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。

清華大學出版社

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。

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

产品编号：062431-01

编 委 会

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、开展关键技术研究和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术操作，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

第2章 数据组织存储技术

第3章 NoSQL

第4章 Hadoop和MapReduce

第5章 数据查询和分析的高级技术

第6章 数据挖掘技术

第7章 数据分析语言R

第8章 大数据用于预测和决策

第9章 大数据与市场营销

第10章 大数据应用案例

第11章 大数据应用的主流解决方案