“十二五”国家重点图书出版规划项目

大数据技术与应用

丛书主编
朱扬勇 吴俊伟

Big Data
Technology and Application Series

张绍华 潘 蓉 宗宇伟
主编

大数据治理与服务

上海科学技术出版社

大数据技术与应用

大数据治理与服务

张绍华　潘　蓉　宗宇伟
主编

上海科学技术出版社

图书在版编目(CIP)数据

大数据治理与服务 / 张绍华，潘蓉，宗宇伟主编.
—上海：上海科学技术出版社，2016.1
（大数据技术与应用）
ISBN 978-7-5478-2838-0

Ⅰ.①大… Ⅱ.①张… ②潘… ③宗… Ⅲ.①数据管理—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 247648 号

大数据治理与服务
张绍华 潘 蓉 宗宇伟 主编

上海世纪出版股份有限公司
上海科学技术出版社 出版
（上海钦州南路 71 号 邮政编码 200235）
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路 193 号 www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092 1/16 印张 16.25
字数 420 千字
2016 年 1 月第 1 版 2016 年 1 月第 1 次印刷
ISBN 978-7-5478-2838-0/TP·37
定价：65.00 元

内容提要

本书从大数据治理的基本概念和现状出发，提出了大数据治理的框架及治理的关键要素，分析了大数据环境下企业面临的挑战、战略转型、组织职能分配，创造性地提出大数据架构，介绍了大数据环境下的数据质量、数据安全特点和应对方案，以及基于数据生命周期的风险、特点和管理方案，最后给出了大数据治理实施的方法论和基于服务的大数据治理价值展现。

本书立足于大数据环境下的数据治理，既有治理视角的战略价值、风险合规，也有管理视角的数据资产、数据服务，模型与案例结合，条理清晰，易于使用，适合高校师生、企业 IT 人员、数据治理从业人员阅读，也可供高层决策人员参考。

大数据技术与应用

学术顾问

大数据技术与应用

编撰委员会

本书编委会

主　编

张绍华　潘　蓉　宗宇伟

编　委

范颖捷　薛君傲　郑大庆　湛家扬　宋俊典　杨　琳　刘　晨
李　鸣　张明英　郑晨光　车春雷　叶俊峰　杨泽明　俞文平
尹立庆　宋跃武　刘小茵　梁育刚

丛书序

我国各级政府非常重视大数据的科研和产业发展，2014 年国务院政府工作报告中明确指出要“以创新支撑和引领经济结构优化升级”，并提出“设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展”。2015 年 8 月 31 日，国务院印发了《促进大数据发展行动纲要》，明确提出将全面推进我国大数据发展和应用，加快建设数据强国。前不久，党的十八届五中全会公报提出要实施“国家大数据战略”，这是大数据第一次写入党的全会决议，标志着大数据战略正式上升为国家战略。

上海的大数据研究与发展在国内起步较早。上海市科学技术委员会于 2012 年开始布局，并组织力量开展大数据三年行动计划的调研和编制工作，于 2013 年 7 月 12 日率先发布了《上海推进大数据研究与发展三年行动计划（2013—2015 年）》，又称“汇计划”，寓意“汇数据、汇技术、汇人才”和“数据‘汇’聚、百川入‘海’”的文化内涵。

“汇计划”围绕“发展数据产业，服务智慧城市”的指导思想，对上海大数据研究与发展做了顶层设计，包括大数据理论研究、关键技术突破、重要产品开发、公共服务平台建设、行业应用、产业模式和模式创新等大数据研究与发展的各个方面。近两年来，“汇计划”针对城市交通、医疗健康、食品安全、公共安全等大型城市中的重大民生问题，逐步建立了大数据公共服务平台，惠及民生。一批新型大数据算法，特别是实时数据库、内存计算平台在国内独树一帜，有企业因此获得了数百万美元的投资。

为确保行动计划的实施，着力营造大数据创新生态，“上海大数据产业技术创新战略联盟”（以下简称“联盟”）于 2013 年 7 月成立。截至 2015 年 8 月底，联盟共有 108 家成员单位，既有从事各类数据应用与服务的企业，也有行业协会和专业学会、高校和研究院所、大数据技术和产品装备研发企业，更有大数据领域投资机构、产业园区、非 IT

领域的数据资源拥有单位，显现出强大的吸引力，勾勒出上海数据产业的良好生态。同时，依托复旦大学筹建成立了“上海市数据科学重点实验室”，开展数据科学和大数据理论基础研究、建设数据科学学科和开展人才培养、解决大数据发展中的基础科学问题和技术问题、开展大数据发展战略咨询等工作。

在“汇计划”引领下，由联盟、上海市数据科学重点实验室、上海产业技术研究院和上海科学技术出版社于2014年初共同策划了《大数据技术与应用》丛书。本丛书第一批已于2015年初上市，包括了《汇计划在行动》《大数据评测》《数据密集型计算和模型》《城市发展的数据逻辑》《智慧城市大数据》《金融大数据》《城市交通大数据》《医疗大数据》共八册，在业界取得了广泛的好评。今年进一步联合北京中关村大数据产业联盟共同策划本丛书第二批，包括《大数据挖掘》《制造业大数据》《航运大数据》《海洋大数据》《能源大数据》《大数据治理与服务》等。从大数据的共性技术概念、主要前沿技术研究和当前的成功应用领域等方面向读者做了阐述，作者希望把上海在大数据领域技术研究的成果和应用成功案例分享给大家，希望读者能从中获得有益启示并共同探讨。第三批的书目也已在策划、编写中，作者将与大家分享更多的技术与应用。

大数据对科学研究、经济建设、社会发展和文化生活等各个领域正在产生革命性的影响。上海希望通过“汇计划”的实施，同时也是本丛书希望带给大家一个理念：大数据所带来的变革，让公众能享受到更个性化的医疗服务、更便利的出行、更放心的食品，以及在互联网、金融等领域创造新型商业模式，让老百姓享受到科技带来的美好生活，促进经济结构调整和产业转型。

上海市科学技术委员会副主任

2015年11月

序（一）

当前，随着以互联网、物联网为代表的信息技术的不断发展和应用，大数据如幽灵般来到了我们的身边。与一个组织相关的日常经营活动、客户/供应商等合作伙伴的行为和活动、消费者的行为轨迹和生活片段、外部自然环境、政治经济环境等无时无刻不在被数字化、被记录下来，形成了大量不断堆积的“数据面包屑”。这些“数据面包屑”被收集、整理、分析、加工处理，其结果可能会无限接近真实的世界。正如徐宗本院士所说：“大数据是指反映真实世界的数据，其量已达到可以从一定程度上反映真实面貌的程度。”

展望大数据时代，数据必将成为除了人力、土地、财务、技术之外的另一种重要的资产。作为一种资产，企业利用大数据，可以更加敏锐地感知周边的变化，更加深邃地洞察客户/消费者以及合作伙伴们的行为和变化趋势，更加精准地优化企业的运营，更加和谐地和商业伙伴一起开展协同创新。大数据正在重塑企业，重新定义行业，正成为跨界的驱动力。当然，大数据更大的商业价值尚未得到体现，有待更多的企业去挖掘、去发现。数据作为一种资产，需要与其他的资产相互组合、相互补充，或此消彼长，或相得益彰。企业需要对数据资产进行管理，直面由此带来的数据治理这一重大课题。

半年以前，在一次有关大数据的行业论坛上，我提到了大数据治理的问题，当时就听说作者团队在编写《大数据治理与服务》一书，我一直期待着看到这本书。今天拿到全书的样稿，内心非常高兴。首先，这本书是国内实践者和学者在这个方面进行的第一次有意义的尝试。这本书的作者团队不但牵头参与国内、国际数据治理标准的研制，而且具有丰富的数据治理和大数据行业应用实践经验，融合了国内外在数据治理、IT 治理、大数据应用等方面的最新实践成果，从梳理最基本的概念开始，建立起一个自洽的大数据治理框架，并且从三个方面——大数据治理关键域、大数据治理的实施和监督、

大数据服务，全面阐述了大数据治理需要关注的问题。

其次，本书对当前大数据治理中的一些关键领域，如大数据组织、大数据架构、大数据安全、大数据环境下的隐私保护、大数据合规管理、大数据质量管理、大数据服务管理等均提出了作者们的看法。这些重大的问题，目前无论在理论上还是实践中，均有待深入研究和探索。尽管如此，本书这部分内容从具体操作的层面给读者带来一些思考和启示。

最后，整本书的逻辑结构合理，既包括了一般性的大数据治理工作的描述，也包括从项目实施层面讨论大数据治理涉及的实际问题，以及从大数据服务提交的角度阐述的重点问题。我相信这本书对于组织的高层管理人员，以及从事大数据治理的专业人员都具有一定的借鉴意义。也殷切地希望本书的作者团队能够再接再厉，在中国大数据治理的领域不断探索，走出一条符合中国特色的大数据治理之路，帮助企业尽快地把散乱在各处的“数据面包屑”加工成醇香可口、营养丰富的“数据面包”，真正从大数据中获取最大的利益。

复旦大学管理学院教授
国务院学位委员会管理科学与工程学科评议组成员
教育部电子商务教学指导委员会成员
黄丽华
2015 年 11 月 1 日

序（二）

推荐这本书给大家，我觉得是自己作为大数据行业一份子的责任。现在很多人都在谈大数据，其中有传统行业，有银行，也有医院。但我注意到大部分企业都在关注如何用数据进行创新，却很少听到大数据作为原材料应该怎么管理。你可能会说，银行业、通信业等不是早就在做数据管理了吗？的确数据管理并不新鲜，20年前就有人在做了。但大数据的含义不仅指数据的大小，还包括数据内容的广泛来源、非结构性及实时连接性等。大数据的定义其实在不断更新中。我们不禁会问，以往的数据管理思路能适应新形势的需求吗？我敢大胆地说，自上而下的管理方法已经过时了。大数据的本质就是来自开放的力量、频繁的数据更新、更丰富的数据种类、更快速的数据流动，但这些都对中央式的管理方式造成了极大的挑战。我们必须意识到，数据治理不等同于数据管理，绝非仅依靠自上而下的贯彻执行便可解决。相反，数据治理需要每个人的参与和协同，要求大家都有意识去治理好数据，做到“人人为我，我为人人”。今天不把数据管好，日后对数据的依赖愈深，便愈容易出现问题。数据治理的新思路，不仅是指组织结构上要从由上而下变成全体协同，而且要在技术上创新，用数据去助力大数据治理，帮助大家提高数据质量，保护数据安全，以及有效控制数据成本。

最后我想说，大数据时代的数据治理，一定是将无形的管理策略化成有形的工作流程，从一纸命令变成根植在每个人心中的信念和下意识的习惯。我们要用大数据的思维方式，用数据治理数据。还要感谢作者对数据治理的坚持，这本书得来不易。

阿里巴巴集团副总裁、数据委员会会长

车品觉

前 言

大数据是移动互联网、云计算和物联网等技术发展的必然趋势，是分析决策方式、科学研究范式和创新思维模式的重要突破，已经渗透到各行业和应用领域，成为组织发展的生产因素和未来竞争的核心要素，必将引领新一轮信息技术产业的发展和新一波生产率增长浪潮的到来。

从数据中发现问题到解决问题，从业务支撑到业务创新，从商业智能到指引决策，数据与业务相伴相生，数据带来的机遇与风险共存，大数据治理的需求应运而生。大数据治理不仅关注大数据相关的战略、组织和架构，也关注安全、隐私与合规以及业务过程中的数据质量，更关注大数据的价值和价值实现蓝图。

本书在中国 IT 治理标准和数据治理标准研制以及金融、电信和互联网等行业应用实践分析的基础上，融合了 ISO38500、COBIT、DAMA、DGI、IBM、CMMI 等国内外研究成果，针对大数据治理这一崭新的领域，构建了大数据治理模型和治理域，提出了面向大数据生命周期的治理实施方法，探讨了大数据治理审计、大数据服务与创新等，希望能从治理的角度在大数据的研究、应用和服务领域为读者呈现一个崭新的视角。

本书分为十章，从基本概念和模型框架的描述，到数据治理的通用要素和实施过程的分析，最后通过大数据服务实现大数据的价值创造。大数据战略章节通过数据支持到驱动的战略路径，结合案例分析了如何从战略层面帮助企业转型，并列举了常见的组织类型。大数据架构章节提出了不同视角下的参考模型，不仅涉及 IT 技术的架构，还包括流程化管理与工具等。大数据安全章节囊括了合规、风险、数据安全，并附上了大量安全趋势与工具介绍。大数据质量章节从大数据的特性出发，重新定义了质量的概念，并指出了大数据环境下的质量管理与小数据环境下的质量管理的异同，通过案例介绍了数据质量管理的方法。大数据生命周期章节根据不同阶段的管理特点，给出了具

第1章

大数据治理概述

在大数据与IT环境相互融合的大趋势下，数据治理的体系和方法发生了深刻的变化，数据治理的理论和实践正在向大数据治理聚焦。本章首先介绍了大数据治理的相关概念，以及概念间的相互关系；然后，在综述数据治理理论和实践进展的基础上，讨论了大数据治理这一发展新趋势；最后，阐述了大数据治理的意义和作用。

1.1 大数据治理相关概念

大数据治理包含很多相关概念，概念之间存在比较复杂的关系。厘清这些概念和关系对于理解后面章节中大数据治理的框架、关键域、实施等核心内容是十分必要和重要的。图1-1展示了大数据治理相关概念的逻辑关系和演化路径。本章将按自下而上、从左到右的顺序逐一对概念和概念组进行介绍、比较和分析。

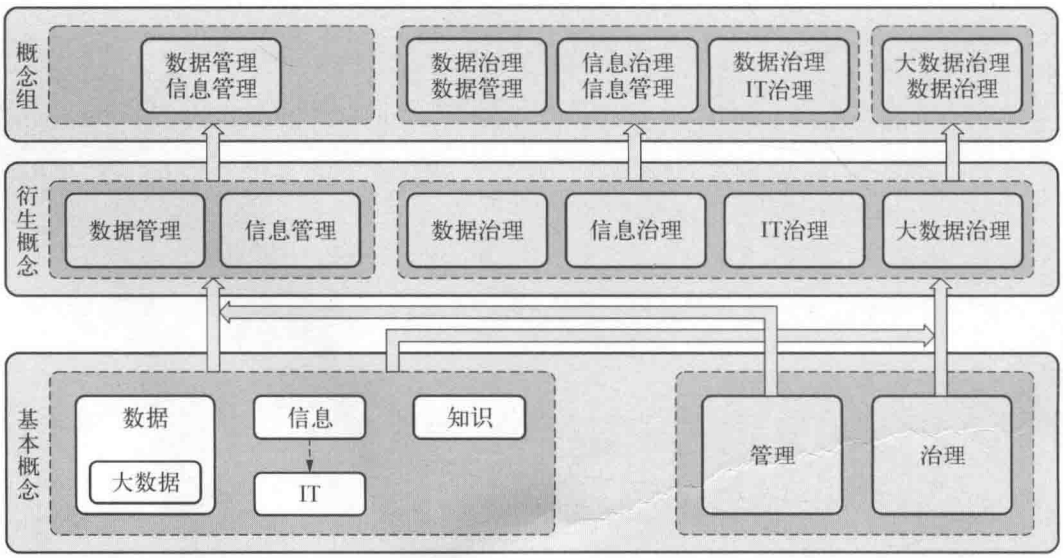

图1-1 大数据治理相关概念的关系

1.1.1 背景知识

1.1.1.1 数据、信息与知识

数据是客观事实经过获取、存储和表达后得到的结果，通常以文本、数字、图形、图像、

声音和视频等表现形式存在[1]。

信息(Information)是包含上下文语境的数据(Data with Context),没有上下文的数据是毫无意义的,人们通过解释上下文来创造有意义的信息。元数据(Metadata),即描述数据的数据(包括数据的各种属性和描述信息),可以帮助创建上下文,所以管理元数据对提高信息质量有直接帮助。上下文通常包括[2]:

(1) 数据元素和相关术语的业务含义。

(2) 数据表达的格式。

(3) 数据所处的时间范围。

(4) 数据与特定用法的相关性。

知识是对情境的理解、意识、认知和识别,以及对其复杂性的把握。知识的获取涉及许多复杂的过程:感知、交流、分析和推理等,它可能是关于理论的,也可能是关于实践的[3]。知识是构成人类智慧的最根本因素。

上面的概念比较抽象,下面举个小例子。例如,从超市买了一瓶15元的酸奶,瓶上标明了酸奶的价格、容量、成分和保质期等。单拿"15"来说,它是一个数值型数据,因为没有上下文语境,所以没有任何意义,但加上"一瓶酸奶的价格"这个上下文后,"15"就变成这瓶酸奶的价格,它就成为一个有意义的信息。接下来,发现酸奶的成分中标有双歧杆菌,同时知道它是肠道有益菌,这两个信息经过分析处理就得到一个知识:常吃酸奶有益于肠道健康。

数据、信息和知识的关系就蕴含在上面的概念表述中,总结如下:信息是一种特殊类型的数据,数据是信息的基本构成元素;知识是一种特殊类型的信息,信息是知识的基本构成元素;信息和知识本质上都是数据,数据是信息和知识的基本构成元素和基础,如图1-2所示。数据、信息与知识的概念与关系见表1-1。

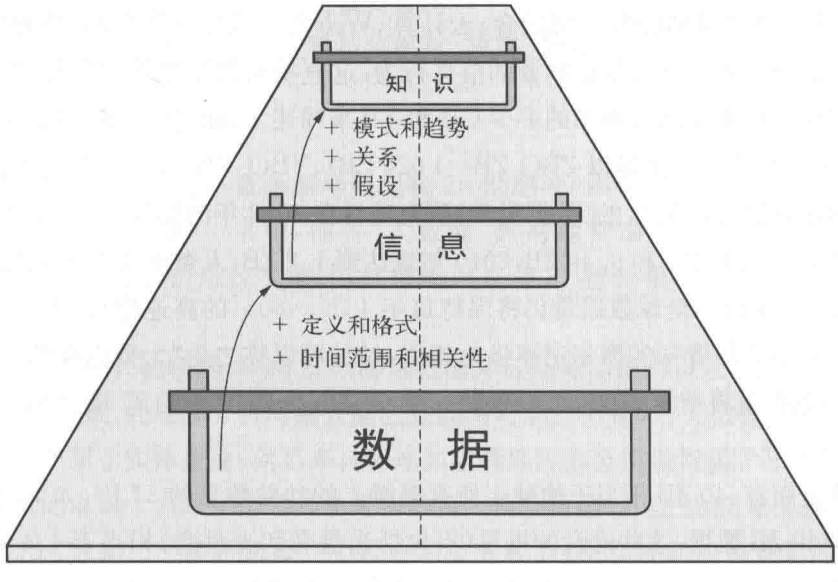

图1-2 数据、信息与知识的关系

表 1-1 数据、信息与知识的概念与关系

类 别		要 点
概念	数据	(1) 数据是客观事实经过获取、存储和表达后得到的结果； (2) 数据可能以多种形式存在，如文本、数字、图形、图像、声音和视频等
	信息	(1) 信息是包含上下文语境的数据； (2) 人们通过解释上下文来创造有意义的信息； (3) 上下文包括数据元素的业务含义、格式、时间范围和相关性
	知识	(1) 知识是对情境的理解、意识、认知、识别，以及对其复杂性的把握； (2) 知识是构成人类智慧的最根本因素
关系	数据、信息与知识	(1) 信息是一种特殊类型的数据，数据是信息的基本构成元素； (2) 知识是一种特殊类型的信息，信息是知识的基本构成元素； (3) 信息和知识本质上都是数据，数据是信息和知识的基本构成元素和基础

上述概念及其关系说明了数据对于人类社会发展的极端重要性，以及所起到的基础性作用。当今是信息经济时代，数据已成为一个企业、机构、政府乃至国家的宝贵资产。如果一个企业没有高质量的数据，并且不能理解“管理数据就像管理有形资产一样重要”，那么它就很难做出正确、及时和有前瞻性的决策，效率和效益无从谈起，市场竞争力也必将受到严重削弱。

1.1.1.2 大数据

1) 大数据时代来了

近年来，随着以电子商务、社交网络、位置服务为代表的新型信息发布方式的不断涌现，以及移动互联网、物联网、三网融合、云计算、Web 2.0 等技术的兴起，各种终端设备每时每刻都在记录着人类社会复杂频繁的信息行为，这直接引发了数据的爆炸式增长。现在数据量的增长已经不是以所熟知的多少 GB 和 TB 来描述了，而是以 PB(1 PB=1 024 TB)、EB(1 EB=1 024 PB)，甚至是以 ZB(1 ZB=1 024 EB)、YB(1 YB=1 024 ZB)为计量单位。

根据国际数据公司 IDC 的《数据世界》研究项目在 2012 年的统计[4]，2005 年和 2008 年全球数据量只有 0.13 ZB 和 0.49 ZB，2010 年就达到 1.2 ZB，人类正式进入 ZB 时代。更为惊人的是，2020 年以前全球数据量仍将保持每年 40%～60%的高速增长，大约每两年就翻一倍，这与 IT 界人尽皆知的摩尔定律极为相似，姑且可以称之为“大数据爆炸定律”。预计 2015 年全球数据量将增至 8 ZB，2020 年将达到 40 ZB，是 2010 年的 33 倍、2008 年的 82 倍、2005 年的 307 倍。

单就数量而言，40 ZB 相当于地球上所有海滩上的沙粒数量的 57 倍；如果用蓝光 DVD 保存所有这 40 ZB 数据，这些光盘的重量(不包括光盘套和光盘盒)相当于 424 艘尼米兹级航空母舰(排水量约 10 万 t)；或者相当于世界上每个人拥有 5 247 GB 的数据[5]。无疑，人