《大数据基础编程、实验和案例教程（第2版）》

教材官网：

http://dblab.xmu.edu.cn/post/bigdatapractice2/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第4章 HDFS操作方法和基础编程

（PPT版本号：2020年12月版本）

扫一扫访问教材官网

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

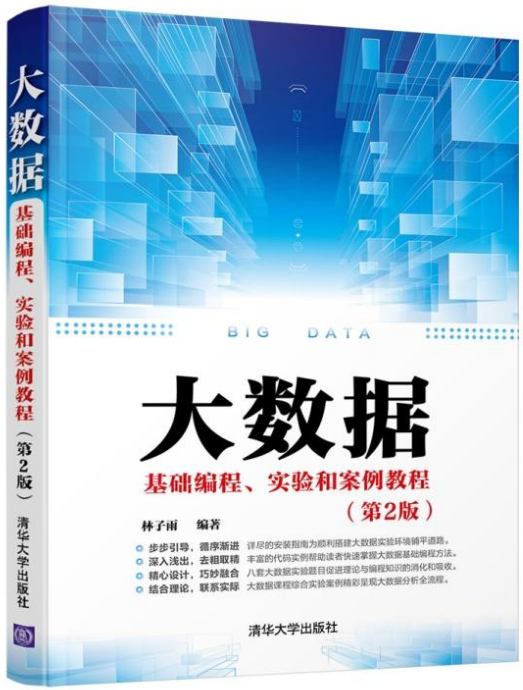
主页：http://dblab.xmu.edu.cn/linziyu

教材简介

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元，2020年10月第2版
教材官网： http://dblab.xmu.edu.cn/post/bigdatapractice2/

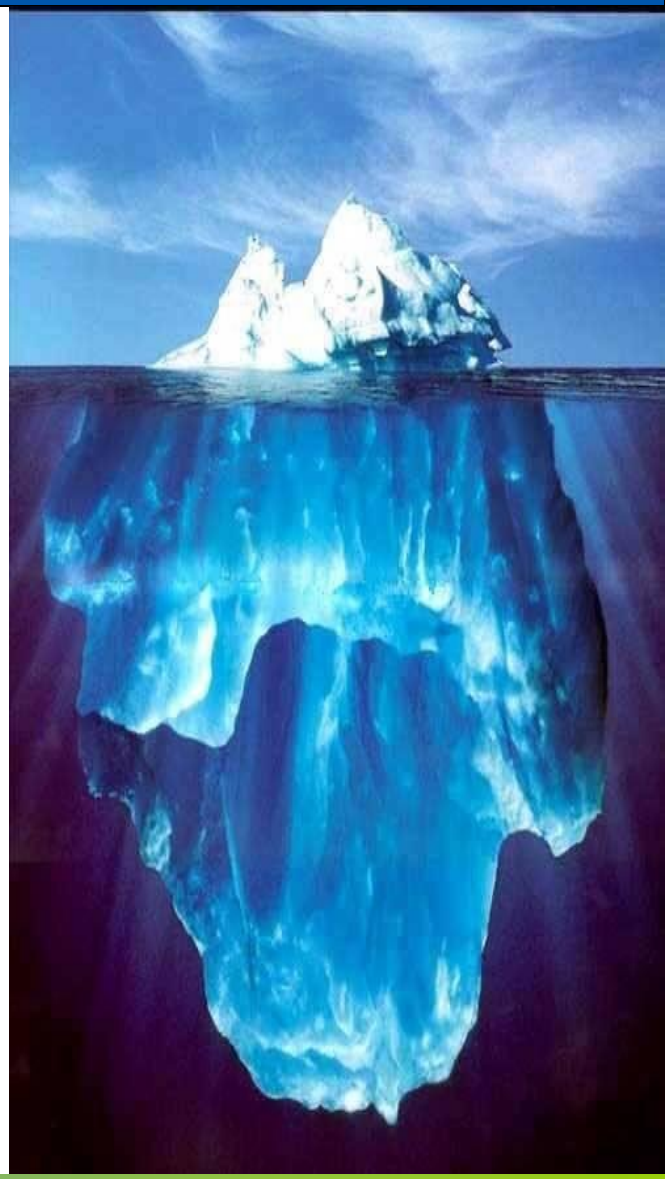
扫一扫访问
教材官网



- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

提纲

4.1 HDFS操作常用Shell命令
4.2 利用HDFS的Web管理界面
4.3 HDFS编程实践



高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

4.1 HDFS操作常用Shell命令

4.1.1 查看命令使用方法
4.1.2 HDFS目录操作

4.1.1 查看命令使用方法

请登录Linux系统，打开一个终端，首先启动Hadoop，命令如下：

```
$ cd /usr/local/hadoop
$ ./sbin/start-dfs.sh
```

可以在终端输入如下命令，查看hdfs dfs总共支持哪些操作：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs
```

可以查看某个命令的作用，比如，当需要查询put命令的具体用法时，可以采用如下命令：

```
$ ./bin/hdfs dfs –help put
```

4.1.2 HDFS目录操作

1. 目录操作

需要注意的是，Hadoop系统安装好以后，第一次使用HDFS时，需要首先在HDFS中创建用户目录。本教程全部采用hadoop用户登录Linux系统，因此，需要在HDFS中为hadoop用户创建一个用户目录，命令如下：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs –mkdir –p /user/hadoop
```

“/user/hadoop”目录就成为hadoop用户对应的用户目录，可以使用如下命令显示HDFS中与当前用户hadoop对应的用户目录下的内容：

```
$ ./bin/hdfs dfs –ls .
```

上面的命令和下面的命令是等价的：

```
$ ./bin/hdfs dfs –ls /user/hadoop
```

4.1.2 HDFS目录操作

如果要列出HDFS上的所有目录，可以使用如下命令：

```
$ ./bin/hdfs dfs –ls
```

下面，可以使用如下命令创建一个input目录：

```
$ ./bin/hdfs dfs –mkdir input
```

如果要在HDFS的根目录下创建一个名称为input的目录，则需要使用如下命令：

```
$ ./bin/hdfs dfs –mkdir /input
```

可以使用rm命令删除一个目录，比如，可以使用如下命令删除刚才在HDFS中创建的“/input”目录（不是“/user/hadoop/input”目录）：

```
$ ./bin/hdfs dfs –rm –r /input
```

4.1.2 HDFS目录操作

2. 文件操作

首先，使用vim编辑器，在本地Linux文件系统的“/home/hadoop/”目录下创建一个文件myLocalFile.txt，里面可以随意输入一些单词，比如，输入如下三行：

```
Hadoop
Spark
XMU DBLAB
```

然后，可以使用如下命令把本地文件系统的“/home/hadoop/myLocalFile.txt”上传到HDFS中的当前用户目录的input目录下，也就是上传到HDFS的“/user/hadoop/input/”目录下：

```
$ ./bin/hdfs dfs -put /home/hadoop/myLocalFile.txt  input
```

可以使用ls命令查看一下文件是否成功上传到HDFS中，具体如下：

```
$ ./bin/hdfs dfs -ls input
```

4.1.2 HDFS目录操作

下面使用如下命令查看HDFS中的myLocalFile.txt这个文件的内容：

```
$ ./bin/hdfs dfs –cat input/myLocalFile.txt
```

下面把HDFS中的myLocalFile.txt文件下载到本地文件系统中的“/home/hadoop/下载/”这个目录下，命令如下：

```
$ ./bin/hdfs dfs -get input/myLocalFile.txt  /home/hadoop/下载
```

可以使用如下命令，到本地文件系统查看下载下来的文件myLocalFile.txt：

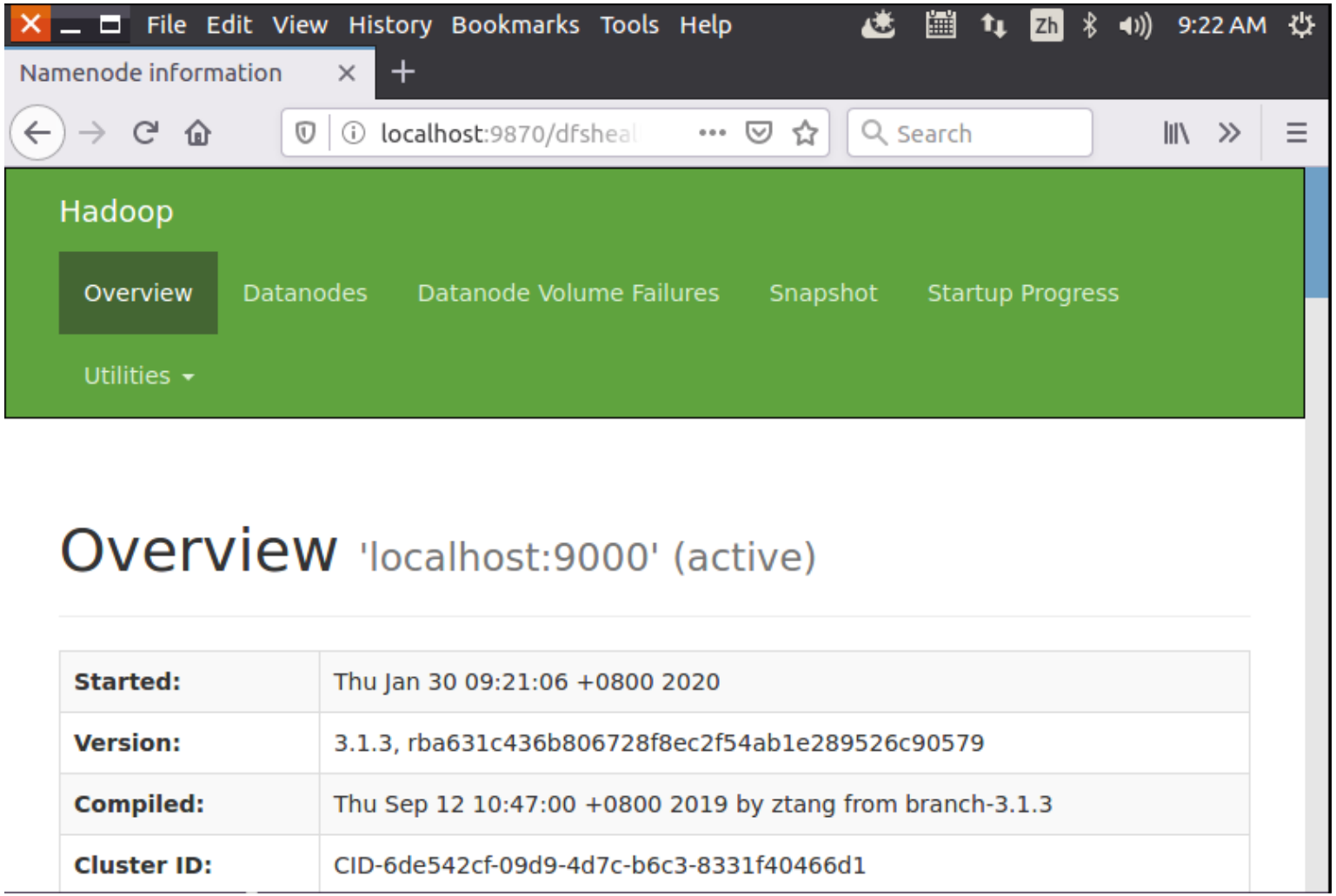
```
$ cd ~
$ cd 下载
$ ls
$ cat myLocalFile.txt
```

4.1.2 HDFS目录操作

最后，了解一下如何把文件从HDFS中的一个目录拷贝到HDFS中的另外一个目录。比如，如果要把HDFS的“/user/hadoop/input/myLocalFile.txt”文件，拷贝到HDFS的另外一个目录“/input”中（注意，这个input目录位于HDFS根目录下），可以使用如下命令：

```
$ ./bin/hdfs dfs -cp input/myLocalFile.txt  /input
```

4.2 利用HDFS的Web管理界面

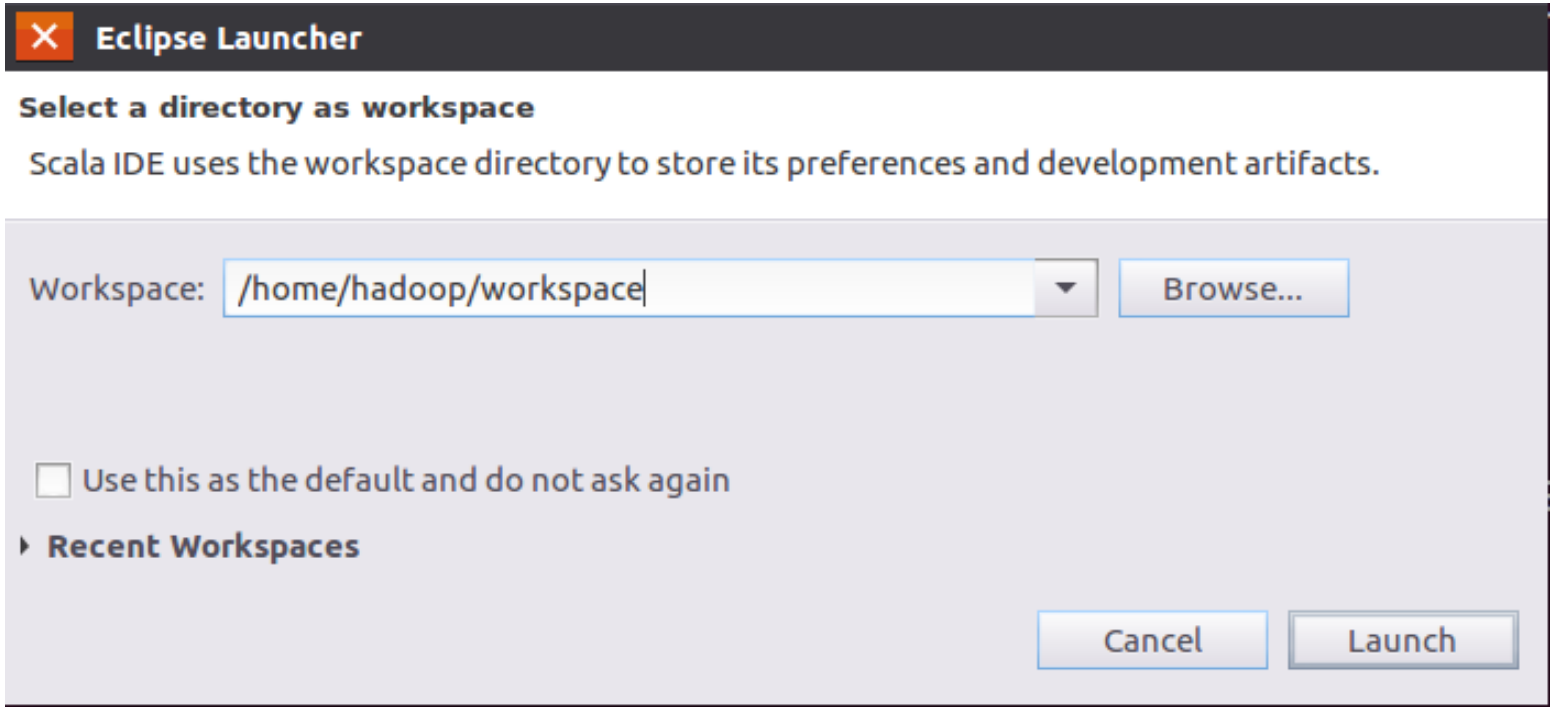


4.3 HDFS编程实践

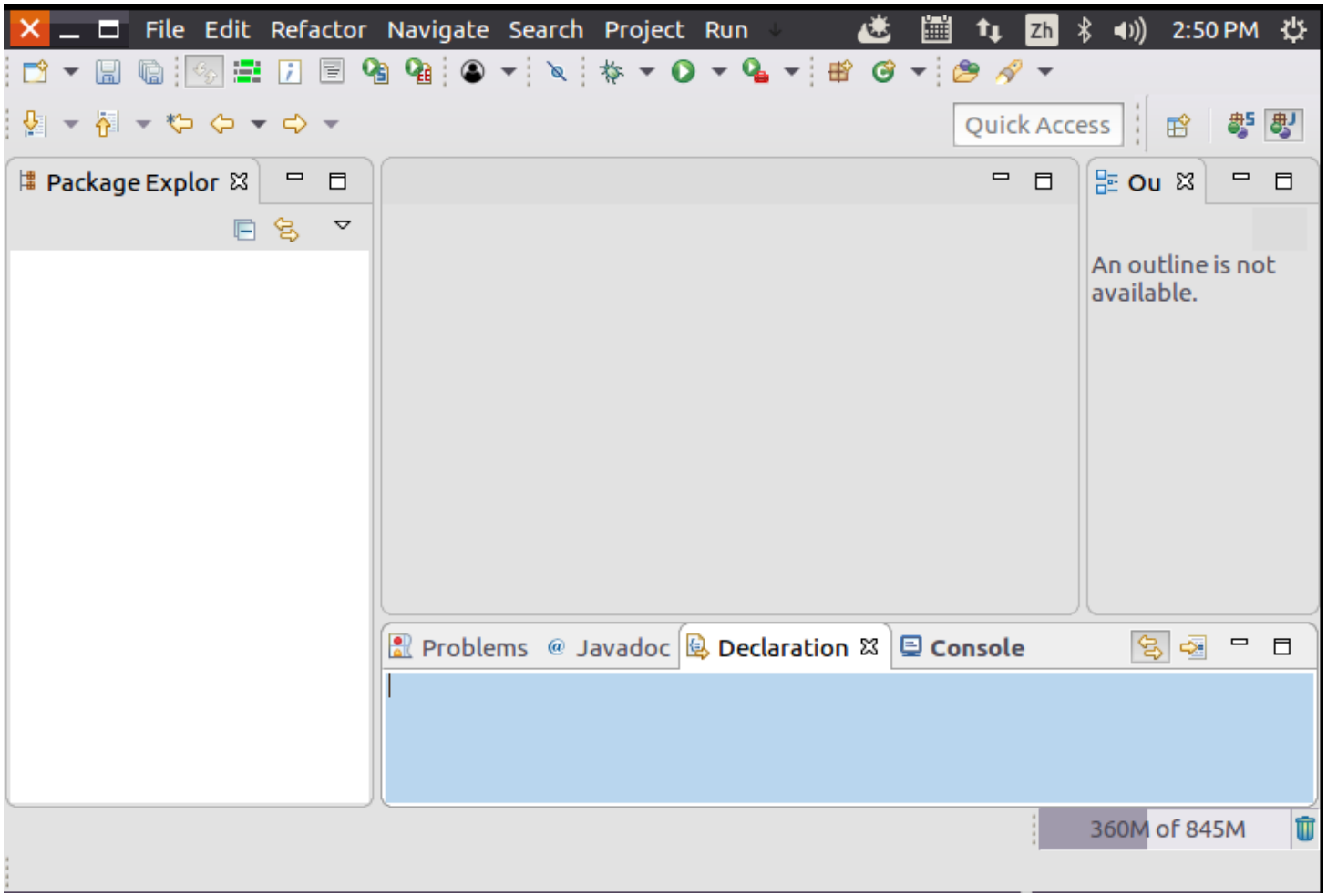
现在要执行的任务是：假设在目录“hdfs://localhost:9000/user/hadoop”下面有几个文件，分别是file1.txt、file2.txt、file3.txt、file4.abc和file5.abc，这里需要从该目录中过滤出所有后缀名不为“.abc”的文件，对过滤之后的文件进行读取，并将这些文件的内容合并到文件“hdfs://localhost:9000/user/hadoop/merge.txt”中。

4.3.1 在Eclipse中创建项目
4.3.2为项目添加需要用到的JAR包
4.3.3 编写Java应用程序
4.3.4 编译运行程序
4.3.5应用程序的部署

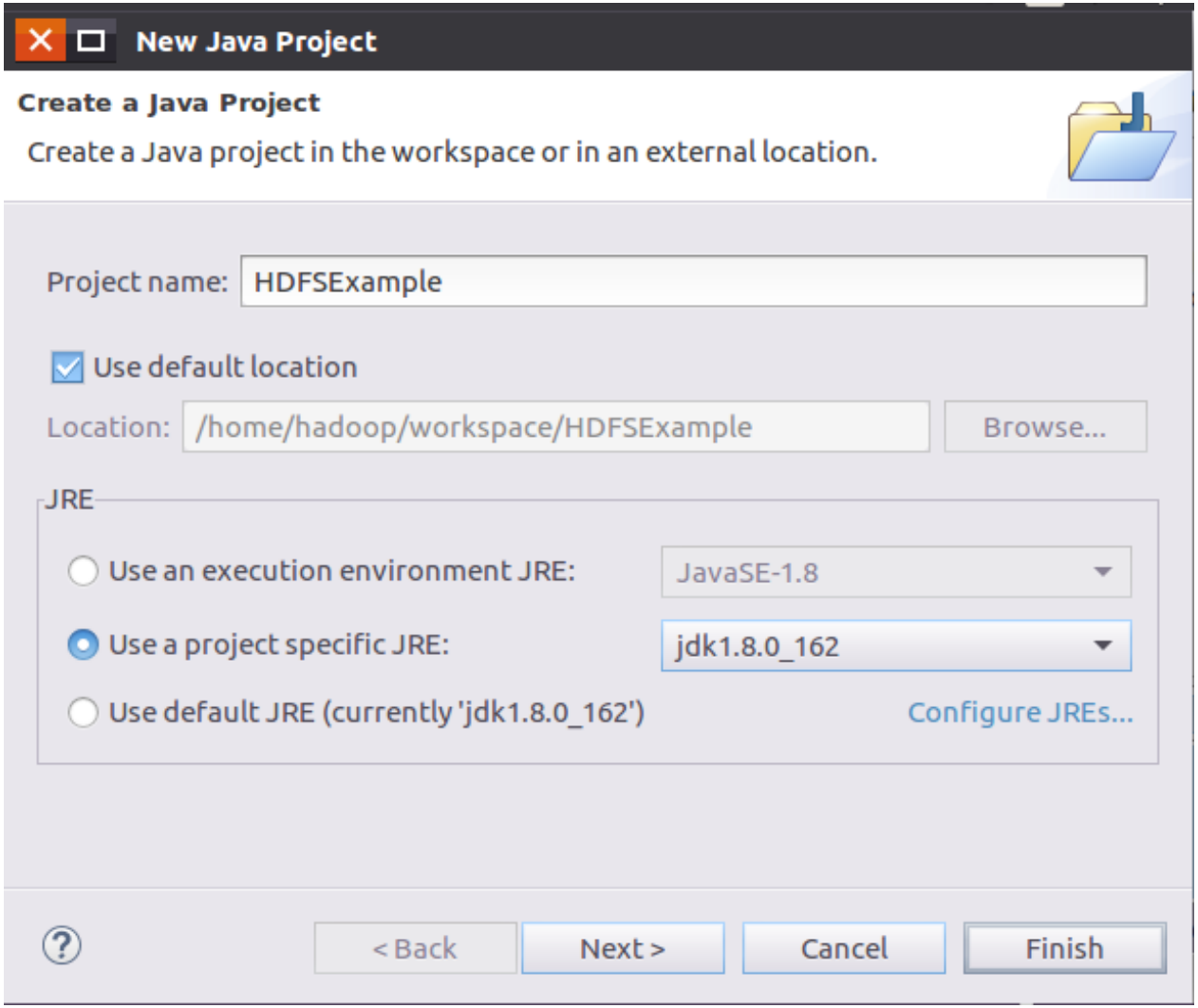
4.3.1 在Eclipse中创建项目



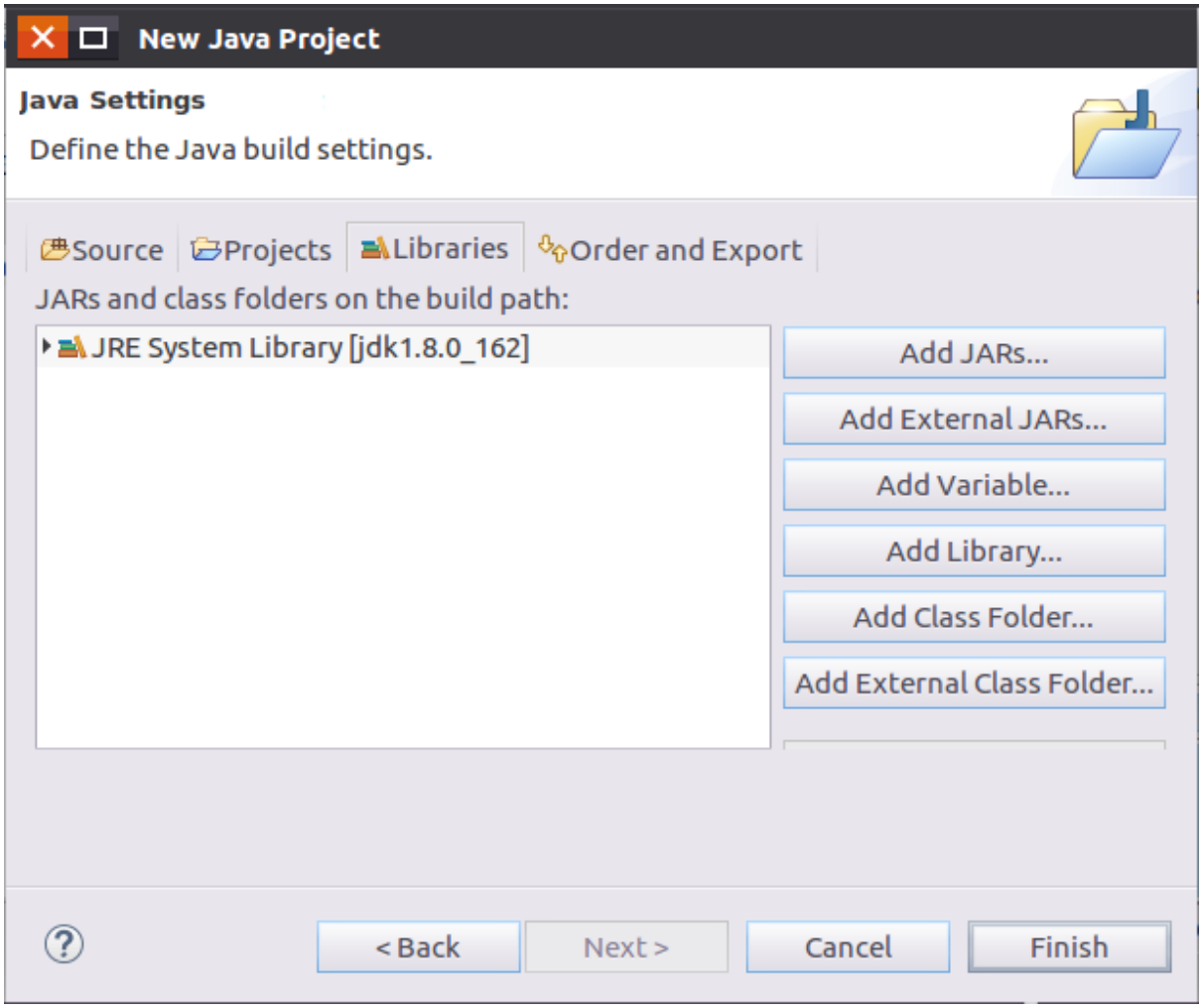
4.3.1 在Eclipse中创建项目



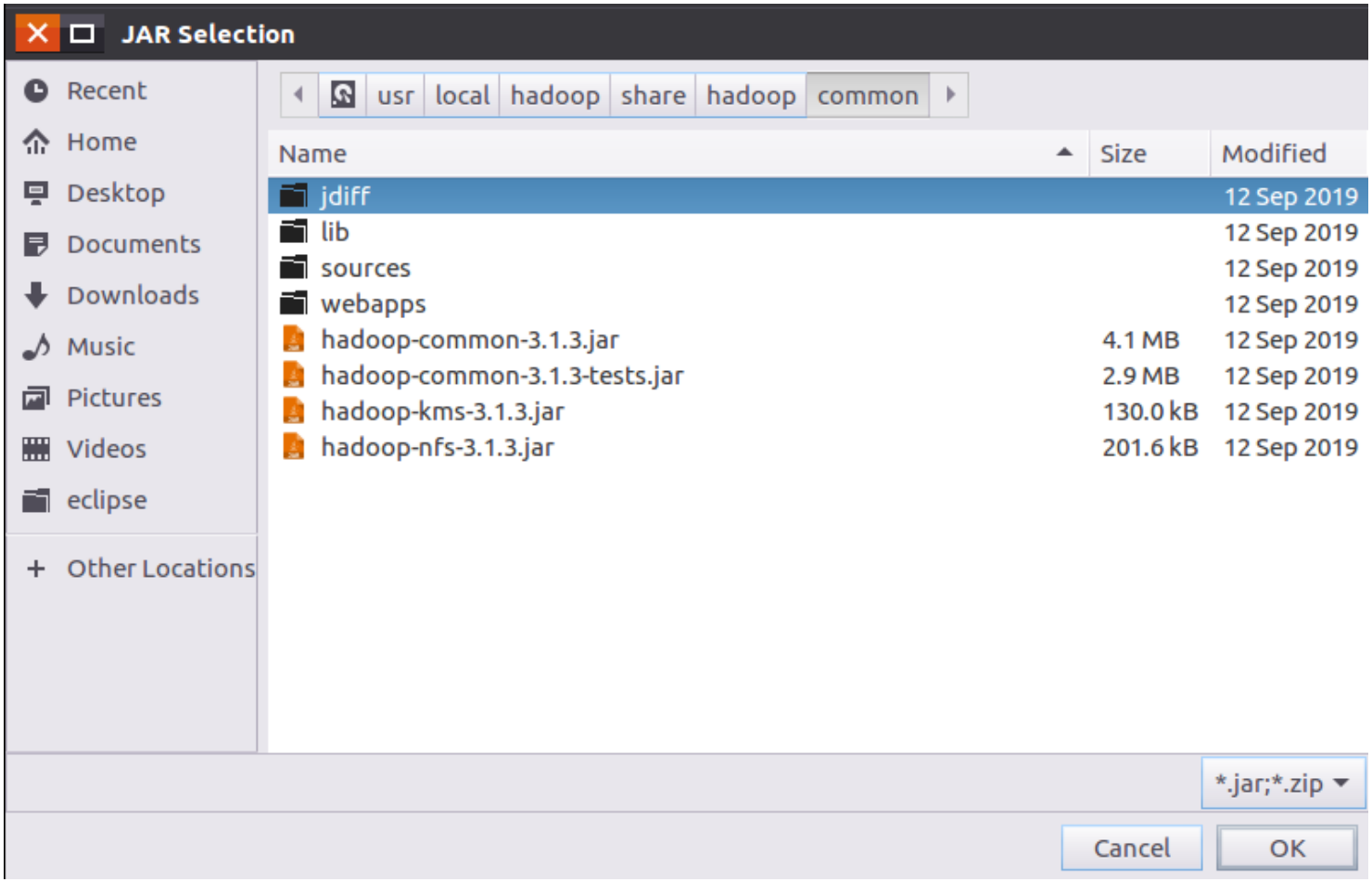
4.3.1 在Eclipse中创建项目



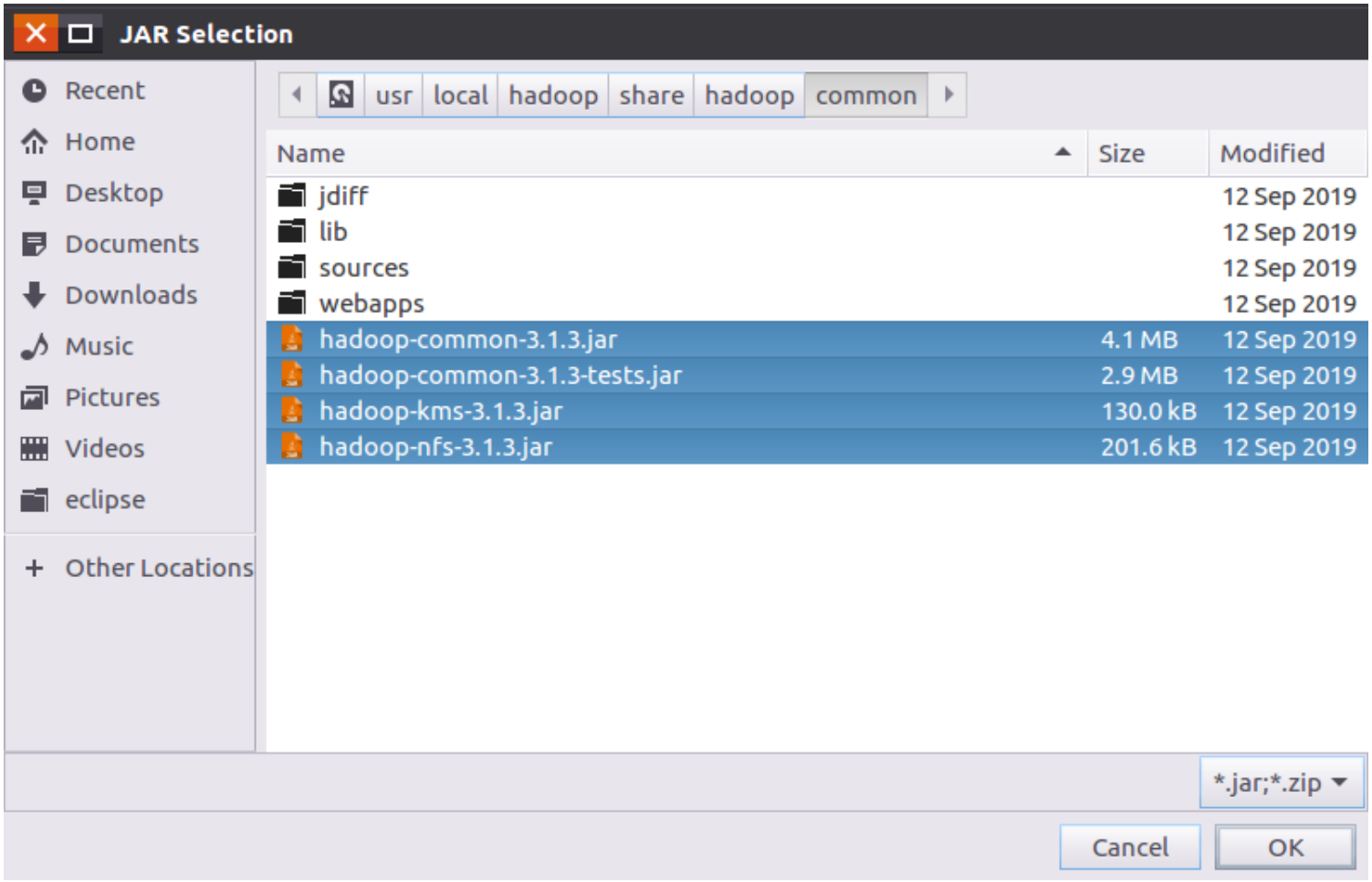
4.3.2为项目添加需要用到的JAR包



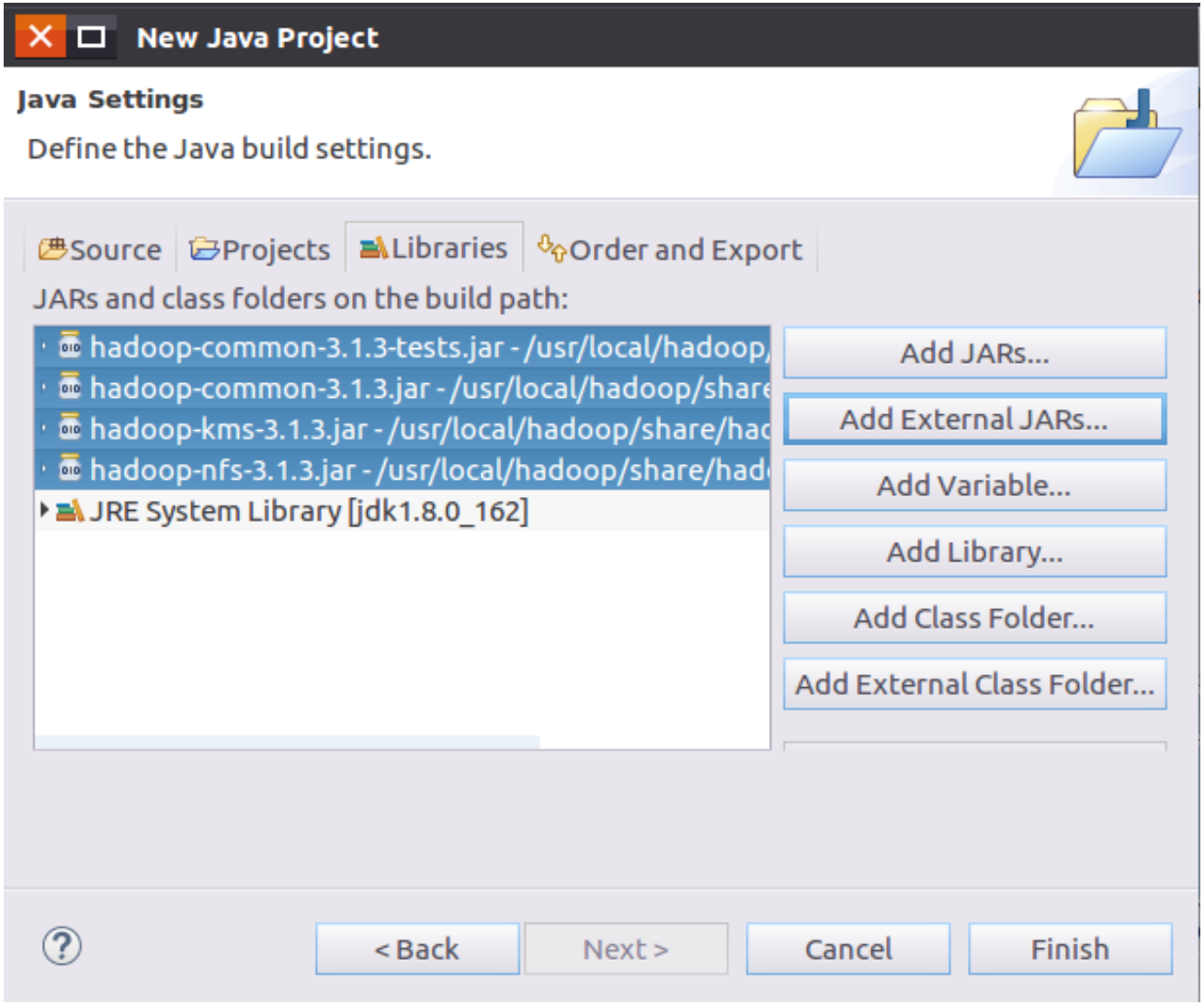
4.3.2为项目添加需要用到的JAR包



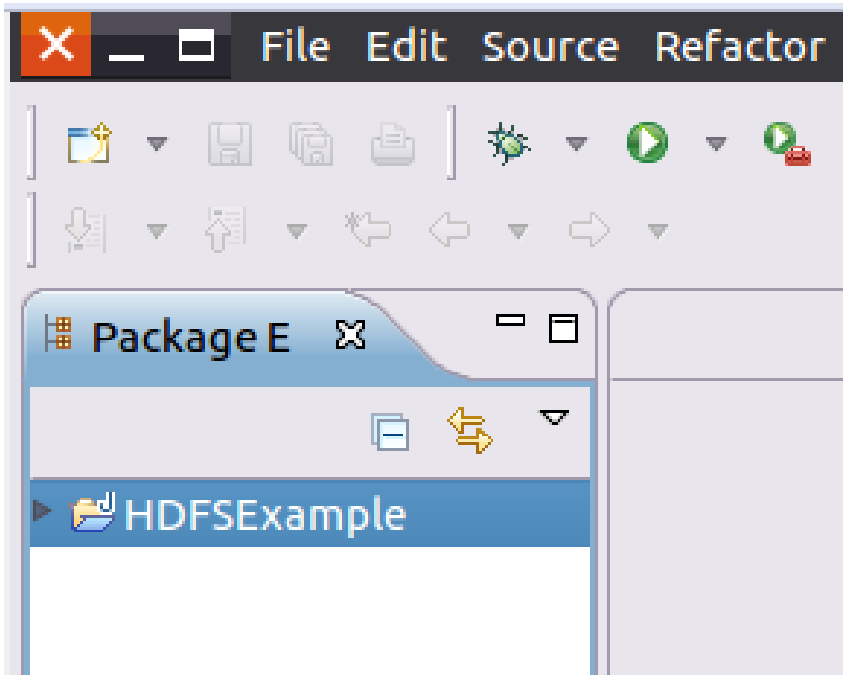
4.3.2为项目添加需要用到的JAR包



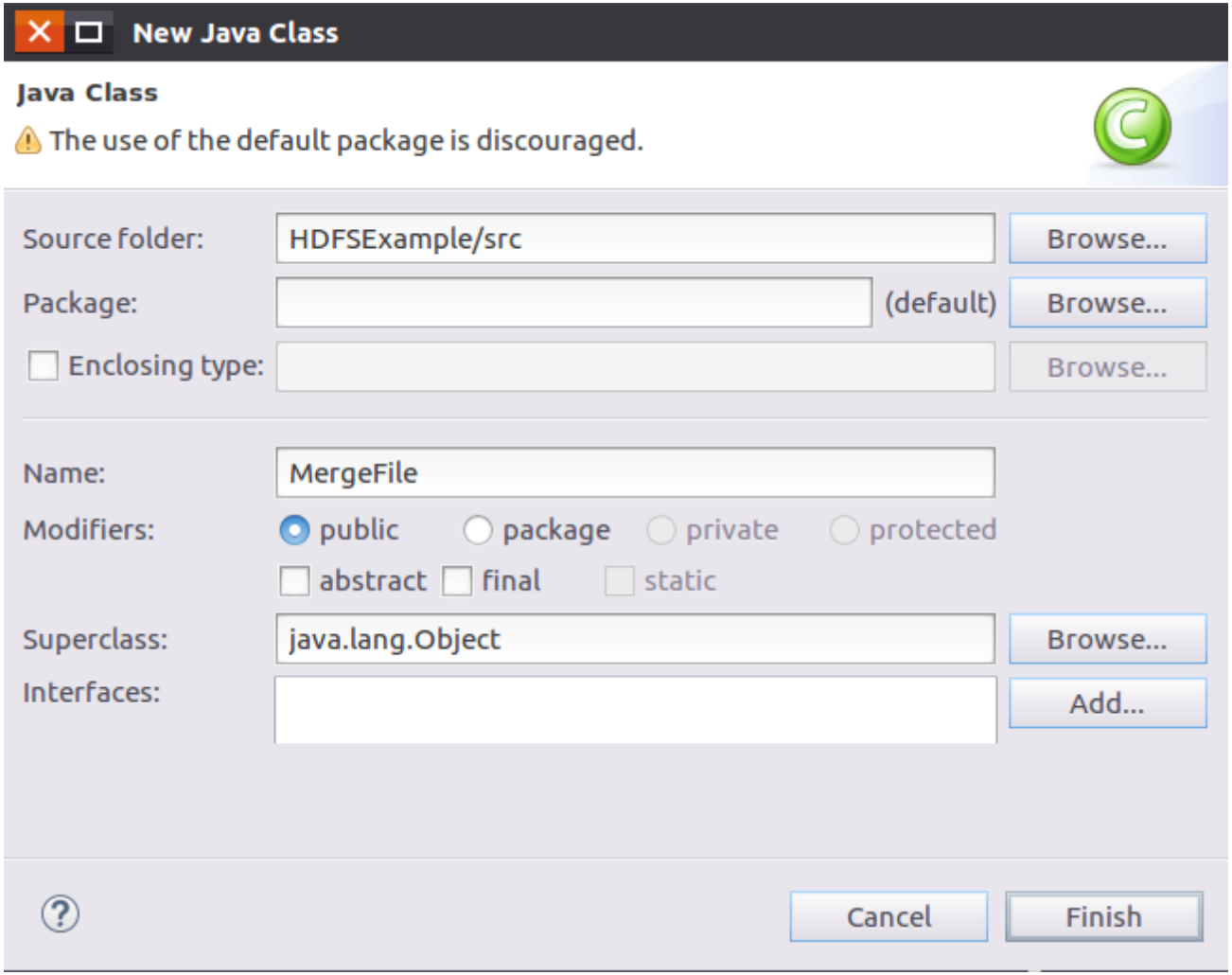
4.3.2为项目添加需要用到的JAR包



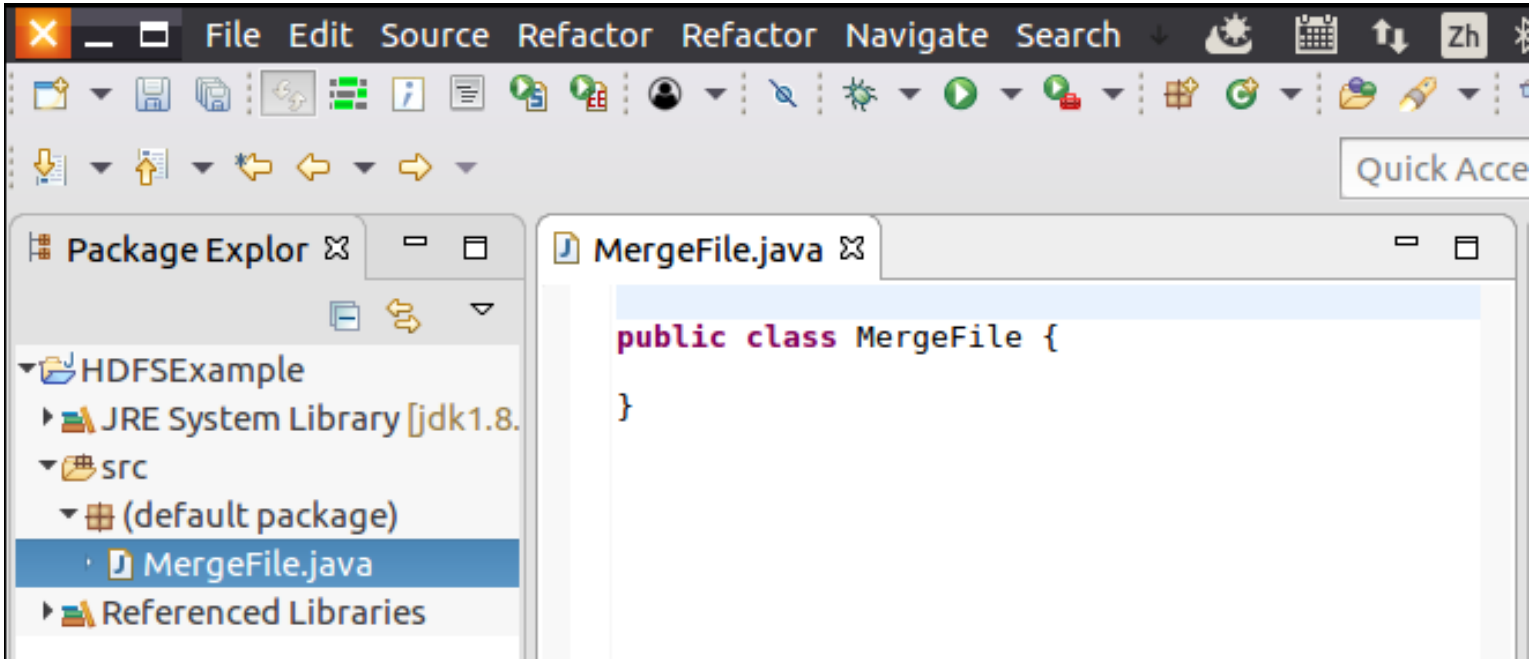
4.3.3 编写Java应用程序



4.3.3 编写Java应用程序



4.3.3 编写Java应用程序



4.3.3 编写Java应用程序

```
import java.io.IOException;
import java.io.PrintStream;
import java.net.URI;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.*;

/**
 * 过滤掉文件名满足特定条件的文件
 */
class MyPathFilter implements PathFilter {
    String reg = null;
    MyPathFilter(String reg) {
        this.reg = reg;
    }
    public boolean accept(Path path) {
                        if (!(path.toString().matches(reg)))
                                return true;
                        return false;
                }
}
```

4.3.3 编写Java应用程序

```
/***
 * 利用FSDataOutputStream和FSDataInputStream合并HDFS中的文件
 */
public class MergeFile {
        Path inputPath = null; //待合并的文件所在的目录的路径
        Path outputPath = null; //输出文件的路径
        public MergeFile(String input, String output) {
                this.inputPath = new Path(input);
                this.outputPath = new Path(output);
        }
        public void doMerge() throws IOException {
                Configuration conf = new Configuration();
                conf.set("fs.defaultFS","hdfs://localhost:9000");

conf.set("fs.hdfs.impl","org.apache.hadoop.hdfs.DistributedFileSystem");
                FileSystem fsSource =
FileSystem.get(URI.create(inputPath.toString()), conf);
```

4.3.3 编写Java应用程序

```
FileSystem fsDst = FileSystem.get(URI.create(outputPath.toString()), conf);
                                        //下面过滤掉输入目录中后缀为.abc的文件
                FileStatus[] sourceStatus = fsSource.listStatus(inputPath,
                                        new MyPathFilter(".*\\.abc"));
                FSDataOutputStream fsdos = fsDst.create(outputPath);
                PrintStream ps = new PrintStream(System.out);
                //下面分别读取过滤之后的每个文件的内容，并输出到同一个文件中
                for (FileStatus sta : sourceStatus) {
                        //下面打印后缀不为.abc的文件的路径、文件大小
                        System.out.print("路径：" + sta.getPath() + "    文件大小：" +
sta.getLen()
                                                + "   权限：" + sta.getPermission() + "   内容：
");
                        FSDataInputStream fsdis = fsSource.open(sta.getPath());
                        byte[] data = new byte[1024];
                        int read = -1;

                        while ((read = fsdis.read(data)) > 0) {
                                ps.write(data, 0, read);
                                fsdos.write(data, 0, read);
                        }
                        fsdis.close();
                }
                ps.close();
                fsdos.close();
        }
```

4.3.3 编写Java应用程序

```
        public static void main(String[] args) throws IOException {
                MergeFile merge = new MergeFile(
                                "hdfs://localhost:9000/user/hadoop/",
                                "hdfs://localhost:9000/user/hadoop/merge.txt");
                merge.doMerge();
        }
}
```

4.3.4 编译运行程序

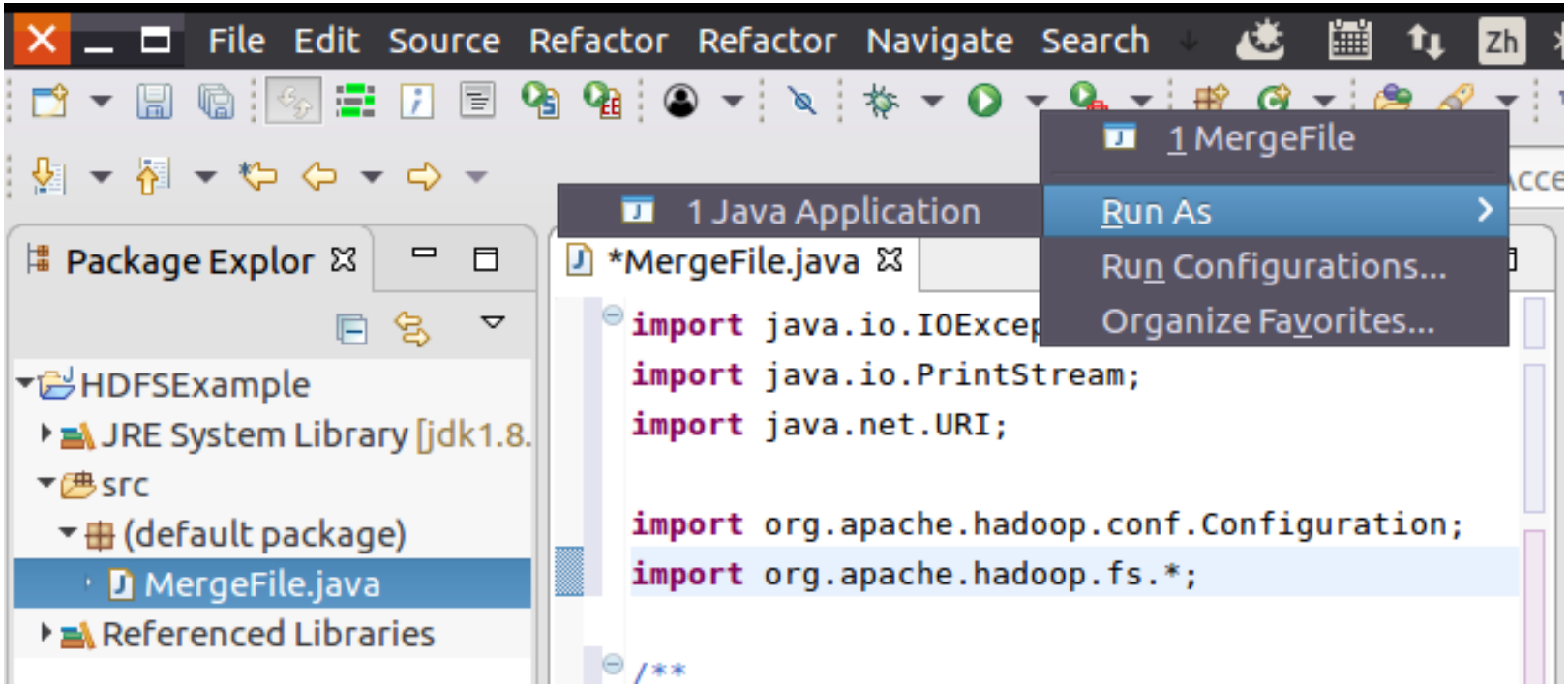
在开始编译运行程序之前，请一定确保Hadoop已经启动运行，如果还没有启动，需要打开一个Linux终端，输入以下命令启动Hadoop：

```
$ cd /usr/local/hadoop
$ ./sbin/start-dfs.sh
```

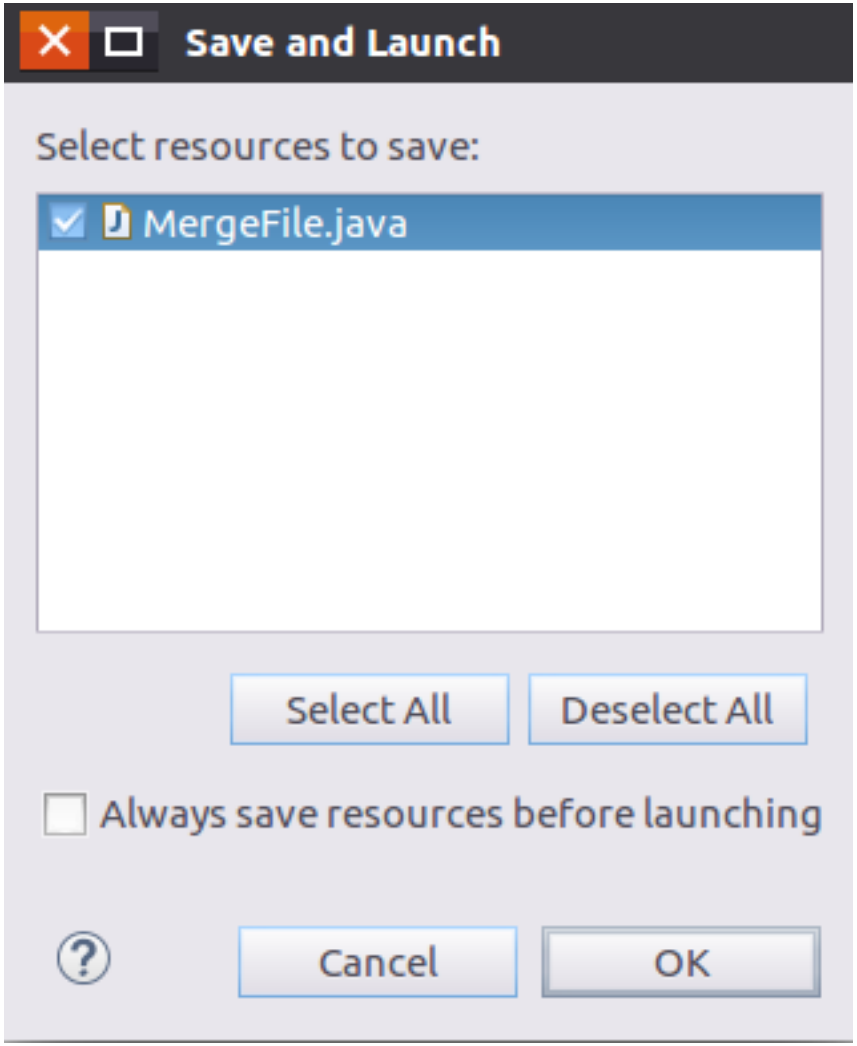
表4-1 HDFS系统中的文件内容

文件名称	文件内容
file1.txt	this is file1.txt
file2.txt	this is file2.txt
file3.txt	this is file3.txt
file4.abc	this is file4.abc
file5.abc	this is file5.abc

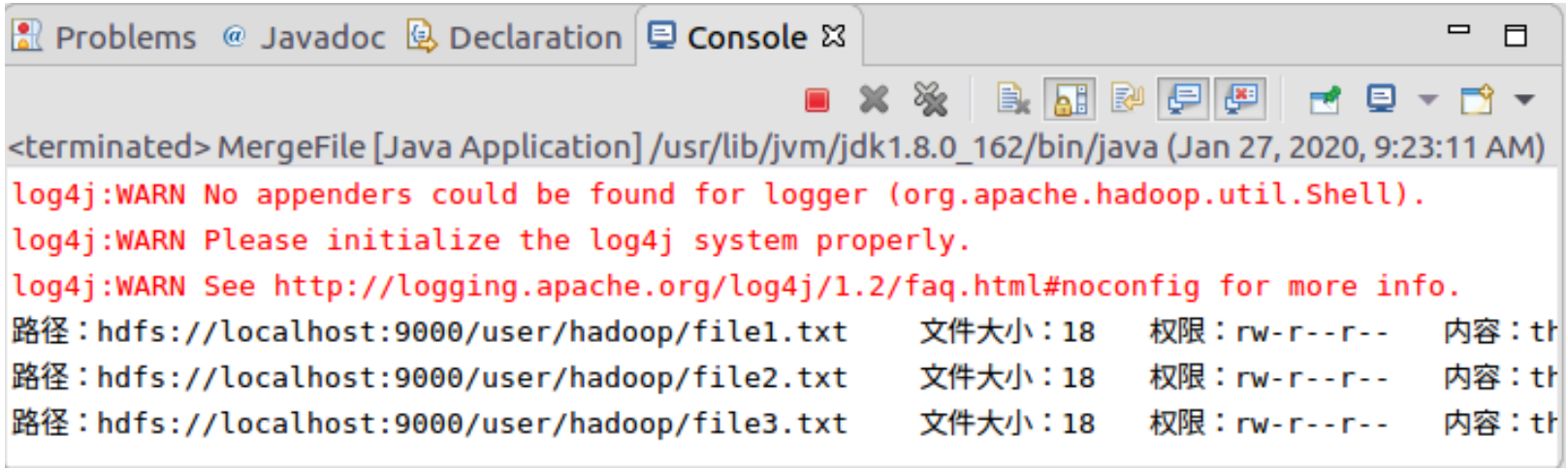
4.3.4 编译运行程序


File Edit Source Refactor Refactor Navigate Search
1 MergeFile
1 Java Application
Run As
Run Configurations...
Organize Favorites...
Package Explor
*MergeFile.java
HDFSExample
JRE System Library [jdk1.8.
src
(default package)
MergeFile.java
Referenced Libraries
import java.io.IOExcep
import java.io.PrintStream;
import java.net.URI;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.*;
/**

4.3.4 编译运行程序


Save and Launch
Select resources to save:
MergeFile.java
Select All
Deselect All
Always save resources before launching
Cancel
OK

4.3.4 编译运行程序


Problems
Javadoc
Declaration
Console
<terminated> MergeFile [Java Application] /usr/lib/jvm/jdk1.8.0_162/bin/java (Jan 27, 2020, 9:23:11 AM)
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
路径：hdfs://localhost:9000/user/hadoop/file1.txt 文件大小：18 权限：rw-r--r-- 内容：th
路径：hdfs://localhost:9000/user/hadoop/file2.txt 文件大小：18 权限：rw-r--r-- 内容：th
路径：hdfs://localhost:9000/user/hadoop/file3.txt 文件大小：18 权限：rw-r--r-- 内容：th

4.3.4 编译运行程序

如果程序运行成功，这时，可以到HDFS中查看生成的merge.txt文件，比如，可以在Linux终端中执行如下命令：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -ls /user/hadoop
$ ./bin/hdfs dfs -cat /user/hadoop/merge.txt
```

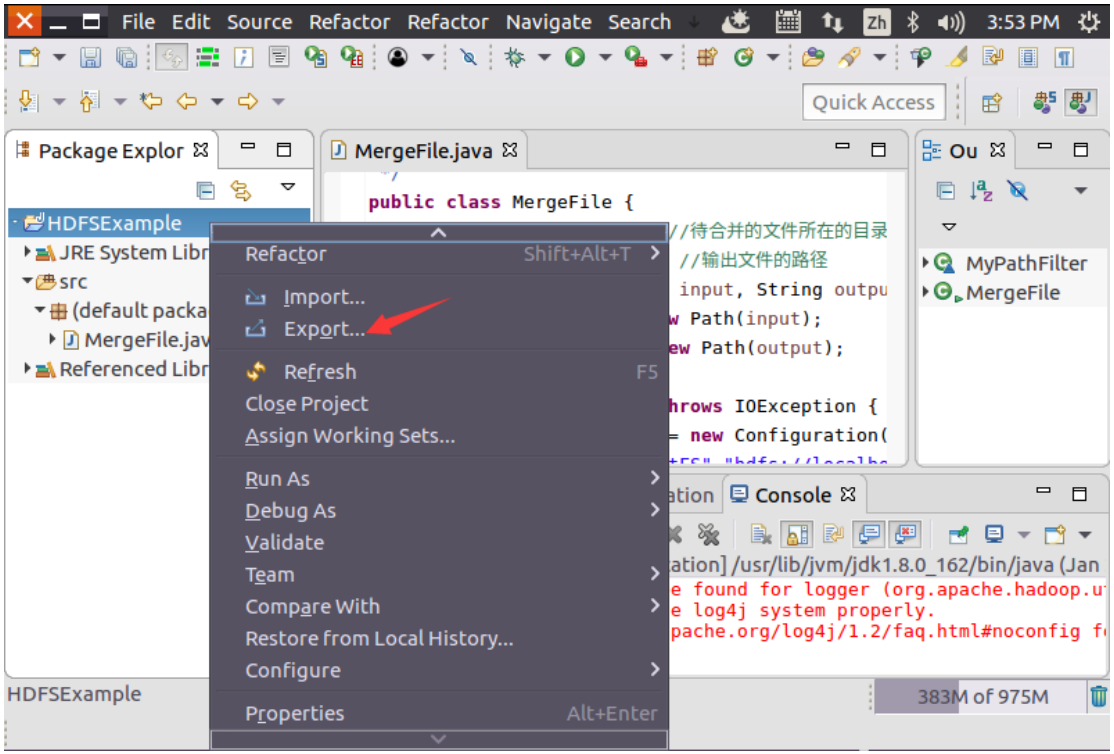
可以看到如下结果：

this is file1.txt
this is file2.txt
this is file3.txt

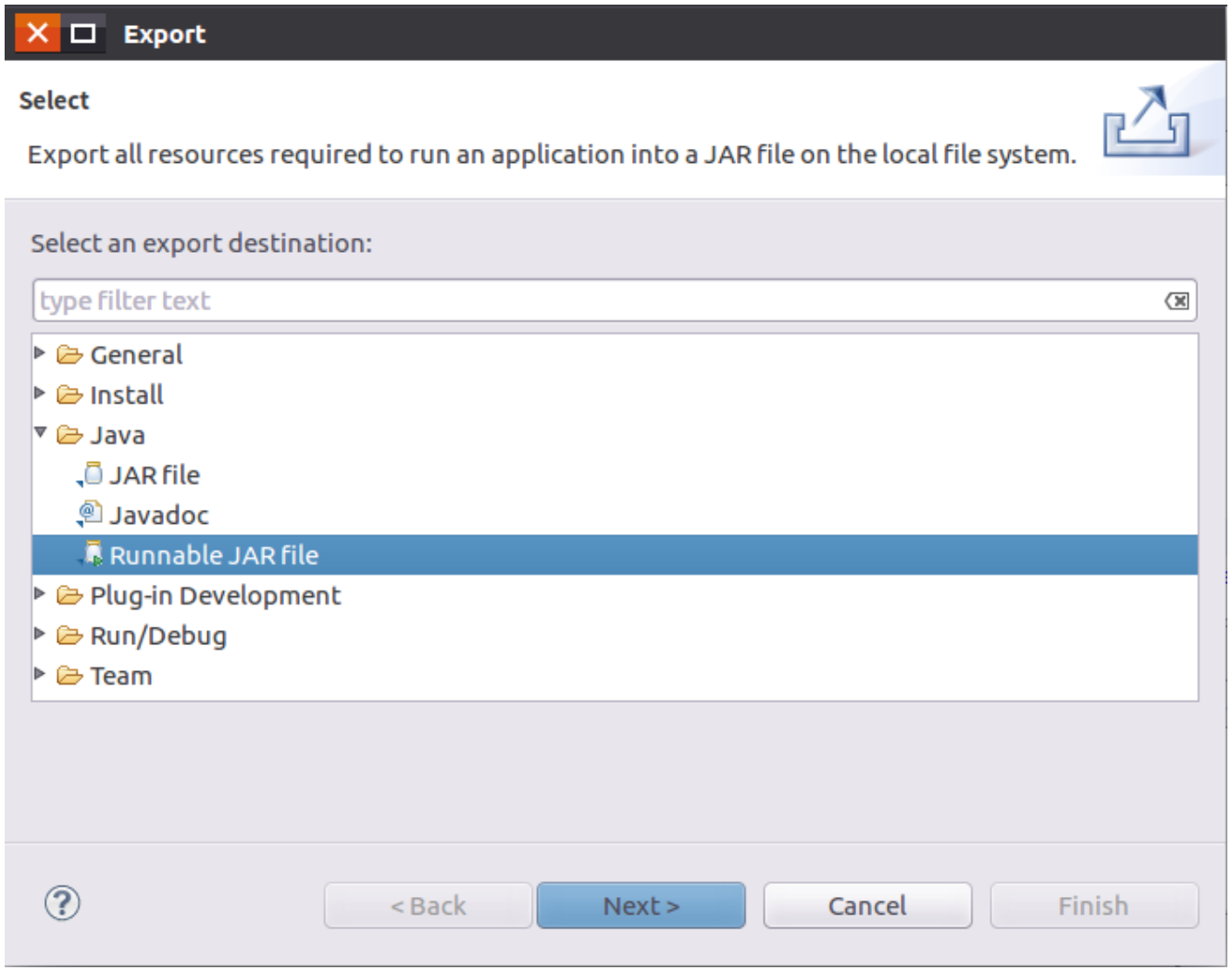
4.3.5应用程序的部署

首先，在Hadoop安装目录下新建一个名称为myapp的目录，用来存放我们自己编写的Hadoop应用程序，可以在Linux的终端中执行如下命令：

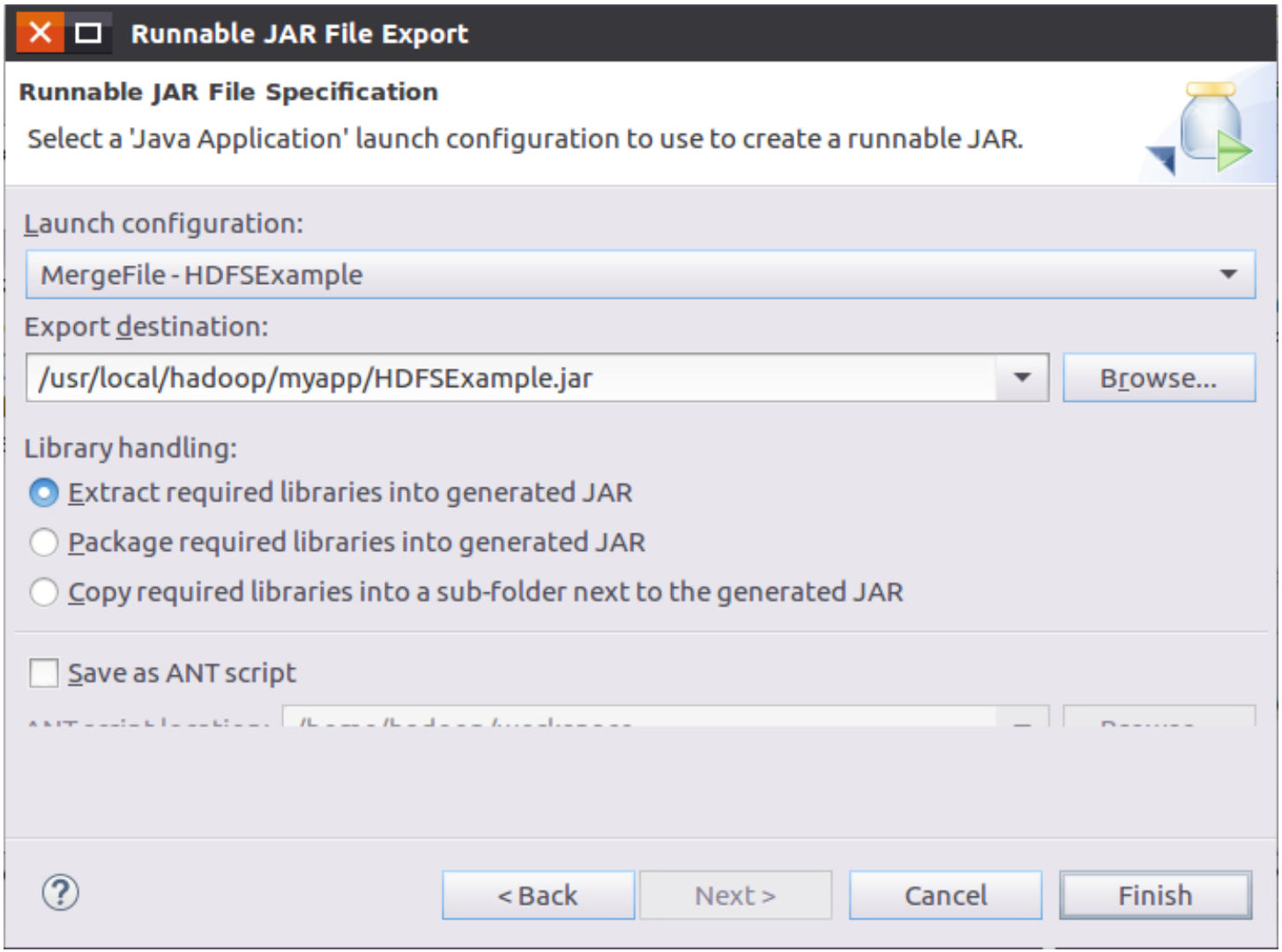
```
$ cd /usr/local/hadoop
$ mkdir myapp
```



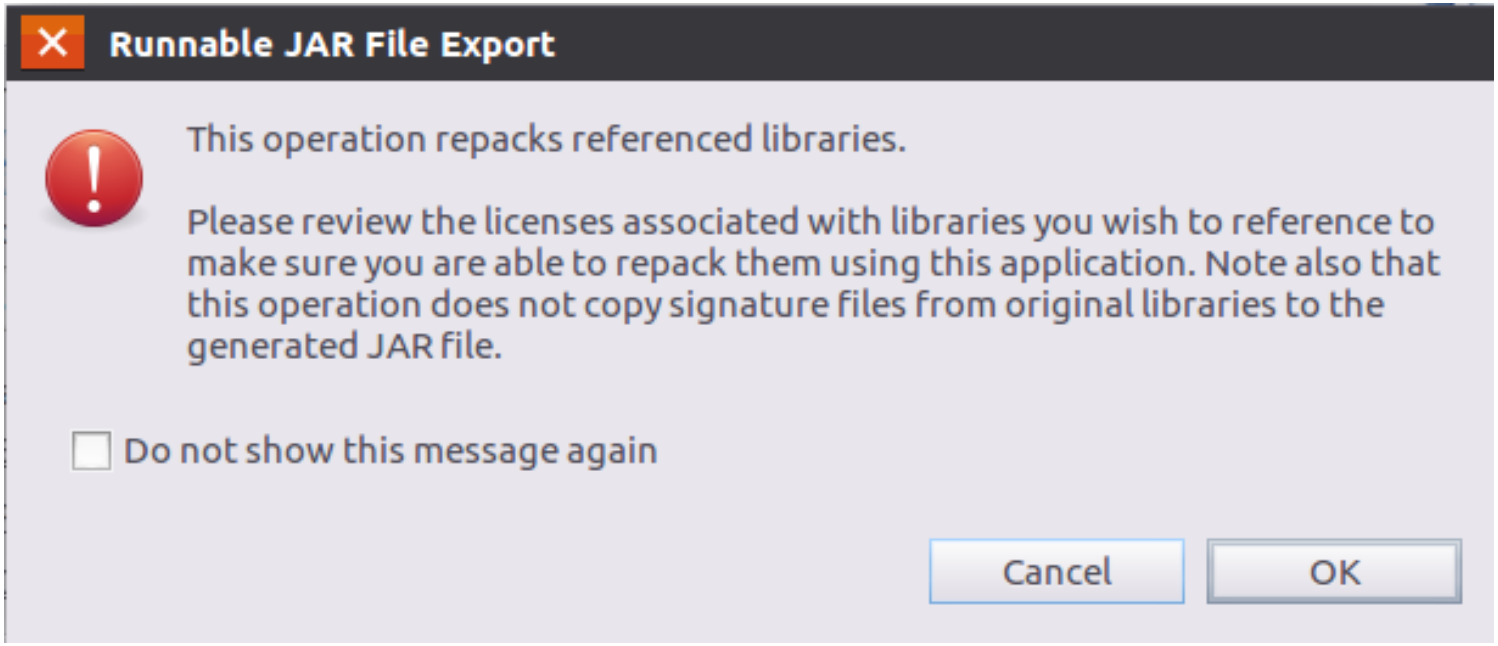
4.3.5应用程序的部署


Export
Select
Export all resources required to run an application into a JAR file on the local file system.
Select an export destination:
type filter text
General
Install
Java
JAR file
Javadoc
Runnable JAR file
Plug-in Development
Run/Debug
Team
< Back
Next >
Cancel
Finish

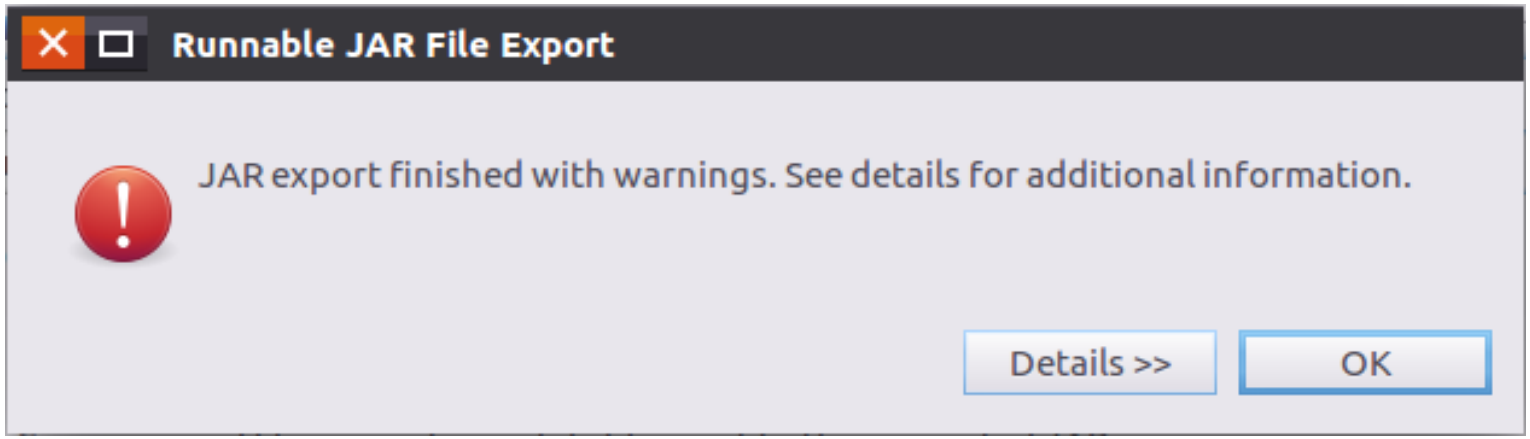
4.3.5应用程序的部署


Runnable JAR File Export
Runnable JAR File Specification
Select a 'Java Application' launch configuration to use to create a runnable JAR.
Launch configuration:
MergeFile - HDFSExample
Export destination:
/usr/local/hadoop/myapp/HDFSExample.jar
Browse...
Library handling:
Extract required libraries into generated JAR
Package required libraries into generated JAR
Copy required libraries into a sub-folder next to the generated JAR
Save as ANT script
< Back
Next >
Cancel
Finish

4.3.5应用程序的部署


Runnable JAR File Export
This operation repacks referenced libraries.
Please review the licenses associated with libraries you wish to reference to make sure you are able to repack them using this application. Note also that this operation does not copy signature files from original libraries to the generated JAR file.
Do not show this message again
Cancel
OK

4.3.5应用程序的部署


Runnable JAR File Export
JAR export finished with warnings. See details for additional information.
Details >>
OK

4.3.5应用程序的部署

可以到Linux系统中查看一下生成的HDFSExample.jar文件，可以在Linux的终端中执行如下命令：

```
$ cd /usr/local/hadoop/myapp
$ ls
```

可以看到，“/usr/local/hadoop/myapp”目录下已经存在一个HDFSExample.jar文件。
由于之前已经运行过一次程序，已经生成了merge.txt，因此，需要首先执行如下命令删除该文件：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -rm /user/hadoop/merge.txt
```

4.3.5应用程序的部署

现在，就可以在Linux系统中，使用hadoop jar命令运行程序，命令如下：

```
$ cd /usr/local/hadoop
$ ./bin/hadoop jar ./myapp/HDFSExample.jar
```

上面程序执行结束以后，可以到HDFS中查看生成的merge.txt文件，比如，可以在Linux终端中执行如下命令：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -ls /user/hadoop
$ ./bin/hdfs dfs -cat /user/hadoop/merge.txt
```

可以看到如下结果：

this is file1.txt
this is file2.txt
this is file3.txt

4.4 本章小结

大数据时代必须解决海量数据的高效存储问题，为此，谷歌开发了分布式文件系统GFS，通过网络实现文件在多台机器上的分布式存储，较好地满足了大规模数据存储的需求。HDFS是针对GFS的开源实现，它是Hadoop两大核心组成部分之一。

在很多情形下，需要使用Shell命令来操作HDFS，因此，本章介绍了HDFS操作常用的Shell命令，包括目录操作命令和文件操作命令等。同时，还介绍了如何利用HDFS的Web管理界面，以可视化的方式查看HDFS的相关信息。最后，本章详细介绍了如何使用Eclipse开发操作HDFS的Java应用程序。本章介绍的Eclipse开发方法，为后续章节的编程开发提供了很好的借鉴。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过200万次，累计访问量超过1000万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元

教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用》教材

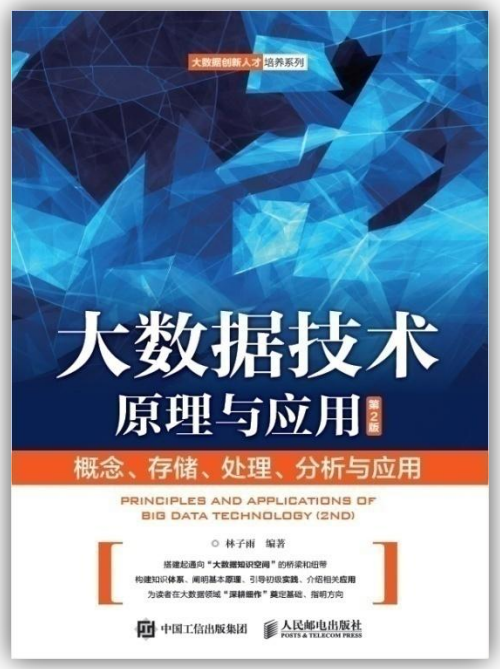
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

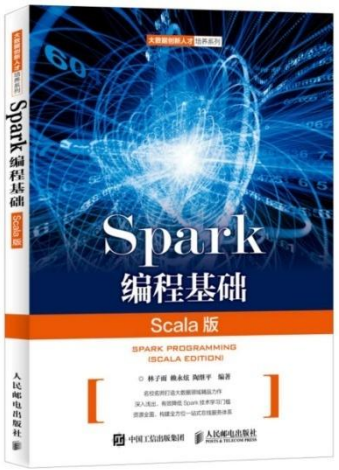
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

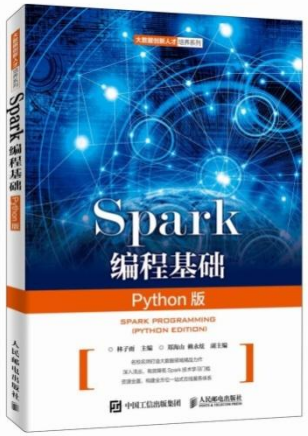
厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3
教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

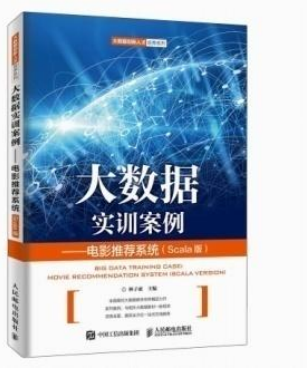
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

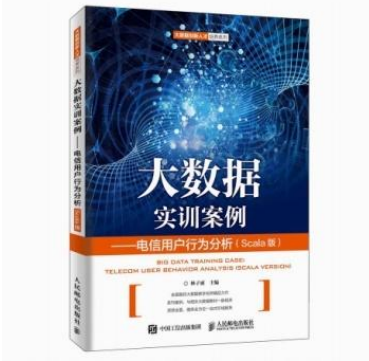
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!