

《深入理解大数据：大数据处理与编程实践》

图书基本信息

书名：《深入理解大数据：大数据处理与编程实践》

13位ISBN编号：9787111473256

出版时间：2014-8

作者：主编：黄宜华,副主编：苗凯翔

页数：520

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu111.com

《深入理解大数据：大数据处理与编程实践》

内容概要

【内容简介】

本书在总结多年来MapReduce并行处理技术课程教学经验和成果的基础上，与业界著名企业Intel公司的大数据技术和产品开发团队和资深工程师联合，以学术界的教学成果与业界高水平系统研发经验完美结合，在理论联系实际的基础上，在基础理论原理、实际算法设计方法以及业界深度技术三个层面上，精心组织材料编写而成。

全书的主要内容包括：

- 大数据处理技术与Hadoop MapReduce简介
- Hadoop系统的安装和操作管理
- 大数据分布式文件系统HDFS
- Hadoop MapReduce并行编程模型、框架与编程接口
- 分布式数据库HBase
- 分布式数据仓库Hive
- Intel Hadoop系统优化与功能增强
- MapReduce基础算法程序设计
- MapReduce高级程序设计技术
- MapReduce机器学习与数据挖掘基础算法
- 大数据处理算法与应用编程案例

本书中算法设计章节的程序源码可在南京大学PASA大数据实验室（PASA：Parallel Algorithms，Systems，and Applications）网站上下载：

<http://pasa-bigdata.nju.edu.cn/links.html>

Intel Hadoop系统免费试用版下载地址：

<http://www.intel.cn/idh>

本书反馈意见发送邮箱：

feedback_bigdata@163.com。

【编辑推荐】

学术界与业界完美结合的结晶，从原理剖析到系统化算法设计与编程实践
多年来系统性教学实践和成果总结，一系列业界产品增强功能深度技术剖析
一系列大赛获奖算法、优秀课程设计以及来自科研课题及业界应用的实战案例

【媒体推荐】

从计算技术的角度看，大数据处理是一种涉及到几乎所有计算机技术层面的综合性计算技术，涉及到计算机软硬件技术的方方面面。大数据研究和应用已成为产业升级与新产业崛起的重要推动力量。作为国内第一本经过多年课堂教学实践总结而成的大数据并行处理和编程技术书籍，本书全面地介绍了大数据处理相关的基本概念和原理，着重讲述了Hadoop MapReduce大数据处理系统的组成结构、工作原理和编程模型，分析了基于MapReduce的各种大数据并行处理算法和程序设计的思想方法。适合高等院校作为MapReduce大数据并行处理技术课程的教材，同时也很适合作为大数据处理应用开发和编程专业技术人员的参考手册。

我很高兴地看到，该书已纳入了教育部计算机类专业教学指导委员会制定的计算机类专业系统能力培养计划。大数据处理是一门综合性、最能体现计算机系统能力培养的课程。把大数据处理纳入计算机类专业系统能力培养课程体系第三层次的核心课程，作为一门起到一定“收官”作用的综合性课程，这是在计算机系统能力培养方面的一个很好的尝试。

——中国工程院院士、中国计算机学会大数据专家委员会主任 李国杰

作为国内最早从事大数据技术研究和教学的团队之一，南京大学黄宜华教授和他的大数据实验室同仁们在大数据技术领域已经进行了多年系统深入的研究工作，取得了卓有成效的研究成果。英特尔作为一家全球领先的计算技术公司，长期以来始终以计算技术的创新为己任。在大数据处理技术方面，我们也竭尽全力发挥出我们在软硬件平台的组合优势引领大数据技术的全面发展和推广。

这本《深入理解大数据》的力作正是我们双方在大数据领域共同努力的结晶，是以学术界和业界完美

《深入理解大数据：大数据处理与编程实践》

结合的方式，在融合了学术界系统化的研究教学工作和业界深度的系统和应用研发工作基础上，成功打造出的一本大数据技术佳作。相信这是一本适合软件技术人员和 IT 行业管理人员理解和掌握大数据技术的不可多得的技术书籍，也是一本适合于在校大学生和研究生学习和掌握大数据处理和编程技术的好教材。

—— 英特尔亚太研发有限公司总经理 何京翔

《深入理解大数据：大数据处理与编程实践》

作者简介

黄宜华博士，南京大学计算机科学与技术系教授、PASA大数据实验室学术带头人。中国计算机学会大数据专家委员会委员、副秘书长，江苏省计算机学会大数据专家委员会主任。于1983、1986和1997年获得南京大学计算机专业学士、硕士和博士学位。主要研究方向为大数据并行处理、云计算以及Web信息挖掘等，发表学术论文60多篇。2010年在Google公司资助下在本校创建并开设了“MapReduce大数据并行处理技术”课程，成为全国最早开设该课程的院校之一。因在该课程教学和人才培养方面的出色成绩获得2012年Google奖教金。目前正在开展系统化的大数据并行处理技术研究工作，主持国家和省部级科研项目以及与美国Intel公司等业界的合作研究项目多项。

苗凯翔 (Kai X. Miao) 博士，英特尔中国大数据首席技术官，中国计算机学会大数据专家委员会委员。曾担任英特尔中国区系统集成部总监、信息技术研究部门亚洲地区总监、英特尔北美地区解决方案首席架构师。于2009荣获英特尔公司首席工程师职称。在加入英特尔以前，曾在美国Rutgers与 DeVry大学任教。获得北方交通大学（北京）通信学士学位、美国辛辛那提大学电机工程硕士和博士学位。发表期刊和会议研究论文多篇，并拥有21项美国专利，在各种会议上发表过上百次主题演讲，曾参与IETF、ITU和MIT CFP等工业标准的制定，并于2006年担任IEEE通信杂志的联合编辑。

书籍目录

推荐序一	
推荐序二	
推荐序三	
丛书序言	
前 言	
第一部分 Hadoop系统	
第1章 大数据处理技术简介 2	
1.1 并行计算技术简介 2	
1.1.1 并行计算的基本概念 2	
1.1.2 并行计算技术的分类 6	
1.1.3 并行计算的主要技术问题 10	
1.2 大数据处理技术简介 13	
1.2.1 大数据的发展背景和研究意义 13	
1.2.2 大数据的技术特点 16	
1.2.3 大数据研究的主要目标、基本原则和基本途径 17	
1.2.4 大数据计算模式和系统 18	
1.2.5 大数据计算模式的发展趋势 21	
1.2.6 大数据的主要技术层面和技术内容 22	
1.3 MapReduce并行计算技术简介 25	
1.3.1 MapReduce的基本概念和由来 25	
1.3.2 MapReduce的基本设计思想 26	
1.3.3 MapReduce的主要功能和技术特征 28	
1.4 Hadoop系统简介 30	
1.4.1 Hadoop的概述与发展历史 30	
1.4.2 Hadoop系统分布式存储与并行计算构架 31	
1.4.3 Hadoop平台的基本组成与生态系统 33	
1.4.4 Hadoop的应用现状和发展趋势 37	
第2章 Hadoop系统的安装与操作管理 39	
2.1 Hadoop系统安装方法简介 39	
2.2 单机和单机伪分布式Hadoop系统安装基本步骤 39	
2.2.1 安装和配置JDK 40	
2.2.2 创建Hadoop用户 40	
2.2.3 下载安装Hadoop 40	
2.2.4 配置SSH 41	
2.2.5 配置Hadoop环境 42	
2.2.6 Hadoop的运行 43	
2.2.7 运行测试程序 43	
2.2.8 查看集群状态 44	
2.3 集群分布式Hadoop系统安装基本步骤 44	
2.3.1 安装和配置JDK 44	
2.3.2 创建Hadoop用户 45	
2.3.3 下载安装Hadoop 45	
2.3.4 配置SSH 45	
2.3.5 配置Hadoop环境 46	
2.3.6 Hadoop的运行 48	
2.3.7 运行测试程序 48	
2.3.8 查看集群状态 49	

- 2.4 Hadoop MapReduce程序开发过程 49
- 2.5 集群远程作业提交与执行 53
 - 2.5.1 集群远程作业提交和执行过程 53
 - 2.5.2 查看作业执行结果和集群状态 53
- 第3章 大数据存储——分布式文件系统HDFS 56
 - 3.1 HDFS的基本特征与构架 56
 - 3.1.1 HDFS的基本特征 57
 - 3.1.2 HDFS的基本框架与工作过程 57
 - 3.2 HDFS可靠性设计 60
 - 3.2.1 HDFS数据块多副本存储设计 60
 - 3.2.2 HDFS可靠性的设计实现 61
 - 3.3 HDFS文件存储组织与读写 63
 - 3.3.1 文件数据的存储组织 63
 - 3.3.2 数据的读写过程 65
 - 3.4 HDFS文件系统操作命令 68
 - 3.4.1 HDFS启动与关闭 68
 - 3.4.2 HDFS文件操作命令格式与注意事项 69
 - 3.4.3 HDFS文件操作命令 69
 - 3.4.4 高级操作命令和工具 77
 - 3.5 HDFS基本编程接口与示例 83
 - 3.5.1 HDFS编程基础知识 83
 - 3.5.2 HDFS基本文件操作API 84
 - 3.5.3 HDFS基本编程实例 87
- 第4章 Hadoop MapReduce并行编程框架 91
 - 4.1 MapReduce基本编程模型和框架 91
 - 4.1.1 MapReduce并行编程抽象模型 91
 - 4.1.2 MapReduce的完整编程模型和框架 93
 - 4.2 Hadoop MapReduce基本构架与工作过程 96
 - 4.2.1 Hadoop系统构架和MapReduce程序执行过程 96
 - 4.2.2 Hadoop MapReduce执行框架和作业执行流程 98
 - 4.2.3 Hadoop MapReduce作业调度过程和调度方法 102
 - 4.2.4 MapReduce执行框架的组件和执行流程 106
 - 4.3 Hadoop MapReduce主要组件与编程接口 107
 - 4.3.1 数据输入格式InputFormat 107
 - 4.3.2 输入数据分块InputSplit 109
 - 4.3.3 数据记录读入RecordReader 110
 - 4.3.4 Mapper类 112
 - 4.3.5 Combiner 114
 - 4.3.6 Partitioner 115
 - 4.3.7 Sort 116
 - 4.3.8 Reducer类 119
 - 4.3.9 数据输出格式OutputFormat 120
 - 4.3.10 数据记录输出RecordWriter 122
- 第5章 分布式数据库HBase 123
 - 5.1 HBase简介 123
 - 5.1.1 为什么需要NoSQL数据库 123
 - 5.1.2 HBase的作用和功能特点 125
 - 5.2 HBase的数据模型 126
 - 5.2.1 HBase的基本数据模型 126

- 5.2.2 HBase的查询模式 128
- 5.2.3 HBase表设计 129
- 5.3 HBase的基本构架与数据存储管理方法 132
 - 5.3.1 HBase在Hadoop生态中的位置和关系 132
 - 5.3.2 HBase的基本组成结构 133
 - 5.3.3 HBase Region 133
 - 5.3.4 Region Server 135
 - 5.3.5 HBase的总体组成结构 138
 - 5.3.6 HBase的寻址和定位 139
 - 5.3.7 HBase节点的上下线管理 142
- 5.4 HBase安装与操作 145
 - 5.4.1 安装一个单机版的HBase 145
 - 5.4.2 HBase Shell操作命令 146
 - 5.4.3 基于集群的HBase安装和配置 149
- 5.5 HBase的编程接口和编程示例 152
 - 5.5.1 表创建编程接口与示例 152
 - 5.5.2 表数据更新编程接口与示例 153
 - 5.5.3 数据读取编程接口与示例 155
 - 5.5.4 HBase MapReduce支持和编程示例 157
- 5.6 HBase的读写操作和特性 161
 - 5.6.1 HBase的数据写入 161
 - 5.6.2 HBase的数据读取 171
- 5.7 其他HBase功能 173
 - 5.7.1 Coprocessor 173
 - 5.7.2 批量数据导入Bulk Load 176
- 第6章 分布式数据仓库Hive 179
 - 6.1 Hive的作用与结构组成 179
 - 6.2 Hive的数据模型 181
 - 6.2.1 Hive的数据存储模型 181
 - 6.2.2 Hive的元数据存储管理 182
 - 6.2.3 Hive的数据类型 183
 - 6.3 Hive的安装 184
 - 6.3.1 下载Hive安装包 184
 - 6.3.2 配置环境变量 184
 - 6.3.3 创建Hive数据文件目录 185
 - 6.3.4 修改Hive配置文件 185
 - 6.4 Hive查询语言——HiveQL 188
 - 6.4.1 DDL语句 188
 - 6.4.2 DML语句 189
 - 6.4.3 SELECT查询语句 190
 - 6.4.4 数据表操作语句示例 190
 - 6.4.5 分区的使用 192
 - 6.4.6 桶的使用 193
 - 6.4.7 子查询 194
 - 6.4.8 Hive的优化和高级功能 194
 - 6.5 Hive JDBC编程接口与程序设计 196
- 第7章 Intel Hadoop系统优化与功能增强 200
 - 7.1 Intel Hadoop系统简介 200
 - 7.1.1 Intel Hadoop系统的主要优化和增强功能 200

- 7.1.2 Intel Hadoop的系统构成与组件 201
- 7.2 Intel Hadoop系统的安装和管理 202
- 7.3 Intel Hadoop HDFS的优化和功能扩展 202
 - 7.3.1 HDFS的高可用性 203
 - 7.3.2 Intel Hadoop系统高可用性配置服务 204
 - 7.3.3 Intel Hadoop系统高可用性配置服务操作 206
 - 7.3.4 自适应数据块副本调整策略 208
- 7.4 Intel Hadoop HBase的功能扩展和编程示例 211
 - 7.4.1 HBase大对象存储（LOB） 211
 - 7.4.2 加盐表 212
 - 7.4.3 HBase跨数据中心大表 213
- 7.5 Intel Hadoop Hive的功能扩展和编程示例 216
 - 7.5.1 开源Hive的不足 216
 - 7.5.2 Intel Hadoop “Hive over HBase” 优化设计 216
 - 7.5.3 Hive over HBase的架构 216
- 第二部分 MapReduce的编程和算法设计
- 第8章 MapReduce基础算法程序设计 220
 - 8.1 WordCount 220
 - 8.1.1 WordCount算法编程实现 220
 - 8.2 矩阵乘法 223
 - 8.2.1 矩阵乘法原理和实现思路 223
 - 8.2.2 矩阵乘法的MapReduce程序实现 224
 - 8.3 关系代数运算 227
 - 8.3.1 选择操作 227
 - 8.3.2 投影操作 228
 - 8.3.3 交运算 229
 - 8.3.4 差运算 230
 - 8.3.5 自然连接 231
 - 8.4 单词共现算法 233
 - 8.4.1 单词共现算法的基本设计 233
 - 8.4.2 单词共现算法的实现 234
 - 8.4.3 单词共现算法实现中的细节问题 235
 - 8.5 文档倒排索引 237
 - 8.5.1 简单的文档倒排索引 237
 - 8.5.2 带词频等属性的文档倒排索引 239
 - 8.6 PageRank网页排名算法 242
 - 8.6.1 PageRank的简化模型 243
 - 8.6.2 PageRank的随机浏览模型 244
 - 8.6.3 PageRank的MapReduce实现 245
 - 8.7 专利文献分析算法 249
 - 8.7.1 构建专利被引用列表 250
 - 8.7.2 专利被引用次数统计 251
 - 8.7.3 专利被引用次数直方图统计 252
 - 8.7.4 按照年份或国家统计专利数 254
- 第9章 MapReduce高级程序设计技术 256
 - 9.1 简介 256
 - 9.2 复合键值对的使用 257
 - 9.2.1 把小的键值对合并成大的键值对 257
 - 9.2.2 巧用复合键让系统完成排序 259

- 9.3 用户定制数据类型 262
 - 9.3.1 Hadoop内置的数据类型 263
 - 9.3.2 用户自定义数据类型的实现 263
- 9.4 用户定制数据输入输出格式 264
 - 9.4.1 Hadoop内置的数据输入格式与RecordReader 265
 - 9.4.2 用户定制数据输入格式与RecordReader 265
 - 9.4.3 Hadoop内置的数据输出格式与RecordWriter 269
 - 9.4.4 用户定制数据输出格式与RecordWriter 269
 - 9.4.5 通过定制数据输出格式实现多集合文件输出 270
- 9.5 用户定制Partitioner和Combiner 271
 - 9.5.1 用户定制Partitioner 272
 - 9.5.2 用户定制Combiner 273
- 9.6 组合式MapReduce计算作业 274
 - 9.6.1 迭代MapReduce计算任务 274
 - 9.6.2 顺序组合式MapReduce作业的执行 275
 - 9.6.3 具有复杂依赖关系的组合式MapReduce作业的执行 275
 - 9.6.4 MapReduce前处理和后处理步骤的链式执行 276
- 9.7 多数据源的连接 278
 - 9.7.1 基本问题数据示例 279
 - 9.7.2 用DataJoin类实现Reduce端连接 279
 - 9.7.3 用全局文件复制方法实现Map端连接 285
 - 9.7.4 带Map端过滤的Reduce端连接 287
 - 9.7.5 多数据源连接解决方法的限制 288
- 9.8 全局参数/数据文件的传递与使用 288
 - 9.8.1 全局作业参数的传递 288
 - 9.8.2 查询全局的MapReduce作业属性 290
 - 9.8.3 全局数据文件的传递 291
- 9.9 关系数据库的连接与访问 292
 - 9.9.1 从数据库中输入数据 292
 - 9.9.2 向数据库中输出计算结果 292
- 第10章 MapReduce数据挖掘基础算法 295
 - 10.1 K-Means聚类算法 295
 - 10.1.1 K-Means聚类算法简介 295
 - 10.1.2 基于MapReduce的K-Means算法的设计实现 297
 - 10.2 KNN最近邻分类算法 300
 - 10.2.1 KNN最近邻分类算法简介 300
 - 10.2.2 基于MapReduce的KNN算法的设计实现 301
 - 10.3 朴素贝叶斯分类算法 303
 - 10.3.1 朴素贝叶斯分类算法简介 303
 - 10.3.2 朴素贝叶斯分类并行化算法的设计 304
 - 10.3.3 朴素贝叶斯分类并行化算法的实现 306
 - 10.4 决策树分类算法 310
 - 10.4.1 决策树分类算法简介 310
 - 10.4.2 决策树并行化算法的设计 313
 - 10.4.3 决策树并行化算法的实现 317
 - 10.5 频繁项集挖掘算法 327
 - 10.5.1 频繁项集挖掘问题描述 327
 - 10.5.2 Apriori频繁项集挖掘算法简介 328
 - 10.5.3 Apriori频繁项集挖掘并行化算法的设计 329

- 10.5.4 Apriori频繁项集挖掘并行化算法的实现 331
- 10.5.5 基于子集求取的频繁项集挖掘算法的设计 335
- 10.5.6 基于子集求取的频繁项集挖掘并行化算法的实现 336
- 10.6 隐马尔科夫模型和最大期望算法 340
 - 10.6.1 隐马尔科夫模型的基本描述 340
 - 10.6.2 隐马尔科夫模型问题的解决方法 341
 - 10.6.3 最大期望算法概述 345
 - 10.6.4 并行化隐马尔科夫算法设计 345
 - 10.6.5 隐马尔科夫算法的并行化实现 348
- 第11章 大数据处理算法设计与应用编程案例 352
 - 11.1 基于MapReduce的搜索引擎算法 352
 - 11.1.1 搜索引擎工作原理简介 353
 - 11.1.2 基于MapReduce的文档预处理 354
 - 11.1.3 基于MapReduce的文档倒排索引构建 356
 - 11.1.4 建立Web信息查询服务 363
 - 11.2 基于MapReduce的大规模短文本多分类算法 365
 - 11.2.1 短文本多分类算法工作原理简介 365
 - 11.2.2 并行化分类训练算法设计实现 366
 - 11.2.3 并行化分类预测算法设计实现 369
 - 11.3 基于MapReduce的大规模基因序列比对算法 371
 - 11.3.1 基因序列比对算法简介 371
 - 11.3.2 并行化BLAST算法的设计与实现 373
 - 11.4 基于MapReduce的大规模城市路径规划算法 379
 - 11.4.1 问题背景和要求 379
 - 11.4.2 数据输入 380
 - 11.4.3 程序设计要求 384
 - 11.4.4 算法设计总体框架和处理过程 385
 - 11.4.5 并行化算法的设计与实现 386
 - 11.5 基于MapReduce的大规模重复文档检测算法 396
 - 11.5.1 重复文档检测问题描述 396
 - 11.5.2 重复文档检测方法和算法设计 397
 - 11.5.3 重复文档检测并行化算法设计实现 401
 - 11.6 基于内容的并行化图像检索算法与引擎 404
 - 11.6.1 基于内容的图像检索问题概述 404
 - 11.6.2 图像检索方法和算法设计思路 405
 - 11.6.3 并行化图像检索算法实现 407
 - 11.7 基于MapReduce的大规模微博传播分析 412
 - 11.7.1 微博分析问题背景与并行化处理过程 413
 - 11.7.2 并行化微博数据获取算法的设计实现 414
 - 11.7.3 并行化微博数据分析算法的设计实现 416
 - 11.8 基于关联规则挖掘的图书推荐算法 422
 - 11.8.1 图书推荐和关联规则挖掘简介 422
 - 11.8.2 图书频繁项集挖掘算法设计与数据获取 423
 - 11.8.3 图书关联规则挖掘并行化算法实现 425
 - 11.9 基于Hadoop的城市智能交通综合应用案例 432
 - 11.9.1 应用案例概述 432
 - 11.9.2 案例一：交通事件检测 433
 - 11.9.3 案例二：交通流统计分析功能 435
 - 11.9.4 案例三：道路旅行时间分析 435

- 11.9.5 案例四：HBase实时查询 436
- 11.9.6 案例五：HBase Endpoint快速统计 437
- 11.9.7 案例六：利用Hive高速统计 439
- 附 录
- 附录A OpenMP并行程序设计简介 442
- 附录B MPI并行程序设计简介 448
- 附录C 英特尔Apache Hadoop*系统安装手册 457
- 参考文献 486

《深入理解大数据：大数据处理与编程实践》

精彩短评

1、勉强算是读完了吧，以后也懒得碰了。书很烂，要么就是介绍安装配置，要么就是贴代码。并没有深入介绍Hadoop的系统设计（介绍的也都是官方文档上的内容），不利于读者理解。另外，Hadoop一代已经过时了啊，Intel的Hadoop发行版已经放弃维护了吧？

2、推荐研究生看看

《深入理解大数据：大数据处理与编程实践》

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu111.com