open Perspective”，弹出一个窗体，从中选择“Map/Reduce”选项即可进行切换。

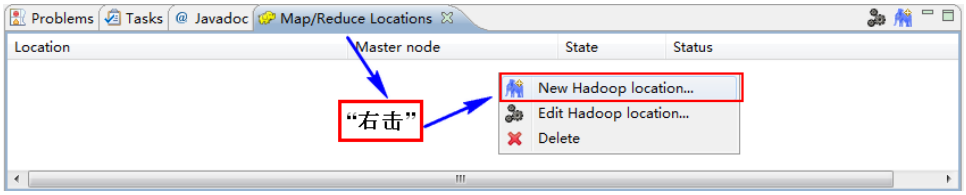


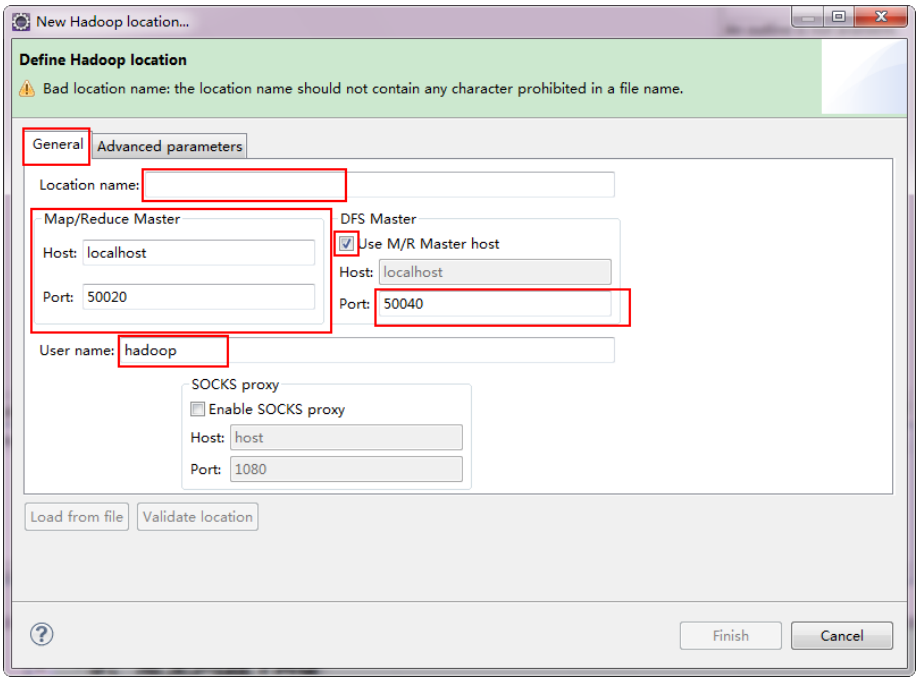
2）在 Eclipse 软件的**右上角**，点击图标“  ”中的“  ”，点击“Other”选项，也可以弹出上图，从中选择“Map/Reduce”，然后点击“OK”即可确定。

切换到“Map/Reduce”工作目录下的界面如下图所示。



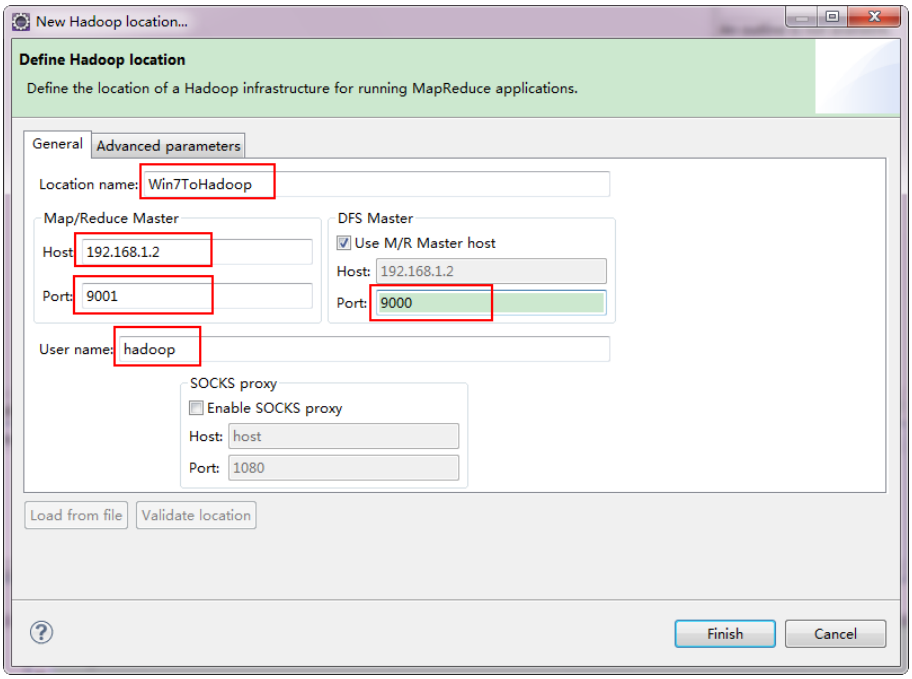
第四步：建立与 Hadoop 集群的连接，在 Eclipse 软件下面的“**Map/Reduce Locations**”进行**右击**，弹出一个选项，选择“**New Hadoop Location**”，然后弹出一个窗体。





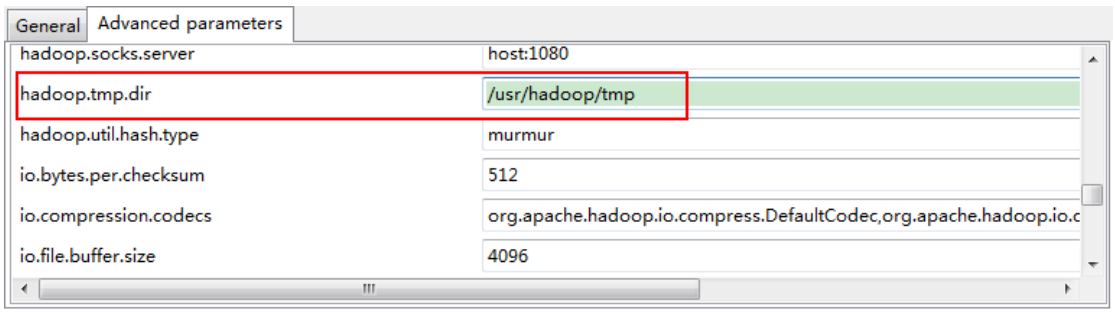
注意上图中的红色标注的地方，是需要我们关注的地方。

- Location Name：可以任意其，标识一个“Map/Reduce Location”
- Map/Reduce Master
  Host：192.168.1.2（**Master.Hadoop** 的 IP 地址）
  Port：9001
- DFS Master
  Use M/R Master host：前面的**勾上**。（因为我们的 NameNode 和 JobTracker 都在一个机器上。）
  Port：9000
- User name：hadoop（默认为 Win 系统管理员名字，因为我们之前改了所以这里就变成了 hadoop。）

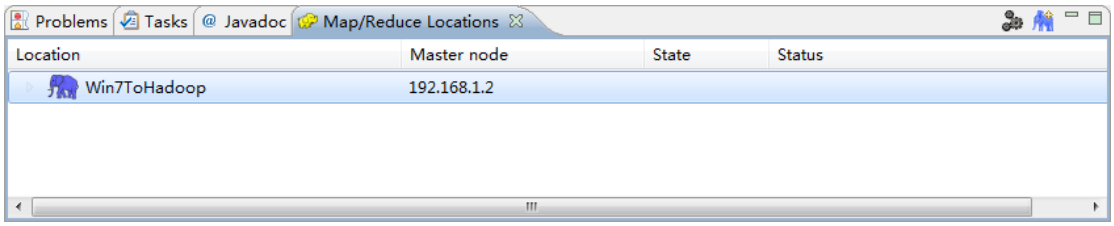


备注：这里面的 Host、Port 分别为你在 mapred-site.xml、core-site.xml 中配置的地址及端口。不清楚的可以参考“**Hadoop 集群_第 5 期_Hadoop 安装配置_V1.0**”进行查看。

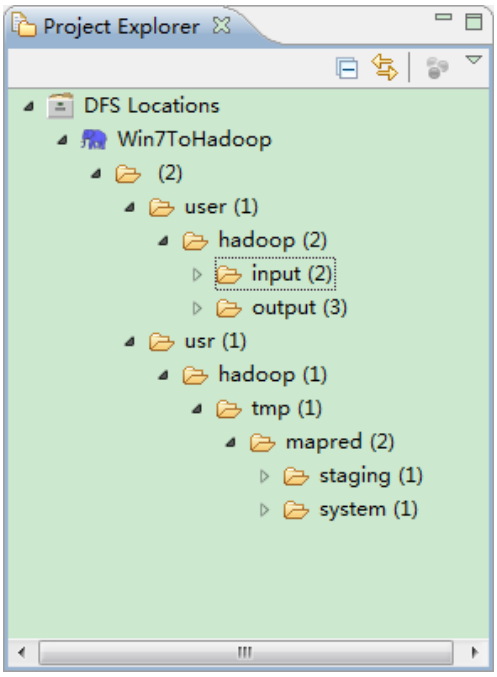
接着点击“Advanced parameters”从中找见“hadoop.tmp.dir”，修改成为我们 Hadoop 集群中设置的地址，我们的 Hadoop 集群是“**/usr/hadoop/tmp**”，这个参数在“core-site.xml”进行了配置。



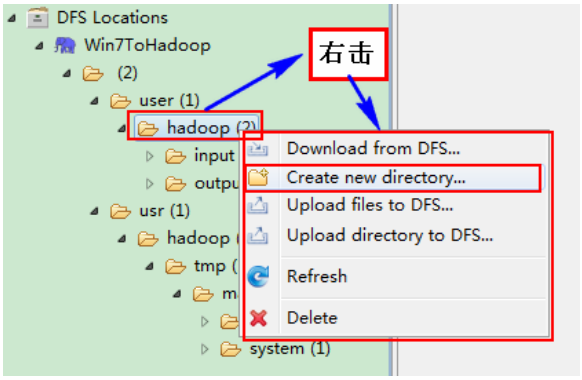
点击“finish”之后，会发现 Eclipse 软件下面的“Map/Reduce Locations”出现一条信息，就是我们刚才建立的“Map/Reduce Location”。



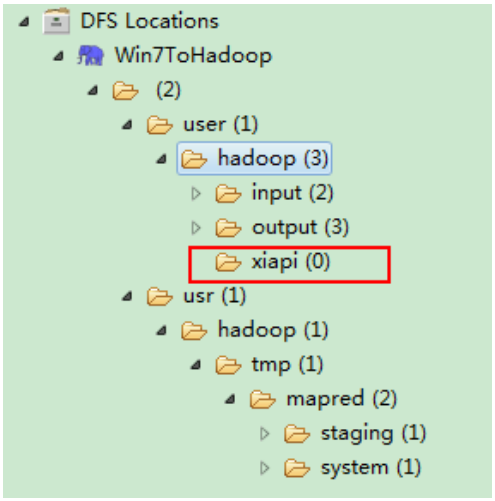
第五步：查看 HDFS 文件系统，并尝试建立文件夹和上传文件。点击 Eclipse 软件左侧的“DFS Locations”下面的“Win7ToHadoop”，就会展示出 HDFS 上的文件结构。



右击“Win7ToHadoop→user→**hadoop**”可以尝试建立一个“文件夹--xiapi”，然后右击刷新就能查看我们刚才建立的文件夹。



创建完之后，并刷新，显示结果如下：



用 SecureCRT 远程登录“Master.Hadoop”服务器，用下面命令查看是否已经建立一个“xiapi”的文件夹。

```
hadoop　fs　-ls
```

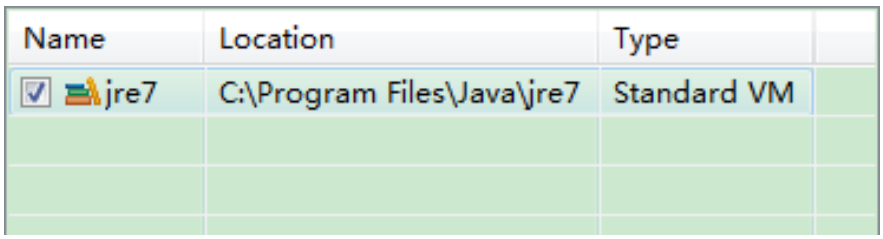
```
[hadoop@Master ~]$ hadoop fs -ls
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2012-03-02 05:45 /user/hadoop/input
drwxr-xr-x   - hadoop supergroup          0 2012-03-02 06:08 /user/hadoop/output
drwxr-xr-x   - hadoop supergroup          0 2012-03-06 01:11 /user/hadoop/xiapi
[hadoop@Master ~]$
```

到此为止，我们的 Hadoop Eclipse 开发环境已经配置完毕，不尽兴的同学可以上传点本地文件到 HDFS 分布式文件上，可以互相对比意见文件是否已经上传成功。

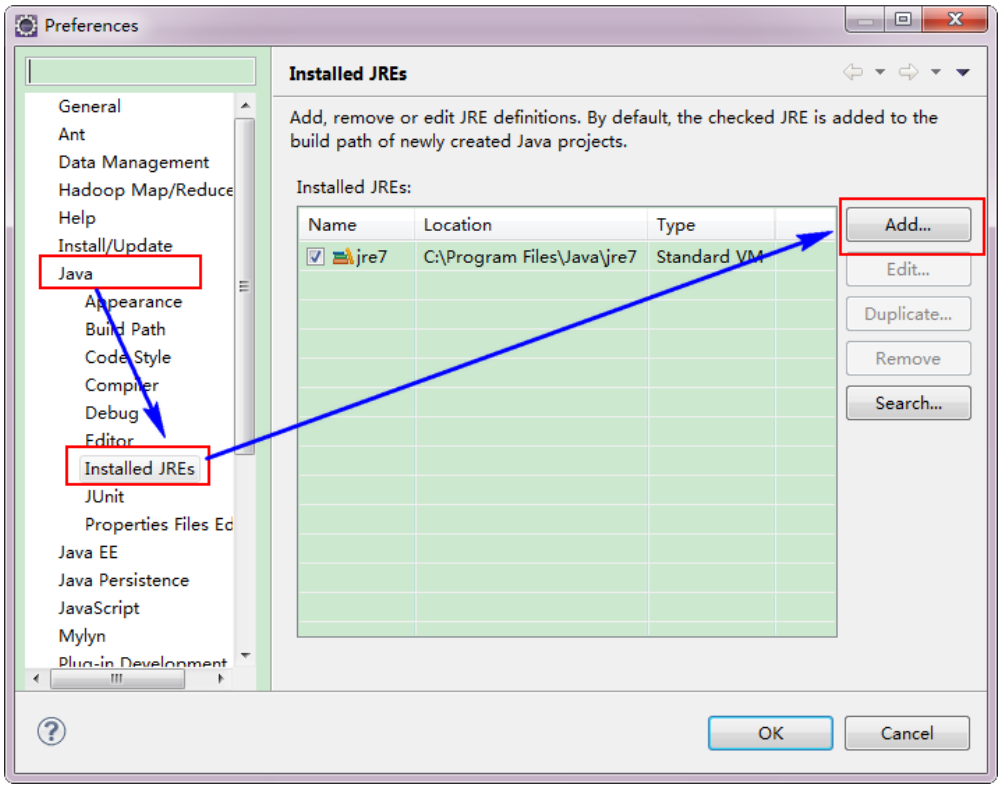
3、Eclipse运行WordCount程序

3.1 配置Eclipse的JDK

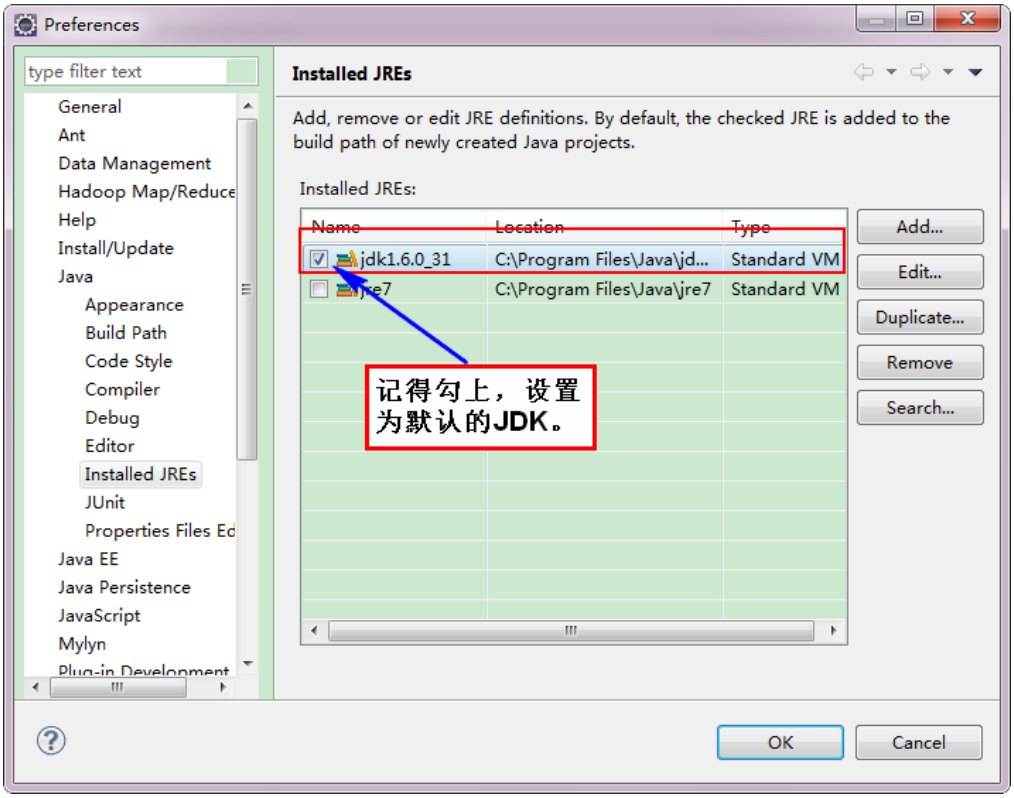
如果电脑上不仅仅安装的 JDK6.0，那么要确定一下 Eclipse 的平台的默认 JDK 是否 6.0。从“Window”菜单下选择“**Preference**”，弹出一个窗体，从窗体的左侧找见“Java”，选择“Installed JREs”，然后添加 JDK6.0。下面是我的默认选择 JRE。



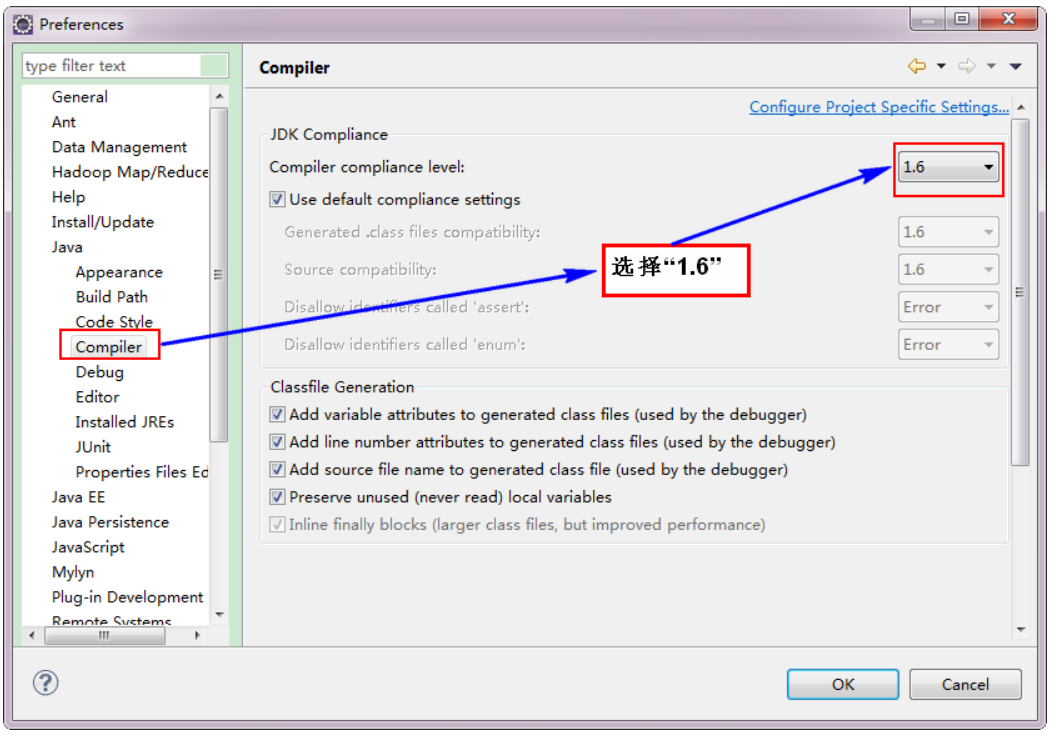
下面是没有添加之前的设置如下：



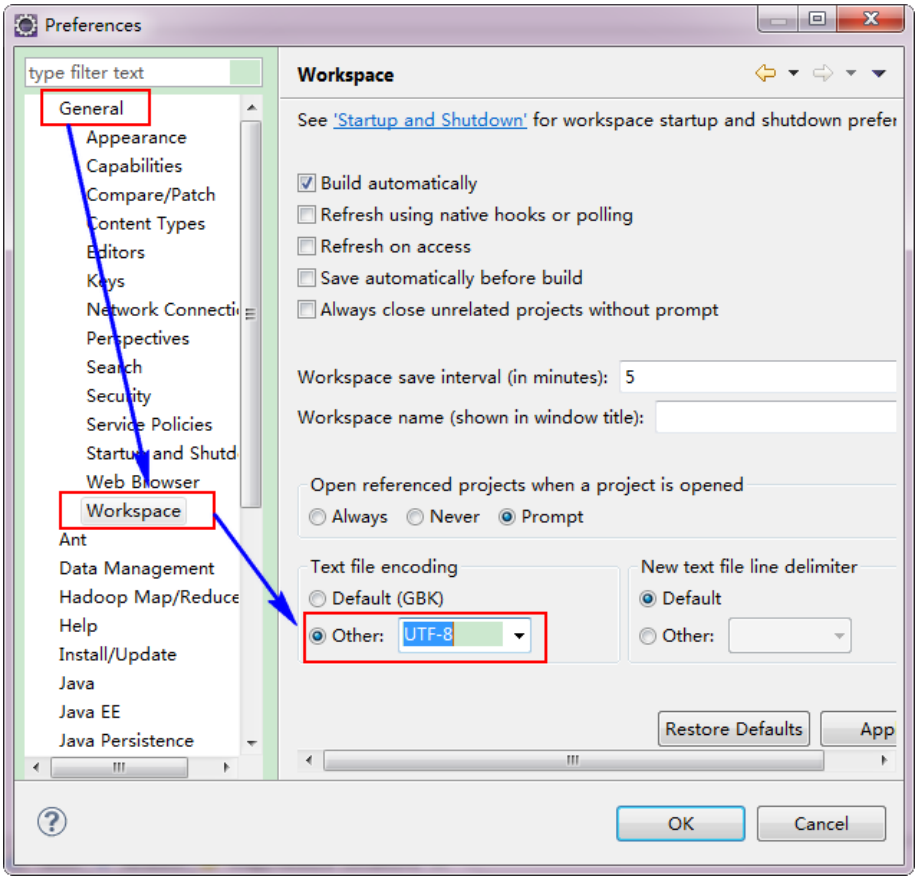
下面是添加完 JDK6.0 之后结果如下：



接着设置 Complier。

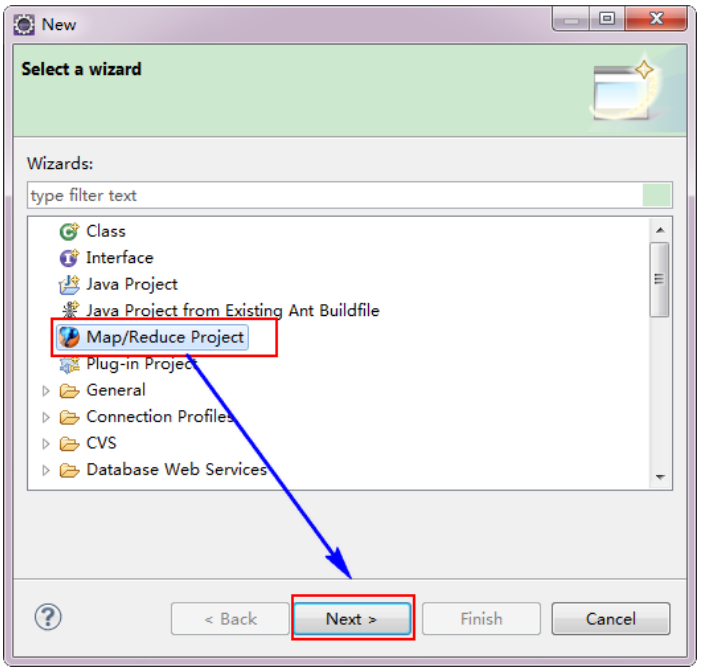


3.2 设置Eclipse的编码为UTF-8

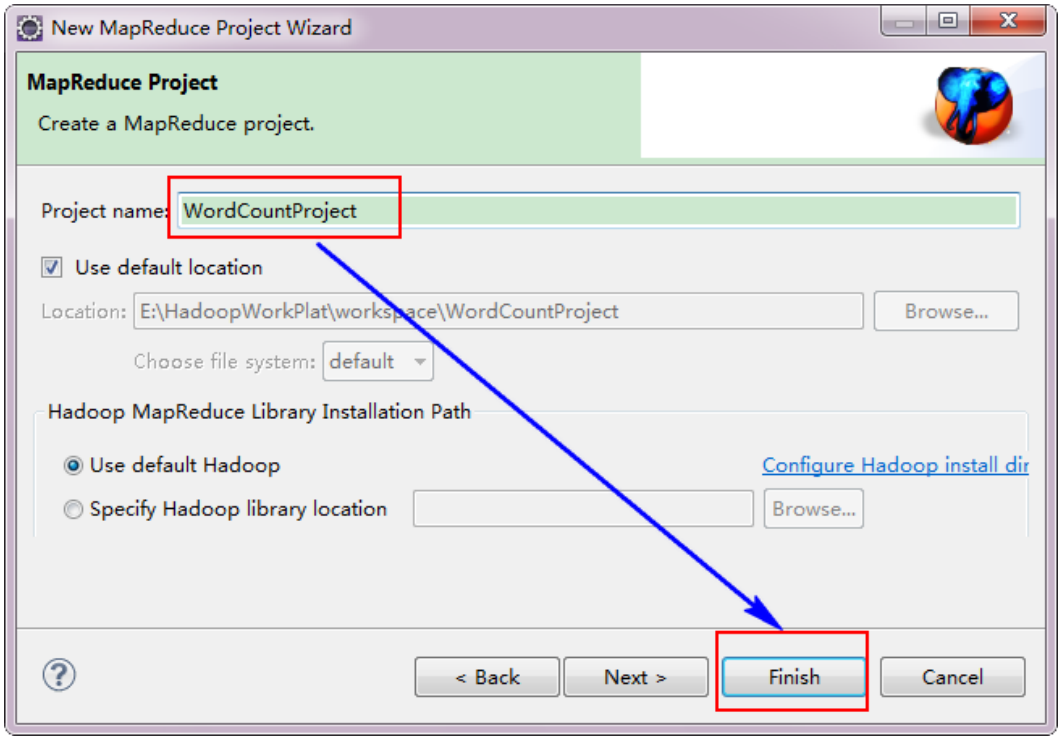


3.3 创建MapReduce项目

从“File”菜单，选择“Other”，找到“**Map/Reduce Project**”，然后选择它。

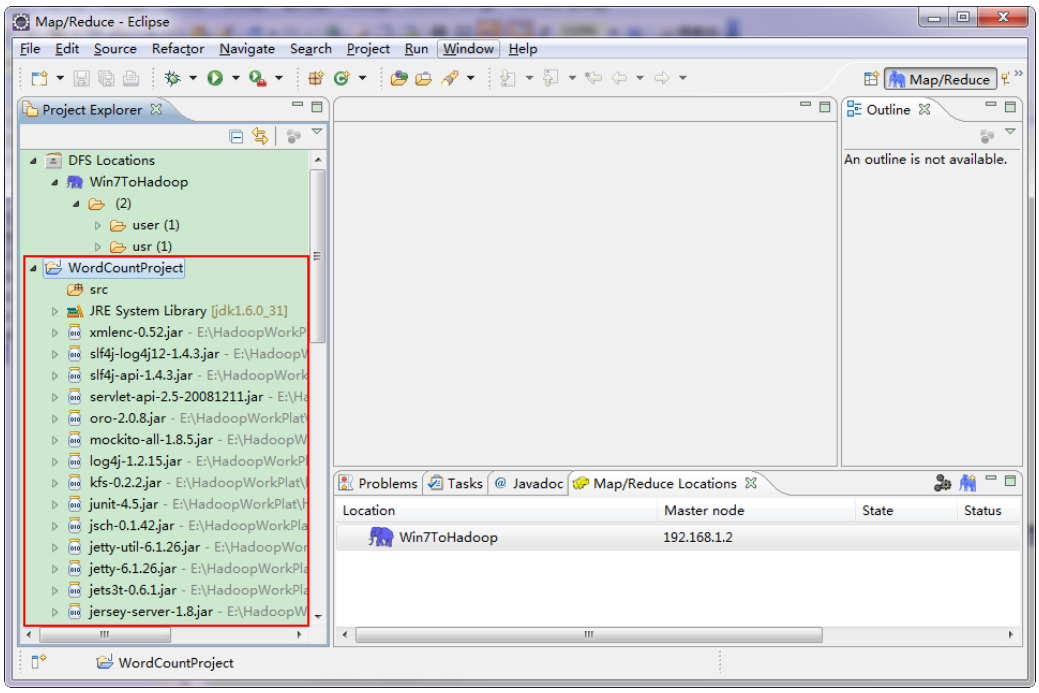


接着，填写 MapReduce 工程的名字为“WordCountProject”，点击“finish”完成。



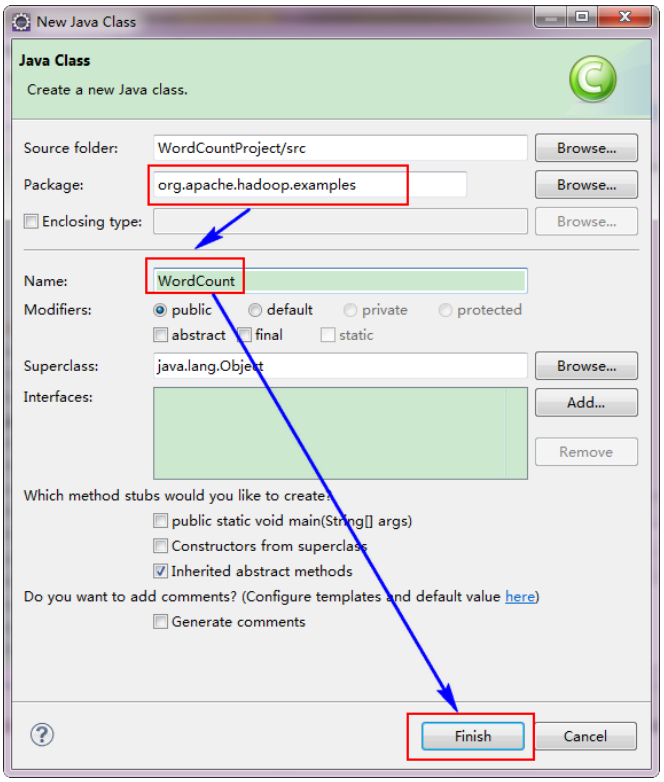
目前为止我们已经成功创建了 MapReduce 项目，我们发现在 Eclipse 软件的左侧多了我

们的刚才建立的项目。



3.4 创建WordCount类

选择“WordCountProject”工程，右击弹出菜单，然后选择“New”，接着选择“Class”，然后填写如下信息：



因为我们直接用 Hadoop1.0.0 自带的 WordCount 程序，所以报名需要和代码中的一致为“org.apache.hadoop.examples”，类名也必须一致为“WordCount”。这个代码放在如下的结构中。

```
hadoop-1.0.0
    |---src
        |---examples
            |---org
                |---apache
                    |---hadoop
                        |---examples
```

从上面目录中找见“**WordCount.java**”文件，用记事本打开，然后把代码复制到刚才建立的 java 文件中。当然源码有些变动，变动的红色已经标记出。

```
package org.apache.hadoop.examples;

import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class WordCount {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
```

```
            context.write(word, one);          }
        }
    }

    public static class IntSumReducer
            extends Reducer<Text,IntWritable,Text,IntWritable> {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values,
                           Context context
                           ) throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            result.set(sum);
            context.write(key, result);
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        conf.set("mapred.job.tracker", "192.168.1.2:9001");
        String[] ars=new String[]{"input","newout"};
        String[] otherArgs = new GenericOptionsParser(conf, ars).getRemainingArgs();
        if (otherArgs.length != 2) {
            System.err.println("Usage: wordcount <in> <out>");
            System.exit(2);
        }
        Job job = new Job(conf, "word count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
        FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

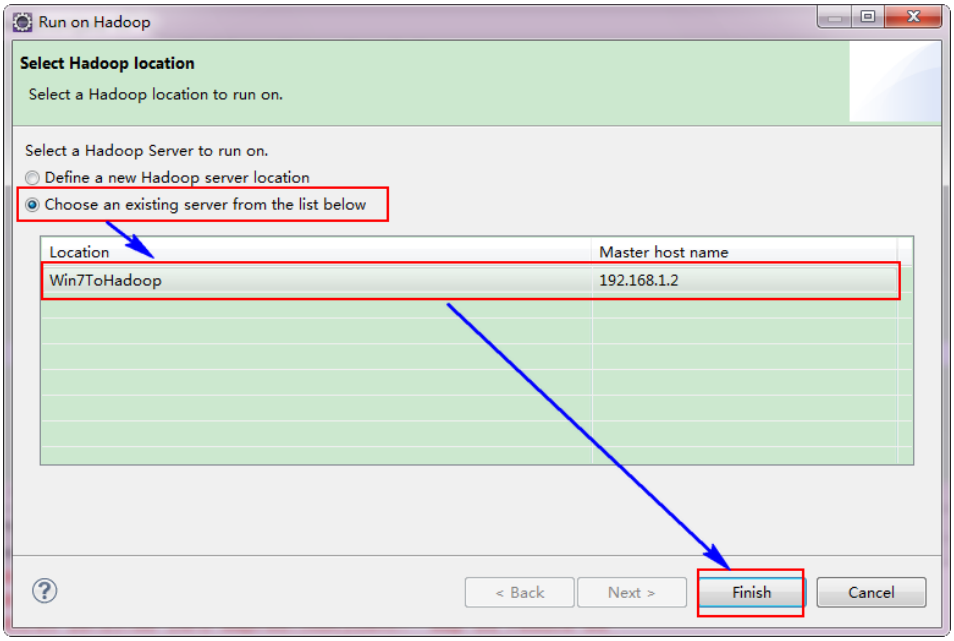
上面的红色就是与文件不一样的地方，其他的没有做任何改变。

备注：如果不加“**conf.set("mapred.job.tracker", "192.168.1.2:9001");**”，将提示你的权

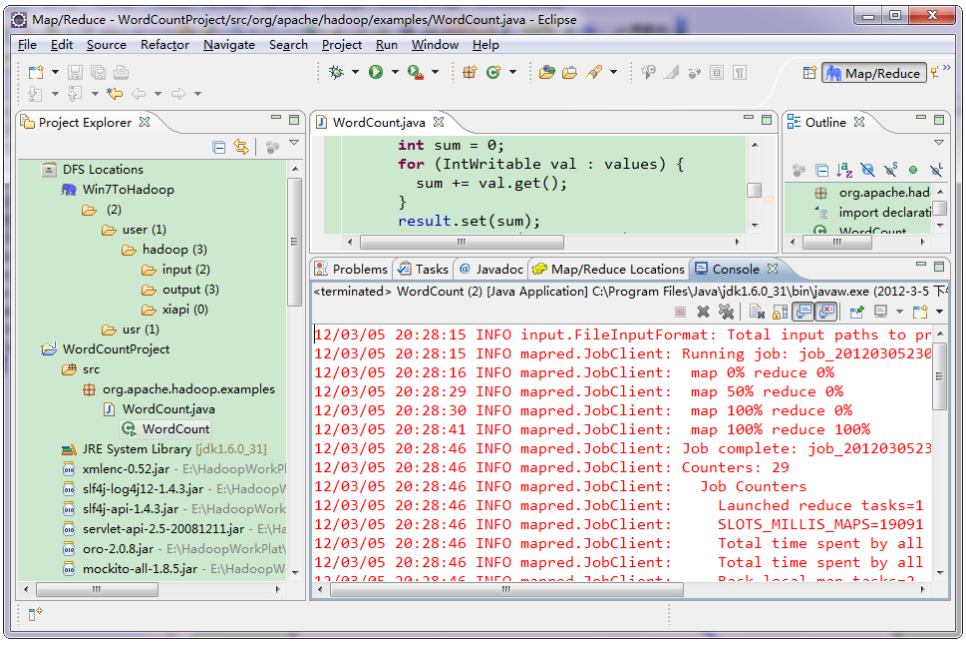
限不够，其实照成这样的原因是刚才设置的“Map/Reduce Location”其中的配置不是完全起作用，而是在本地的磁盘上建立了文件，并尝试运行，显然是不行的。我们要让 Eclipse 提交作业到 Hadoop 集群上，所以我们这里手动添加 Job 运行地址。详细参考“常见问题 FAQ_3”。

3.5 运行WordCount程序

选择“Wordcount.java”程序，右击一次按照“Run AS→Run on Hadoop”运行。然后会弹出如下图，按照下图进行操作。



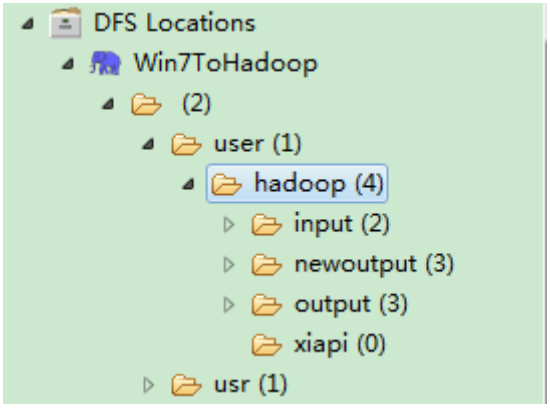
运行结果如下：



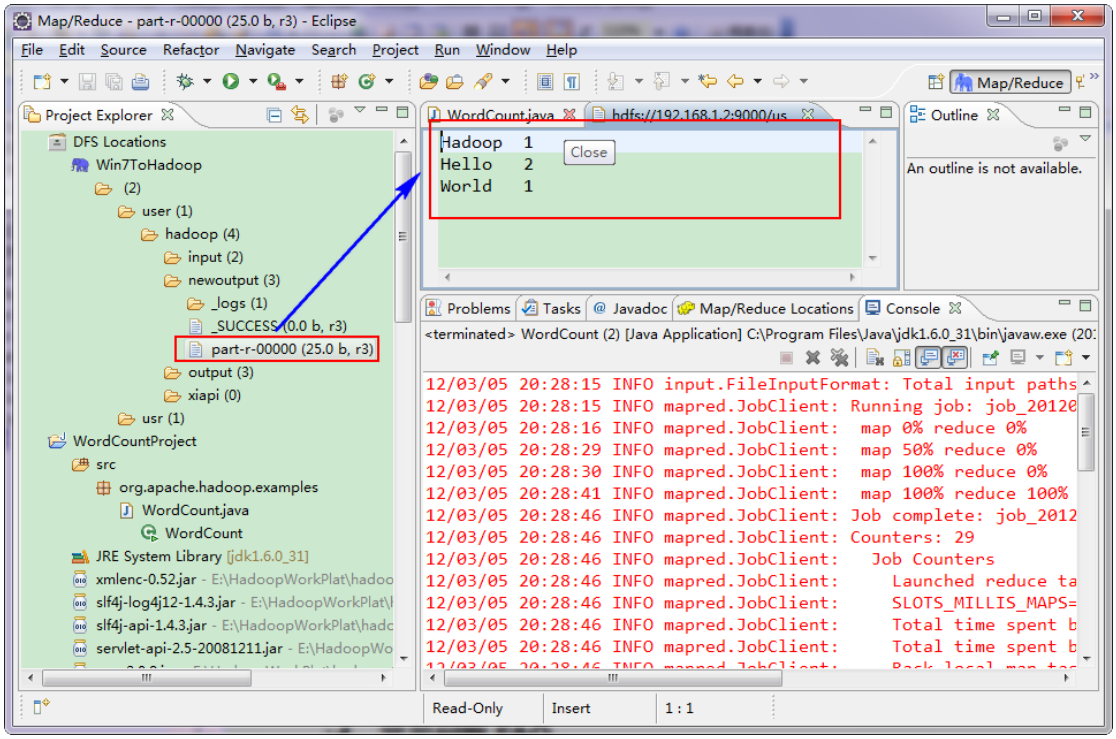
从上图中我们得知我们的程序已经运行成功了。

3.6 查看WordCount运行结果

查看 Eclipse 软件左侧，右击“DFS Locations→Win7ToHadoop→user→hadoop”，点击刷新按钮“Refresh”，我们刚才出现的文件夹“newoutput”会出现。记得“newoutput”文件夹是运行程序时自动创建的，如果已经存在相同的的文件夹，要么程序换个新的输出文件夹，要么删除 HDFS 上的那个重名文件夹，不然会出错。



打开“newoutput”文件夹，打开“**part-r-00000**”文件，可以看见执行后的结果。



到此为止，Eclipse 开发环境设置已经完毕，并且成功运行 Wordcount 程序，下一步我们真正开始 Hadoop 之旅。

4、常见问题FAQ

4.1 “error: failure to login”问题

下面以网上找的“hadoop-0.20.203.0”为例，我在使用“V1.0”时也出现这样的情况，原因就是那个“hadoop-eclipse-plugin-1.0.0_V1.0.jar”，是直接把源码编译而成，故而缺少相应的 Jar 包。具体情况如下

详细地址：http://blog.csdn.net/chengfei112233/article/details/7252404

在我实践尝试中，发现 hadoop-0.20.203.0 版本的该包如果直接复制到 eclipse 的插件目录中，在连接 DFS 时会出现错误，提示信息为： "error: failure to login"。

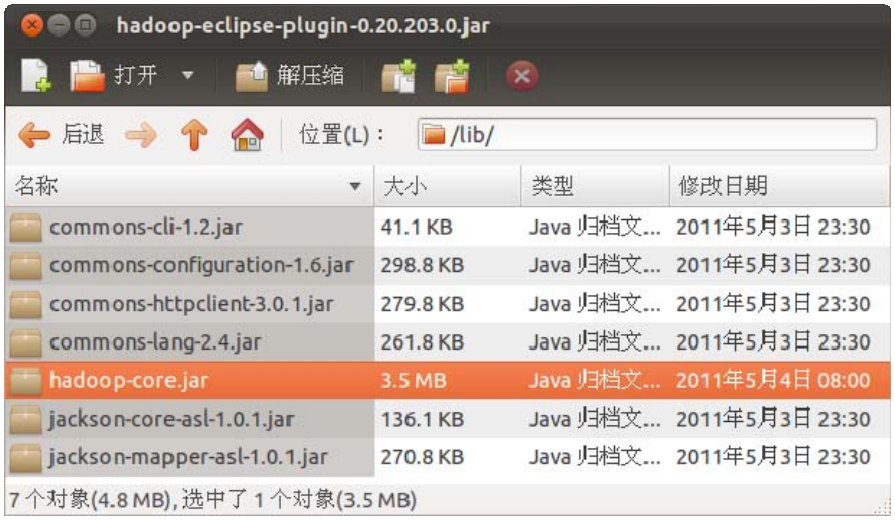
弹出的错误提示框内容为"**An internal error occurred during: "Connecting to DFS hadoop". org/apache/commons/configuration/Configuration**". 经过察看 Eclipse 的 log，发现是缺少 jar 包导致的。进一步查找资料后，发现直接复制 hadoop-eclipse-plugin-0.20.203.0.jar，该包中 lib 目录下**缺少**了 jar 包。

经过网上资料搜集，此处给出正确的安装方法：

首先要对 hadoop-eclipse-plugin-0.20.203.0.jar 进行修改。用归档管理器打开该包，发现只有 commons-cli-1.2.jar 和 hadoop-core.jar 两个包。将 hadoop/lib 目录下的：

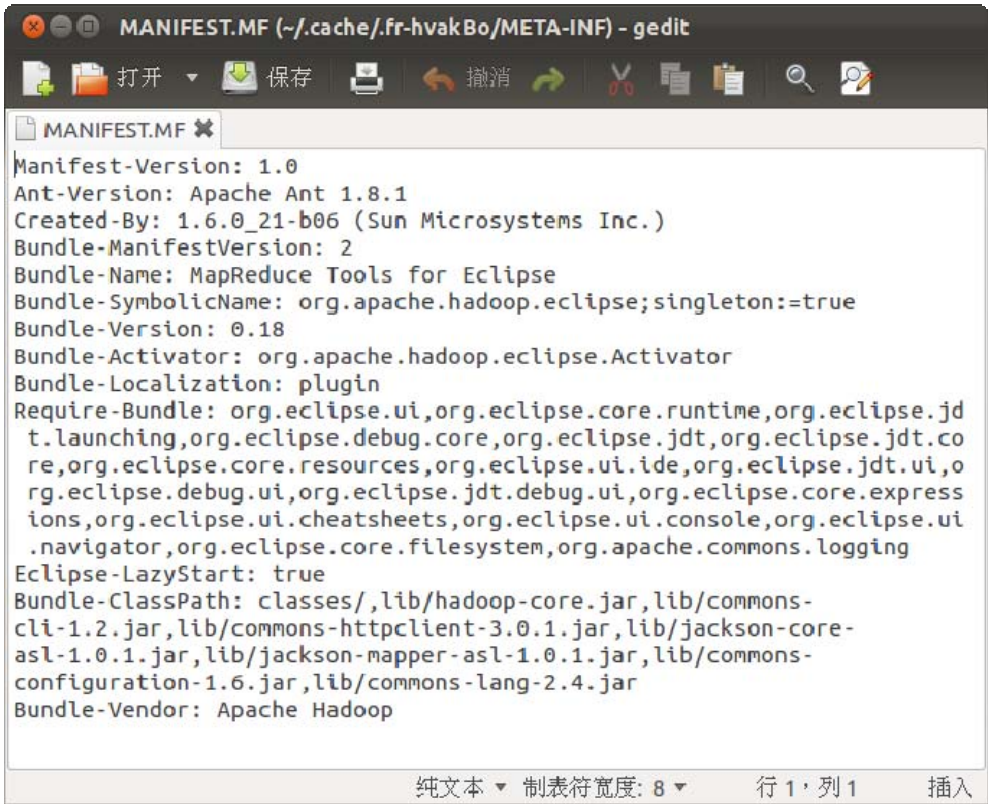
- commons-configuration-1.6.jar ,
- commons-httpclient-3.0.1.jar ,
- commons-lang-2.4.jar ,
- jackson-core-asl-1.0.1.jar
- jackson-mapper-asl-1.0.1.jar

一共 5 个包复制到 hadoop-eclipse-plugin-0.20.203.0.jar 的 lib 目录下，如下图：



然后，修改该包 META-INF 目录下的 **MANIFEST.MF**，将 classpath 修改为一下内容：

Bundle-ClassPath:classes/,lib/hadoop-core.jar,lib/commons-cli-1.2.jar,lib/commons-httpclient-3.0.1.jar,lib/jackson-core-asl-1.0.1.jar,lib/jackson-mapper-asl-1.0.1.jar,lib/commons-configuration-1.6.jar,lib/commons-lang-2.4.jar



这样就完成了对 hadoop-eclipse-plugin-0.20.203.0.jar 的修改。

最后，将 hadoop-eclipse-plugin-0.20.203.0.jar 复制到 Eclipse 的 plugins 目录下。

备注：上面的操作对“hadoop-1.0.0”一样适用。

4.2 “Permission denied”问题

网上试了很多，有提到“hadoop fs -chmod 777 /user/hadoop ”，有提到“dfs.permissions 的配置项，将 value 值改为 false”，有提到“hadoop.job.ugi”，但是通通没有效果。

参考文献：

地址 1：http://www.cnblogs.com/acmy/archive/2011/10/28/2227901.html

地址 2：http://sunjun041640.blog.163.com/blog/static/256268322010617518252 92/

错误类型：org.apache.hadoop.security.AccessControlException: org.apache.hadoop.security.AccessControlException: Permission denied: user=**********, access=WRITE, inode="hadoop":hadoop:supergroup:rwxr-xr-x

解决方案：

我的解决方案直接把系统管理员的名字改成你的 Hadoop 集群运行 hadoop 的那个用户。

4.3 “Failed to set permissions of path”问题

参考文献：https://issues.apache.org/jira/browse/HADOOP-8089
错误信息如下：

ERROR security.UserGroupInformation: PriviledgedActionException as:
hadoop cause:java.io.IOException Failed to set permissions of path:
\usr\hadoop\tmp\mapred\staging\hadoop753422487\.staging to 0700
Exception in thread "main" java.io.IOException: Failed to set permissions of path:
\usr\hadoop\tmp \mapred\staging\hadoop753422487\.staging to 0700

解决方法：

```
Configuration conf = new Configuration();
conf.set("mapred.job.tracker", "[server]:9001");
```

“**[server]:9001**”中的“**[server]**”为 Hadoop 集群 Master 的 IP 地址。

4.4 “hadoop mapred执行目录文件权”限问题

参考文献：http://blog.csdn.net/azhao_dn/article/details/6921398
错误信息如下：
job Submission failed with exception 'java.io.IOException(The ownership/permissions on the staging directory /tmp/hadoop-hadoop-user1/mapred/staging/hadoop-user1/.staging is not as expected. It is owned by hadoop-user1 and permissions are rwxrwxrwx. The directory must be owned by the submitter hadoop-user1 or by hadoop-user1 and permissions must be rwx------)
修改权限：

```
[hadoop@Master ~]$ hadoop fs -chmod -R 700  /usr/hadoop/tmp
```

这样就能解决问题。

编者简介

基本信息

姓　　名：　解耀伟　　性　　别：　男

笔　　名：　虾皮　　民　　族：　汉

学　　历：　研究生　　专　　业：　计算机应用技术

电子信箱：　xieyaowei1986@163.com

学　　校：　河北工业大学（211 工程）

求职意向

希望在 IT 行业从事软件研发等工作。

编程语言

Java、C#、C、ExtJS、Flex、汇编、PHP、VB，熟练程度由左到右逐级减弱。

个人经历

大学期间

1）担任职务：学生会生活部部长、生活委员、团支书

2）获得奖项：二等奖学金（2 次）、三好学生（1 次）

研究生期间

1）担任职务：班长

2）获得奖项：优秀班干部（1 次）

工作经历

实验室项目：国家 863 计划项目 1 项；国家技术基础专项 2 项；河北省技术专项 1 项。

研究生课题：基于 Hadoop 分布式搜索引擎研究

个人评价

性格开朗，善于与人沟通，上进心强，品德优秀，吃苦耐劳，喜欢团队合作，能积极服从上级的安排。

寄　言

相信您的信任与我的能力将为我们带来共同的成功。

参考文献

感谢以下文章的编写作者，没有你们的铺路，我或许会走得很艰难，参考不分先后，贡献同等珍贵。

【1】Hadoop 实战——陆嘉恒——机械工业出版社

【2】实战 Hadoop——刘鹏——电子工业出版社

【3】基于 Eclipse 的 Hadoop 应用开发环境配置

地址：http://www.cnblogs.com/flyoung2008/archive/2011/12/09/2281400.html

【4】eclipse hadoop 开发环境配置

地址：http://blog.csdn.net/cybercode/article/details/7084603

【5】Eclipse 运行 hadoop（解决错误）

地址：http://sunjun041640.blog.163.com/blog/static/25626832201061751825292/

【6】Permission denied

地址：http://www.cnblogs.com/acmy/archive/2011/10/28/2227901.html

【7】hadoop-eclipse 开发环境搭建及 error: failure to login 错误

地址：http://blog.csdn.net/chengfei112233/article/details/7252404

【8】hadoop mapred(hive)执行目录 文件权限问题

地址：http://blog.csdn.net/azhao_dn/article/details/6921398

【9】cannot submit job from Eclipse plugin running on Windows

地址：https://issues.apache.org/jira/browse/HADOOP-8089

【10】Hadoop 初学者可能会遇到的问题

地址：http://bbs.hadoopor.com/thread-3967-1-1.html