厦门大学林子雨编著

《大数据技术原理与应用》

教材配套

机房上机实验指南

实验5

第七章 MapReduce编程初级实践

（版本号：2016年5月14日版本）



主讲教师：林子雨

厦门大学数据库实验室

二零一六年五月

目录

[**1.** **实验目的** 1](#_Toc451240062)

[**2.** **实验平台** 1](#_Toc451240063)

[**3.** **实验内容和要求** 1](#_Toc451240064)

[1.编程实现文件合并和去重操作 1](#_Toc451240065)

[2. 编写程序实现对输入文件的排序 2](#_Toc451240066)

[3. 对给定的表格进行信息挖掘 3](#_Toc451240067)

[**4.** **实验报告** 4](#_Toc451240068)

[**附录1:任课教师介绍** 4](#_Toc451240069)

[**附录2：课程教材介绍** 4](#_Toc451240070)

[**附录3：中国高校大数据课程公共服务平台介绍** 6](#_Toc451240071)

厦门大学林子雨编著《大数据技术原理与应用》

教材配套机房上机实验指南

实验5

第七章 MapReduce编程初级实践

**主讲教师：林子雨**

E-mail: ziyulin@xmu.edu.cn 个人主页：<http://www.cs.xmu.edu.cn/linziyu>

1. **实验目的**

1.通过实验掌握基本的MapReduce编程方法；

2.掌握用MapReduce解决一些常见的数据处理问题，包括数据去重、数据排序和数据挖掘等。

1. **实验平台**

已经配置完成的Hadoop伪分布式环境。

1. **实验内容和要求**

## 1.编程实现文件合并和去重操作

对于两个输入文件，即文件A和文件B，请编写MapReduce程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新的输出文件C。下面是输入文件和输出文件的一个样例供参考。

 输入文件A的样例如下：

|  |
| --- |
|  20150101 x 20150102 y 20150103 x 20150104 y 20150105 z20150106 x |

输入文件B的样例如下：

|  |
| --- |
| 20150101 y20150102 y20150103 x20150104 z20150105 y |

根据输入文件A和B合并得到的输出文件C的样例如下：

|  |
| --- |
| 20150101 x20150101 y20150102 y20150103 x20150104 y20150104 z20150105 y 20150105 z20150106 x |

## 2. 编写程序实现对输入文件的排序

现在有多个输入文件，每个文件中的每行内容均为一个整数。要求读取所有文件中的整数，进行升序排序后，输出到一个新的文件中，输出的数据格式为每行两个整数，第一个数字为第二个整数的排序位次，第二个整数为原待排列的整数。下面是输入文件和输出文件的一个样例供参考。

输入文件1的样例如下：

|  |
| --- |
| 33371240 |

输入文件2的样例如下：

|  |
| --- |
| 416395 |

输入文件3的样例如下：

|  |
| --- |
| 14525 |

根据输入文件1、2和3得到的输出文件如下：

|  |
| --- |
| 1 12 43 54 125 166 257 338 379 3910 4011 45 |

## 3. 对给定的表格进行信息挖掘

下面给出一个child-parent的表格，要求挖掘其中的父子辈关系，给出祖孙辈关系的表格。

输入文件内容如下：

|  |
| --- |
|  child parent Steven Lucy Steven Jack Jone Lucy Jone Jack Lucy Mary Lucy Frank Jack Alice Jack Jesse David Alice David Jesse Philip David Philip Alma Mark DavidMark Alma |

输出文件内容如下：

|  |
| --- |
|  grandchild grandparent Steven Alice Steven Jesse Jone Alice Jone Jesse Steven Mary Steven Frank Jone Mary Jone Frank Philip Alice Philip Jesse Mark AliceMark Jesse |

1. **实验报告**

|  |
| --- |
| 《大数据技术原理与应用》 实验报告 |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： |
| 解决问题的思路： |
| 实验内容与完成情况： |
| 出现的问题： |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： |

**附录1:任课教师介绍**

|  |  |
| --- | --- |
| linziyu2014 | 林子雨(1978－),男,博士,厦门大学计算机科学系助理教授,主要研究领域为数据库,实时主动数据仓库,数据挖掘.主讲课程：《大数据技术基础》办公地点：厦门大学海韵园科研2号楼E-mail: ziyulin@xmu.edu.cn个人主页：http://www.cs.xmu.edu.cn/linziyu数据库实验室网站：http://dblab.xmu.edu.cn |

**附录2：课程教材介绍**



《大数据技术原理与应用——概念、存储、处理、分析与应用》，由厦门大学计算机科学系教师林子雨博士编著，是中国高校第一本系统介绍大数据知识的专业教材。本书定位为大数据技术入门教材，为读者搭建起通向“大数据知识空间”的桥梁和纽带，以“构建知识体系、阐明基本原理、引导初级实践、了解相关应用”为原则，为读者在大数据领域“深耕细作”奠定基础、指明方向。

全书共有13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、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3：中国高校大数据课程公共服务平台介绍**



中国高校大数据课程公共服务平台，由中国高校首个“数字教师”的提出者和建设者——林子雨老师发起，由厦门大学数据库实验室全力打造，由厦门大学云计算与大数据研究中心、海峡云计算与大数据应用研究中心携手共建。这是国内第一个服务于高校大数据课程建设的公共服务平台，旨在促进国内高校大数据课程体系建设，提高大数据课程教学水平，降低大数据课程学习门槛，提升学生课程学习效果。平台服务对象涵盖高校、教师和学生。平台为高校开设大数据课程提供全流程辅助，为教师开展教学工作提供一站式服务，为学生学习大数据课程提供全方位辅导。平台重点打造“9个1工程”，即1本教材（含官网）、1个教师服务站、1个学生服务站、1个公益项目、1堂巡讲公开课、1个示范班级、1门在线课程、1个交流群（QQ群、微信群）和1个保障团队。

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



扫一扫访问平台主页