

大数据导论

关键技术与行业应用最佳实践

INTRODUCTION TO **BIG DATA**

深圳国泰安教育技术股份有限公司大数据事业部群

中科院深圳先进技术研究院—国泰安金融大数据研究中心 编著

揭开大数据的神秘面纱，全面解读大数据领域的
应用现状、原理、热门技术、前沿工具和解决方案。

清华大学出版社



大数据导论

关键技术与行业应用最佳实践

深圳国泰安教育技术股份有限公司大数据事业部群
中科院深圳先进技术研究院—国泰安金融大数据研究中心 编著



清华大学出版社

内容简介

本书全面阐释了大数据的概念、相关的技术和应用的现状,使读者对大数据的相关技术、应用和产业链能有一个比较清晰的认识。

全书共 11 章,主要内容包括大数据概论、数据组织存储技术、NoSQL、Hadoop 和 MapReduce、数据查询和分析高级技术、数据挖掘技术、数据分析语言 R、大数据用于预测和决策、大数据与市场营销、大数据应用案例、大数据应用主流解决方案等。

本书在内容的选择上进行了深入的思考,不论是大数据领域的初学者还是具备一定相关专业知识的读者都能从书中得到一定的收获或启发,同时,本书还适合高等院校的计算机相关专业的本专科生、研究生以及 IT 行业的从业人员,和所有对大数据感兴趣的人士阅读。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。
版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

大数据导论:关键技术与行业应用最佳实践 / 深圳国泰安教育技术股份有限公司大数据事业部群,中科院深圳先进技术研究院——国泰安金融大数据研究中心 编著. —北京:清华大学出版社,2015

ISBN 978-7-302-39271-2

I. ①大… II. ①深… ②中… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 024762 号

责任编辑:杨如林

装帧设计:深圳国泰安教育技术股份有限公司

责任校对:徐俊伟

责任印制:杨艳

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社总机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:河北新华第一印刷有限责任公司

经 销:全国新华书店

开 本:190mm×260mm 印 张:24.25 字 数:590千字

版 次:2015年3月第1版 印 次:2015年3月第1次印刷

印 数:1~4000

定 价:57.00元

编委会

编撰单位

深圳国泰安教育技术股份有限公司大数据事业部群

中科院深圳先进技术研究院——国泰安金融大数据研究中心

主 编

陈工孟 深圳国泰安教育技术股份有限公司董事长，上海交通大学教授、博导

须成忠 中科院深圳先进技术研究院数字所所长，云计算研究中心主任、教授、博导

执行副主编

凌宗平 深圳国泰安教育技术股份有限公司大数据事业部群常务副总经理、中国量化投资研究院助理院长

姜义平 中科院深圳先进技术研究院——国泰安金融大数据研究中心秘书长、深圳国泰安教育技术股份有限公司大数据事业部群副总经理

编写人员

吴德辉 周阳锦 麻海煜 黄子平 陈 淋 宋国平 周 珺 刘 瑶 余 丹

杨子荀 朱 清 胡 强 李泽璇

总序

大数据一词，最早出现于20世纪90年代。随着云计算和物联网的不断发展，大量数据源的出现导致了非结构化和半结构化数据的迅速增长，数据单位也由TB级别跨越到了ZB级别，大量信息源产生的这些数据已远远超越目前人力所能处理的范围，人们在思索如何对这些数据进行管理及使用时，逐渐探索出一个新的领域。

大数据的“大”不仅指其容量，还体现在多样性、处理速度和复杂度等方面。无论人们是否关注过，海量的数据已如决堤之洪流涌入人们的生活，大数据的时代已然到来了。可以目睹的是，大数据的激流已经给个人生活、企业经营乃至国家和社会的全面发展带来了新的机遇与挑战。2012年，世界经济论坛年会的重要议题之一是“大数据、大影响”；美国也开始从开放政府数据、开展关键技术研究 and 推动大数据应用三个方面来布局其大数据产业；2011年以来，中国计算机学会、中国通信学会先后成立了大数据委员会——研究大数据中的科学与工程问题；中国《“十二五”国家战略性新兴产业发展规划》也提出了支持海量数据存储、处理技术的研发与产业化……大数据正以不可抵挡之势席卷全球。

随着大数据技术和市场的快速发展，驾驭大数据的呼声渐涨，蕴含在大数据中的价值使得大数据已经成为IT信息产业中最具潜力的蓝海，这也使得学习及掌握国际前沿的大数据处理工具和解决方案中的核心技术显得十分迫切。从全球角度来看，对大数据的认识、研究和应用还都处于初期阶段，特别是对我国来说，大数据真正落地，还需要一个长期的过程。由于大数据领域的研究和分析方法综合了云计算、数据仓库、统计学、数据挖掘、机器学习、数据可视化等学科知识，因此编写一套系统的大数据技术与应用的丛书，作为学习和掌握大数据相关理论与方法的开端，无疑是一件意义非凡的事。

为了满足国内大数据领域学术界和产业界系统学习和掌握大数据理论及分析方法的迫切需要，同时也为解决国内大数据相关专业，尤其是本科生、研究生教材过于零散不够系统等问题，深圳国泰安教育技术股份有限公司大数据事业部群和中科院深圳先进技术研究院——国泰安金融大数据研究中心合作，组织该领域的专家、学者编写了“大数据技术与应用丛书”。这套丛书包括《大数据导论：关键技术与行业应用最佳实践》《Datawatch在各行业的应用》等。这套丛书是我们结合了大数据发展技术、发展现状，并调查了国内学术界和产业界的实际需求后精心编写的。希望这套丛书的出版，能为中国大数据学术界和产业界吸纳国内外先进的研究理念和研究方法，为国内大数据领域的学术研究、学术发展和实务运作提供

支持与帮助。

编写出版这套丛书是一项长期的工程，从2014年初开始策划编写这套丛书，包括咨询专家、选择书目、组织编写、校对内容、图书出版，从策划者、编写者到校对者和出版者都投入了大量心血和时间。在选择书目时，我们主要考虑所选书目要尽量保证理论体系的完整性，涵盖了大数据领域最新的理论研究和技術操作，内容编排循序渐进，方法详尽且操作步骤简单明了。

由于时间关系，书中难免存在不妥和疏漏之处，敬请读者给予批评指正。

深圳国泰安教育技术股份有限公司董事长、
上海交通大学教授、博导

陈工孟

2015年1月

前言

随着云计算、物联网等技术的不断发展和应用，海量数据在生产经营、商务活动、社交生活等领域不断产生。世界处在信息时代，美国奥巴马政府将大数据提升到国家战略层面，启动了“大数据研究和发展计划”；企业将大数据作为提高自身竞争力的重要手段；最热的IT词汇中“大数据”必有一席之地……数据思维深刻影响着人们的工作和生活，这让人们真切地感受到大数据时代已经到来。

大数据即将带来一场颠覆性的革命，它将推动社会生产取得全面进步，助推政府、金融、医疗、教育、零售、制造业、能源和交通等行业产生根本性的变革。大数据是一个事关国家、社会发展全局的产业，围绕产业链的上下游，大数据将带动智能终端、服务器和信息服务业等产业发展，有效减少社会运行成本，提高社会和经济运行效率。

为了在信息时代立于不败之地，了解大数据相关的知识是必要的。本书全面阐释了大数据的概念、相关技术和应用现状，使读者对大数据的相关技术、应用和产业链有一个比较清晰的认识。本书适合高等院校计算机相关专业的本专科生、研究生、IT行业的从业人员和对大数据感兴趣的人士阅读。本书在内容的选择上进行了深入的研究，使得不论是大数据领域的初学者还是具备一定相关知识的专业人员都能从书中得到一定的收获和启发。

本书共11章，内容涵盖大数据的基本概念、关键技术和行业应用解决方案。对技术感兴趣的读者可以重点阅读第2~7章，这几章全面介绍了数据的存储和分析技术；对大数据应用感兴趣的读者可以优先阅读第8~10章；第11章对那些想了解大数据产业链的读者会很有帮助。

第1章简要介绍大数据的概念、大数据与商业智能的关系和大数据的相关技术及发展趋势；第2章介绍数据组织和存储的关键技术；第3章重点介绍了NoSQL；第4章介绍了Hadoop和MapReduce相关技术；第5章阐明了数据查询和分析技术，对常用的分析工具进行了介绍；第6章对数据挖掘技术的概念、挖掘算法和数据挖掘的发展趋势进行了分析；第7章重点介绍数据分析语言R；第8章论述大数据在预测和决策方面的作用，并阐述了商业和政府决策管理的机遇和挑战；第9章包括大数据与市场营销的联系和大数据时代的营销模式创新等问题；第10章简述了大数据在金融、医疗、互联网和影视等行业的应用案例；第11章主要介绍大数据产业链，介绍了新兴科技企业（如Cloudera、深圳国泰安教育技术股份有限公司等）和传统IT巨头（如IBM等）在大数据领域的主流解决方案。

新思想、新技术的不断涌现，推动着大数据的成熟与应用，也有效地推动着科技和社会

的进步。本书结合了深圳国泰安教育技术股份有限公司的产品和解决方案，为读者呈现了大数据领域的全景图。在成书过程中，我们参考了大量国内外学者的研究成果和业界的产品及解决方案，资料来源列在每章参考文献中，在此对各位学者和专业人士表示敬意和感谢！

由于作者水平和时间有限，书中难免存在疏漏和错误之处，恳请读者批评指正。

编者
2015年1月

目 录

第1章 大数据概论

1.1 什么是大数据	1
1.1.1 大数据的概念	2
1.1.2 大数据的特征	2
1.1.3 大数据的产生	4
1.1.4 数据的量级	5
1.1.5 大数据的数据类型	6
1.1.6 大数据的潜在价值	8
1.1.7 大数据的挑战	8
1.2 大数据与商业智能	9
1.2.1 商业智能的概念	9
1.2.2 商业智能的架构体系	10
1.2.3 商业智能的核心技术	11
1.2.4 商业智能的研究内容和发展方向	13
1.2.5 商业智能与大数据的关系	14
1.2.6 商业智能与大数据的结合应用	15
1.3 大数据相关技术与应用概况	17
1.3.1 大数据的相关技术	17
1.3.2 大数据的应用概况	19
1.4 大数据热点问题与发展趋势介绍	21
1.4.1 大数据的热点问题	21
1.4.2 大数据的发展趋势	23
1.5 练习	25
参考文献	25

第2章 数据组织存储技术

2.1 数据存储概述	27
2.1.1 数据存储介质	27
2.1.2 数据存储模式	28
2.1.3 大数据存储存在的问题	30
2.2 数据存储技术研究现状	32
2.2.1 传统关系型数据库	32
2.2.2 新兴的数据存储系统	33
2.3 海量数据存储的关键技术	36
2.3.1 数据划分	37
2.3.2 数据一致性与可用性	37
2.3.3 负载均衡	38
2.3.4 容错机制	39
2.3.5 虚拟存储技术	40
2.3.6 云存储技术	41
2.4 数据仓库	42
2.4.1 数据仓库的相关概念	42
2.4.2 数据仓库体系结构	50
2.4.3 数据仓库设计与实施	51
2.4.4 数据抽取、转换和装载	54
2.4.5 联机分析处理	57
2.5 练习	64
参考文献	64

第3章 NoSQL

3.1 NoSQL简介	66
3.1.1 什么是NoSQL	66
3.1.2 什么是关系型数据库	68
3.1.3 NoSQL数据库与关系型数据库的比较	68
3.2 NoSQL的三大基石	70
3.2.1 CAP	71
3.2.2 BASE	73
3.2.3 最终一致性	74
3.3 key-value数据库	78

3.3.1	Redis.....	78
3.4	Column-oriented数据库.....	80
3.4.1	Bigtable.....	80
3.4.2	Apache Cassandra.....	81
3.4.3	HBase.....	81
3.5	图存数据库.....	89
3.5.1	Neo4j.....	89
3.6	文档数据库.....	93
3.6.1	CouchDB.....	93
3.6.2	MongoDB.....	95
3.7	NewSQL数据库.....	96
3.7.1	NewSQL数据库简介.....	96
3.7.2	MySQL Cluster.....	97
3.7.3	VoltDB.....	99
3.8	分布式缓存系统.....	100
3.9	练习.....	103
	参考文献.....	103

第4章 Hadoop和MapReduce

4.1	Hadoop简介.....	104
4.2	Hadoop的体系结构.....	105
4.2.1	HDFS的体系结构.....	105
4.2.2	MapReduce的体系结构.....	106
4.2.3	其他组件.....	106
4.2.4	Hadoop的I/O操作.....	107
4.2.5	Hadoop与分布式开发.....	111
4.3	Hadoop的安装与配置.....	112
4.3.1	在Windows上安装与配置Hadoop.....	112
4.3.2	在Linux上安装与配置Hadoop.....	120
4.4	Hadoop应用案例.....	126
4.4.1	Last.fm.....	126
4.4.2	Facebook.....	128
4.5	MapReduce模型概述.....	130
4.5.1	Map和Reduce函数.....	132

4.5.2	MapReduce工作流程	132
4.5.3	并行计算的实现	136
4.6	实例分析：WordCount	138
4.6.1	WordCount设计思路	140
4.6.2	WordCount代码	141
4.6.3	过程解释	144
4.7	练习	146
	参考文献	146

第5章 数据查询和分析的高级技术

5.1	SQL on Hadoop查询技术	148
5.1.1	Hive：基本的查询技术	149
5.1.2	Hive的优化和升级	153
5.1.3	实时交互式SQL查询	155
5.1.4	基于PostgreSQL的SQL on Hadoop	157
5.2	数据分析的方法与技术	158
5.2.1	基本分析方法	159
5.2.2	高级分析方法	164
5.2.3	可视化技术	174
5.3	常用分析工具介绍	179
5.3.1	统计分析工具	179
5.3.2	数据挖掘工具	182
5.3.3	可视化设计工具	185
5.4	练习	188
	参考文献	189

第6章 数据挖掘技术

6.1	数据挖掘简介	190
6.2	关联分析	192
6.2.1	基本概念	193
6.2.2	经典频集算法	194
6.2.3	FP Growth	194
6.2.4	多层关联规则	195
6.2.5	多维关联规则	195

6.3 分类与回归	195
6.3.1 基本概念	196
6.3.2 决策树	197
6.3.3 贝叶斯分类算法	199
6.3.4 人工神经网络	201
6.3.5 支持向量机	204
6.3.6 其他分类方法	206
6.3.7 回归	209
6.4 聚类分析	211
6.4.1 基本概念	211
6.4.2 划分方法	212
6.4.3 层次方法	213
6.4.4 基于密度的方法	215
6.4.5 基于网格的方法	215
6.4.6 基于模型的方法	216
6.4.7 双聚类方法	217
6.5 离群点检测	219
6.5.1 基本概念	219
6.5.2 基于统计的离群点检测	220
6.5.3 基于距离的离群点检测	220
6.5.4 基于偏差的离群点检测	221
6.6 复杂数据类型挖掘	222
6.7 数据挖掘的研究前沿和发展趋势	223
6.7.1 数据挖掘的应用	224
6.7.2 数据挖掘中的隐私问题	225
6.7.3 数据挖掘的发展趋势	225
6.8 练习	227
参考文献	227

第7章 数据分析语言R

7.1 R概述	229
7.1.1 R是什么	229
7.1.2 R的获取与安装	230
7.1.3 R的使用	231
7.1.4 R包	233

7.2	R的数据操作	234
7.2.1	数据结构	234
7.2.2	数据输入	236
7.3	绘图功能简介	240
7.3.1	管理绘图	240
7.3.2	绘图函数	242
7.3.3	绘图参数	244
7.3.4	基本图形	246
7.4	R的初级数据分析	250
7.4.1	描述性统计分析	252
7.4.2	频数表和列联表	255
7.4.3	相关分析	258
7.4.4	t检验	261
7.4.5	回归分析	262
7.4.6	方差分析	268
7.5	R的高级数据分析	271
7.5.1	广义线性模型	271
7.5.2	聚类分析	274
7.5.3	判别分析	276
7.5.4	主成分分析	277
7.5.5	因子分析	279
7.6	R在大数据处理中的应用	284
7.6.1	R处理大数据	284
7.6.2	R与Hadoop交互	286
7.7	练习	287
	参考文献	288

第8章 大数据用于预测和决策

8.1	利用分析技术作决策的发展历史和展望	289
8.1.1	利用分析技术作决策的发展历程	289
8.1.2	大数据决策的展望	291
8.2	统计预测和决策概述	292
8.2.1	统计预测的作用及方法	292
8.2.2	统计决策的概述及方法	294
8.3	大数据预测决策的关键	295

8.4	大数据分析用于商业的预测决策	297
8.4.1	乐购——分析客户消费信息	297
8.4.2	Netflix——了解客户的真正需求	297
8.4.3	哈拉斯——使用客户数据	298
8.4.4	大通银行——决策树方法分析按揭数据	298
8.4.5	好事达——采用高级预测分析技术	299
8.5	大数据时代给政府决策管理带来的机遇与挑战	299
8.5.1	大数据提升政府的决策管理能力	299
8.5.2	大数据浪潮中政府面临的挑战	301
8.5.3	政府以变革来顺应大数据时代	303
8.6	大数据时代的跨界与颠覆	305
8.6.1	大数据时代，颠覆浪潮席卷传统产业	305
8.6.2	大数据时代，全新的投资理念和巨大的投资机会	308
8.7	练习	309
	参考文献	309

第9章 大数据与市场营销

9.1	大数据时代的营销模式创新	311
9.1.1	营销模式的突出优势	311
9.1.2	营销模式的创新之举	313
9.2	大数据时代下的网络化精准营销	315
9.2.1	精准营销概述	315
9.2.2	网络精准营销模式	316
9.3	大数据应用与商业机会	318
9.3.1	车载信息服务数据在汽车保险业中的价值	318
9.3.2	RFID数据在零售制造业中的价值	319
9.3.3	大数据在医疗行业中的价值	319
9.3.4	社交网络数据在电信业及其他行业中的价值	320
9.3.5	遥测数据在视频游戏中的价值	321
9.4	大数据时代的商业变革	321
9.4.1	大数据时代商业思维的变革	322
9.4.2	大数据时代管理的变革	323
9.4.3	大数据时代营销的变革	324
9.4.4	大数据时代产业链的变革	325
9.5	大数据提高企业竞争力	326

9.6 练习.....	329
参考文献	330

第10章 大数据应用案例

10.1 大数据在金融行业中的应用案例	331
10.1.1 摩根大通信贷市场分析	331
10.1.2 奥马哈外汇风险敞口和实时数据分析	332
10.1.3 瑞士银行集合风险分析	333
10.1.4 汇丰银行多维度的历史数据分析和异常值快速分析	334
10.1.5 对冲基金选择Datawatch来观察实时的市场流数据	335
10.1.6 衍生品交易公司的交易活动的浏览与分析	336
10.1.7 跨国保险公司连接多个数据库来进行风险分析	336
10.2 大数据在医疗行业中的应用案例	337
10.2.1 美国糖尿病患者分布情况分析	337
10.2.2 医疗机构病房的实时监控	339
10.2.3 流行病学研究	341
10.3 大数据在互联网企业中的应用案例	344
10.3.1 亚马逊	344
10.3.2 淘宝网	345
10.3.3 Facebook	346
10.4 大数据在影视行业中的应用案例	346
10.4.1 大数据分析节目收视特征和用户喜好	346
10.4.2 大数据分析电影票房	348
10.5 练习	350
参考文献	350

第11章 大数据应用的主流解决方案

11.1 Cloudera大数据解决方案	352
11.2 Hortonworks大数据解决方案	352
11.3 MapR大数据解决方案	354
11.4 亚马逊大数据解决方案	355
11.5 IBM大数据解决方案	357
11.6 甲骨文大数据解决方案	359

11.7	EMC大数据解决方案.....	360
11.8	英特尔大数据解决方案.....	362
11.9	SAP大数据解决方案.....	363
11.10	Teradata大数据解决方案.....	365
11.11	微软大数据解决方案.....	366
11.12	国泰安大数据解决方案.....	368
11.13	练习.....	370
	参考文献.....	370