大数据工程技术与应用

数据质量管理与安全管理

[韩] 金　范——主编

上海科学技术出版社

大数据工程技术与应用

数据质量管理与安全管理

[韩] 金　范　主编

上海科学技术出版社

图书在版编目(CIP)数据

数据质量管理与安全管理 /（韩）金范主编. —上海：上海科学技术出版社，2016.10
（大数据工程技术与应用）
ISBN 978-7-5478-3247-9

Ⅰ.①数… Ⅱ.①金… Ⅲ.①关系数据库系统—质量管理②关系数据库系统—安全管理 Ⅳ.①TP311.138

中国版本图书馆 CIP 数据核字(2016)第 216455 号

数据质量管理与安全管理
[韩]金　范　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 8.5
字数 170 千字
2016 年 10 月第 1 版　2016 年 10 月第 1 次印刷
ISBN 978-7-5478-3247-9/TP·45
定价：34.00 元

内容提要

本书重点介绍数据质量管理与安全管理的理论及应用，强调提高数据质量不仅可提高信息系统的质量，还可提高经营活动的质量。本书给出了数据质量的指标以及对应的管理流程，划分了5个能力成熟度的等级，界定了从管理者到执行者等各个环节的质量管理活动和责任，帮助组织制定数据质量管理计划或执行具体的质量管理活动。书中提出了多项数据质量管理主要技术和各国应用案例，还进一步在Orange数据库中实践了数据质量诊断流程。本书对数据质量管理与安全管理的介绍全面系统、条理清晰，涵盖理论方法、关键技术、实践操作等多个方面，可供致力于提高数据质量及安全性的研究人员及技术人员参考阅读。

大数据工程技术与应用

编撰委员会

丛书序

“数支配着宇宙”——毕达哥拉斯。

大数据技术，使这句2000多年前的哲言如此形象、如此真切；大数据技术，正以前所未有的发展速度变革着人类的认知、产业和生活。

当前，我国正处于创新驱动发展、产业全面转型升级的关键阶段，大数据既是新的经济增长点，更是推进创新与发展的利器。上海产业技术研究院以服务于成果转化和产业化为使命，较早开始了大数据的应用研究和服务工作，构建了大数据应用技术平台，针对重点行业开展了一批大数据应用研究，涉及数据建模、数据分析、数据安全和数据库管理相关软件开发、测试、评价等多个方面。本丛书的出版既是前期工作探索的分享，更是进一步服务于成果转化和产业化的一个尝试。

大数据应用与产业化急需大量的工程应用技术人才。本丛书主要面向大数据工程应用的广大科技人员，在内容上汇聚了不同国家和区域、不同专业和领域的专家智慧，侧重大数据工程化知识、最佳实践和实用技巧，力求可操作性、实用性。

由于大数据技术研究和应用是一个新兴领域，发展方兴未艾，本丛书在编撰过程中因编者的知识和经验局限，必然存在许多不当之处，敬请广大读者提出宝贵意见。

2016年9月

前言

智能机器的发展、IOT 时代的到来、人工智能(AI)时代的序幕均意味着数据的爆发式增长,与单纯数值上的增长相比,管理的重要性正渐渐兴起。能否在这样的大数据时代中取得成功,取决于如何实现数据质量和安全管理。

数据质量差不仅会让数据检查、精炼及调整消耗必要的和不必要的时间及资源,而且还会降低对整个系统的信任程度。为了克服这种情况,人们虽然采取了手动数据收集及修改,但由此也产生了其他的附加费用。不仅如此,拖延决策或错误决策对业务成果产生了消极影响。面临现在的大数据时代,诸如此类的问题随着数据量急增而日益突出,对于所有的企业与政府部门,如何确保数据质量成为一个重要的课题。

随着数据量的急速增长,数据安全同样也上升成为重要的管理课题。最近,以公共服务部门、金融单位为代表的整个产业界,不仅是个人信息,企业经营机密的泄露事件也正在逐渐增多,恶意使用此事件而引起的第二阶段、第三阶段犯罪也随之产生,人们对安全的重视程度不断提高。在一般的系统网络等中,时而有必要越过不被认可控制安全的行为,从而实现对数据的直接安全管理。各种数据泄露事故,不仅是外部黑客所致,而由内部的许可人员造成的事故也正在急速增加。为了最终的数据安全,有必要实现数据访问控制、加密、操作审批、漏洞分析等多方位、精密的数据安全管理。

假如没有数据的质量管理与安全管理,管理数据的国家和企业的风险无疑将不断增加。为了提高累积的数据质量,质量管理与安全管理成为当今不可避免的重要课题。本系列丛书对数据质量的各种评估方法与指南,以及从数据安全的计划、构建再到运营的全面业界技术进行了整理,希望能够帮助每个实务人员提高对这些知识的理解程度。

本书由金范编写,周兆明、张青对本书进行了认真校对,王一帆、邱雯等参与了资料

的收集、整理、录入等工作。此外，本书的编写得到了上海产业技术研究院大数据专家委员会的大力支持和指导，上海产业技术研究院的组织协调也使本书得以顺利出版，在此一并表示衷心感谢。

金　范

2016 年 6 月

目 录

第1章

数据标准化

1.1 数据标准化的必要性

现如今数据应用的问题主要集中在数据重复、系统间数据不一致、数据定义模糊等方面，其原因一方面是数据质量和标准存在问题，另一方面，同时开发多个信息系统、企业缺乏数据管理意识、企业缺乏数据管理人员、企业缺乏数据标准管理工具等也是造成这些困扰的重要因素。随着数据在企业战略性决策中的重要性逐步加大，企业数据标准化的需求也愈发迫切。数据实现标准化后，可以通过统一的命名方式使交流更顺畅，通过统一的数据格式和规则提高数据质量。

数据标准化指的是确立分散在各个系统中的信息要素的定义、名称、格式和规则后将其推广应用到全公司的活动。数据标准化的构成要素包括数据名称、数据定义、数据格式、数据规则等。

定义数据名称时，应考虑数据的唯一性、业务上观点的普遍性、含义传达的充分性等。描述数据定义时，应确保用户完全理解其含义。若仅靠叙述难以表达其含义时，可加上相关计算公式或示例以便于理解。对数据的默认值、允许值、允许范围等进行定义时，通过对数据项使用域，可使格式类似的项保持一致。通过预定义可能会出现的数据值，可提高数据的一致性和完整性。

应用数据标准管理的对象包括标准用语、标准词、标准域、代码等。标准用语可通过日常业务中使用的业务用语和信息系统中使用的技术用语进行定义。

数据管理员的基本作用是定义数据的政策和标准、设计数据结构、管理数据模型(data modeling)等，为了管理企业数据标准及支持整体流程，应考虑使用数据标准系统。一般来说，数据标准管理系统的基本功能包括标准管理、结构管理和流程管理，但为了支持企业中存在的各种系统标准及未来对企业标准化进行整合，也需有多重标准功能。

1.2 数据管理现状和问题

为保证数据的质量，必须要进行数据标准化。然而在实际应用数据时，以下问题会成为向用户及时传达正确信息的主要障碍。

(1) 发生数据重复，各组织、业务或系统的数据不一致。因数据标准政策不完善，导致

信息系统在开发和运营过程中相同含义的数据以不同名称进行重复管理，或相同名称的数据在不同系统中以不同的含义进行不同的逻辑运算。

(2) 不能及时了解数据含义，从而无法及时提供信息。因数据名称、数据定义缺乏标准管理，在获取新信息点或信息点变更所需数据时浪费大量时间，而无法及时向用户提供正确信息。

(3) 数据整合困难。存在采用或不采用以单位系统为主的数据标准的情况，因此在以企业数据整合信息点为基础建立系统时(如建立企业数据仓库)，很难确认数据的含义或确认数据是否重复。

(4) 难以变更或维护信息系统。因数据标准政策不完善，变更或维护信息系统时难以确认数据的含义，反映新信息点时也难以确认基本数据是否可用，因此在维护时需要花费很多精力。

上述问题是由过去开发和运行信息系统时固有的以下几点因素造成的。

(1) 同时开发多个信息系统。最近信息系统的开发项目加大了系统间的相互关联性，相比于以单个系统为主的开发，同时开发多个相关信息系统的趋势越来越明显。在这种开发环境下，并没有制定企业数据标准规范，仅根据以单个系统为主的标准规范，将开发项目的重点放在单个系统的业务功能实现上。

(2) 缺乏企业数据管理意识。数据的管理主体以单个系统的开发者、运营者为中心，重点放在单个业务的支持上。最近信息化项目不仅使用单个系统的数据，灵活地组合使用多系统数据的情况也越来越多，因此需要形成对企业数据进行系统化管理的意识。

(3) 缺乏企业数据管理人员。在信息系统开发阶段，通过组织对开发人员的品质管理达到标准化管理的目的。但在维护阶段，没有在开发阶段负责制定标准和监督标准实行情况的数据管理人员参与进来，仅依赖个别维护人员。

(4) 缺乏企业数据标准管理工具。数据标准管理需要多个自动化系统的支持，其功能包括制定数据标准、监督数据标准的实行、查询和应用数据标准等。虽然在开发信息系统时手工执行了数据标准套用和合规检查，但在运营阶段仍然沿用类似的手工标准管理方法会产生很多难题。

1.3 数据管理改善方案和标准化效果

数据作为企业战略决策的核心要素，为了实现数据整合、达到数据质量要求，需对企业数据实行标准化管理。

首先应设置数据标准化、规格化的基本方针，为了共享企业信息，导出要保留的共同数据元素，需要构建企业数据元素登记和管理体系。据此进行信息系统开发和维护时，通过

使用被认可的数据元素，提高系统开发的效率和数据的通用性，实现有效的数据管理。

企业数据实现标准化后，当前用户即可使用正确的数据，做出正确的决策。这对确保企业的竞争力有很大推动作用。

(1) 通过统一的名称实现更加明确的交流。对相同数据使用相同名称，可在开发人员与当前业务、运营人员与当前业务以及运营人员等多种阶层间实现明确迅速的交流。

(2) 减少掌握所需数据素材而花费的时间和精力。出现新的信息请求时，通过使用标准化的数据可快速掌握数据的含义和数据的位置等，在数据使用人员期望的时间内提供正确的信息。

(3) 通过使用一致的数据格式和规则来提高数据质量。在数据标准中引用数据格式和规则，可防止数据错误输入，提高数据质量。此外，按照标准使用数据可以减少因数据应用导致的决策错误。

(4) 减少信息系统的数据接口间数据转换和整理的费用。在数据整合项目或个别系统中需要其他系统的数据时，因为按照企业数据标准管理数据，无需执行任何转换或整理作业操作即可直接应用，因此不会产生额外费用。

1.4 数据标准化定义

数据标准化指的是确立分散在各个系统中的信息要素的定义、名称、格式和规则后将其推广应用到全公司的活动。数据标准化不仅有助于确认数据的正确含义，还可据此调整对数据的审视角度。

数据名称是企业内唯一区分数据的名称，因此数据名称标准化需要对同音异义词和异音同义词进行定义。数据名称通常应符合以下原则。

(1) 唯一性。数据名称应是唯一区分特定概念的称号。为使所有用户对同一概念使用统一的用语，只允许使用同一名称。例如“客户账号”、“客户账户号”应统一为“客户账号”，“邮箱地址”、“邮箱”应统一为“邮箱地址”。

(2) 业务观点的普遍性。数据名称应该是从业务观点出发具有普遍认知的名称。通常企业和组织内部成员指定相应概念的名称时，都希望使用最常用的业务用语。

(3) 含义转达的充分性。数据名称应起到看到名字就能掌握数据含义和范围的作用。如果不同的业务或用户观点可能会造成不同的含义时，建议使用修饰词等具体表达方式。

数据定义应指定相应数据含义的范围和资格条件。针对仅通过名称难以向用户传达准确信息的其他事项，为了使用户完全理解数据的含义，应明确指出业务观点的范围和资格条件。此外，数据定义应成为数据所有者的决定标准。描述数据定义时应考虑以下事项。

(1) 为了使数据用户完全理解数据的含义，应从不了解相关业务的第三者的立场出发

进行描述。

(2) 只通过叙述式定义难以表达数据的含义时，应同时提供实际可能产生的数据值的描述。

(3) 尽量不要按原样描述数据名称或使用缩写和专业用语描述定义。

(4) 数据格式。

使用数据格式，可通过定义数据表现形态将数据输入错误和控制风险降到最低。数据格式的定义应与业务规则和使用目的一致。

- 数值(numeric)
- 文本(text)
- 日期(date)
- 字符(char)
- 时间戳(timestamp)
- 数据长度和小数点位数

定义数据格式时应考虑以下事项。

(1) 通过定义域并应用到数据标准，统一性质类似的数据之间的数据形式。

(2) 数据的最大值或最大长度不固定时，定义时应留有余量。

(3) 对特殊数据类型(CLOB、long、raw 等)执行数据查询、备份、实行等操作时存在很多限制，尽量不要使用。

使用数据规则，通过预定义可能出现的数据值，将数据输入错误和控制风险降到最低。数据规则可提高数据的一致性和完整性。数据规则的类型有以下几种。

(1) 默认值。用户在画面或应用程序中未输入任何值时，根据数据类型输入预定义的默认值。也就是说在未输入数据值时自动输入数据值。例如，Numeric 类型的项目自动输入默认值“0”，Char 类型的项目自动输入默认值“空格”。

(2) 允许值。为了与业务规则保持一致性，限制可以输入的数据值，标准代码中预定义了各数据项目相应的代码值。例如，标准代码中定义的允许值为 01、02、03、04……10，而特定数据项目中可能出现的允许值为 01、03、05 等一部分值。

(3) 允许范围。为了与业务规则保持一致性，限制可以输入的数据值的范围。例如，若特定数据项目中定义的允许范围为 1～5 时，则预先限制输入 1～5 以外的值。

1.5 数据标准化

为促进企业数据标准化而制定的标准化构成要素包括数据标准、数据管理组织和数据标准化步骤。