

大数据创新人才培养系列

大数据技术 原理与应用

第2版

概念、存储、处理、分析与应用

PRINCIPLES AND APPLICATIONS OF
BIG DATA TECHNOLOGY (2ND)

◎ 林子雨 编著

搭建起通向“大数据知识空间”的桥梁和纽带
构建知识体系、阐明基本原理、引导初级实践、介绍相关应用
为读者在大数据领域“深耕细作”奠定基础、指明方向

 中国工信出版集团

 人民邮电出版社
POSTS & TELECOM PRESS

大数据创新人才培养系列

大数据技术 原理与应用

第2版

概念、存储、处理、分析与应用

PRINCIPLES AND APPLICATIONS OF
BIG DATA TECHNOLOGY (2ND)

◎ 林子雨 编著

本书第1版出版后，广受好评，得到了大量的读者来信，对本书提出了许多宝贵的改进意见和建议，深感荣幸和感谢。同时，笔者还参加了多期全国高校大数据课程教师培训交流班和全国高校大数据高峰论坛，开展了全国高校大数据公开课巡讲计划与各大高校开展深度合作，进一步了解了当前国内高校

人民邮电出版社

北京

图书在版编目(CIP)数据

大数据技术原理与应用：概念、存储、处理、分析与应用 / 林子雨编著. — 2版. — 北京：人民邮电出版社，2017.1

(大数据创新人才培养系列)

ISBN 978-7-115-44330-4

I. ①大… II. ①林… III. ①数据处理 IV. ①TP274

中国版本图书馆CIP数据核字(2016)第313809号

内 容 提 要

本书系统介绍了大数据的相关知识，分为大数据基础篇、大数据存储与管理篇、大数据处理与分析篇、大数据应用篇。全书共15章，内容包含大数据的基本概念、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、MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学领域和其他行业的应用。本书在Hadoop、HDFS、HBase、MapReduce和Spark等重要章节安排了入门级的实践操作，以便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、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。

-
- ◆ 编 著 林子雨
 - 责任编辑 吴 婷
 - 责任印制 沈 蓉 彭志环
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
 邮编 100164 电子邮件 315@ptpress.com.cn
 网址 <http://www.ptpress.com.cn>
 固安县铭成印刷有限公司印刷
 - ◆ 开本：787×1092 1/16
 印张：18.75 2017年1月第2版
 字数：487千字 2017年1月河北第1次印刷
-

定价：49.80元

读者服务热线：(010)81055256 印装质量热线：(010)81055316

反盗版热线：(010)81055315

前言 (第2版)

《大数据技术原理与应用》第1版于2015年8月出版,虽然距今仅有一年左右的时间,但是在过去一年里,大数据技术发展迅猛,诸如 Spark 等新技术迅速崛起,开始改变 Hadoop 一枝独秀的市场格局。因此,我们及时对第1版内容进行了补充和修订,以适应大数据技术的快速发展,保持本书的先进性和实用性。

本书依然沿用第1版的篇章设计,共分四大部分,包括大数据基础篇、大数据存储与管理篇、大数据处理与分析篇和大数据应用篇。在大数据基础篇中,第1章介绍大数据的基本概念和应用领域,并阐述大数据、云计算和物联网的相互关系;第2章介绍大数据处理架构 Hadoop,并补充介绍了 Hadoop 版本演化。在大数据存储与管理篇中,第3章介绍了分布式文件系统 HDFS,在编程实践部分根据最新版本的 API 进行了修订;第4章介绍了分布式数据库 HBase,在编程实践部分根据最新版本的 API 进行了修订;第5章介绍了 NoSQL 数据库;第6章介绍了云数据库。在大数据处理与分析篇中,首先在第7章介绍了分布式并行编程模型 MapReduce,然后在新增的第8章中对 Hadoop 进行了再探讨,介绍了 Hadoop 的发展演化和一些新特性,并在新增的第9章中介绍了当前比较热门的、基于内存的分布式计算框架 Spark,在第10章和第11章分别介绍了两种典型的大数据分析技术——流计算和图计算,最后在第12章简单介绍了可视化技术。在大数据应用篇中,用3章(第13章~第15章)内容介绍了大数据在互联网、生物医学领域和其他行业的典型应用。

本书第1版于2015年8月出版后,厦门大学数据库实验室建设了与本书配套的“中国高校大数据课程公共服务平台”(<http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/>),为教师教学和学生学大数据课程提供 PPT 讲义、学习指南、备课指南、上机习题、实验指南、技术资料、授课视频等全方位、一站式免费服务,并提供面向全国高校的大数据实验平台建设方案和大数据课程师资培训服务。

本书是厦门大学计算机科学系大数据课程的配套教材,根据近几年的教学实践,建议安排 32 学时理论课,16 个教学周,每周 2 学时。每章的具体学时分配如下:第1、3、4、5、6、8、10、11、12、13 章每章安排 2 学时;第2、7、9 章每章安排 4 学时;第14、15 章这两章内容由学生自学完成。已经建设大数据教学实验室的高校,可以增加 16 学时上机实践课,分成 4 次上机,每次连续 4 节课,“中国高校大数据课程公共服务平台”的“教师服务站”为本书提供了配套的上机实验指南。

本书第1版出版后,笔者收到了大量的读者来信,对本书提出了许多宝贵的改进意见和建议,这里表示衷心的感谢。同时,笔者举办了多期全国高校大数据课程教师培训交流班和全国高校大数据教学论坛,开展了全国高校大数据公开课巡讲计划与辅助国内高校开设大数据课程公益项目,建立了大数据课程教师交流群,与全国高校大数据课程教师进行了广泛的接触、沟通和交流,更好地了解了当前国内高校

大数据课程教学发展需求和前进方向,这也为本书第2版撰写奠定了很好的基础。这里向参与交流的全国高校大数据课程教师表示衷心的感谢!

本书由林子雨执笔。在撰写第2版过程中,厦门大学计算机科学系硕士研究生蔡珉星、李雨倩、谢荣东、罗道文、邓少军、阮榕城、薛倩、魏亮、曾冠华等做了大量辅助性工作,在此,向他们的辛勤工作表示衷心的感谢。

大数据技术发展日新月异,在今后的工作中,笔者以及厦门大学数据库实验室将持续跟踪大数据技术发展趋势,把大数据最新技术和本书相关补充资料及时发布到“中国高校大数据课程公共服务平台”,方便本书读者通过网络及时免费获取相关信息。由于笔者能力有限,书中难免存在不足之处,望广大读者不吝赐教。

林子雨

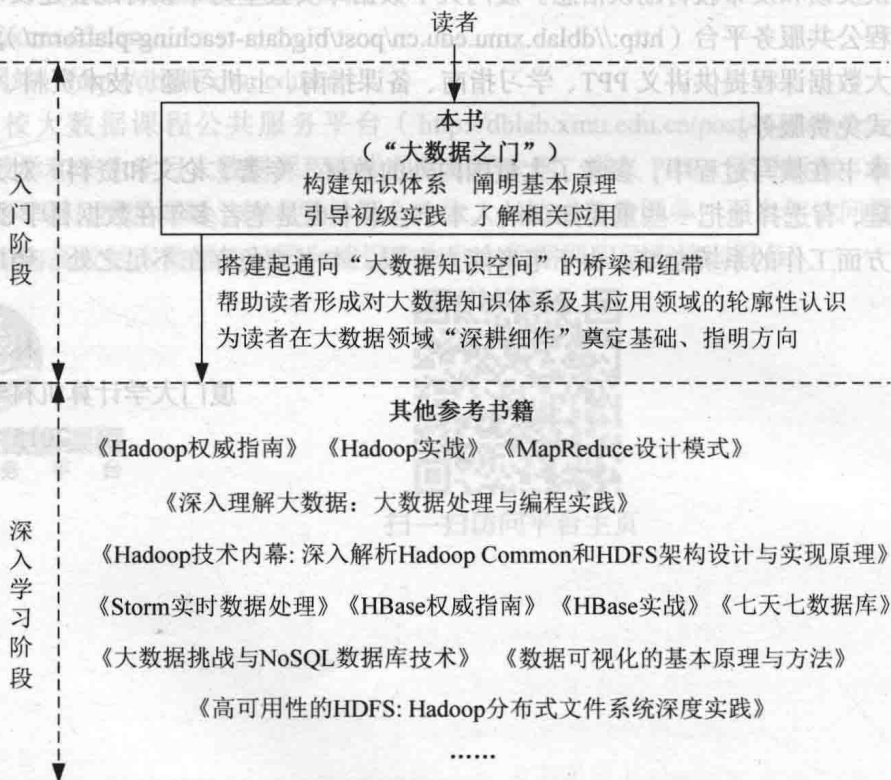
厦门大学计算机科学系数据库实验室

2016年9月

前言 (第1版)

大数据作为继云计算、物联网之后 IT 行业又一颠覆性的技术, 备受人们关注。大数据无处不在, 包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业, 都融入了大数据的印迹, 大数据对人类的社会生产和生活必将产生重大而深远的影响。

大数据时代的到来, 迫切需要高校及时建立大数据技术课程体系, 为社会培养和输送一大批具备大数据专业素养的高级人才, 满足社会对大数据人才日益旺盛的需求。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。本书将系统梳理、总结大数据相关技术, 介绍大数据技术的基本原理和大数据的主要应用, 帮助读者形成对大数据知识体系及其应用领域的轮廓性认识, 为读者在大数据领域“深耕细作”奠定基础、指明方向。在本书的基础上, 感兴趣的读者可以通过其他诸如《Hadoop 权威指南》等工具书, 继续深入学习和实践大数据相关技术。



本书紧紧围绕“构建知识体系, 阐明基本原理, 引导初级实践, 了解相关应用”的指导思想, 对大数据知识体系进行系统梳理, 做到“有序组织、去粗取精、由浅入深、渐次展开”。本书共分四大部分, 包括大数据基础篇、大数据存储篇、大数据处理与分析篇和大数据应用篇。在大数据基础篇中, 第1章介绍大数据的基本概念

和应用领域，并阐述大数据、云计算和物联网的相互关系；第2章介绍大数据处理架构 Hadoop，由于 Hadoop 已经成为应用最广泛的大数据技术，因此，本书的大数据相关技术主要围绕 Hadoop 展开，包括 Hadoop MapReduce、HDFS 和 HBase，本章是第3、4、7章的基础。在大数据存储篇中，用4章（第3~6章）的内容介绍了大数据存储相关技术的概念与原理，包括分布式文件系统（HDFS）、分布式数据库（HBase）、NoSQL 数据库和云数据库。在大数据处理与分析篇，首先在第7章介绍了大数据处理和分析的核心技术——分布式并行编程模型 MapReduce，然后，在第8章和第9章分别介绍了大数据时代两种新兴的数据分析技术——流计算和图计算，最后在第10章简单介绍了可视化技术。在大数据应用篇，用3章（第11章~第13章）内容介绍了大数据在互联网、生物医学领域和其他行业的典型应用。

本书面向高校计算机和信息管理等相关专业的学生，可以作为专业必修课或选修课的教材。在教学过程中，建议安排32学时，16个教学周，每周2学时。每章的具体学时分配如下：第1、2、5、6、8、10、11章每章安排2学时；第3、4、9章每章安排4学时；第7章安排6学时；第12、13章这两章内容由学生自学完成。

本书由林子雨执笔。在撰写过程中，厦门大学计算机科学系硕士研究生刘颖杰、叶林宝、蔡珉星、李雨倩、谢荣东、罗道文以及本科生黄梓铭、李燦等做了大量辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。

本书官方网站是 <http://dblab.xmu.edu.cn/post/bigdata>，提供教学 PPT 和相关资料的下载，并接受错误反馈和发布教材勘误信息。厦门大学数据库实验室为本教材配套建设了国内高校首个大数据课程公共服务平台（<http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/>），为教师教学和学生学习大数据课程提供讲义 PPT、学习指南、备课指南、上机习题、技术资料、授课视频等全方位、一站式免费服务。

本书在撰写过程中，参考了大量国内外的教材、专著、论文和资料，对大数据知识进行了系统梳理，有选择地把一些重要知识纳入本书。本书也是笔者多年在数据科学领域从事教学、科研、产业方面工作的系统总结。由于笔者能力有限，本书难免存在不足之处，望广大读者不吝赐教。

林子雨

厦门大学计算机科学系数据库实验室

2015年3月

作者简介



林子雨 (1978 -), 男, 博士, 厦门大学计算机科学系助理教授, 厦门大学云计算与大数据研究中心创始成员, 海峡云计算与大数据应用研究中心副主任。于 2001 年获得福州大学水利水电专业学士学位, 2005 年获得厦门大学计算机专业硕士学位, 2009 年获得北京大学计算机专业博士学位。中国高校首个“数字教师”提出者和建设者 (<http://www.cs.xmu.edu.cn/linziyu>), 2009 年至今, “数字教师”大平台累计向网络免费发布超过 100 万字高价值的教学和科研资料, 累计网络访问量超过 100 万次。

主要研究方向为数据库、数据仓库、数据挖掘、大数据和云计算, 发表期刊和会议学术论文多篇, 并作为课题组负责人承担了国家自然科学基金和福建省自然科学基金项目。曾作为志愿者翻译了 Google Spanner、BigTable 和《Architecture of a Database System》等大量英文学术资料, 与广大网友分享, 深受欢迎。2013 年在厦门大学开设大数据课程, 并因在教学领域的突出贡献和受到学生的认可, 成为 2013 年度厦门大学教学类奖教金获得者。

主讲课程:《大数据处理技术》。

个人主页: <http://www.cs.xmu.edu.cn/linziyu>。

E-mail: ziyulin@xmu.edu.cn。

数据库实验室网站: <http://dblab.xmu.edu.cn>。

建设了中国高校大数据课程公共服务平台 (<http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/>), 为教师教学和学生学习大数据课程提供包括教学大纲、讲义 PPT、学习指南、备课指南、实验指南、上机习题、授课视频、技术资料等全方位、一站式免费服务, 平台年访问量超过 50 万次, 同时提供面向高校的大数据实验平台建设方案和大数据课程师资培训服务。



中国高校大数据课程
公共服务平台



扫一扫访问平台主页

1.2 大数据对传统方式的影响	11
1.3 大数据对传统应用的影响	11
1.3.1 大数据对传统应用的影响	11
1.3.2 大数据对传统应用的影响	11
1.3.3 大数据对传统应用的影响	11
1.3.4 大数据对传统应用的影响	12
1.4 大数据的应用	14
1.5 大数据关键技术	14
1.6 大数据计算模式	15
1.6.1 批处理计算	16
1.6.2 流计算	16
1.6.3 图计算	16
1.6.4 查询分析计算	17
1.7 大数据产业	17
1.8 大数据与云计算、物联网	18
1.8.1 云计算	18
1.8.2 物联网	21

2.1 Hadoop 的概述	28
2.1.1 Hadoop 的概述	28
2.1.2 Hadoop 的架构	29
2.1.3 Hadoop 的特性	29
2.1.4 Hadoop 的应用现状	29
2.1.5 Hadoop 的版本	30
2.2 Hadoop 的部署	32
2.2.1 Hadoop 的部署	32
2.2.2 Hadoop 的部署	32
2.2.3 Hadoop 的部署	32
2.2.4 Hadoop 的部署	32
2.2.5 Hadoop 的部署	32
2.2.6 Hadoop 的部署	32
2.2.7 Hadoop 的部署	32
2.2.8 Hadoop 的部署	32
2.2.9 Hadoop 的部署	32
2.2.10 Hadoop 的部署	33
2.3 Hadoop 的安装与使用	33
2.3.1 创建 Hadoop 用户	33
2.3.2 Java 的依赖	34
2.3.3 SSH 登录权限设置	34
2.3.4 安装单机 Hadoop	34
2.3.5 Hadoop 的分布式安装	35
2.4 本章小结	37
2.5 习题	38
实验 1 安装 Hadoop	38

目 录

第一篇 大数据基础	
第1章 大数据概述2	1.8.3 大数据与云计算、物联网的关系.....25
1.1 大数据时代.....2	1.9 本章小结.....26
1.1.1 第三次信息化浪潮.....2	1.10 习题.....26
1.1.2 信息科技为大数据时代提供 技术支持.....3	第2章 大数据处理架构 Hadoop28
1.1.3 数据产生方式的变革促成大数据 时代的来临.....5	2.1 概述.....28
1.1.4 大数据的发展历程.....6	2.1.1 Hadoop 简介.....28
1.2 大数据的概念.....7	2.1.2 Hadoop 的发展简史.....28
1.2.1 数据量大.....7	2.1.3 Hadoop 的特性.....29
1.2.2 数据类型繁多.....8	2.1.4 Hadoop 的应用现状.....29
1.2.3 处理速度快.....9	2.1.5 Hadoop 的版本.....30
1.2.4 价值密度低.....9	2.2 Hadoop 生态系统.....30
1.3 大数据的影响.....9	2.2.1 HDFS.....31
1.3.1 大数据对科学研究的影响.....10	2.2.2 HBase.....31
1.3.2 大数据对思维方式的影响.....11	2.2.3 MapReduce.....31
1.3.3 大数据对社会发展的影响.....11	2.2.4 Hive.....32
1.3.4 大数据对就业市场的影响.....12	2.2.5 Pig.....32
1.3.5 大数据对人才培养的影响.....13	2.2.6 Mahout.....32
1.4 大数据的应用.....14	2.2.7 Zookeeper.....32
1.5 大数据关键技术.....14	2.2.8 Flume.....32
1.6 大数据计算模式.....15	2.2.9 Sqoop.....32
1.6.1 批处理计算.....16	2.2.10 Ambari.....33
1.6.2 流计算.....16	2.3 Hadoop 的安装与使用.....33
1.6.3 图计算.....16	2.3.1 创建 Hadoop 用户.....33
1.6.4 查询分析计算.....17	2.3.2 Java 的安装.....34
1.7 大数据产业.....17	2.3.3 SSH 登录权限设置.....34
1.8 大数据与云计算、物联网.....18	2.3.4 安装单机 Hadoop.....34
1.8.1 云计算.....18	2.3.5 Hadoop 伪分布式安装.....35
1.8.2 物联网.....21	2.4 本章小结.....37
	2.5 习题.....38
	实验1 安装 Hadoop.....38

第二篇 大数据存储与管理

第3章 分布式文件系统 HDFS42	4.2 HBase 访问接口.....65
3.1 分布式文件系统.....42	4.3 HBase 数据模型.....66
3.1.1 计算机集群结构.....42	4.3.1 数据模型概述.....66
3.1.2 分布式文件系统的结构.....43	4.3.2 数据模型的相关概念.....66
3.1.3 分布式文件系统的设计需求.....44	4.3.3 数据坐标.....67
3.2 HDFS 简介.....44	4.3.4 概念视图.....68
3.3 HDFS 的相关概念.....45	4.3.5 物理视图.....69
3.3.1 块.....45	4.3.6 面向列的存储.....69
3.3.2 名称节点和数据节点.....46	4.4 HBase 的实现原理.....71
3.3.3 第二名称节点.....47	4.4.1 HBase 的功能组件.....71
3.4 HDFS 体系结构.....48	4.4.2 表和 Region.....71
3.4.1 概述.....48	4.4.3 Region 的定位.....72
3.4.2 HDFS 命名空间管理.....49	4.5 HBase 运行机制.....74
3.4.3 通信协议.....49	4.5.1 HBase 系统架构.....74
3.4.4 客户端.....50	4.5.2 Region 服务器的工作原理.....76
3.4.5 HDFS 体系结构的局限性.....50	4.5.3 Store 的工作原理.....77
3.5 HDFS 的存储原理.....50	4.5.4 HLog 的工作原理.....77
3.5.1 数据的冗余存储.....50	4.6 HBase 编程实践.....78
3.5.2 数据存取策略.....51	4.6.1 HBase 常用的 Shell 命令.....78
3.5.3 数据错误与恢复.....52	4.6.2 HBase 常用的 Java API 及 应用实例.....80
3.6 HDFS 的数据读写过程.....53	4.7 本章小结.....90
3.6.1 读数据的过程.....53	4.8 习题.....90
3.6.2 写数据的过程.....54	实验3 熟悉常用的 HBase 操作.....91
3.7 HDFS 编程实践.....55	第5章 NoSQL 数据库94
3.7.1 HDFS 常用命令.....55	5.1 NoSQL 简介.....94
3.7.2 HDFS 的 Web 界面.....56	5.2 NoSQL 兴起的原因.....95
3.7.3 HDFS 常用 Java API 及应用实例.....57	5.2.1 关系数据库无法满足 Web 2.0 的需求.....95
3.8 本章小结.....60	5.2.2 关系数据库的关键特性在 Web 2.0 时代成为“鸡肋”.....96
3.9 习题.....61	5.3 NoSQL 与关系数据库的比较.....97
实验2 熟悉常用的 HDFS 操作.....61	5.4 NoSQL 的四大类型.....98
第4章 分布式数据库 HBase63	5.4.1 键值数据库.....99
4.1 概述.....63	5.4.2 列族数据库.....100
4.1.1 从 BigTable 说起.....63	5.4.3 文档数据库.....100
4.1.2 HBase 简介.....63	5.4.4 图数据库.....101
4.1.3 HBase 与传统关系数据库的 对比分析.....64	

5.5 NoSQL 的三大基石	101	6.2.2 Amazon 的云数据库产品	113
5.5.1 CAP	101	6.2.3 Google 的云数据库产品	114
5.5.2 BASE	103	6.2.4 微软的云数据库产品	114
5.5.3 最终一致性	104	6.2.5 其他云数据库产品	115
5.6 从 NoSQL 到 NewSQL 数据库	105	6.3 云数据库系统架构	115
5.7 本章小结	107	6.3.1 UMP 系统概述	115
5.8 习题	107	6.3.2 UMP 系统架构	116
第 6 章 云数据库	108	6.3.3 UMP 系统功能	118
6.1 云数据库概述	108	6.4 云数据库实践	121
6.1.1 云计算是云数据库兴起的基础	108	6.4.1 阿里云 RDS 简介	121
6.1.2 云数据库的概念	109	6.4.2 RDS 中的概念	121
6.1.3 云数据库的特性	110	6.4.3 购买和使用 RDS 数据库	122
6.1.4 云数据库是个性化数据存储需求的理想选择	111	6.4.4 将本地数据库迁移到云端 RDS 数据库	126
6.1.5 云数据库与其他数据库的关系	112	6.5 本章小结	127
6.2 云数据库产品	113	6.6 习题	127
6.2.1 云数据库厂商概述	113	实验 4 熟练使用 RDS for MySQL 数据库	128
第三篇 大数据处理与分析			
第 7 章 MapReduce	132	7.4.3 矩阵-向量乘法	144
7.1 概述	132	7.4.4 矩阵乘法	144
7.1.1 分布式并行编程	132	7.5 MapReduce 编程实践	145
7.1.2 MapReduce 模型简介	133	7.5.1 任务要求	145
7.1.3 Map 和 Reduce 函数	133	7.5.2 编写 Map 处理逻辑	146
7.2 MapReduce 的工作流程	134	7.5.3 编写 Reduce 处理逻辑	147
7.2.1 工作流程概述	134	7.5.4 编写 main 方法	147
7.2.2 MapReduce 的各个执行阶段	135	7.5.5 编译打包代码以及运行程序	148
7.2.3 Shuffle 过程详解	136	7.6 本章小结	150
7.3 实例分析: WordCount	139	7.7 习题	151
7.3.1 WordCount 的程序任务	139	实验 5 MapReduce 编程初级实践	152
7.3.2 WordCount 的设计思路	139	第 8 章 Hadoop 再探讨	155
7.3.3 WordCount 的具体执行过程	140	8.1 Hadoop 的优化与发展	155
7.3.4 一个 WordCount 执行过程的实例	141	8.1.1 Hadoop 的局限与不足	155
7.4 MapReduce 的具体应用	142	8.1.2 针对 Hadoop 的改进与提升	156
7.4.1 MapReduce 在关系代数运算中的应用	142	8.2 HDFS2.0 的新特性	156
7.4.2 分组与聚合运算	144	8.2.1 HDFS HA	157
		8.2.2 HDFS 联邦	158
		8.3 新一代资源管理调度框架 YARN	159

8.3.1	MapReduce1.0 的缺陷	159	10.1.1	静态数据和流数据	194
8.3.2	YARN 设计思路	160	10.1.2	批量计算和实时计算	195
8.3.3	YARN 体系结构	161	10.1.3	流计算的概念	196
8.3.4	YARN 工作流程	163	10.1.4	流计算与 Hadoop	196
8.3.5	YARN 框架与 MapReduce1.0 框架的对比分析	164	10.1.5	流计算框架	197
8.3.6	YARN 的发展目标	165	10.2	流计算的处理流程	197
8.4	Hadoop 生态系统中具有代表性的 功能组件	166	10.2.1	概述	197
8.4.1	Pig	166	10.2.2	数据实时采集	198
8.4.2	Tez	167	10.2.3	数据实时计算	198
8.4.3	Kafka	169	10.2.4	实时查询服务	199
8.5	本章小结	170	10.3	流计算的应用	199
8.6	习题	170	10.3.1	应用场景 1: 实时分析	199
第 9 章	Spark	172	10.3.2	应用场景 2: 实时交通	200
9.1	概述	172	10.4	开源流计算框架 Storm	200
9.1.1	Spark 简介	172	10.4.1	Storm 简介	201
9.1.2	Scala 简介	173	10.4.2	Storm 的特点	201
9.1.3	Spark 与 Hadoop 的对比	174	10.4.3	Storm 的设计思想	202
9.2	Spark 生态系统	175	10.4.4	Storm 的框架设计	203
9.3	Spark 运行架构	177	10.4.5	Storm 实例	204
9.3.1	基本概念	177	10.5	Spark Streaming	206
9.3.2	架构设计	177	10.5.1	Spark Streaming 设计	206
9.3.3	Spark 运行基本流程	178	10.5.2	Spark Streaming 与 Storm 的 对比	207
9.3.4	RDD 的设计与运行原理	179	10.6	本章小结	208
9.4	Spark 的部署和应用方式	184	10.7	习题	208
9.4.1	Spark 三种部署方式	184	第 11 章	图计算	210
9.4.2	从“Hadoop+Storm”架构转向 Spark 架构	185	11.1	图计算简介	210
9.4.3	Hadoop 和 Spark 的统一部署	186	11.1.1	传统图计算解决方案的 不足之处	210
9.5	Spark 编程实践	186	11.1.2	图计算通用软件	211
9.5.1	启动 Spark Shell	187	11.2	Pregel 简介	211
9.5.2	Spark RDD 基本操作	187	11.3	Pregel 图计算模型	212
9.5.3	Spark 应用程序	189	11.3.1	有向图和顶点	212
9.6	本章小结	192	11.3.2	顶点之间的消息传递	212
9.7	习题	193	11.3.3	Pregel 的计算过程	213
第 10 章	流计算	194	11.3.4	实例	214
10.1	流计算概述	194	11.4	Pregel 的 C++ API	216
			11.4.1	消息传递机制	217

11.4.2 Combiner	217	第 12 章 数据可视化	230
11.4.3 Aggregator	218	12.1 可视化概述	230
11.4.4 拓扑改变	218	12.1.1 什么是数据可视化	230
11.4.5 输入和输出	218	12.1.2 可视化的发展历程	230
11.5 Pregel 的体系结构	219	12.1.3 可视化的重要作用	231
11.5.1 Pregel 的执行过程	219	12.2 可视化工具	233
11.5.2 容错性	220	12.2.1 入门级工具	233
11.5.3 Worker	221	12.2.2 信息图表工具	234
11.5.4 Master	221	12.2.3 地图工具	235
11.5.5 Aggregator	222	12.2.4 时间线工具	236
11.6 Pregel 的应用实例	222	12.2.5 高级分析工具	236
11.6.1 单源最短路径	222	12.3 可视化典型案例	237
11.6.2 二分匹配	223	12.3.1 全球黑客活动	237
11.7 Pregel 和 MapReduce 实现 PageRank 算法的对比	224	12.3.2 互联网地图	237
11.7.1 PageRank 算法	224	12.3.3 编程语言之间的影响力关系图	238
11.7.2 PageRank 算法在 Pregel 中的 实现	225	12.3.4 百度迁徙	239
11.7.3 PageRank 算法在 MapReduce 中的实现	225	12.3.5 世界国家健康与财富之间的 关系	239
11.7.4 PageRank 算法在 Pregel 和 MapReduce 中实现的比较	228	12.3.6 3D 可视化互联网地图 APP	239
11.8 本章小结	228	12.4 本章小结	240
11.9 习题	228	12.5 习题	240
第四篇 大数据应用			
第 13 章 大数据在互联网领域的 应用	242	对比	248
13.1 推荐系统概述	242	13.3 协同过滤实践	248
13.1.1 什么是推荐系统	242	13.3.1 实践背景	248
13.1.2 长尾理论	243	13.3.2 数据处理	249
13.1.3 推荐方法	243	13.3.3 计算相似度矩阵	249
13.1.4 推荐系统模型	244	13.3.4 计算推荐结果	250
13.1.5 推荐系统的应用	244	13.3.5 展示推荐结果	250
13.2 协同过滤	245	13.4 本章小结	251
13.2.1 基于用户的协同过滤	245	13.5 习题	251
13.2.2 基于物品的协同过滤	246	第 14 章 大数据在生物医学 领域的应用	252
13.2.3 UserCF 算法和 ItemCF 算法的		14.1 流行病预测	252

14.1.1 传统流行病预测机制的不足	252	15.3.2 市场情绪分析	269
14.1.2 基于大数据的流行病预测	253	15.3.3 信贷风险分析	270
14.1.3 基于大数据的流行病预测的 重要作用	253	15.4 大数据在汽车行业中的应用	271
14.1.4 案例: 百度疾病预测	254	15.5 大数据在零售行业中的应用	272
14.2 智慧医疗	255	15.5.1 发现关联购买行为	272
14.3 生物信息学	256	15.5.2 客户群体细分	273
14.4 案例: 基于大数据的综合健康服务 平台	257	15.5.3 供应链管理	273
14.4.1 平台概述	257	15.6 大数据在餐饮行业中的应用	274
14.4.2 平台业务架构	258	15.6.1 餐饮行业拥抱大数据	274
14.4.3 平台技术架构	258	15.6.2 餐饮 O2O	274
14.4.4 平台关键技术	259	15.7 大数据在电信行业中的应用	276
14.5 本章小结	260	15.8 大数据在能源行业中的应用	276
14.6 习题	261	15.9 大数据在体育和娱乐领域中的 应用	277
第 15 章 大数据的其他应用	262	15.9.1 训练球队	277
15.1 大数据在物流领域中的应用	262	15.9.2 投拍影视作品	278
15.1.1 智能物流的概念	262	15.9.3 预测比赛结果	279
15.1.2 智能物流的作用	263	15.10 大数据在安全领域中的应用	280
15.1.3 智能物流的应用	263	15.10.1 大数据与国家安全	280
15.1.4 大数据是智能物流的关键	263	15.10.2 应用大数据技术防御 网络攻击	280
15.1.5 中国智能物流骨干网——菜鸟	264	15.10.3 警察应用大数据工具 预防犯罪	281
15.2 大数据在城市管理中的应用	266	15.11 大数据在政府领域中的应用	282
15.2.1 智能交通	266	15.12 大数据在日常生活中的应用	283
15.2.2 环保监测	267	15.13 本章小结	284
15.2.3 城市规划	268	15.14 习题	284
15.2.4 安防领域	269	参考文献	285
15.3 大数据在金融行业中的应用	269		
15.3.1 高频交易	269		

第一篇

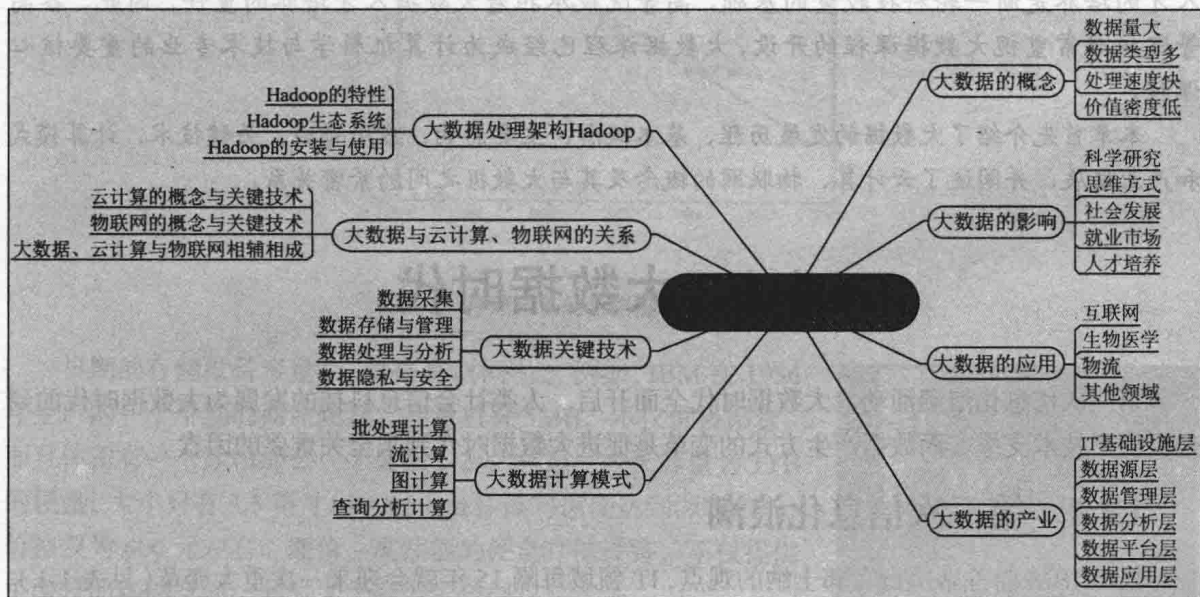
大数据基础

本篇内容

本篇介绍大数据 (Big Data) 的基本概念、影响和应用领域, 并阐述大数据、云计算和物联网的相互关系, 还介绍了大数据处理架构 Hadoop。由于 Hadoop 已经成为应用最广泛的大数据技术, 因此本书的大数据相关技术主要围绕 Hadoop 展开, 包括 Hadoop、MapReduce、HDFS 和 HBase。本篇内容是学习后续内容的基础。

本篇包括 2 章。第 1 章介绍大数据的概念和应用, 分析了大数据、云计算和物联网的相互关系; 第 2 章介绍大数据处理架构 Hadoop。

知识地图



重点与难点

重点为理解大数据的概念, 大数据对科学研究、思维方式和社会发展的影响, 以及大数据处理架构 Hadoop。难点为掌握 Hadoop 的安装与使用方法。

14.1.1 传统流行病预测机制的不足	252	15.3.2 市场情绪分析	269
14.1.2 基于大数据的流行病预测	253	15.3.3 信贷风险分析	270
14.1.3 基于大数据的流行病预测的		15.4 大数据在汽车行业中的应用	271

第1章

大数据概述

大数据时代悄然来临,带来了信息技术发展的巨大变革,并深刻影响着社会生产和人民生活的方方面面。全球范围内,世界各国政府均高度重视大数据技术的研究和产业发展,纷纷把大数据上升为国家战略加以重点推进。企业和学术机构纷纷加大技术、资金和人员投入力度,加强对大数据关键技术的研发与应用,以期在“第三次信息化浪潮”中占得先机、引领市场。大数据已经不是“镜中花、水中月”,它的影响力和作用力正迅速触及社会的每个角落,所到之处,或是颠覆,或是提升,都让人们深切感受到了大数据实实在在的威力。

对于一个国家而言,能否紧紧抓住大数据发展机遇,快速形成核心技术和应用参与新一轮的全球化竞争,将直接决定未来若干年世界范围内各国科技力量博弈的格局。大数据专业人才的培养是新一轮科技较量的基础,高等院校承担着大数据人才培养的重任,因此,各高等院校非常重视大数据课程的开设,大数据课程已经成为计算机科学与技术专业的重要核心课程。

本章首先介绍了大数据的发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展,并阐述了云计算、物联网的概念及其与大数据之间的紧密关系。

1.1 大数据时代

第三次信息化浪潮涌动,大数据时代全面开启。人类社会信息科技的发展为大数据时代的到来提供了技术支撑,而数据产生方式的变革是促进大数据时代到来至关重要的因素。

1.1.1 第三次信息化浪潮

根据 IBM 前首席执行官郭士纳的观点,IT 领域每隔 15 年就会迎来一次重大变革(见表 1-1)。1980 年前后,个人计算机(PC)开始普及,使得计算机走入企业和千家万户,大大提高了社会生产力,也使人类迎来了第一次信息化浪潮,Intel、IBM、苹果、微软、联想等企业是这个时期的标志。随后,在 1995 年前后,人类开始全面进入互联网时代,互联网的普及把世界变成“地球村”,每个人都可以自由徜徉于信息的海洋,由此,人类迎来了第二次信息化浪潮,这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔 15 年,在 2010 年前后,云计算、大数据、物联网的快速发展,拉开了第三次信息化浪潮的大幕,大数据时代已经到来,也必将涌现出一批新的市场标杆企业。

表 1-1 三次信息化浪潮

信息化浪潮	发生时间	标志	解决的问题	代表企业
第一次浪潮	1980 年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995 年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010 年前后	物联网、云计算和大数据	信息爆炸	亚马逊、谷歌、IBM、VMWare、Palantir、Hortonworks、Cloudera、阿里云等

1.1.2 信息科技为大数据时代提供技术支撑

信息科技需要解决信息存储、信息传输和信息处理 3 个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

1. 存储设备容量不断增加

数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中，随着科学技术的不断进步，存储设备的制造工艺不断升级，容量大幅增加，速度不断提升，价格却在不断下降（见图 1-1）。

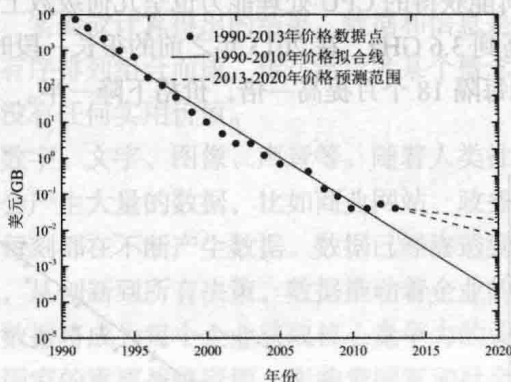


图 1-1 存储设备的价格随时间变化的情况

早期的存储设备容量小、价格高、体积大，例如，IBM 在 1956 年生产的一个早期的商业硬盘，容量只有 5MB，不仅价格昂贵，而且体积有一个冰箱那么大（见图 1-2）。相反，今天容量为 1TB 的硬盘，大小只有 3.5 英寸（约 8.89cm），读写速度达到 200MB/s，价格仅为 400 元左右。廉价、高性能的硬盘存储设备，不仅提供了海量的存储空间，同时大大降低了数据存储成本。

与此同时，以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存是一种新兴的半导体存储器，从 1989 年诞生第一款闪存产品开始，闪存技术不断获得新的突破，并逐渐在计算机存储产品市场中确立了自己的重要地位。闪存是一种非易失性存储器，即使发生断电也不会丢失数据；因此，可以作为永久性存储设备，它具有体积小、质量轻、能耗低、抗振性好等优良特性。

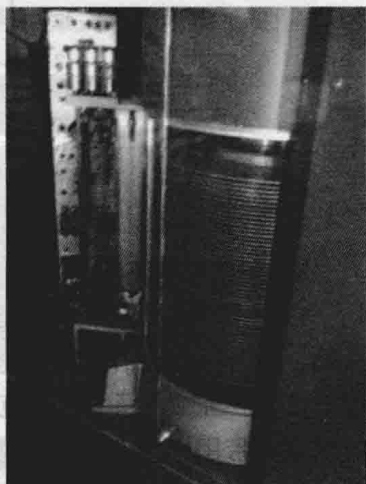


图 1-2 IBM 在 1956 年生产的一个早期的商业硬盘