大数据丛书系列之七

总主编◎曾　羽　龙奋杰

大数据治理及数据仓库模型设计

DASHUJU ZHILI
JI SHUJU CANGKU MOXING SHEJI

主　编◎曾　凯　高　亮　王新颖

电子科技大学出版社

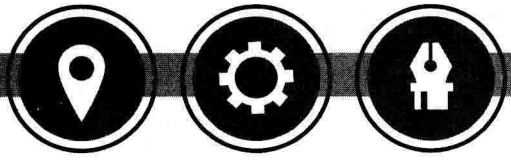
大数据丛书系列之七

总主编◎曾　羽　龙奋杰

大数据治理及数据仓库模型设计

DASHUJU ZHILI
JI SHUJU CANGKU
MOXING SHEJI

主　编◎曾　凯　高　亮　王新颖



电子科技大学出版社

图书在版编目(CIP)数据

大数据治理及数据仓库模型设计 / 曾凯，高亮，王新颖主编. -- 成都：电子科技大学出版社，2017.7
ISBN 978-7-5647-4817-3

Ⅰ. ①大… Ⅱ. ①曾… ②高… ③王… Ⅲ. ①数据库系统－研究 Ⅳ. ①TP311.13

中国版本图书馆CIP数据核字（2017）第177073号

大数据治理与数据仓库模型设计

曾 凯 高 亮 王新颖 主编

策划编辑 杨仪玮 李燕芩
责任编辑 杨仪玮

出版发行 电子科技大学出版社
成都市一环路东一段159号电子信息产业大厦 邮编 610051
主 页 www.uestcp.com.cn
服务电话 028-83203399
邮购电话 028-83201495

印 刷 成都市火炬印务有限公司
成品尺寸 165mm × 240mm
印 张 13.25
字 数 245千字
版 次 2017年7月第一版
印 次 2017年7月第一次印刷
书 号 ISBN 978-7-5647-4817-3
定 价 48.00元

序　言

大数据是当前IT领域较为火热的研究方向。它能够为企业、政府以及社会组织的发展提供决策依据并创造社会价值。目前，大数据已经逐渐应用在各个行业和领域，并被称为新能源。伴随着技术的不断进步，大数据的发展也逐渐出现瓶颈。数据治理技术的出现能够弥补大数据发展的短板，更快更好地推进大数据的实际应用，使其创造更大的价值。当前数据治理已经成为大数据研究较为活跃的企业积极开拓的新领域。

本书立足于数据治理框架体系，从主数据、元数据、数据仓库、数据治理实施等多个维度对其进行详细阐述，并进一步对如何做数据仓库模型进行了讨论。本书可作为大数据方向的科普读物，更可以当作数据仓库实际应用的指导书。具体章节编排为：第一章阐述了数据治理的基本概念和框架，第二章介绍了主数据的概念和应用，第三章介绍了数据治理中元数据的相关内涵，第四章介绍了数据质量的相关内容，第五章介绍了数据库原理，第六章介绍了数据集成的相关背景和内涵，第七章介绍了数据治理实施的方法，第八章介绍了数据治理审计的相关内容，第九章主要介绍了数据治理企业组织团队的相关内容，第十章、第十一章主要阐述了数据仓库概念以及数据仓库模型设计的相关技术和原理。

各章节的作者分别是：曾凯负责第一章、第四章、第八章、第九章、第十一章，王新颖负责第二章、第三章、第十章，高亮负责第五章、第六章、第七章。

在本书的编纂过程中，受到了曾羽教授、龙奋杰教授、蒋学勤教授的指导和建议，在此表示感谢。同时对本书参考资料的作者们表示由衷谢意，本书的完成离不开你们已积累的丰硕成果。本书作者水平有限，疏漏之处在所难免，欢迎广大读者批评指正！

目　录

第一章　数据治理概述

在IT技术不断发展的今天，大数据成为人们讨论和关注的焦点，其根本原因是数据能够为人类创造越来越多的价值。基于大数据的相关技术，推动企业业务乃至整个产业的发展并实现盈利，已经成为信息化建设的重要目标。目前无论新兴的IT行业还是传统领域都在积极开展大数据的技术研究和相关IT建设，但是取得真正意义的成果相对较少，创造的价值还远没有达到大家的预期，大数据的潜力还并没有得到真正的开发。造成这种局面的关键原因是数据质量问题导致很多大数据项目无法得到实施，而数据治理正是数据质量的有力保障手段。在大数据技术发展较为迅速的行业，数据治理已经成为各大企业正在积极开拓的新领域。

1.1　数据治理背景

1.1.1　什么是数据

数据通常以多种形式表达，例如文字、语音、图像、文本等，它们是对客观事物的具体描述并以一定形式存储的结果。

数据本体以及上下文环境描述，就形成了信息。其中用于描述数据问题的上下文环境可以称为元数据（metadata），即用来描述数据的数据。很多情况下，这些数据比数据本体更为重要。例如，一个存储系统中，人们并不会真正和存储介质打交道，我们也并不关心数据具体存在了磁盘哪个扇区，这些用于描述具体文件的信息都由文件系统来管理，一旦这些描述信息（元数据）丢失，则标志着在磁盘上的数据彻底丢失[1]。

从信息中挖掘出来的真正有意义的数据，则可以称为知识，只是有的需要从学习、交流、分析以及推理而获得。因此知识的获取更为复杂，它包含了理论和实践两个过程，也是构成人类智慧的根本因素。

举一个例子就可以很好地阐述数据、信息、知识三者的关系。在学校的数据库系统中存在一个数字“20”，这个数据单独看来并没有任何意义，仅仅是一个数据而已。如果加上其上下文的描述，例如学生数据库的列属性信息标识“年龄”这个字段，则表示这是一个学生的年龄，这些“数据”啮合在一起则组成了信息。我们还可以进一步分析这些“信息”，例如学生数据库里有学生的年龄、性别、选课记录等，我们通过一些数据挖掘的手段，能够分析出20岁左右的男生喜欢选IT相关的课程（如计算机网络、C语言

等），因此学校可以适当增开一些信息技术的选修课来满足学生的需要。“20岁左右的男生喜欢IT课程”这种规则就是我们所说的“知识”。

从上面的分析过程可以看出，数据、信息和知识之间有着必然的联系，其中数据是基本构成元素，而信息和知识从本质上说也都是数据，是更加“高级”的数据，如图1-1所示。

图1-1　数据、信息与知识的关系

随着信息技术的不断发展，数据已经逐渐成为企业、政府乃至国家的新的能源。如果企业部门的数据质量不高，则等同于能源质量不高，那么数据本身就无法真正推动企业的业务发展。

1.1.2　大数据

近些年来，伴随着移动社交、电子商务、O2O等新型IT服务模式的涌现和不断更新，催生了以大数据、云计算、物联网等为代表的技术概念。其中大数据的出现是由于大IT环境下，各种终端、服务端等设备每时每刻都在产生各种数据，这些数据客观描述了人和事物的各种信息。这也导致了另一问题：数据爆炸式增长。早些年人们还认为GB是个很大的数字，这些年以TB（1 TB=1024 GB）、PB（1 PB=1024 TB）甚至ZB（1 EB=1024 PB）来定义数据的规模已经不再是什么新鲜事了，这也标志着大数据时代真正到来了。

各个企业、政府和高校科研团队都对大数据有着不同的定义。一方面，从字面角度来理解，大数据是指数据过于庞大，在传统的单计算节点的环境下无法处理的数据，亚马逊和麦肯锡等国际著名科技公司和机构曾给出过这种简明的定义。另一方面，我们认为大数据不仅规模庞大，更重要的是它应该具有价值，能够为企业和政府提供科学的决策指导和帮助。当前与普通百姓相关的大数据更多来源于大家熟悉的互联网行业，例如微博、微信等各种社交平台，电商购物、视频音乐等内容服务；除此之外，企业各种商业行为也会产生大量的数据。只有将这些数据更加科学地保存和利用，才会为社会创造更多的价值。

严格来说大数据并没有一个十分准确的科学定义，而是对当前海量数据现状以及数据爆炸式增长情形的客观描述。事实上，在10年前，数据规模不断增大，当时的IT手段已逐渐无力应对时，人们更多地以“海量数据”或者“大规模数据”等概念来描述。这就说明当时人们对“大数据”的认识还仅

停留在其本身的数据规模上。而现如今，数据规模庞大依然是主旋律，但是人们提出“大数据”的概念，标志了人们对数据本身有了更加深刻的认识。数据不光能够表达信息，更能够创造价值。在“海量数据”和“大规模数据”时代，还没有能够完全体现出“数据”这种新生能源的作用。因此“大数据”这一概念的提出，立马得到了大家的认同。企业、政府以及科研团队也都纷纷开展了“大数据”或者“数据创造价值”等课题的研究。

面对“大数据”这一新兴学科，我们应首先了解其关键的特征。通常来看，大数据具有四个特征：大量（Volume）、多样（Variety）、时效（Velocity）、价值（Value），简称4V特征。由此可见，数量巨大仅仅是大数据的特点之一。除此之外，大数据还应该具有多种输入源、数据变化快和价值潜力巨大等特点。

（1）大量（Volume）：规模巨大是大数据的首要特点。随着时代的进步，现如今人们所谓数据的“大”，往往是指PB级别以上的数据，如EB、ZB等。导致数据规模日益庞大的原因大概有三个方面。①近些年来互联网行业发展迅速，越来越多的人可以通过终端特别是移动终端参与各种互联网活动。这些由大量人群参与的互联网活动会产生各种各样的行为数据，这些活动如发表微博、点赞、下载视频等。②物联网概念的兴起也成为大量数据产生的源头。物联网指物物相连，这种连接比互联网更加底层。例如RFID、NFC等，这些设备和设备之间的对话也会产生大量的数据，甚至数据规模比互联网更加庞大。③人们对“数据”概念的认识发生变化，根本上导致数据量激增。早些年人们认为数据仅仅能够描述客观事物，因此通常对其生命周期设定较短。例如系统的日志数据仅仅用于表明系统在正常运行，当系统不出故障的情况下则会很快删除掉。而现如今人们已经意识到这些数据还能够通过合理分析和挖掘创造出更多的价值。同时，当今存储技术的发展，其成本不断下降，也为这些数据长久保存创造了基本条件，由此造成了规模巨大的数据出现。

（2）多样（Variety）：大数据类型多样，而且结构差异性较大。这是由于大数据的输入源具有多模性，例如在同一个数据分析的场景下，可能同时存在声音数据、文本数据、图像数据等。从数据的结构上看，我们可以把大数据分成结构化数据（Structured Data）、半结构化数据（Semi-structured Data）和非结构化数据（Unstructured Data）。

结构化数据是较为传统、在小数据时代也经常会用到的数据结构。结构化数据通常以二维表形式表达。在传统的小数据时代，这些数据往往存放于关系型数据库中，而大数据则通常存放在文件中。结构化数据中的每一行可以称为一个样本，纵向列为样本的描述属性。非结构化数据的表达方式则与

结构化数据完全不同，其产生后的形式很难用二维表这种形式表达。例如声音、影像、图片等。这些数据的特点是异构性，但是又具有可变化性，并没有什么特殊的结构来规范和描述这类数据。半结构化数据可以看作是介于结构化数据和非结构化数据之间的一类数据。首先说半结构化数据也不是以二维表的形式存在，而是有其自身的结构特点。例如HTML、XML文档，它们并不是二维表形式，但是必须由自身语义定义的首位标识符来表达和约束其关键内容。因此从严格意义上将，笔者认为半结构化数据更像非结构化数据，甚至可以规划到非结构化数据的范畴。

目前大家经常使用和产生的数据大多都是非结构化数据，例如在社交媒体分享的自拍，发的微博、微信、语言文字、声音等。可见当前人们能够接触到的大数据大多在非结构化数据的范畴。结构化数据处理起来相对容易，非结构化数据由于其多模性和异构性则较难处理。例如通常需要把非结构化数据转化成为结构化后，才能够真正用于分析和处理。

（3）时效（Velocity）：时效性是大数据的又一重要特征。在当今社会，互联网发展迅速，数据的产生和传播能力十分强大，这就需要对实时的大数据进行有效处理。由于数据量巨大（TB、PB级），其处理方式也有别于传统的方法，由此也出现了大数据的实时流处理等技术。

我们最能够直接接触到的这种场景是电子商务领域。例如天猫、京东等大型电商平台都纷纷推出了“双11”等大型促销活动。这些业务活动会直接影响到大数据的特征，例如数据会突然产生、快速流动，也有可能会很快消失，同时在某一时间段内会达到峰值又有可能很快回落。这就要求系统必须在极短的时间内满足用户的请求，毕竟谁都希望自己的“秒杀”请求得到迅速回应，否则会大大降低用户体验。同时，企业在满足数据时效性的同时也必须考虑自己的成本，例如可以通过增加服务器来提高数据的相应能力，但是这种时效的大数据并不会一直存在，企业也必须保证自己的IT架构具有充足的弹性，这就要求必须有云计算能力的保障。

由此可见，传统的小数据往往以离线分析为主，而大数据时代需要强调数据的实时处理能力，又要考虑企业获得这种能力的成本。这就是大数据带给企业IT解决方案的变化。

（4）价值（Value）：价值是大数据的特点，也是其存在的重要意义，更是大数据这一学科产生的根本原因。通过大数据的分析和挖掘，能够为企业业务发展提供很好的辅助支撑作用，甚至创造新的商业模式，推动社会的进一步发展。

目前，在新兴的IT领域，很多公司都纷纷开展了大数据的研究和应用工作。例如国际著名的电商公司亚马逊早在2003年就尝试通过用户的历史消费

记录，向其推荐可能感兴趣的产品，这一模式不但增加了销售额，同时也大大提高了用户的消费体验。中国大数据研究的起步较晚，但是发展却十分迅速。Gartner的一份分析报告指出，到2015年使用先进大数据管理系统的企业将比未使用的企业盈利能力高出20%。

大数据虽然具有很高的价值，但是也具有其自身固有缺点，就是数据的价值密度较低。这是由于绝大部分数据都是以非结构化的形式存在，但是具有价值的数据仅仅是其中的很小一部分，这也为充分发挥数据的价值制造了难度。例如城市交通监控系统需要对城市的主要街道进行24小时的监控，但是真正能够利用的监控视频一天内可能只有几分钟甚至没有，为了记录这短暂的瞬间却需要浪费大量的资源做监控、存储等工作。为此也产生了许多相应的解决方案，例如在存储层面的重删、压缩等技术也逐渐被提出。

伴随着互联网时代的到来，非结构化数据出现了大幅度增长的趋势。有相关统计指出，大数据大部分是由社交网络、传感器等终端和网络平台产生，非结构化数据将占到大数据的75%以上。因此在各行业中，如何能够更好更高效地利用大数据则成了重要问题。在这个背景下，数据治理的概念被逐渐提出。

1.1.3 管理与治理

我们提到“治理”一词，往往会联想到另外一个概念“管理”。但这两者之间是有许多不同点的[1]，相关标准化组织也对两者做出了定义和描述。1996年信息系统审计与控制协会公布了COBIT（Control Objectives for Information and related Technology），目前已经发布了5.0版本（COBIT 5）并逐渐成为国际上最通用的IT治理标准。这一标准体系为商业环境中IT技术的实施提供了统一规范和定义，并在全世界100多个国家的企业和组织中得到应用，指导企业和组织有效利用信息资源，更有效地管理信息和控制风险。关于“治理”和“管理”的定义，COBIT 5都有相关的描述。

（1）COBIT 5的“治理”定义：治理（Governance）是指评估利益相关者的需求、条件和选择以达成平衡一致的企业目标，通过优先排序和决策机制来设定方向，然后根据方向和目标来监督绩效与合规。根据这一定义，我们可以总结出治理应包含评估、指导和监督三个层次的含义，要确保治理的结果符合最初治理实施的动因，并达到了预期的目标。

（2）COBIT 5的“管理”定义：管理（Management）是指按照治理机构设定的方向开展计划、建设、运营和监控活动，以实现企业目标。基于此定义，管理包含计划、建设、运营和监控四个关键活动，并确保活动符合治理机构所设定的方向和目标。

从上述定义可以看出，治理和管理主要存在以下三方面的不同。

（1）治理和管理的具体工作内容有所不同。治理包括评估、指导和监督三方面内容，而管理包括含计划、建设、运营和监控四个相关内容。由此可分析，治理更侧重结果是否满足预期，并对执行过程提供指导建议；管理则侧重执行任务过程本身。

（2）治理和管理有着不同控制领域或者目的。治理的目的是确保收益达标、降低和控制风险、优化各方面资源、保障相关者利益。管理则关注4个领域：APO（调整、计划和组织）、BAI（建立、获取和实施）、DSS（交付、服务和支持）、MEA（监视、评价和评估），每个领域都包含若干个流程。

（3）从实施者角度上看，通常治理的责任主体是在董事会主席领导下的董事会，而管理的责任主体是CEO领导下的执行管理层。

通过以上分析，治理和管理虽有交叉，但其根本内容和目标却有着较大不同。为了更好地控制和管理企业及组织，治理和管理必须相互作用，尤其是在过程、信息、组织架构等方面。

1.1.4 数据管理与数据治理

类比与管理和治理，相关标准化组织也同时给出了数据管理和数据治理的相关描述和定义。

DMBOK关于数据管理的定义：数据管理（Data Management，DM）是指通过策划与实施相关的方针、活动和项目，以获取、控制、保护、交付和提高数据资产价值。

我们可以从几个方面来理解这一定义：①数据管理包括管理的仿真、相关活动以及管理项目的实施和实施人的相关职责定义；②数据管理需要制定一系列的标准规范，确保数据管理的过程能够得到有效推进；③数据管理团队可能由多个部门或者业务的专家领导组成管理团队，确保数据管理实施过程的规范性。

我们进一步对上述三点来总结，数据管理的三个关键点为职责、过程以及规范。职责的定义是企业数据管理业务能否开展的关键点，如果企业不能够清晰透彻地定义和分配职责，则还会造成员工不健康履职，进而导致企业资金浪费、设备损毁等情况的出现。过程描述了数据管理的全部生命周期。严格来说数据管理的过程不会有严格的重点，因为数据管理是企业长期需要执行的工作。规范明确了数据管理全部内容执行的标准和范畴。数据管理的一切活动必须严格遵守相关规范规定，确保数据管理政策被顺利执行。

与数据管理向并行的，还有另外一个概念：信息管理。实际上，两者在企业应用中并无太大不同。我们首先来看COBIT 5对信息管理的定义。

信息管理（Information Management，IM）是指根据信息治理机构设定的方向，对与获取、控制、保护、交付和提升信息资产价值相关的实践、项目和功能等方面，进行全面的计划、建设、运营和监控。

从字面上理解，COBIT 5定义的信息管理与DMBOK的数据管理定义本质上是相同的。我们在前文曾分析过数据和信息的关系。在存在描述上下文的情况下，数据则转化成了信息。由此可以看出，信息也是数据的一种表现形式，数据则是更加广义上的信息。因此，信息和数据的管理的概念、范围、业务流程等是高度相似甚至是相同的。

数据管理体系能够很好地使企业IT系统健康运转，使数据创造更大的价值。但是数据管理或者说信息管理还需要科学的企业组织体系、完善的评估机制和迭代改进方法论的配合。这就引申出来另一概念：数据治理。例如企业或者高校的财务系统负责管理金融资产，这就属于数据管理的范畴。同时，金融数据管理的效果需要审计员的评估，确保数据得到有效的管理，这就是数据治理的范畴。

目前数据治理还是较为新的领域，各行各业都对其有不同的看法和定义。其中以DMBOK、COBIT 5、DGI和IBM这些权威机构提出的定义认可程度相对较高。本章节主要介绍这些标准化组织和机构对数据治理的定义。其中，COBIT 5给出的是信息治理的定义。由于数据和信息的概念本质上是一致的，因此我们不再具体区分数据治理和信息治理之间的区别。

（1）DMBOK数据治理定义：数据治理（Data Governance，DG）是指对数据资产管理行使权力和控制的活动集合（计划、监督和执行）。

（2）COBIT 5数据治理定义。信息治理（Information Governance，IG）包含以下三个方面的内容：①确保信息利益相关者的需求、条件和选择得到评估，以达成平衡一致的企业目标，这些目标通过信息资源的获取和管理来实现；②确保通过优先排序和决策机制为信息管理职能设定方向；③确保基于达成一致的方向和目标对信息资源的绩效和合规进行监督。

（3）DGI数据治理定义：数据治理是指针对信息相关过程的决策权和职责体系，这些过程遵循“在什么时间和情况下、用什么方式、由谁、对哪些数据、采取哪些行动”的方法来执行。

（4）IBM数据治理委员会给出的数据治理定义：数据治理是针对数据管理的质量控制规范，它将严密性和纪律性植入企业的数据管理、利用、优化和保护过程中。

上述4家结构的定义从字面理解上有些抽象晦涩，我们可以从以下几个方面来具体解读。

（1）数据治理要达到什么效果。数据治理是为了保证数据策略的正确