**

**HUNAN UNIVERSITY**

《数据仓库技术与应用》

实验报告四

|  |  |
| --- | --- |
| **报告名称：** | MapReduce实验 |
| **学生姓名：** | 杨超然 |
| **学生学号：** | 202106060220 |
| **专业班级：** | 电商2102班 |
| **学 院：** | 工商管理学院 |
| **指导老师：** | 李信儒 |
| **日 期：** | 2023.3.13 |

目录

[一、实验过程 3](#_Toc908536128)

[（一）在Ubuntu中安装Eclipse 3](#_Toc1157973126)

[（二） 安装Hadoop-Eclipse-Plugin 4](#_Toc1557519568)

[（三） 配置Hadoop-Eclipse-Plugin 6](#_Toc1553206093)

[（四） 在Eclipse中操作HDFS文件 8](#_Toc2071075766)

[（五） 在Eclipse中创建MapReduce项目 9](#_Toc7964939)

[（六） 在Eclipse中运行MapReduce 12](#_Toc722743659)

[二、易错点总结 14](#_Toc985169381)

[（一）因路径不正确而无法安装Hadoop-Eclipse-Plugins 14](#_Toc642868097)

[（二） 因权限问题而无法安装Hadoop-Eclipse-Plugins 15](#_Toc693878222)

[（三） 版本不兼容而无法安装 16](#_Toc1175073944)

[（四） 路径不一致而无法复制配置文件 16](#_Toc1208158996)

[（五） 未配置相关文件而报错 17](#_Toc1070363387)

[三、学习小记 17](#_Toc126934390)

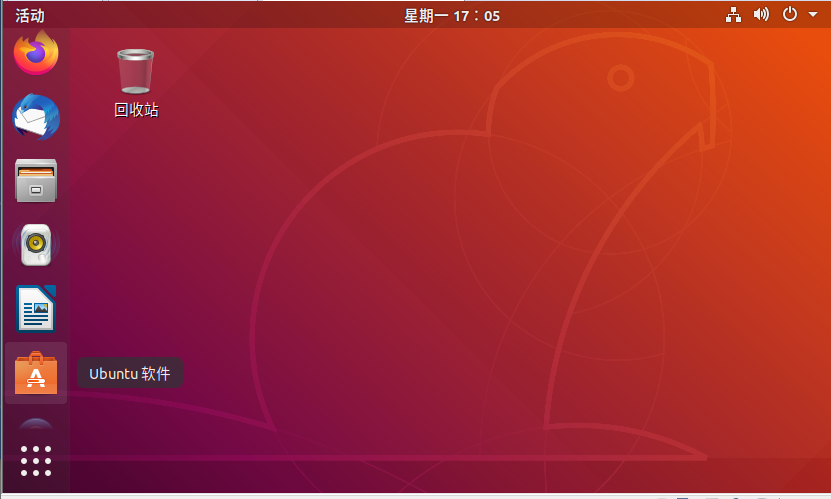
[（1）MapReduce相较于传统分布式并行编程框架的优势 17](#_Toc935031259)

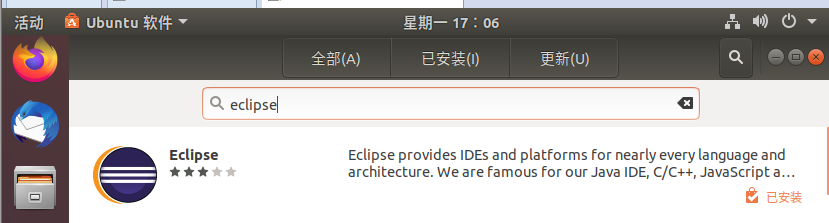
[（2）MapReduce执行全过程 18](#_Toc1932524914)

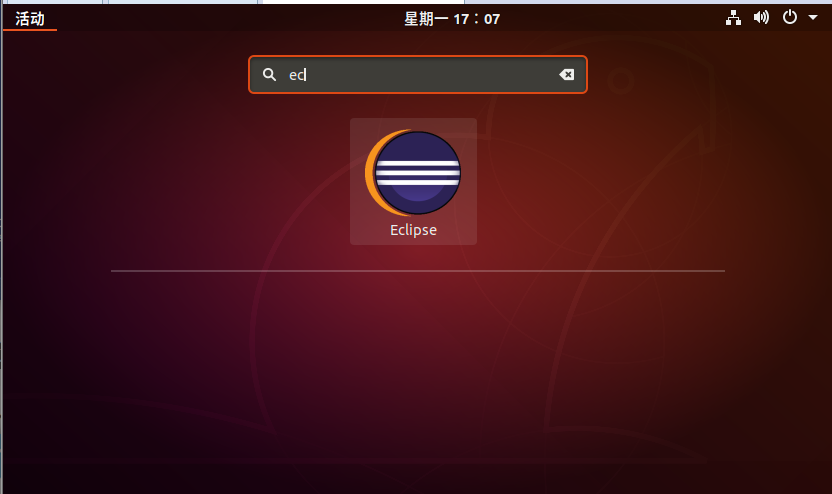
[（3）MapReduce的具体应用 18](#_Toc1403552370)

# 一、实验过程

## （一）在Ubuntu中安装Eclipse

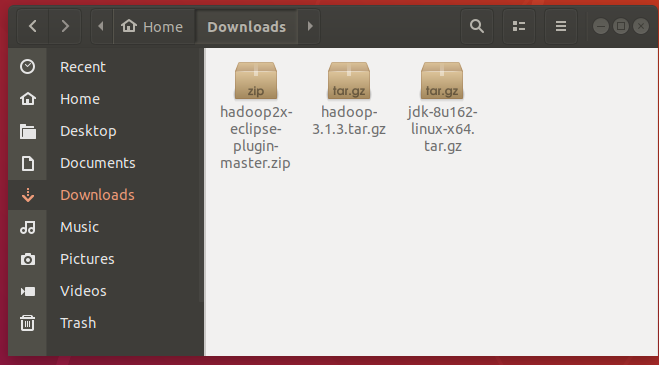




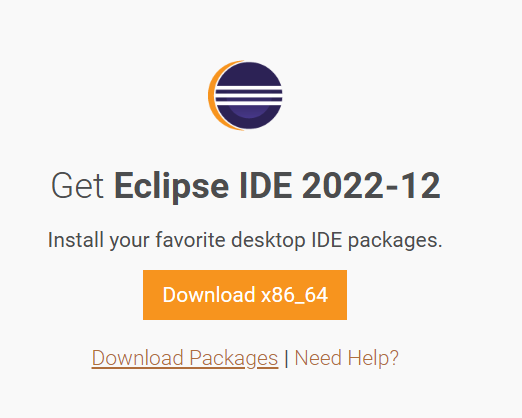


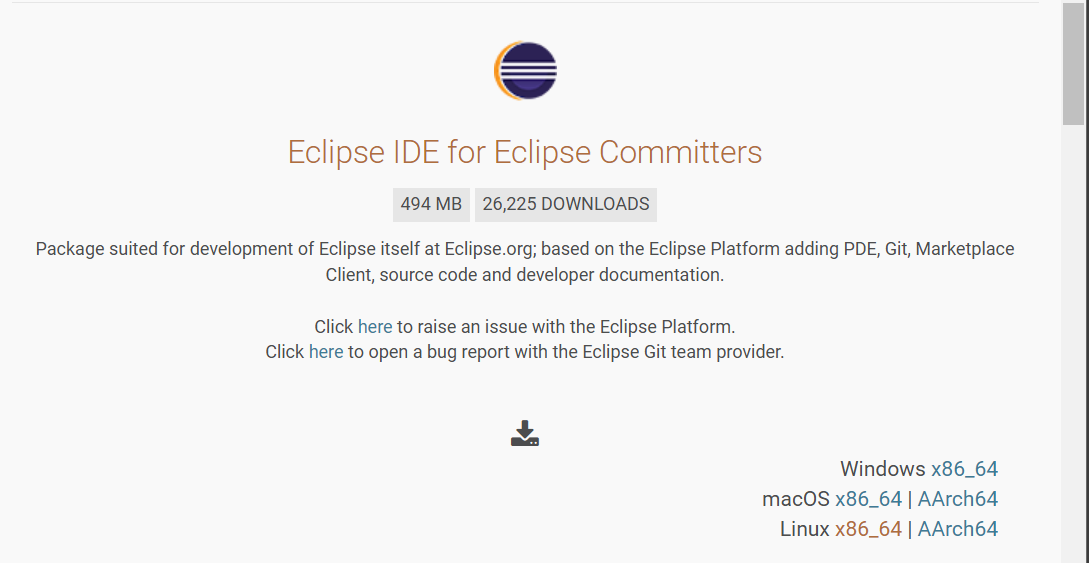
## 安装Hadoop-Eclipse-Plugin

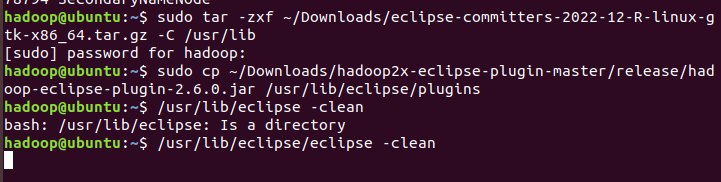
从Github上下载hadoop2x-eclipse-plugin

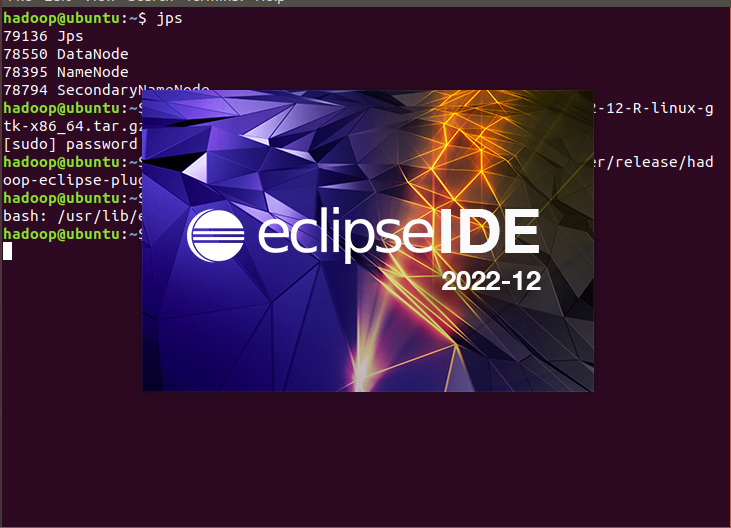


出现默认路径不匹配问题，可以在eclipse官网上下载对应的安装包，通过终端指令安装在指定路径下，就可以按照指南步骤正确配置Hadoop-Eclipse-Plugin了



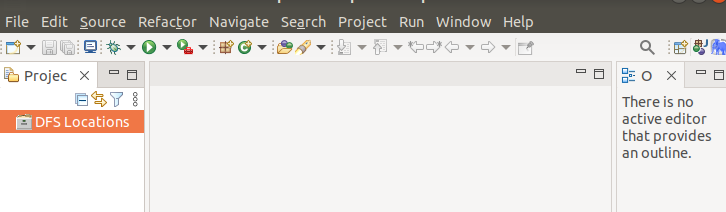


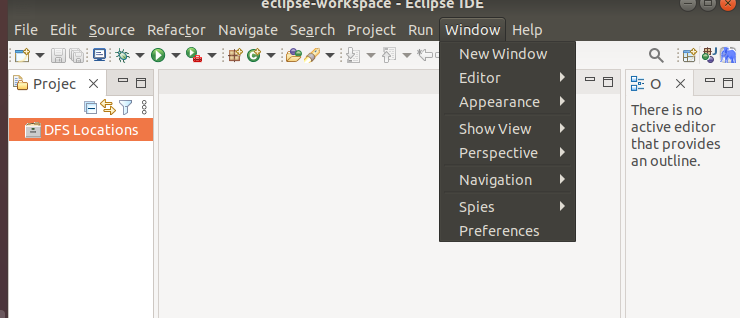


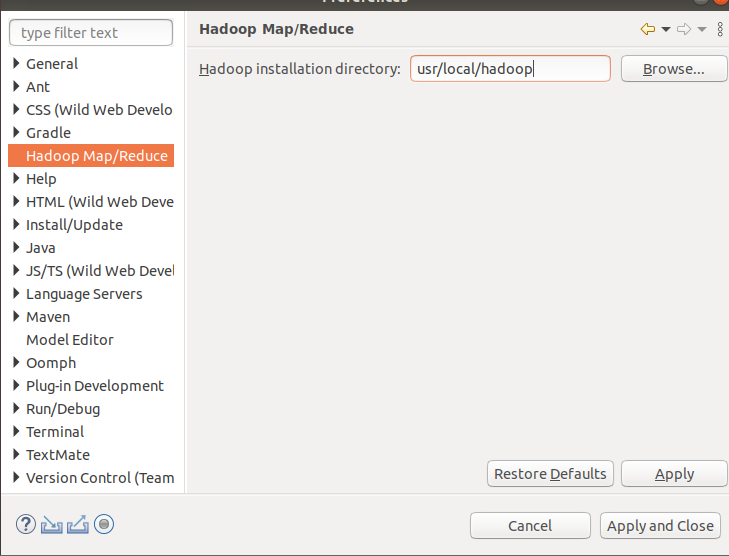


## 配置Hadoop-Eclipse-Plugin

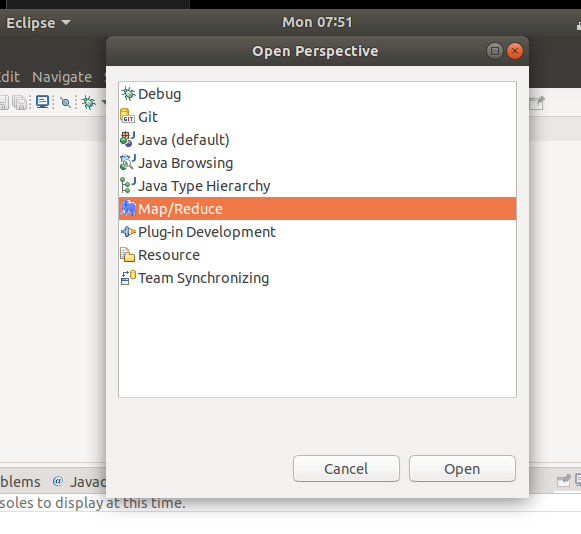
1.选择Window下的Preference



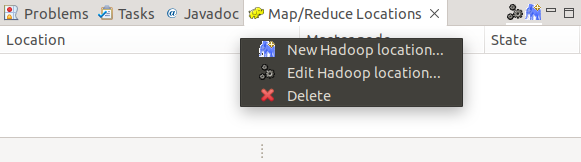


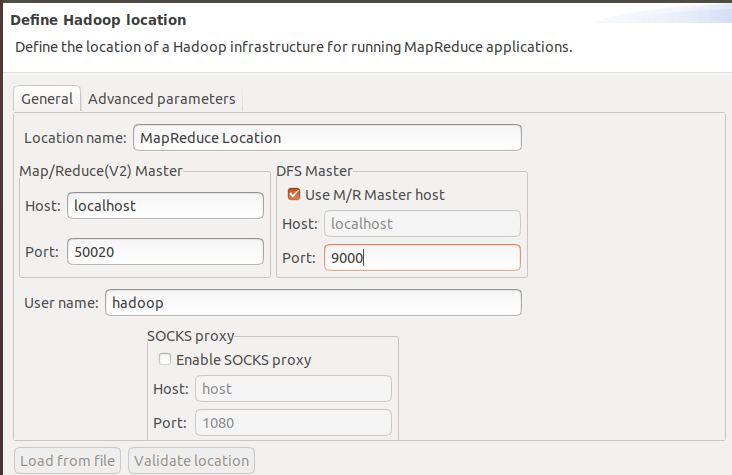


2.切换Map Reduce开发视图

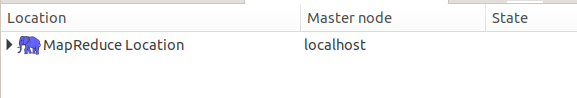


3.建立与Hadoop集群的连接



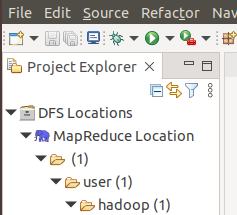


创建成功的MapReduce Location如下

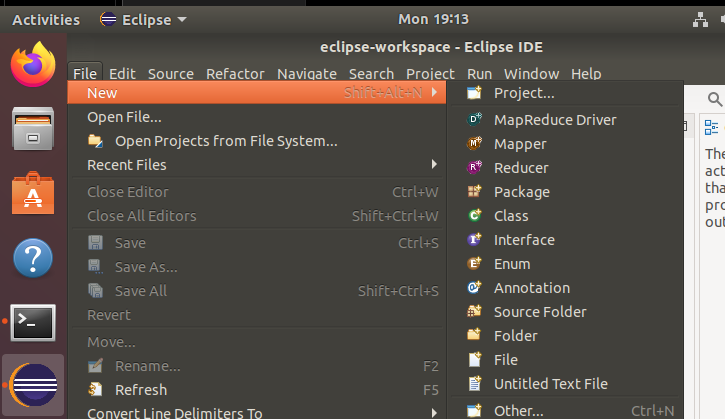


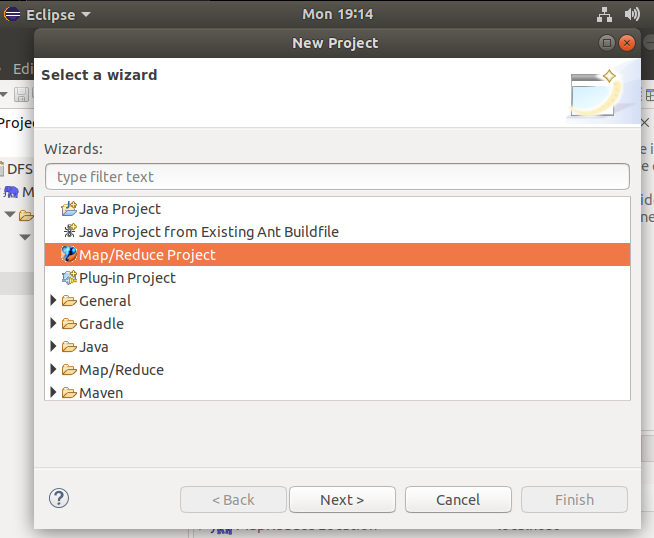
## 在Eclipse中操作HDFS文件

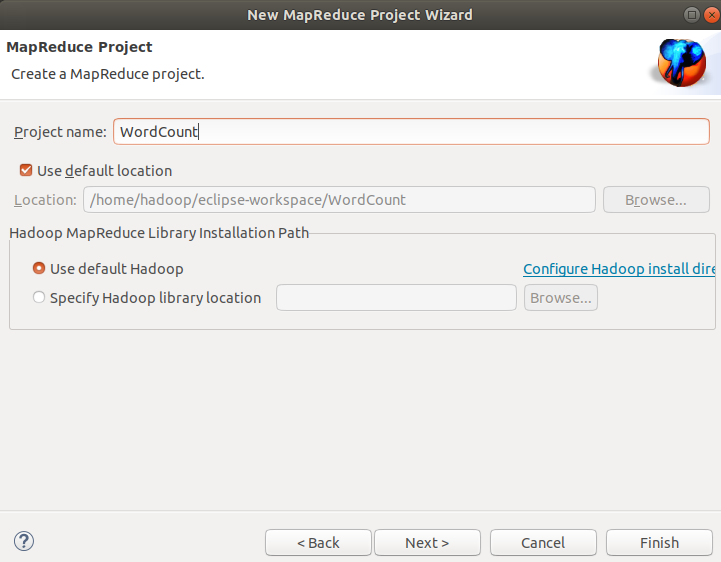
配置好后，点击左侧Project Explorer中的MapReduce Location就能就能查看HDFS中的文件列表，双击可以查看内容，右键可以上传、下载、删除HDFS文件。

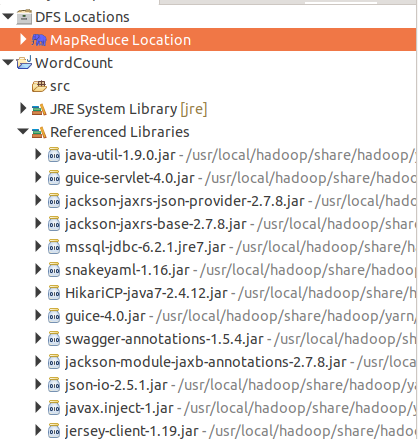


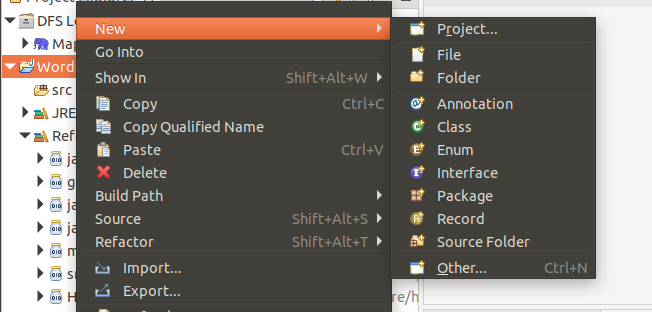
## 在Eclipse中创建MapReduce项目

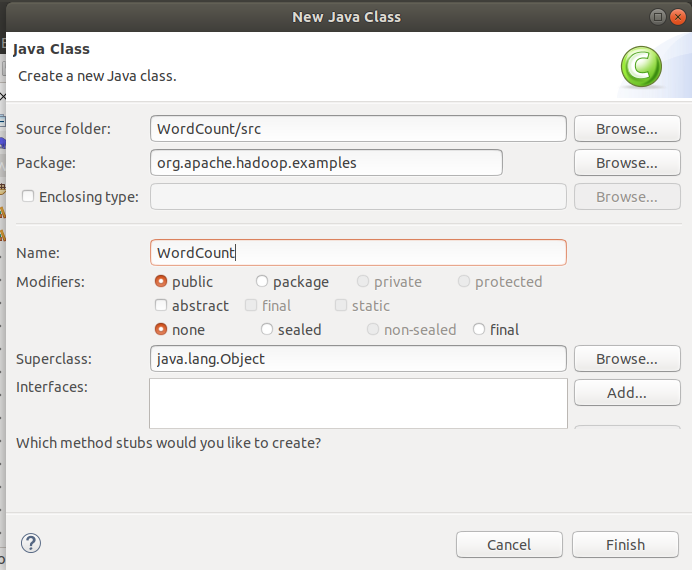


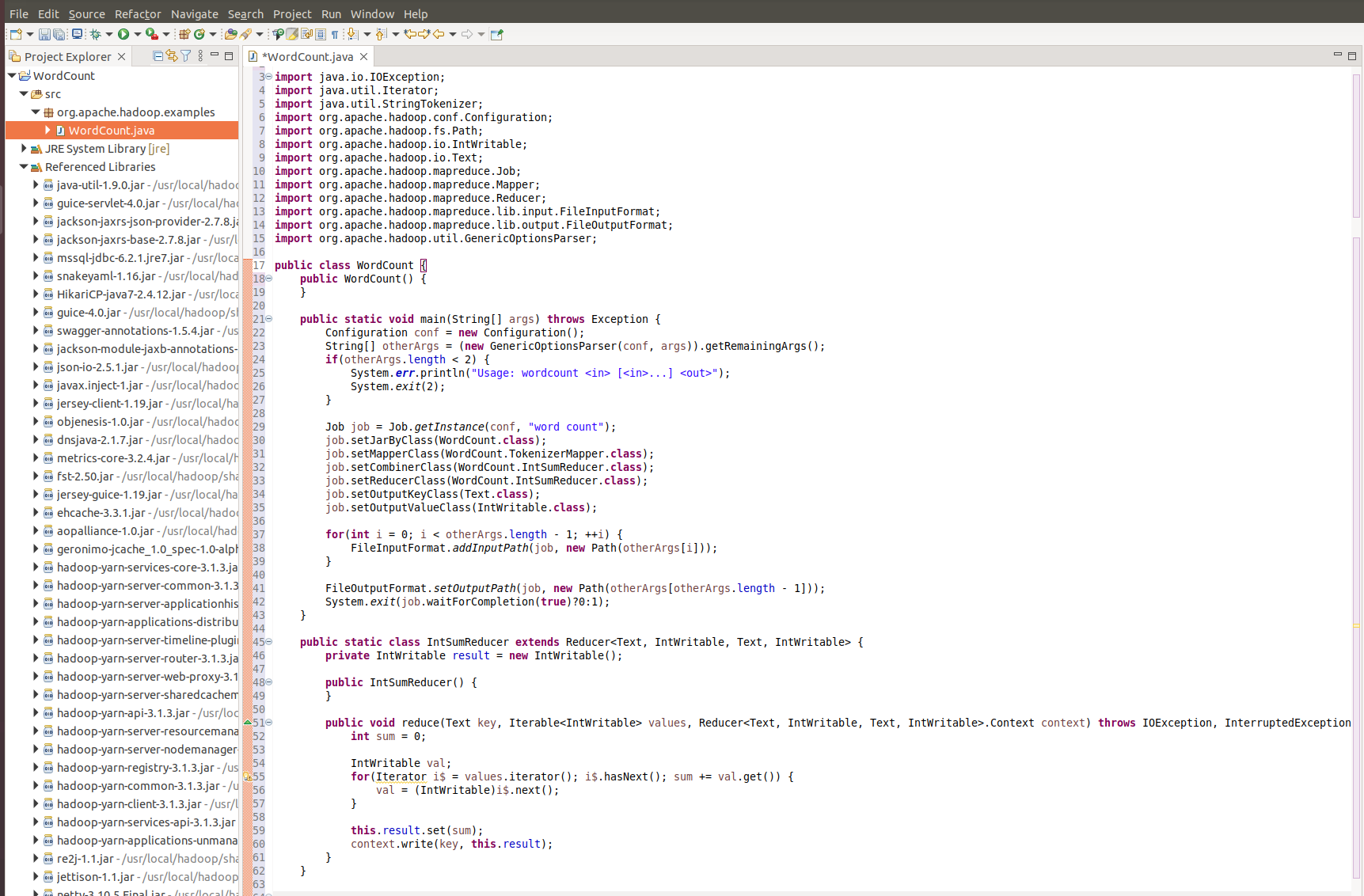






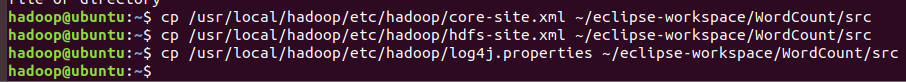


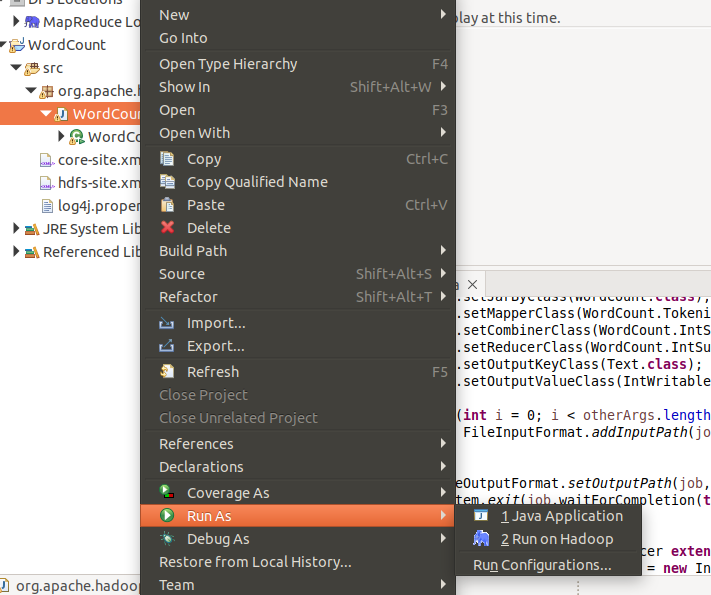


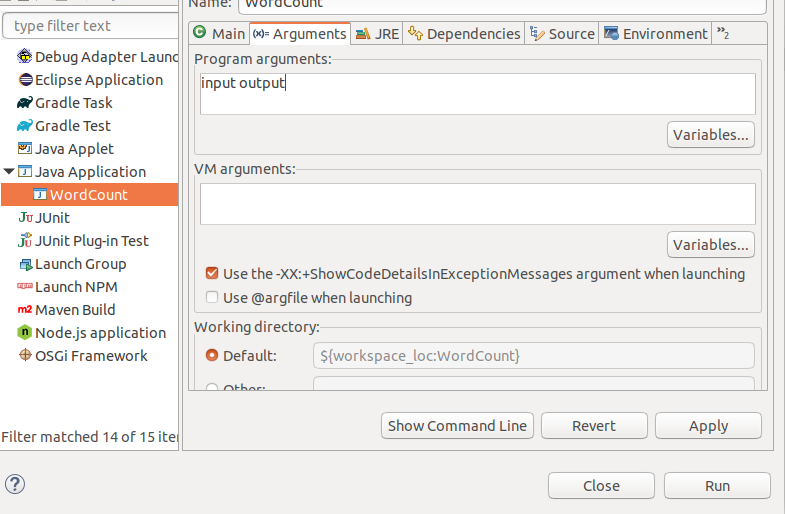


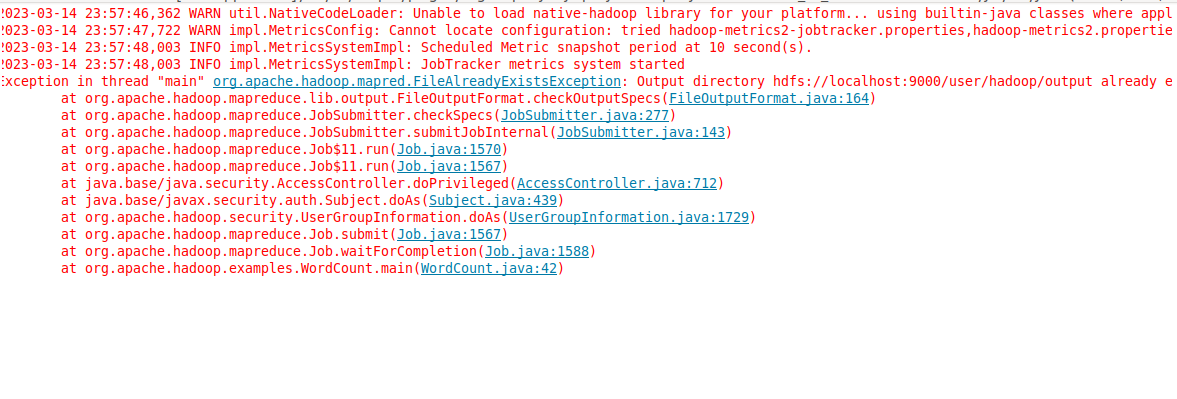
## 在Eclipse中运行MapReduce

在运行MapReduce程序前，还需要执行一项重要操作，即通过复制配置文件解决参数问题。



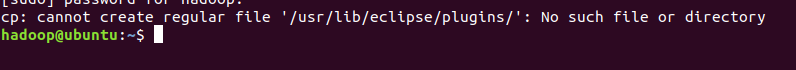






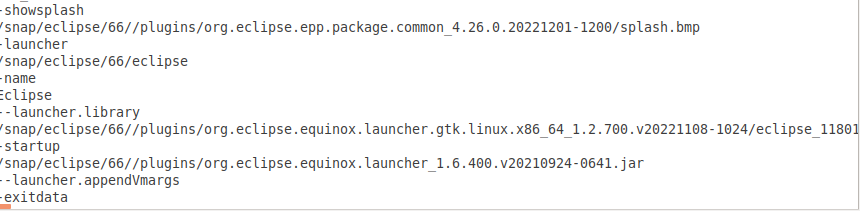
# 二、易错点总结

## （一）因路径不正确而无法安装Hadoop-Eclipse-Plugins

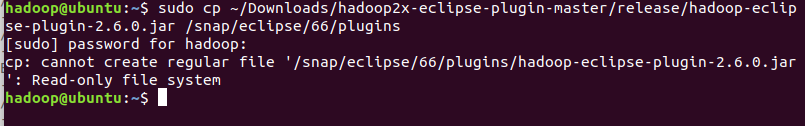


指南中曾提到过Eclipse的默认安装目录为：usr/lib/eclipse，但在实际实验过程中发现通过Ubuntu软件商城中eclipse的安装目录并非上述目录，因此会报错

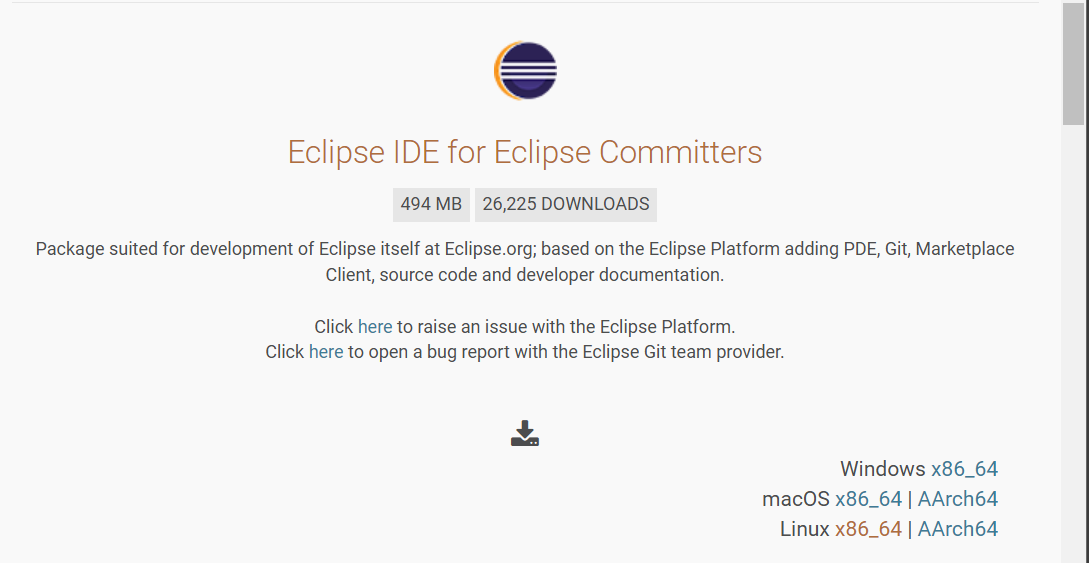
**解决方案：**进入Eclipse-Help中查看eclipse的安装路径

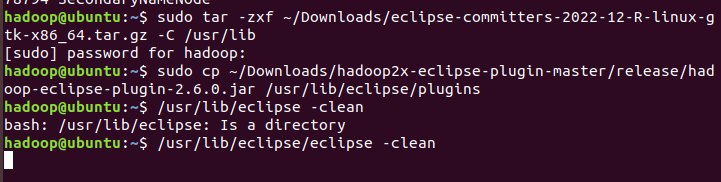


## 因权限问题而无法安装Hadoop-Eclipse-Plugins

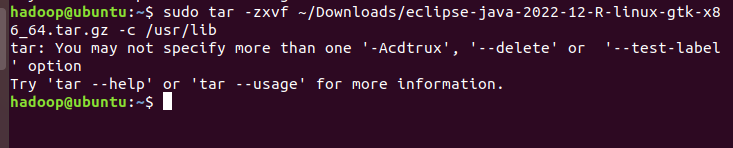


**解决方案：**不通过Ubuntu软件商城安装Eclipse，而是进入Eclipse官网下载对应的安装包（我下载的是Eclipse IDE for Eclipse Committers），再进入终端将安装包解压并安装到指定的目录中。



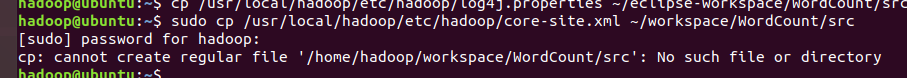


## 版本不兼容而无法安装



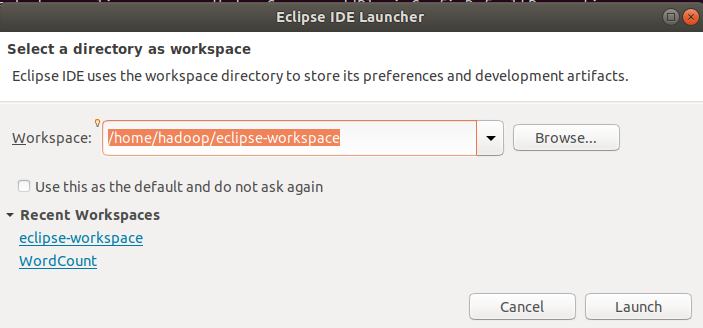
**解决方案：**可能是安装包不兼容，我换了个安装包就可以正常解压了

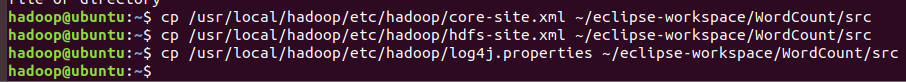
## 路径不一致而无法复制配置文件



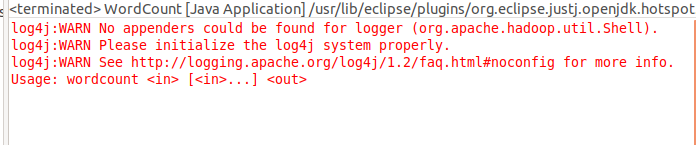
该错误产生的原因是workspace的路径与指南中的路径不一致（连名字都不一致，实际我的工作台叫eclipse-workspace）

**解决方案：**又一次告诉了我们不要照抄代码，workspace的路径可以在打开eclipse之前看到路径，然后根据路径执行cp命令



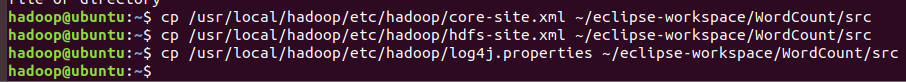


## 未配置相关文件而报错



出现该警告的原因是没有将 log4j.properties 这个配置文件复制到项目中，虽然不影响程序的正确运行的，但程序运行时无法看到任何提示消息（只能看到出错信息）。

**解决方案：**在终端中复制配置文件以解决参数问题



# 三、学习小记

## （1）MapReduce相较于传统分布式并行编程框架的优势

自2005年之后，随着需要处理的数据量呈指数式爆炸上升，预测CPU性能大概每隔18个月翻一番的“摩尔定律”已经不再适用，因此人们开始使用分布式并行编程，在大规模计算机集群上运行分布式程序来并行执行大规模数据处理任务，获得海量计算能力。

MapReduce作为谷歌公司提出的分布式并行模型，在其之前已经有了比较成熟的并行计算框架（如MPI），那么MapReduce出现的必要性是什么，相较于传统的并行计算框架有什么优势呢？

·集群架构与容错性方面：传统并行计算框架使用共享式集群架构（共享内存/共享存储），而MapReduce采用非共享式，具有较高的容错性。

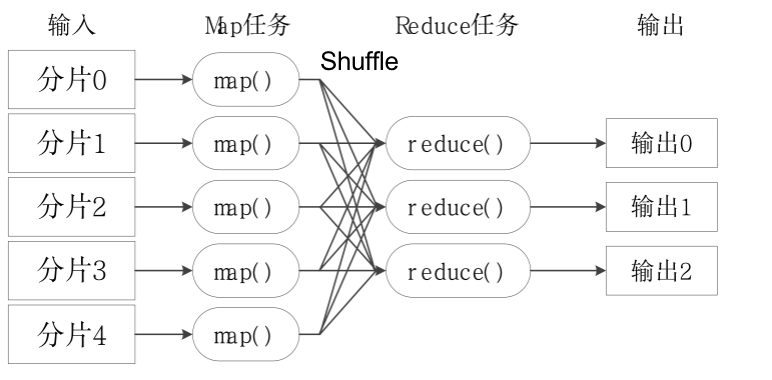
·硬件、价格与扩展性方面：传统并行计算框架需要刀片服务器、高速网与SAN。价格贵，扩展性差。而MapReduce只需要普通PC机，便宜且扩展性好。

·学习难度方面：MapReduce的编程与学习难度较为简单。

综合上述方面，可见传统并行计算框架适用于实时、细粒度计算，属于计算密集型，而MapReduce适用于非实时的批处理、属于数据密集型。

## （2）MapReduce执行全过程

MapReduce的执行全过程包括以下几个主要阶段：1.从分布式文件系统读入数据；2.执行Map任务输出中间结果；3.通过Shuffle阶段把中间结果分区并写入分布式文件系统



注：1.不同的Map任务之间不会进行通信

1. 不同的Reduce任务之间也不会发生任何信息交换
2. 用户不能显式地从一台机器向另一台机器发送消息
3. 所有的数据交换都是通过MapReduce框架自身去实现的

## （3）MapReduce的具体应用

MapReduce可以很好的应用于各种计算问题：

1. 关系代数运算（选择、投影、并、交、差、连接等）
2. 分组与聚合运算
3. 矩阵-向量乘法
4. 矩阵乘法