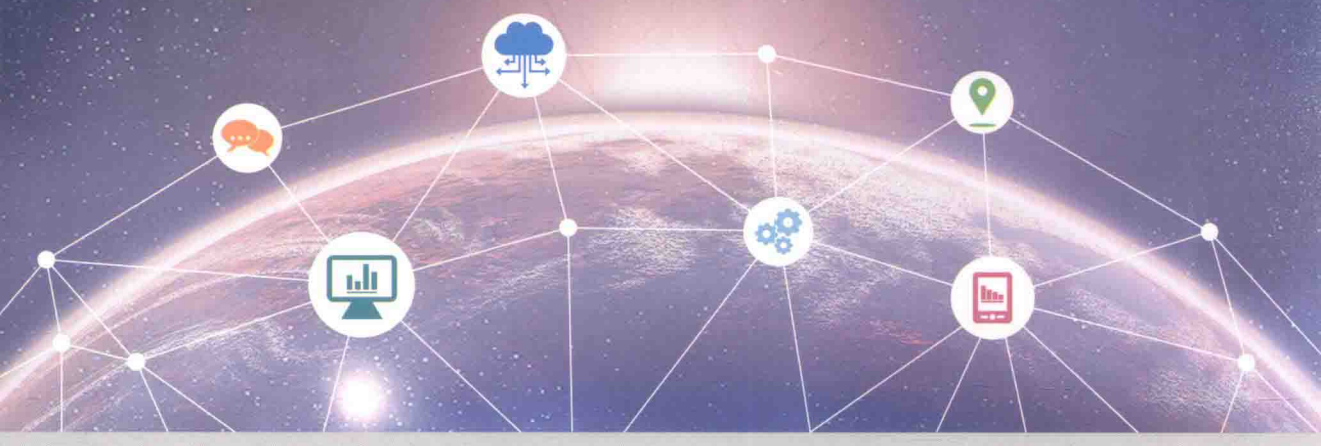
清華
科技大讲堂
前沿·科技·分享

大数据

数据管理与数据工程

◎ 赵眸光 赵勇 编著

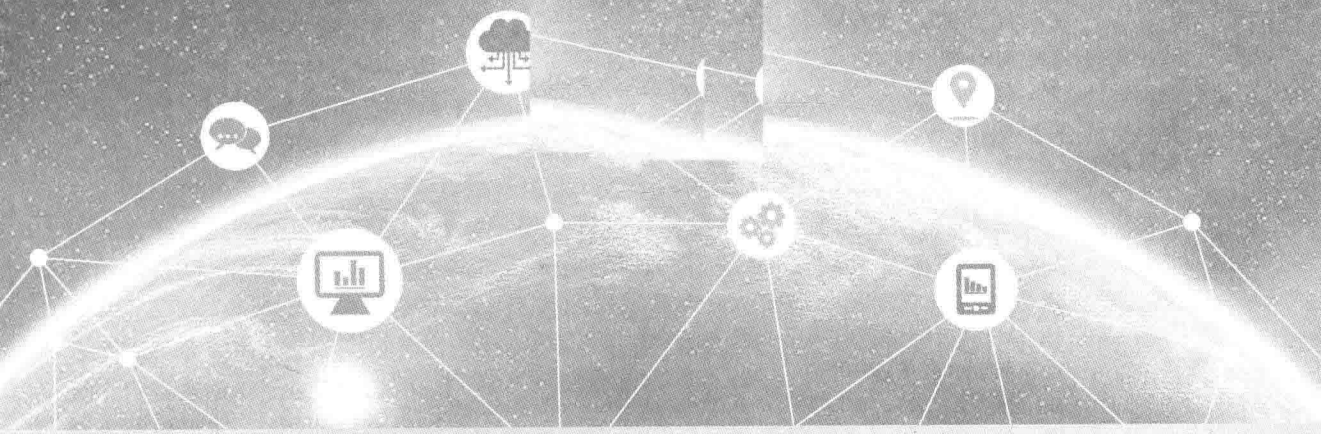


清華大學出版社

大数据

数据管理与数据工程

◎ 赵眸光 赵勇 编著



清華大學出版社
北京

内 容 简 介

大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物，必将对物联网产业产生深远的影响。大数据应用也将对社会的组织结构、经济运行机制、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活、工作和思维方式等产生深远的影响。

本书由两大部分组成，第一部分介绍大数据管理理论框架和生态系统，包括大数据概述；大数据战略和商业模式变革；大数据平台的架构体系；大数据的数据整合、交换与交易；大数据管理和治理；最后提出大数据创新方法论。第二部分介绍数据科学和数据工程，包括数据科学理论和工具；医疗健康大数据解决方案、环保行业大数据解决方案、移动社交行业大数据解决方案、金融大数据解决方案、中国制造大数据解决方案和大数据工程保障体系建设。

大数据是综合性较高的交叉学科，本书全面、系统地阐述了大数据管理和技术、大数据科学和工程，具有很强的理论指导性和实践意义。本书可以供企业管理者、数据科学研究工作者、首席信息官等作为参考资料，也可以作为企业管理、计算机、软件工程等相关专业学生的教材使用。

图书在版编目(CIP)数据

大数据·数据管理与数据工程/赵眸光，赵勇编著. —北京：清华大学出版社，2017
ISBN 978-7-302-46928-5

Ⅰ. ①大…　Ⅱ. ①赵… ②赵　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 074347 号

责任编辑：郑寅堃　薛　阳
封面设计：刘　键
责任校对：焦丽丽
责任印制：宋　林

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
课 件 下 载：http://www.tup.com.cn，010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：26.25　　**字　　数**：632 千字
版　　次：2017 年 7 月第 1 版　　**印　　次**：2017 年 7 月第1 次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：070944-01

序一

PREWORD

大数据是信息领域的前沿技术。大数据时代的来临，使人类有可能在浩如烟海的技术领域中，通过使用各种数据，发现和探索自然世界的规律。大数据时代的物理科学、计算科学、生命科学、社会科学及其他许多科学门类都将发生本质上的变化和发展，进而对人类的生产方式、生活方式和学习方式产生深刻的影响。

信息技术与经济社会的交汇融合引发了数据的迅猛增长，大数据已成为国家基础性战略资源，发展大数据及其相关技术研究更是重塑国家竞争优势的新机遇。国务院在2015年印发的《促进大数据发展行动纲要》中就强调了发展新兴产业大数据和工业大数据管理应用的重要性。通过大数据这种创新方式来解决我国在教育、交通、医疗和工程行业现代化所面临的种种问题，创建新的产业群，对实现由“中国制造”到“中国智造”再到“中国创造”有着重大意义。

云计算、物联网、移动互联网等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长。大数据具有Volume(大量)、Velocity(高速)、Variety(多样)、Value(低价值密度)、Veracity(真实性)这“5V”特性。对制造企业而言，大数据技术的战略意义不仅在于掌握庞大的数据信息，更在于对数据的“加工能力”，即对大量数据进行专业化处理的能力，使之转化成为对企业有价值的信息。制造企业如果能够在工业环境中建立起大数据平台，提高工厂对不同设备收集的海量信息进行数据挖掘的能力，提高企业信息系统的计算能力和数据处理能力，实现对企业的产品数据、运营数据、销售数据、客户数据的实时而有针对性的分析，用于洞察市场先机、客户需求，优化生产与管理流程，降低成本、提高运营效率、实现精准营销等，使得企业能够在成本有效控制的条件下，实现智能化生产、协同化组织和个性化服务。

赵晔光和赵勇博士多年在大数据理论、技术与应用等方面深入研究，取得了一系列成就。本书重点围绕大数据管理和大数据工程两方面进行了系统化的阐述，研究了大数据平台的体系架构和数据整合、交换与交易技术，通过对大数据的管理，总结出大数据创新方法论。此外，本书详细介绍了数据科学理论与工具，包括数据仓库、数据挖掘和知识发现等，对于医疗行业、移动社交、工业制造等几个热点行业数据工程的实践，进行了有针对性的阐述。全书内容系统，论述充分，为高校、研究院所科技研究人员和企业工程技术人员、管理人员从事大数据研究、应用和培训提供了一本极好的参考书。特此推荐。

中国工程院　院士
中国大数据技术与应用联盟　理事长
浙江大学　教授

2017年4月

序二
PREWORD

数据自古存在。

乌龟壳、树皮、绸缎、竹简都曾是记录数据的媒介，留声机、磁带机也曾经风靡过，就连现在的信息技术，像个人电脑、智能手机、iPad 在不远的将来也将会退出舞台，唯有数据，虽然不断地变换表现形态，却将一直伴随人类走向未来。

物联网本质上是器物层面的技术，从大数据的视角而言，是采集数据的终端。云计算本质上是传统计算机和网络技术发展融合的产物。物联网和云计算都是信息技术发展的一定阶段的自然延伸，依然属于信息技术范畴。而大数据其实是传统数据发生的质变。大数据超越信息技术，使人们重新界定国家竞争的主战场，重新审视政府治理水平，重新认识科学研究的新范式，重新审视产业变迁的驱动因素，重新理解投资的决策依据，重新思考公司的战略和组织。总之，大数据是推动经济发展、保障国家安全和社会治理的永恒主题。

大数据蕴含巨大价值，是国家意志和主权不可分割的部分。

2012 年 3 月，奥巴马发布美国版的《大数据发展计划》时，我曾经写过一段点评："国家层面大数据技术领域的竞争事关一国的安全和未来。国家数字主权体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。"在这篇点评中，明确提出数字主权的概念，坦言大数据必须上升为国家意志，落实为国家战略。2014 年 5 月 1 日，美国白宫发布了《美国白宫：2014 年全球"大数据"》白皮书，阐述了大数据带来的机遇与挑战。2014 年 8 月，联合国开发计划署首次携手科技企业共建大数据实验室。我国 2015 年 9 月《促进大数据发展行动纲要》出台，赋予了大数据作为建设数据强国、提升政府治理能力和推动经济转型升级的战略地位。

保护国家层面的数据安全，恰恰是以数据开放为基础的。开放是一种态度，更是一项能力。一些重大基础数据开放，可以构成社会的数据基础，按照大数据定律之一"数据之和的价值远远大于数据价值的和"来推断，来自不同领域的数据聚合在一起，开放给社会，将会产生类似核聚变一样的价值发现效应。

开放的数据是基础，促使信息产业繁荣，才能诞生真正的数据驱动的企业，这些企业反过来在数据领域的技术进步，才是确保国家数据安全的长治久安之策。很难想象，如果没有谷歌、亚马逊、Facebook、苹果这样的公司，单凭美国政府一己之力能够实施如此庞大的"棱镜"计划吗？所以制定国家大数据战略，需要重新思考传统的所谓的"国家机密"和国家安全的关系。应当把消除部门数据格局，建立公开、透明、共享的数据公共平台作为长期的战略目标。

大数据将成为政府治理、企业管理、产业价值发现的重要工具。

大数据将打开各行各业的数据宝藏。政府治理、社交网络、医疗、教育、环保、金融、智能制造等，都会受益于大数据而被挖掘出更多的价值。在政府治理领域，通过让海量、动态、多

样的数据有效集成为有价值的信息资源，推动政府转变管理理念和治理模式，进而加快治理体系和治理能力现代化。还有推动政府治理决策精细化和科学化。如何将海量数据对行业进行管理决策、产品设计、精准营销、客户个性化服务等？如何对行业大数据进行商业模式设计？如何进行大数据平台建设？如何发挥大数据性能优势？如何解决安全和隐私？如何对各种大数据进行可视化？达到好的效果？本书解决了大数据从顶层设计到应用落地，从商业模式到技术平台，从数据管理到数据价值发现。本书的出版，正如天降甘露，恰到时机。

数据科学——科学地研究数据，用数据来研究科学。

大数据的五大特征（体量、类型、速度、价值和真实）蕴含了大数据丰富的内涵和外延。学术界在大数据时代有了广阔的舞台。大数据的早期发展是由技术性公司推动起来的，例如谷歌、亚马逊等一线互联网公司。产学研合作也正好是推进大数据发展的最佳途径。学术界有很好的理论基础和算法优势，产业界有很好的支持平台。鄂维南院士呼吁学术界向谷歌公司学习，同时指出："大数据在科学领域的表现是数据科学的兴起，数据科学将成为科研体系中的重要组成部分"，也为数据科学发展指明了方向。

在大数据时代，许多学科表面上看来研究的方向大不相同，但是从数据的视角来看，其实是相通的。比方说自然语言处理和生物大分子模型里之所以都用到隐式马氏过程和动态规划方法，其最根本的原因是它们处理的都是一维的随机信号；再如用于图像处理的算法和用于压缩感知的算法也有着许多共同之处。

图灵奖得主格雷（Jim Gray）提出科学研究的"第四范式"是数据，不同于实验、理论和计算这三种范式。在该范式下，需要"将计算用于数据，而非将数据用于计算"。吴军博士在《数学之美》一书中也讲到了这方面的故事。以自然语言的机器翻译研究为例，最初科学家们都是试图为计算机建立一系列的语法规则，按照语法、词义来翻译成另外一门语言。这个思路非常直观，因为人们就是如此理解学习的语言的。但是在实践中却困难重重，基于语法规则的翻译器，几乎就没有商用过。而当科学家们改弦易张，计算每一个词、每一句话的"合理概率"时，复杂的机器翻译就简化成了文字的概率计算。通俗地说就是："如果大多数人都这么说，就认为是对的。"这种思想在越来越多的领域得到应用。比如宏观尺度研究的天体信息学、社会行为学，微观尺度上分析人类的基因组、追踪物理学家们梦寐以求的"上帝粒子"。

随着大数据应用领域的逐步深入，越来越多的应用在数据层面趋于一致。数据科学在数学、概率模型、统计学等和实际应用之间建立起了直接的桥梁。本书在数据科学理论方面建立起了有效的方法论体系。

数据工程——大数据产业发展支撑体系。

曾经和中关村大数据产业联盟几位专家、总裁一起讨论，大家七嘴八舌地提出"＋大数据"的概念。希望在联盟中培育出各个专家组，把大数据思维嫁接到不同的产业，推动大数据在各行各业落地。大数据产业变革综合运用了大数据相关理论。本书介绍了医疗、环保、社区、金融和智能制造大数据产业分析和系统架构实现，对其他行业的应用也有很好的指导作用。许多行业龙头也开始蠢蠢欲动，应用大数据思维解决产业变革问题。例如农业领域的大北农、教育行业的新东方、玩具领域的奥飞动漫……

给企业家们带来冲击的不仅仅是大数据引起的产业变革，更是一些新兴公司的不可思议的跨界能力。就像本书中指出的那样，行业之间的界限变得越来越模糊，这些新兴的"野

蛮人”采用新的技术、新的模式，大规模采集数据，迅速形成预判，然后就以看似“野蛮”的方式扩张到其他行业。譬如卖农产品的去搞金融服务，做金融业务的帮助企业做采购等等，不一而足。

传统产业的各行各业，都面临在大数据和移动互联网时代如何彻底转型和再造问题。产业整合，也在大数据时代出现了全新的整合逻辑和实现契机……我仿佛看到了一个未来景象：传统产业都可能在大数据和移动互联时代重现生机、焕发青春。当然，与此对应的是，凡是不能跟上这个时代步伐的企业和行业，将会退出历史舞台。

在星空格局之下，公司的竞争力更多体现在“平台＋特种部队”模式。就像美军前线的一个小分队，甚至单兵可以直接指挥后方的导弹、飞机一样。以星空格局作为产业演化的最终形态，以特种部队作为业务竞争的基本单元，整个公司的战略、组织、文化等方面需要彻底的重组。传统公司的确需要重新审视自己的战略，重构组织，再育文化，这也是大数据思维非常重要的原因。

综上所述，不能狭隘地看待大数据，不能将其作为数据挖掘的工具，不能唯技术论。很欣慰看到两位学者编写的《大数据·数据管理与数据工程》一书，不是就技术而谈技术，而是从更宽广的视角阐释大数据带来的冲击、管理理念的变革以及大数据生态系统。尤为重要的是，提出数据工程的概念，奠定了大数据应用领域标准化、工程化的基础。

从大历史观来看，“大数据”的内涵远远超越物联网、云计算等信息技术的概念，它的意义可以比肩活字印刷术的发明。“大数据”将在世界尺度上大范围地消除信息不对称的现象，释放巨大的生产力，深刻改变社会的面貌，革新科学研究的思想，促进产业间的跨界、融合和颠覆，并将极大地促进文明的传播、凝聚和升华。

是以为序！

中关村大数据产业联盟秘书长

2016 年 12 月

序三

PREWORD

建立互联网金融治理体系，应该成为我国金融治理体系和金融治理能力建设的重要内容，大力发展互联网金融，以互联网金融治理推进中国金融治理体系和治理能力现代化，是金融治理创新发展的重要引擎。凯文·凯利(Kevin Kelly)被誉为互联网经济的预言家，他精准预测 Web 2.0 时代的到来和网络经济的运行规律。凯文·凯利预言，未来，大数据、云计算、移动通信三者相结合的技术进步将激发大数据、深度学习、人工智能、P2P、虚拟货币等方面的技术突变，而这些正在成为现实。未来技术改变的世界有四大特征：万物互联、信息交互、数据集成、智能决策，这四大特征正是物联网大数据时代的主要特征，这也正是金融模式创新的基础前提。

从该书大数据市场行业应用分析中可以看出，金融行业在大数据应用可行性和市场成熟度方面都属于优先级比较高的领域，是大数据应用热点。本书提纲挈领、高屋建瓴地从大数据系统科学的角度去认识“大数据”，指出大数据的内涵和研究方向，从而发现大数据的价值，通过全球大数据的战略视角，窥视行业应用商业模式和商业机会。从金融创新来看，数据成为资产、行业垂直整合、平台泛金融化成为商业发展主流趋势，行业产业链条加深加长，促使商业创新模式层出不穷。互联网创造出新的商业模式，塑造新的经济形态，创造新的经济生态空间，加大生产可能性边界，降低生产成本和融资成本，互联网基因已经融入到社会运行的底层物质技术结构之中。大数据时代的金融创新，必将发生像作者在书中提到的种种金融变革。

该书从大数据架构体系、安全和隐私、系统整合、数据管理以及理论创新方面全面系统地提出管理方法和技术工具，通过数据科学理论在金融创新和风险控制方面，在大数据征信贷款、大数据反欺诈、大数据客户管理和精准营销方面做出了分析。例如大数据技术运用于信贷技术前，借款需要很长时间的审核，尤其是线下取证、财务报表、抵押担保、审批流程、领导签批、最后借款等环节。根据内在的大数据信用评估和内控技术，能够实现实时计算借款人的信用额度，在信用额度内实现即时放款。这在传统金融领域是难以想象的，而这种快速借款模式，将成为未来互联网金融时代的标志。

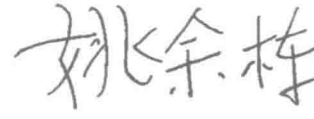
该书体系完整、结构清晰、逻辑严谨，是大数据从战略到战术、理论到实践、产业到模式、标准到工程，具有战略性、系统性、理论性和指导性的大数据百宝箱和重要参考全书。当前，国家大数据战略日渐清晰，产业应用初具规模，大数据技术日趋成熟，本书为大数据从业者和应用机构提供了大数据应用知识地图、全新的认识和决策思路，非常值得一读。

大数据金融创新的数据可视化已经成为经济分析、管理决策、绩效评价等工作的重要工具。金融可视化是利用数学模型、网络技术、数据挖掘、计算机语言等一系列数据科学前沿科技综合应用的重要成果。该书提供了丰富的金融数据可视化展示工具和方法，不仅能够让数据丰富多彩地展示，还原真实世界，得出精准信息，更让人们能够通过数据模型直观地

感受到数据的真实变化。数据使得决策更加科学化、智能化、动态化、实时化，成为决策的重要依据。

从金融业的发展趋势来看，大数据技术将会成为风险管理的最佳工具，云计算为金融业务的高效实时处理做出保障，点对点的资源配置方式充分发挥金融职能，越来越多的传统金融需要这些互联网金融新模式作为技术载体、信息载体和业务载体。互联网金融对现代金融业的塑造主要体现在互联网金融平台上，通过自我创造、自我发展衍生出金融业务交易平台、新兴技术应用平台、风险控制管理平台、金融模式创新平台和普惠金融服务平台。本书在数据工程实现和金融平台建设上提供了技术支持保障。

书中在大数据管理创新和工程实践中提供了全新的视角和系统性思维，在目前大数据领域丛书中，具有更强的指导性。随着应用的不断深入，学习和研究也要与时俱进。互联网金融会成为金融创新发展的必然趋势。新的技术不断涌现、智能搜索引擎、区域链技术、全新的信息通信和物联网技术等必将会对金融业产生革命性的影响，也为互联网金融的发展提供一个良好的契机，可以让金融监管发挥更大的效力。先进的大数据金融信息系统可以及时检测金融市场与企业的动态，而电子化的渠道可有效地降低监管的搜索成本，多渠道的信息数据来源可以降低监管面对的信息不对称难题，而通过机器学习可以构建智能监管监测系统。这些信息化金融监管手段来源于市场，作用于市场，检测于市场。金融是现代经济的核心，推进我国互联网金融治理体系和治理能力现代化，是金融治理创新和经济发展的必由之路。本书一定会成为大数据青睐者和行业践行者的良师益友。



中国人民银行金融研究所所长

序四
PREWORD

信息作为一种资源自古就存在，信息就是物质，信息通过电子化、数字化无限增值。1800年，伏特发明了世界上第一块电池；1946年，人类发明第一台电脑。伴随电脑、互联网时代的到来，信息成为可生产、交换、传播的商品。个人电脑、互联网、浏览器、搜索引擎、智能手机、社交网络、可穿戴设备、3D打印比过去基于蔡伦、毕升、古登堡时代，传统印刷更为丰富、多元、有效。不到半个世纪，人类存储的数据量以指数级在增长，数据传输速度从数天缩短到数毫秒，提升达9个数量级，成为全球拥有、共享、传播的大数据海量信息。随着全球大数据、物联网、云计算、移动社交网络等信息网络新技术的普及，推动世界数字经济呈指数增长，人类社会信息化进入大数据时代。

然而，数据规模如此之大、数据结构如此复杂、数据传播如此之快，已经远远超过了目前政府或企业在数据采集、存储、处理和分析、管理和应用方面的能力。企业如何发现数据的价值？如何利用数据产生效益？大多数企业还是手足无措。

本书通过大数据管理理论框架与生态系统、数据科学与数据工程两大部分，基本上覆盖了数据起源、数据架构(基础设施、数据采集、存储、分析处理、可视化、应用、运维、安全和隐私)、数据整合与交换及交易、大数据管理与治理、数据创新与数据科学、重点行业应用等。全面解决了大数据如何应用和价值发现的过程。

大数据成为全球重要的战略资源和核心资产。大数据时代，各国对数据的依赖快速上升，国家竞争焦点已经从资本、土地、人口、资源的争夺转向了对大数据的争夺，对大数据的开发、利用与保护的竞争日趋激烈，制数权成为继制陆权、制海权、制空权之后的新制权。大数据使得强国与弱国不再以经济规模和经济实力论英雄，而是取决于一国大数据能力的优劣。

借助大数据革命，美国等发达国家全球数据监控能力升级，美国先后推出《网络空间国际战略》《网络空间国际行动》等重要战略规划，确保自身在网络和数据空间的主导地位。

中共中央十八届五中全会提出，要拓展发展新空间，实施网络强国战略，实施“互联网+”行动计划，发展分享经济，实施国家大数据战略。国务院通过《关于促进大数据发展的行动纲要》为未来中国的大数据发展指明了方向。

据统计，2015年全球信息社会指数为0.5494，正在从工业社会向信息社会加速转型，专家预计人类2018年进入信息社会。中国互联网经济占GDP比重4.4%，已超过美国、法国和德国，达到全球领先国家水平。要实现两个百年发展目标，2021年中国人均信息消费将接近1000美元，2049年中国人均信息消费将超过3000美元，成为世界最大的信息经济体。2013年中国大数据产业市场规模为34.3亿元，同比增长率超100%，未来一段时间将持续快速增长。2014年7月，麦肯锡全球研究员发布的《中国的数字化转型：互联网对生产力与增长的影响》预测：2013到2025年，互联网将占到中国经济年增长率的0.3%～1.0%，互

联网将可能在中国 GDP 增长总量中贡献 7%～22%，我国正从数据大国向数据强国过渡。

中国作为世界最大的发展中国家，能否吸取工业革命中“落后挨打”的悲剧教训，在全球化信息网络时代跨越中等收入国家陷阱和修昔底德陷阱？中国能否在这次全球信息革命浪潮中抢占先机、立于不败之地？能否实现中华民族伟大复兴的中国梦、两个百年目标？

我国必须要紧抓大数据技术发展机遇，正如本书所述，建立起大数据标准体系、数据科学理论体系、标准化大数据治理体系，实现弯道超车快速崛起，成为全球最大信息经济体的受惠者。

中央人民广播电台高级编辑、央广网副总裁

致读者

TO THE READER

大数据历经几年的发展，在全球已进入了高速发展期。我国“十三五”规划正式将大数据上升为国家战略，当前全国各省市级和地区级城市正在制定大数据发展战略和实施规划，中国正在创造一个万亿级的大数据市场。在此期间，笔者 2014 年编著了《大数据革命——理论、模式与技术创新》，2015 年又出版了大数据的技术教材《架构大数据——大数据技术与算法解析》。在大数据产业发展上，以成都为基地，成立大数据协会和联盟，如四川大数据产业联盟、中国西部互联网与大数据产业协会等，提供大数据人才培训和培养、政府大数据产业规划和企业转型升级咨询。成立第五维国际大数据孵化器，通过和硅谷孵化器合作，为大数据创业团队提供导师、技术、办公场地和资金等全方位的孵化服务。在大数据产品研发上，以清数科技公司为依托，开发了 Neo 大数据一体“傻瓜机”，把数据从采集、存储、处理、分析和挖掘、可视化和应用服务全部集成部署到一体机服务器中，让政府和企业拥有“开箱即得”的大数据分析处理能力，方便了用户的操作使用。

本书正是笔者在大数据产品研发和产业落地基础上的理论升华和管理思考。笔者预测，中国的大数据产业将在明年中期迎来应用的全面爆发，大数据的平台、分析、应用类的产品和服务将供不应求。而大数据交换和交易的市场，随着国务院制定的政府数据开放日程的临近（《大数据发展行动纲要》要求各部委数据在 2018 年底完成开放），也将在两年后成为大数据产业的最大的市场，数据资产、数据产品、数据服务都会带来巨额的财富。本书正是顺应大数据发展趋势，重点阐述了大数据生态系统、大数据管理、数据交换、共享、交易等理论体系，数据科学理论和大数据行业应用实践，以及相应的大数据标准体系；全面系统地阐述了大数据体系建设和工程实践，真正挖掘和实现了大数据的价值。本书内容主要围绕大数据应用热点和重点行业展开分析，如医疗、环保、社交、金融、工业制造等，这些理论实践同时也适用于教育、政务、交通、能源、航空、农业、旅游等行业的发展应用。总结出了大数据管理创新方法论和工程实践经验，为中国大数据产业发展和创新生态链打造奠定了理论和实践基础。

众所周知，从上届美国总统的选举到本届美国总统选举，无疑都是大数据应用的最好例证。本届选举演变成了希拉里和特朗普背后的大数据团队的生死角力。双方都拥有阵容强大的大数据团队，服务于特朗普的 Deep Root Analytics（深根分析）公司和英国的剑桥分析公司采取的是类似于精准广告投放的技术，分析摇摆投票者们的意识形态、价值观以及他们喜欢的信息接收方式和渠道，然后针对他们制定竞选演说、拉票方式和信息传递方式，最终帮助特朗普问鼎美国总统宝座。尽管是在被业界称为投资寒冬的大环境下，大数据以及人工智能还是在美国硅谷和中国的投资圈刮起一股旋风，数百家相关的大数据企业都顺利拿到了投资。大数据应用成为产业聚焦的热点。

本书的编写得到了很多协会和清数的同事们的支持和帮助，尤其是李小龙、张晓东、唐

犀、赵虎、滕雨橦，还有电子科技大学极限网络计算与服务实验室的同学们，他们为本书收集了大量的资料，并提供了很多的内容。我也要感谢我的家人们对我的鼓励和支持，很多节假日都没能陪同她们。

本书由于笔者的知识和经验有限，存在的疏漏敬请读者原谅，也欢迎与我们联系，我们一起为中国的大数据事业贡献力量，谢谢大家。

赵　勇

2016.12

前言
PROFACE

大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物，也必将对网络通信(ICT)和物联网(IOT)产业产生深远的影响。大数据技术的发展与应用，将对社会的组织结构、经济运行机制、社会生活方式、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活、工作和思维方式等产生深远的影响。随着社会网络安全、应急管理、医疗健康、经济金融、交通运输、制造领域、社交社区等各个领域大量数据的使用，对于我们而言，能够及时、有效地了解数据和信息的意义，进而改善决策制定的过程将变得尤为重要。大数据的价值必将对现代企业的管理运作理念、市场营销决策以及消费者行为模式等产生巨大影响，使得企业商务管理决策越来越依赖于数据分析而非经验甚至直觉。因而，大数据也必将对这种传统的商业模式进行近乎彻底的颠覆与模式的重构。

当前，美国、日本、法国、韩国、澳大利亚等国家相继启动了推动大数据产业发展的政策改革，并把大数据产业发展纳入国家发展战略，通过有力的资金和政策支持加强大数据研究，优化其发展环境，抢占大数据产业发展的制高点，使其成为推动国民经济社会发展的新手段。鉴于发达国家对大数据产业的强力推动，大数据在经济、国家安全、社会、科研等方面的巨大价值和适应经济社会发展的要求，中国各级政府和社会各界也纷纷制定相关政策推动大数据产业深入发展，运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势，我国相继制定实施大数据战略性文件，大力推动大数据发展和应用。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

我们认为大数据的发展必将经历三个重要的阶段。①“技术驱动型”。大数据的核心关键技术正在加速发展和快速迭代，技术体系框架也已日趋成熟，基本能够满足产业发展需求，比如 Hadoop 生态框架系统。大数据架构体系分为基础设施、采集、存储、处理、分析、应用、安全和维护几个方面。②“行业驱动型”。各大解决方案服务商围绕电信、环保、金融、交通、医疗、政府、教育、工业、城市管理、社交网络等重点行业领域描绘美好蓝图，力求推动行业应用，如节能环保产业布局了高效储能、节能监测和能源计量；生物医药产业布局了生物资源样本库、基因测序，以及基于物联网的远程健康管理服务等。这一阶段发展虽然还有距离，但这一转变过程正在加速进行。③“模式驱动型”。大数据行业应用深化发展，使得领域和行业边界愈加模糊，商业模式应用创新超越技术本身，企业以独特数据资源进行的整合朝着纵向产业链上下游整合和横向多种产业整合两个方向发展，生产模式向服务化转变，数据作为一种资产资源为企业带来新的商业价值，数据开放为政府治理和个人福祉都带来新的机遇。

从大数据系统论的角度，可以将大数据划分为大数据技术、大数据管理、大数据科学和大数据工程，本书重点围绕大数据管理和大数据工程两部分展开阐述。

第一部分介绍大数据管理理论框架和生态系统，共分为6章，主要内容有：数据时代背景、大数据定义、特征、数据结构、度量价值、数据管理与技术和大数据科学与工程研究方向以及大数据生态系统；国内外大数据战略和大数据应用的商业模式变革；大数据平台架构体系自下而上包括基础设施、数据采集、数据存储、数据处理、数据可视化、大数据应用、运维和数据安全；大数据平台整合、大数据与存储、大数据与网络、大数据与虚拟化技术整合、大数据环境的数据整合、大数据交换和数据交易；大数据流程管理、大数据事务管理、大数据技术管理以及大数据质量管理阐述；最后提出大数据创新理论指标体系、大数据创新重要环节和大数据创新最佳实践。该部分章节框架清晰、结构分明、逻辑严谨、层次有序、概念明确、重点突出、体系完整，形成整个大数据技术管理体系。

第二部分介绍数据科学和数据工程内容，共分为7章，主要内容有：数据科学概念、研究重要角色、生命周期管理、数据仓库、数据挖掘分析方法、知识发现及大数据处理平台，通过建立科学系统的数据分析方法论，指导数据工程实践；在数据工程方面，重点介绍医疗行业大数据、环保行业大数据、移动社交大数据、金融行业大数据和工业制造大数据等几个热点行业数据工程实践，每个行业又侧重大数据应用的不同角度，总体上全面解析大数据应用的多个方面；医疗健康主要包括总体架构(业务架构、技术架构和网络架构)、医疗大数据存储处理、容灾备份解决方案和医疗大数据分析等；环保行业包括环保物联网架构、电力脱硫工作原理、电力脱硫数据分析优化目标以及空气质量大数据分析评价体系；移动社交包括发展趋势、社交理论、社交网络商业模式、社交网络平台以及社交网络数据分析；金融行业包括金融大数据特征、发展机会、总体架构(业务架构、技术架构和网络架构)、金融大数据风险管理平台、大数据征信、大数据反欺诈、大数据精准营销以及大数据带来的产业变革；工业大数据通过回顾全球工业信息化发展历程和现状，提出了中国制造2025发展战略，同时指出工业信息技术集成和协同发展方向，利用工业信息化应用系统搭建工业大数据架构体系(业务架构、技术架构和安全架构)、智能化协同制造架构原理，最终实现智能化协同制造服务。工业是国民经济的基础，工业的未来也是我国经济发展的未来。最后提出大数据工程保障体系建设，包括法律体系建设、标准体系建设、标准化大数据治理体系建设、技术和应用研究、创新平台建设等，该部分章节充分体现了理论性、科学性、创新性、实用性、经济性、社会性、标准性、保障性和完整性，形成了数据科学和数据工程体系。

本书是作者和在大数据研究领域非常有名望的赵勇博士共同编写而成的。书中的第3～6章来源于赵勇博士研究成果，其他是作者多年来对物联网、云计算和大数据的研究、咨询和应用实践经验的智慧结晶，同时也是在清华大学继续教育学院致力于智慧城市规划设计和企业管理咨询工作经验的积累。希望本书将我们多年从事于大数据研究方面的成果展现给读者，本书可以作为企业管理者、数据科学研究工作者、首席信息官等的参考资料，也可以作为企业管理、计算机、软件工程等相关专业学生教材使用。

本书在撰写的过程中，得到了清华大学、北京大学多位老师，清华大学数据研究院和行业同仁的资料提供和支持帮助，在此表示衷心的感谢！也感谢我的家人给予我莫大的支持和鼓励，使我顺利完成写作。大数据发展日新月异，相关技术快速发展，由于我们对大数据的理解和知识水平都有局限，书中疏漏或不足之处在所难免，敬请读者批评指正。

赵眸光

2016年12月于清华园

目 录

CONTENTS

第一部分　大数据管理理论框架与生态系统