CCF 大数据教材系列丛书
CCF 大数据专家委员会 组编

主编 梅宏

大数据导论

高等教育出版社

CCF 大数据教材系列丛书
CCF 大数据专家委员会 组编

主编 梅宏

大数据导论

高等教育出版社·北京

大数据导论

Dashuju Daolun

图书在版编目（CIP）数据

大数据导论 / 梅宏主编. -- 北京 : 高等教育出版社, 2018.11
ISBN 978-7-04-050726-3

Ⅰ. ①大… Ⅱ. ①梅… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第235060号

策划编辑　王勇莉
责任编辑　王勇莉
书籍设计　张申申
插图绘制　于　博
责任校对　李大鹏
责任印制　田　甜

出版发行　高等教育出版社
社址　北京市西城区德外大街4号
邮政编码　100120
购书热线　010-58581118
咨询电话　400-810-0598
网址　http://www.hep.edu.cn
http://www.hep.com.cn
网上订购
http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
印刷　北京信彩瑞禾印刷厂
开本　787mm×1092mm　1/16
印张　21.5
字数　330千字
版次　2018年11月第1版
印次　2018年11月第2次印刷
定价　48.00元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物料号　50726-00

内容提要

本书是计算机学会大数据专业委员会大数据教材编委会编著的大数据系列教材中的第一本，其目的是从技术角度，为大数据相关专业本科生、研究生及科研人员，提供一本全面介绍大数据相关技术的专业通识教材。本书系统地介绍大数据涵盖的内容，包括数据与大数据概述、大数据感知与获取、大数据存储与管理、大数据智能分析、大数据架构与处理技术、大数据分析计算平台、大数据治理、大数据安全与隐私等。除了介绍大数据的技术内容，本书还介绍了部分行业中大数据的典型应用案例，反映了大数据在社会经济生活中的重要价值。

本书既可作为普通高等学校大数据相关专业的教材使用，也可供有关技术人员参考。

编者按

随着大数据的蓬勃发展，大数据领域人才的需求越来越大，大数据人才培养受到了各界的广泛关注。2016年，教育部开始批准设立“数据科学与大数据技术”本科专业，越来越多的高校申请开设“数据科学与大数据技术”专业或开设大数据方向的相关课程，截至2018年3月，已有近三百所高校获批设立“数据科学与大数据技术”专业。虽然大数据专业和大数据方向的课程不断开设，但是，当前我国高校的大数据教学尚处在摸索阶段，尤其缺乏成熟的、系统性和规范性的大数据教学体系和教材。

在此背景下，中国计算机学会大数据专家委员会成立了大数据教材系列丛书编委会，着手编著系列化、规范化的大数据教材。自2017年6月，经编委会多次研讨，形成丛书框架，作者们随即开始紧张的编写工作。编委会和作者间也有多轮的初稿审阅和研讨交流。数易其稿，终于付梓。

大数据教材系列丛书采用“1+3+X”的体系，即以1本《大数据导论》为基础，设置《大数据管理》、《大数据处理》和《大数据分析》3本关键技术教材，以及针对行业领域的X本应用教材。本套教材系列丛书既适合高校大数据专业的本科生以及研究生系统地学习大数据相关知识与技术，也适合从事大数据相关技术的企事业单位研究人员、工程师作为参考用书。

《大数据导论》是一本全面介绍大数据相关知识的专业通识教材，其系统地介绍大数据涵盖的内容，包括数据与大数据、大数据获取与感知、大数据存储与管理、大数据分析、大数据处理、大数据治理、大数据安全与隐私等，同时还介绍了部分行业中大数据的典型应用案例，反映了大数据在社会经济生活中的重要价值。

《大数据管理》首先综述数据管理系统的发展，指出发展大数据管理系统是历史的必然，并沿着数据模型和系统构件两个维度上展开。在数据模型的维度上，主要介绍关系、键值对、图和文档数据模型及其语言；在系统维度上，介绍系统结构、存储与组织、查询处理、事务管理、故障恢复等话题。

《大数据处理》包括大数据处理基础技术、大数据处理编程与典型应用处理、大数据处理系统与优化三个方面。本教材以大数据处理编程为核心，从基础、编程到优化等多个方面对大数据处理技术进行系统介绍，

使得读者能够快速入门，同时体会大数据处理系统的设计理念与优化方法本质。

《大数据分析》包括大数据分析方法和理论、典型大数据分析任务以及大数据分析系统与应用。本教材特色是理论联系实际，本书从基础理论、典型任务以及系统应用多个方面对大数据分析相关知识进行了系统而详细的介绍，使得读者能够快速入门，体会大数据分析技术的本质特征，领略大数据技术带来的创新理念。

“X”系列教材包含面向各行各业的大数据应用的知识与技术，既可面向工程实践，又可面向职业培训，且将随着产业界大数据应用的发展进行更新迭代。

大数据已成为学术界、产业界和政府共同关注的热点，正在开启信息化的新阶段。大数据人才培养也刚刚起步，还需要付出更多努力去探索。通过汇集中国计算机学会大数据专家委员会的智力资源，丛书编委会希望本系列教材能够为我国大数据人才培养尽到绵薄之力，助力我国大数据事业的蓬勃发展。尽管编委会和作者花费了很大精力规划和编写本系列教材，但是囿于对大数据的认识局限和自身能力限制，难免存在疏漏和错误，欢迎读者批评指正，以待再版时修正完善。

大数据教材系列丛书编委会

2018年7月

前言

大数据相关产业的高速发展，带来了大数据人才的严重短缺，大数据人才培养成为当前急迫的任务。2016年起，教育部开始批准高校设立“数据科学与大数据技术”专业，目前已有近三百所高校获批设立该专业。然而，大数据教学体系建设尚处在探索阶段，尤其缺乏系统性、规范性的大数据教材。为此，中国计算机学会大数据专家委员会组建了大数据教材系列丛书编委会（以下简称“编委会”），希望能够编写一套系列教材，应对当前之急需。编委会由大数据领域著名的专家学者、教学经验丰富的一线老师和实践经验丰富的产业界专家构成，我被推选担任编委会主任。我深知，这是一项时间紧、难度大的任务，但又是一项意义大、必须做的事！

有赖于编委会同仁及作者共同的努力，本系列丛书的第一本——《大数据导论》，即将付梓。欣慰之余，作为本书主编，我也借撰写前言的机会，和读者分享若干自己对大数据的认识和感想。

一、认识大数据

大数据是信息技术及其普适应用发展到一定阶段的“自然现象”，源于互联网及其延伸所带来的无处不在的信息技术应用以及信息技术的不断廉价化，其主要驱动力包括：摩尔定律驱动的指数增长模式、技术低成本化驱动的万物数字化、宽带移动泛在互联驱动的人机物广泛联结以及云计算模式驱动的数据大规模汇聚。

从文明之初的“结绳记事”到文字发明后的“文以载道”，再到近现代科学的“数据建模”，数据一直伴随着人类社会的成长变迁。然而，直到以电子计算机为代表的现代信息技术出现后，才使人类掌握数据、处理数据的能力得以空前高速的发展。信息技术及其在社会经济生活方方面面的应用（即信息化）推动数据（信息）成为继物质、能源之后的第三大战略资源。

大数据概念由计算领域发端，之后逐渐波及科学和商业领域，引发了一系列新思潮。1997年，高性能计算企业SGI首席科学家John Mashey就曾指出数据快速增长将成为计算发展的重要趋势，并指出大数据难理解、难获取、难处理、难组织等四个方面的问题，引发计算领域对大数据的思考；2007年，数据库领域的先驱人物Jim Gray提出了“第四范式”的概念，指出大数据为人类提供了基于大数据触摸、理解和

逼近现实复杂系统的可能性，从而使数据密集型科研在实验观察、理论推导和计算仿真之后，成为人类探索未知、求解问题的一种新型科研范式；2012年，牛津大学教授Viktor Mayer-Schönberger在其畅销著作《大数据时代》中，指出数据分析从“随机采样”、“精确求解”和“强调因果”的传统模式将演变为大数据时代的“全体数据”、“近似求解”和“只看关联不问因果”的新模式，引发商业应用领域对大数据方法的广泛思考与探讨。

经历了早期的众说纷纭，当前对大数据这一概念的定义已基本形成共识。一是从技术能力的视角，将大数据定义为“规模超过现有数据库工具获取、存储、管理和分析能力的数据集”，并同时强调并不是超过某个特定数量级的数据集才是大数据；另一是从大数据内涵的视角，称大数据为“具备海量性、高速性、多样性和可变性等特征的多维数据集，需要通过可伸缩的体系结构实现高效的存储、处理和分析”。当然，共识之下，对大数据相关的一些核心观点和命题仍然存在争议，例如：数据“大”与“小”的对立统一，“关联”与“因果”的辩证性，“全数据”的相对性等。

大数据蕴含的巨大应用价值和潜力已被广泛认知：提供了人类认识复杂系统的新思维和新手段，成为促进经济转型增长的新引擎，为政府提升治理能力提供了新途径，更是提升国家综合能力和保障国家安全的新利器。大数据已成为学术界、产业界和政府共同关注的热点，正在开启信息化发展的新阶段。

二、大数据带来信息化第三波浪潮

回顾信息化的发展历程，我们已经经历了两次高速发展的浪潮。20世纪40年代第一台电子计算机出现到20世纪80年代之前，计算机价格昂贵、体积巨大、能耗可观，仅应用在国防、气象和科学探索等领域。20世纪80年代，随着个人计算机的大规模普及应用，第一次信息化浪潮到来，可总结为**以单机应用为主要特征的数字化阶段**（信息化1.0）。这一波浪潮中，信息技术褪去神秘的面纱，走入普通民众的视野，并开始广泛应用到其他领域。受这一波信息化影响而最先发生改变的当属办公环境。数字化办公和计算机信息管理系统逐渐取代了纯手工处理，人类第一次体会到信息化带来的巨大改变。

从20世纪90年代中期始，以美国提出“信息高速公路”建设计划为重要标志，互联网开始了其大规模商用进程，信息化迎来了蓬勃发展的第二次浪潮，即**以联网应用为主要特征的网络化阶段**（信息化2.0）。透过计算机工作的人们，通过互联网实现了高效的连接，人类信息交互、任务协同的规模得到空前的拓展，空间上的距离不再成为制约沟通和协作的障碍。政府和企业利用互联网促进信息交流与异地协作，从而实现业务流程和资源配置的优化，并大幅提高工作效率和产品（服务）质量。另一方面，越来越多的人通过互联网结识好友、交流情感、表达自我、学习娱乐，人们开启了在信息空间中的数字化生存方式。互联网快速发展及延伸，加速了数据的流通与汇聚，促使数据资源体量指数增长，数据呈现出海量、多样、时效、低价值密度等一系列特征，这一现象即被称为“大数据”。

当前，信息化建设的第三波浪潮正扑面而来，信息化正在开启**以数据的深度挖掘和融合应用为主要特征的智能化阶段**（信息化3.0）。随着互联网向物联网（含工业互联网）延伸而覆盖物理世界，“人机物”三元融合的发展态势已然成型，除了人类在使用信息系统的过程中产生数据以外，各种传感器、智能设备也在源源不断地产生数据，并逐渐成为数据最重要的来源。近年来，数据资源的不断丰富、计算能力的快速提升，推动数据驱动的智能快速兴起。大量智能应用通过对数据的深度融合与挖掘，帮助人们采用新的视角和新的手段，全方位、全视角展现事物的演化历史和当前状态，掌握事物的全局态势和细微差别；归纳事物发展的内在规律，预测预判事物的未来状态；分析各种备选方案可能产生的结果，从而为决策提供最佳选项。当然，第三次浪潮还刚刚开启、方兴未艾，大数据理论和技术还远未成熟，智能化应用发展还处于初级阶段。然而，聚集和挖掘数据资源，开发和释放数据蕴含的巨大价值，已经成为信息化新阶段的共识。

纵观信息化发展的三个阶段，数字化、网络化和智能化是三条并行不悖的主线。数字化奠定基础，实现数据资源的获取和积累；网络化构造平台，促进数据资源的流通和汇聚；智能化展现能力，通过多源数据的融合分析呈现信息应用的类人智能，帮助人类更好认知事物和解决问题。三个阶段的“数字化”又各有其特色和重点。信息化的第一阶段是从具有广泛需求且与个人计算机能力最为匹配的办公环境起步，如文字

处理、人事财务物资管理等，大量与组织业务相关的纸质文档和表格被转移到计算机可读的数字化媒体中，“**办公数字化**”是这个阶段的重点。在第二阶段，通信带宽不断增长、覆盖范围日益广泛的互联网成了信息化的基础平台，各种信息系统纷纷接入互联网并与其他系统交换数据。人们不仅依靠互联网协同工作，也借助互联网开展生活中的几乎一切活动，信息化场景从办公室拓展到整个人类社会。人类积累的数据不再仅限于结构化的业务数据，无结构的文本、图片、音视频等用户生成内容（user generated content，UGC）在总量上占比日益增加，数据呈现结构化、非结构化并存，并通过网络大规模交换、共享和聚集的态势。这个阶段的重点可归为“**社会数字化**”。信息化进入新阶段，数字化的重点将是“**万物数字化**”，越来越多物理实体的实时状态被采集、传输和汇聚，从而使数字化的范围蔓延到整个物理世界，物联网数据将成为人类整体掌握的数据集中最主要的组成部分，海量、多样、时效等大数据特征也更加突出。

需要进一步说明的是，在第二阶段，网络化的重点平台是互联网和移动互联网，而在当前的新阶段，网络化的重点平台将是面向各行各业、面向物理世界各类实体的物联网。智能化作为刚刚开启的信息化新阶段的主要特征，通过各类智能化的信息应用帮助人们判断态势、预测趋势和辅助决策，当前仍处于起步期，其本质上还是数据驱动的智能。相信随着信息科技的不断进步，信息应用智能化程度的不断提升，数据资源蕴藏的巨大能量将得以不断释放，从而惠及人类社会。

连续三波浪潮，信息化已经广泛并深刻地影响和改变了人类社会！特别是过去的二十年，以互联网为核心的信息技术深度渗透到现有经济体系中，打乱了原有的社会结构，并逐渐编织起新的工业网络，建立新的基础设施，扩散新的和先进的思维模式和行事方法。二十余年的积累和储备，数据资源大规模聚集，其基础性、战略性凸显。当前，信息技术正从助力经济发展的辅助工具向引领经济发展的核心引擎转变，一种新的经济范式——“数字经济”正在逐渐成形，即将进入信息技术带动经济发展的爆发期和黄金期。数字经济是指以使用数字化的知识和信息作为关键生产要素、以现代信息网络作为重要载体、以信息通信技术的有效使用作为效率提升和经济结构优化的重要推动力的一系列经济活动，是以新一代信息技术和产业为依托，继农业经济、工业经济之后的新经济形态。

三、大数据挑战现有信息技术体系

大数据的出现对现有信息技术体系及产业发展均带来了一系列挑战，诸如：大数据应用需求驱动计算技术体系的重构；数据的交换汇聚需求驱动网络通信向宽带、移动、泛在发展；渐进式技术发展面临极限，亟需基础原理突破；软硬件开源开放催生产业发展新生态等。从计算技术的视角，可观察到如下方面的技术挑战。

在大数据存储与管理方面，梳理传统关系型数据库管理系统的发展，其追求目标大致可归为三个方面：应用的通用性，即凝练沉淀共性的数据管理功能以支持不同的上层应用；数据的一致性，即要求事务执行满足ACID特性（原子性、一致性、隔离性、持久性）；系统的高性能，即查询响应时间快、数据吞吐量高、系统可扩展强。随着大数据的出现，传统的数据库管理系统在上述三个方面都面临巨大挑战。如大数据的特性和应用的多样需求在处理方式和响应时间等方面的不同使得很难有一种通用的数据管理方式应对所有场景；由于数据类型和应用需求的多样化，使得系统不能事先定义数据模式，因此不能有效支持事务特性及数据一致性；关系型数据库中数据通常采用表格存储，对大规模多表关联查询及复杂分析类型的SQL查询，查询性能严重下降。

目前，以NoSQL（非关系型数据库）和NewSQL为代表的新型数据库管理模式正在快速发展，其中NoSQL主要解决大规模数据集和多数据种类带来的挑战，数据模式简单，不保证数据一致性。针对不同数据类型和应用有不同的NoSQL系统方案，如键值（Key-Value）存储数据库、列存储数据库、文档型数据库、图（Graph）数据库。NewSQL是指新的可扩展/高性能数据库，通常采用分布式系统架构，利用基于内存SQL引擎和轻量级的事务支持等来提高性能，保持传统数据库支持ACID和SQL等特性，是一个值得期待的发展方向。

大数据的出现对数据的存储、管理的众多环节提出了全方位的挑战，现有的数据管理和处理技术很难满足大数据的需求。

在大数据分析方面，目标是充分挖掘和利用隐藏于数据中的价值。大数据的出现给现有的数据分析技术带来了新的变化和挑战，表现在：（1）数据分析对象的改变。传统的数据分析方法通常需要对数据进行清洗和预处理，而在大数据环境下，由于缺少了数据预处理的时间，需要对包含大量噪声的原始数据进行分析；传统分析方法通常需要对数据进

行采样和特征筛选处理，而在大数据环境中，数据采样和特征筛选会造成数据价值的损失，因此需要对未经采样和特征筛选的全数据进行分析；传统方法通常针对单一数据源进行分析，而大数据环境下，则需要对更大时空范围内的多数据源进行协同分析，以便发现或逼近数据的本质规律。（2）数据分析需求的改变。大数据为数据分析提供了更加逼近现实世界的数据资源，人们对数据分析结果精确性的预期也随之不断提升；同时，人们也更加期待对数据的深层特征和复杂关联关系进行分析，而不仅仅是对数据的表层特征和直观联系进行分析。（3）数据分析模型能力的变化。大数据环境中数据种类和特征维度都有明显的增多，传统基于低维数据的分析模型在数据表达能力上的限制也越来越明显，而基于高维数据的分析模型则越来越受重视；大数据环境下，要求分析模型能够在更多的场景和数据条件下有效，因此要求模型具有更强的泛化能力；随着分析数据量和复杂度的提升，模型的复杂程度也在不断提升，随之，对计算能力的要求也在不断提升。

大数据的挑战推动了数据分析方法的改进，也促动了新方法和技术的出现。例如，以深度学习为代表的数据分析技术近年来产生了较大的飞跃，众多适用于大数据分析的深度学习模型应运而生，如深度卷积神经网络、深度信念网络、深度自编码网络等，这些技术更适于对原始数据进行分析，具有更强的抽象特征提取和复杂数据关系分析能力。大数据的出现，一方面为深度学习方法提供了必备的训练数据资源，另一方面也为新的深度学习方法提供了展现其价值的应用场景。数据可视化技术也在大数据的促动下快速发展，它利用图形对大数据进行多维呈现，帮助人们理解大数据中蕴含的规律。

在大数据处理方面，传统计算技术架构有可能会面临“革命性”的重构。由于单台计算设备能力的局限，面向海量数据的处理需求，并行就成了不二选择。在数据的并行处理方面，根据大数据的特性和应用需求的不同，当前存在多种不同的大数据并行计算模型。例如，批处理计算模型数据吞吐率较高，适用于海量预存数据的成批处理，典型系统包括支持MapReduce模型的Hadoop平台和支持内存计算模式的Spark平台；流处理计算模型处理时延较短，适用于产生速度快并需及时处理的实时数据流，代表性系统是Storm和S3平台；混合计算模型能够综合批处理与流处理的优点，然而却具有较高的系统复杂度，不易部署和实施，

目前仅Yahoo和Metamarkets等较少企业采用此模型构建数据处理系统；图处理模型则适合处理具有亿万个顶点的大规模图数据，典型系统包括Pregel、Graphlab、X-Stream等。当前，不存在一种具有普遍适用性的并行处理模型和系统。基于“软件定义”技术，具有可伸缩的存储和计算能力的云计算平台，正在成为大数据并行处理领域广受青睐的一个关注点。软件定义通过硬件资源虚拟化和管理功能可编程，实现对云计算平台的灵活配置，将统一的平台“定义”为符合不同需求的虚拟平台，从而满足特定大数据处理应用的个性化需求。

四、亟待构建大数据治理体系

大数据一方面给现有信息技术体系带来了系列挑战，需要研发投入和创新发展，另一方面，还需要营造有利于大数据产业健康有序发展的良好环境，为此，大数据治理的概念受到关注，成为大数据产业生态系统的新热点。

近年来，围绕大数据治理这一主题及其相关问题，国际上已有不少成功的实践和研究探索工作，诸如在国家层面推出的促进数据共享开放、保障数据安全和保护公民隐私的相关政策和法规，针对企业机构的数据管理能力评估和改善，面向数据质量保证的方法与技术，促进数据互操作的技术规范和标准等。然而，纵观当前的研究和实践，仍存在三个方面的主要问题。

一是大数据治理概念的使用相对“狭义”，研究和实践大都以企业组织为对象，仅从一个组织的角度考虑大数据治理的相关问题。有的从大数据类型、行业领域、治理科目等维度定义大数据治理框架，指导企业制定相应的大数据治理计划；有的从原则、范围、实施与评估等维度来规范企业的大数据治理工作。然而，大数据治理的范围仅限于组织内部显然是不够的，多源数据聚集和跨组织、跨领域的数据深度融合挖掘是展现大数据价值的前提，在价值驱动下，各界普遍存在着数据突破组织边界流动的需求。而且，随着数据开放和流通技术及渠道的逐步完善，数据跨组织流动和应用已经发生，并呈现日益普遍的趋势。无疑，这将是涉及行业内和跨行业、区域内和跨区域、全国乃至全球的多个层次，企业组织的大数据治理离不开行业的规范和自律、国家的“上位法”，甚至国家间的约定或协议，多层次协同才可能构成大数据生态建设的基础

性保障。

二是现有研究实践对大数据治理内涵的理解尚未形成共识。不同的人结合大数据的特征，从企业业务和管理流程设计、组织信息治理规划、组织数据管理与应用等不同的视角，给出了大数据治理的不同定义。如有人认为大数据治理是传统IT治理的一部分，有人则认为大数据治理与IT治理独立，是数据管理概念的延伸与扩展。有的将大数据治理定位为描述数据如何在其全生命周期内有用及其经济管理的组织策略或程序，有的定位于企业对数据可获性、可用性、完整性和安全性的措施及其全面管理，有的则着重于制定与大数据有关的数据优化、隐私保护与数据变现的政策。共识的形成尚有待时日！

三是大数据治理相关的研究实践多条线索并行，关联性、完整性和一致性不足。诸如，国家层面的政策法规和法律制定等较少被纳入大数据治理的视角；数据作为一种资产的地位仍未通过法律法规予以确立，难以进行有效的管理和应用；大数据管理已有不少可用技术与产品，但还缺乏完善的多层级管理体制和高效管理机制；如何有机结合技术与标准，建立良好的大数据共享与开放环境仍需要进一步探索；除了不断完善发展相关技术以应对各种新型攻击挑战外，企业安全保障制度、行业自律监管机制和国家通过法律确定的强制手段还有待完善；缺乏系统化设计的已有相关体系的扩展和延伸可能导致数据治理的“碎片化”和一致性缺失等。

大数据治理必须跳出单个组织的边界，从营造国家大数据产业发展环境的视角予以全面、系统化考虑！大数据治理体系建设涉及国家、行业和组织三个层次，包含资产地位确立、管理体制机制、共享开放、安全与隐私保护等四项内容，需要从制度法规、标准规范、应用实践和支撑技术四个方面多管齐下，提供支撑。

国家层次：需要在法律法规层面明确数据的资产地位，奠定数据确权、流通、交易和保护的基础；需要兼顾现状及发展，建设适合国情的良好的数据管控协调体制和相应的管理机制；需要制定促进数据共享开放的政策法规和标准规范，实现政府部门间的数据共享，规范市场主体间的数据流通和交易，建设政府主导的数据开放平台，促进政务数据和行业数据的融合应用；需要出台数据安全与隐私保护的法律法规，保障国家、组织和个人的数据安全。

行业层次：行业大数据治理应在国家相关法律法规框架下，充分考虑本行业中企业的共同利益与长效发展，构建相应的行业大数据治理规则。需要建立规范行业数据管理的组织机构，制定行业内的数据管控制度；需要制定行业内数据共享与开放的规则和技术规范，构建行业数据共享交换平台，为本行业企业提供数据服务，促进行业内数据的融合应用；需要制定行业内数据安全保障制度，确保行业内每个成员单位的数据安全、权益和商业秘密。

组织层次：需要通过组织内部规章将数据确定为其核心资产，以利于有效管理和应用；需要建立适应数据资源完善、价值实现、质量保证等方面的组织结构和过程规范，提升企业对数据全生命周期的管理能力；需要促进企业内部部门间的数据共享，并加强对外的数据流通和交易，充分盘活数据价值；需要结合“上位法”及自身的管理和技术措施，保障企业自身的数据安全及客户的数据安全和隐私信息。

在这个大数据治理体系框架中，三个层次相互关联和支撑。国家层次制定大数据治理的“上位法”，指导和监管行业及组织的大数据治理；行业层次通过行业自治的模式，在自愿原则上形成行业协会或联盟等，作为政府和企业之间的桥梁，在国家法规和政策的指导下，制定并执行行规行约和各类标准，监督企业的行为，并向政府传达企业的共同需求；组织则在国家和行业大数据治理的框架下，针对自身的特点，确定大数据治理的目标，优化对大数据资源的管理，最大化从大数据获得的收益，并为行业和国家大数据发展贡献成功应用实践。大数据治理体系的建设将为营造大数据产业发展的良好环境提供基本遵循。

五、关于本书

《大数据导论》一书是大数据系列教材中的第一本，其目的是从技术层面，为大数据相关专业本科生、研究生及科研人员，提供一本全面介绍大数据相关技术的专业通识教材。本书系统地介绍大数据涵盖的内容，包括数据与大数据概述、大数据感知与获取、大数据存储与管理、大数据分析、大数据处理、大数据治理、大数据安全与隐私等。除了介绍大数据的技术内容，本书还介绍了部分行业中大数据的典型应用案例，反映了大数据在社会经济生活中的重要价值。本书既可作为普通高等学校大数据相关专业的教材使用，也可供专业技术人员参考。

本书共分八章。

第一章首先扼要综述从数到数据到大数据的发展脉络，然后从大数据的多边定义和理解出发，针对不同的价值期望，从应用逻辑、工程管理和实施逻辑三个角度介绍了实践可行的系列方法、思路和策略。最后系统地梳理了大数据生命周期、技术图谱以及本文后续涉及的内容介绍。

第二章首先介绍了数据源的一般分布以及通常意义下的不同数据源的获取方法和策略，然后重点介绍了ETL、表面网数据获取、深网数据获取等关键技术和方法。

第三章详细介绍了大数据存储与管理涉及的分布式文件系统及主流技术HDFS、分布式数据库中的典型技术HBase、非关系型数据库NoSQL、云数据库及大数据分析平台等。

第四章首先从对数据的理解和特征提取谈起，然后重点讲述了基于机器学习的数据分析方法，包括有监督方法、无监督方法等内容。然后，从人工神经网络进展到深度学习算法，对当前主流的卷积神经网络、循环神经网络和对抗式生成网络等深度学习模型进行了介绍，最后，对可视化分析在大数据不同领域的应用以及常用的大数据可视化工具进行了介绍。

第五章首先介绍了传统的集中式计算架构，重点介绍了超级计算机的发展历史和大数据处理特点，然后着重介绍了MapReduce和Spark两种分布式计算架构和近年来出现的针对流式数据的计算架构，最后介绍对大数据计算进行加速的三种技术：GPU、TPU和FPGA。

第六章介绍了大数据治理和管理中的一系列关键的内容，包含大数据治理的基本概念、数据架构管理的定义和参考模型、元数据管理的概念和作用、主数据管理的概念、主数据的架构和应用、数据质量管理的基本技术、数据标准化相关的概念和应用实例、数据资产化方面的内容。

第七章从大数据安全、大数据隐私保护等方面介绍了大数据的安全与隐私问题，然后介绍了大数据技术如何在安全问题中进行应用，最后就大数据安全技术的发展进行了分析与展望。

第八章从大数据在企业营销、交通旅游、物流供应和教育教学等方面分别介绍大数据的应用与意义，这些大数据的应用案例，为人们的生活和国家经济发展带来新的机遇。

参与本书撰写的人员包括：王元卓、王崇骏、方粮、彭绍亮、冯建