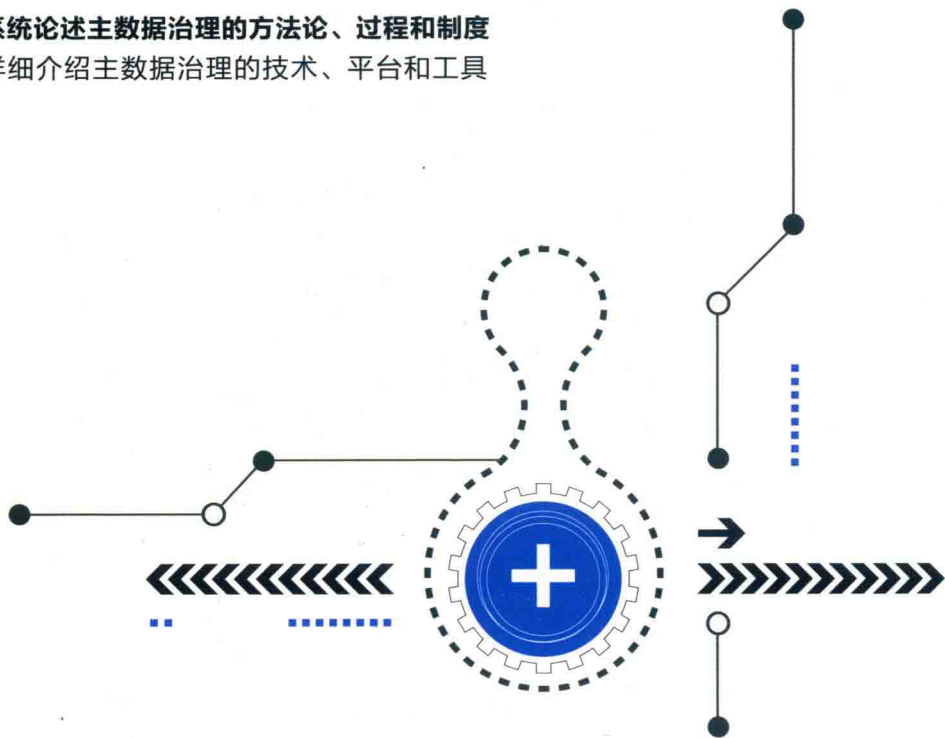


系统论述主数据治理的方法论、过程和制度
详细介绍主数据治理的技术、平台和工具



开发者书库



Data Governance Driven by Master Data
Principles, Technologies and Practices

主数据驱动的数据治理

原理、技术与实践

王兆君 王 钺 曹朝辉◎编著

Wang Zhaojun Wang Yue Cao Zhaohui

唱 伟 中国五矿集团有限公司招标采购中心主任
王红楼 丽珠医药集团股份有限公司首席信息官
梁 旭 三一重工股份有限公司董事长助理/副总经理/首席信息官
朱 亮 中国铝业集团有限公司信息管理部处长
郑 锋 特变电工集团首席信息官

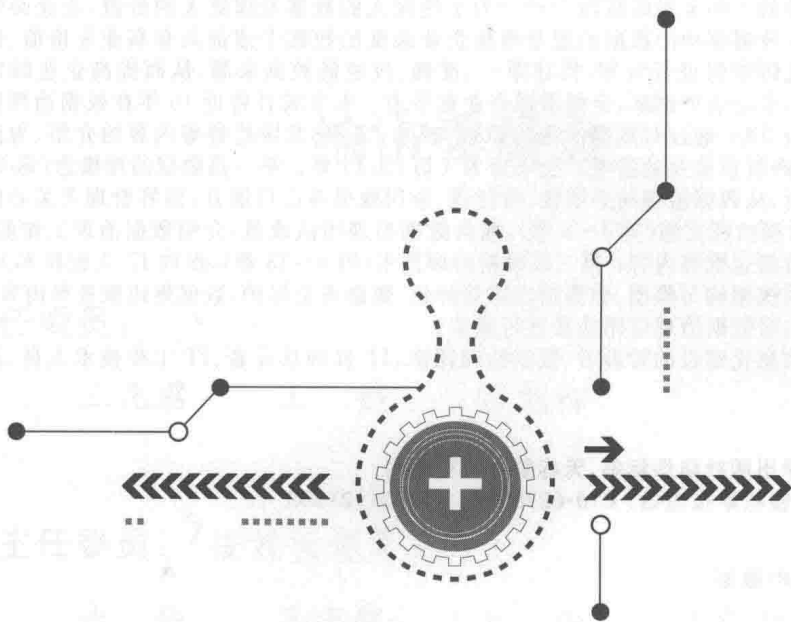
联袂推荐

清华大学出版社



清華

开发者书库



Data Governance Driven by Master Data
Principles, Technologies and Practices

主数据驱动的数据治理

原理、技术与实践

王兆君 王 钺 曹朝辉◎编著
Wang Zhaojun Wang Yue Cao Zhaohui



清华大学出版社
北京

内 容 简 介

“数据”已成为企业的一项宝贵的战略资产。为了使庞大的数据发挥更大的价值,企业必须着眼于数据治理和综合利用。主数据驱动的数据治理是指从企业杂乱的数据中捕捉具有高业务价值、被企业内各业务部门重复使用的关键数据进行管理,构建单一、准确、权威的数据来源,从而提高企业的整体数据质量,提升数据资产价值,推动业务创新,全面增强企业竞争力。本书编者将近10年在数据治理咨询工作中积累的经验 and 知识进行总结,通过对数据治理的原理、技术、案例、发展趋势等内容的介绍,为读者进行数据治理、主数据管理实践提供重要的参考。全书分为4篇,共14章。第一篇数据治理概念(第1~3章),面向数据治理组织管理者,从数据治理的必要性、可行性、应用效果等进行展开,回答管理者关心的数据治理的核心问题;第二篇数据治理实施(第4~8章),面向数据治理团队成员,介绍数据治理工作的前期准备、工作步骤、治理过程、后期运维等内容;第三篇数据治理技术(第9~13章),面向IT工程技术人员,从技术视角展开数据治理的系统架构与模型、数据治理质量评估、数据安全保护、数据集成服务等内容;第四篇数据治理前景(第14章),对数据治理应用前景进行展望。

本书可作为从事信息化建设的管理者、数据治理团队、IT咨询从业者、IT工程技术人员、相关专业在校师生的参考读物。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

主数据驱动的数据治理:原理、技术与实践/王兆君,王钺,曹朝辉编著. —北京:清华大学出版社,2019
(清华开发者书库)

ISBN 978-7-302-52295-9

I. ①主… II. ①王… ②王… ③曹… III. ①数据处理 IV. ①TP274

中国版本图书馆CIP数据核字(2019)第029341号

责任编辑:盛东亮
封面设计:李召霞
责任校对:李建庄
责任印制:董 瑾

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 刷 者:北京富博印刷有限公司

装 订 者:北京市密云县京文制本装订厂

经 销:全国新华书店

开 本:186mm×240mm

印 张:23.5

字 数:528千字

版 次:2019年4月第1版

印 次:2019年4月第1次印刷

定 价:89.00元

产品编号:081386-01

编审委员会

主任委员：

王兆君 王 钱 曹朝辉

副主任委员：（按姓氏笔画排序）

马 俊 王志民 王 乐 汪东升

张 扬 徐成华 隋 娜

委员：

王 波 王 旭 刘 青 李召波

毕旭东 孟 杰 张毅超

序

FOREWORD

让我们从一个简单的问题开始：IBM 有多少雇员？

这个问题看上去非常简单直接，对吗？但是请注意下面这个清单：

International Business Machines Corporation	IBM
IBM Microelectronics Division	IBM Global Services
IBM Global Financing	IBM Global Network
IBM de Columbia, S. A.	Lotus Development Corporation
Software Artistry, Inc.	Dominion Semiconductor Company
MiCRUS	Computing-Tabulating-Recording Co.

这个长长的清单中的每一项都与 IBM 有关系，有全称、缩写、别名、分支机构、全资子公司等，有的公司名称中完全没有 IBM 的字样，但它归属于 IBM，有的公司曾经归属于 IBM，后来又被卖掉了，还有的公司现在已经完全不存在了。

现在我们再来看刚才那个问题——IBM 有多少雇员？还会觉得这个问题简单吗？

事实上，我遇到过一所知名大学信息管理部门的负责人，问他这个关于雇员的问题。他告诉我，这正是让他们头疼的问题。太多不同时期建设的信息系统、不同的编号、不同的命名体系、不同的管理方式，所有信息汇总到一起之后，不知道哪些是重复的，哪些是陈旧的。用不同的方式统计，结果都不相同。

对于拥有众多部门和分支机构的大型企业，这样一种数据管理的困境随处可见。因此，不仅对于人员、物料、市场，而且对于那些与企业运营密切相关的重要信息都存在着管理的挑战。所以，我们需要数据治理，尤其是对企业中最关键的数据资产——主数据进行治理，进而提升数据质量，使数据真正成为管理和决策的可靠依据。

我们正处在历史的转折点上，数据技术在快速变革，大数据成为人们竞相议论的热点。无疑，未来的竞争就是数据的竞争。但是，在这个变革的关键时点上，更多的人将注意力的焦点放在了数据的“量”上，很少有人提及和关注数据的“质”，仿佛只要有了足够大量的数据，一切问题都可以解决。很可惜，真实情况是，海量数据如果未能经过合理的加工和组织，并确保一定的数据质量，它不仅不能解决问题，反而可能制造出更多的麻烦。也许，我们应该尽早从华而不实的喧嚣中抽身出来，通过具体而细致的数据治理工作，切实改善企业的数据环境，让大数据真正从“看”到“用”，真正活跃起来。

本书凝聚了编著者在数据治理和主数据管理领域多年的从业经验，涵盖数据治理和主

数据管理的基本概念、实施过程、关键技术等重要内容,并结合大量实际积累的案例和技术方案,系统地介绍了数据治理这一新兴领域及其应用情况,可作为工作指南为正在或准备开展数据治理工作的 IT 人员提供参考,更能为数据时代的企业管理者提供新的思路、新的方向。

张林 清华大学教授,清华-伯克利深圳学院院长
清华大学物联网与社会物理信息系统实验室主任
2019年1月于北京

前言

PREFACE

在过去的几十年里,对数据的计算和存储能力以及可用性的巨大进步,促成了当今数据驱动型的世界现状。数据正在对整个人类社会产生巨大的积极影响,它不仅在改变着人们生活的各个方面,而且也使得企业的运营更加高效。互联网数据中心(IDC)预测,到2025年,全球数据圈将扩展至163ZB(1ZB相当于1万亿GB),是2016年所产生16.1ZB数据的10倍,这些数据将给个人带来全新的用户体验并且给企业带来更多的商业机会。

虽然已经有部分企业认识到数据资产的重要性,但是随着数据数量、种类以及重要性的不断增加,收集、存储和处理这些数据的难度也越来越大。如何从海量数据中挖掘出对制定决策有价值的信息,成为企业在管理和使用数据过程中面临的主要挑战。

数据治理的核心正是加强对数据资产的管控,通过深化数据服务以持续创造价值,企业领导者必须关注其中最重要的那部分数据,只有识别并充分利用这些至关重要的数据,才能发挥其巨大潜力。主数据管理就是从来源复杂的数据中捕捉关键数据,并且对这些具有高业务价值的、可以在企业内跨越各个业务部门被重复使用的数据进行管理,通过为跨构架、跨平台、跨应用的系统提供一致的、可识别的主数据对象来支持整个企业的业务需求,从而提高企业的整体数据质量,提升数据资产价值,推动业务创新,全面增强企业竞争力。主数据管理是一个全面的战略,涵盖所有需要统一定义的、企业所需的核心数据和数据标准。主数据管理的有效途径是建立一个包括主数据标准体系、主数据管控体系、主数据质量体系和主数据安全体系在内的、完整的主数据体系,建立持续长期的管理机制,这样才能构建企业数据的核心治理能力,合理利用企业数据来寻求竞争优势。

本书编者从事数据治理和主数据管理咨询工作近10年,亲身经历了数据治理和主数据管理在中国企业信息化浪潮中的兴起、演进和实践的过程。目前,为了配合国家信息化发展战略,很多企业把数据治理和主数据管理系统建设项目提上日程,并且开展了部分信息标准化工作。但是,从总体上看,国内企业的主数据体系建设工作仍然处在起步阶段,很多企业管理者对数据治理和主数据管理的概念理解有限,对主数据管理体系建设的重要性认识不足。编者将在数据治理和主数据管理领域的从业经验和知识积累进行总结,与大家分享和探讨,并希望能回答什么是数据治理和主数据管理、为什么需要数据治理以及如何进行主数据管理等问题。

本书坚持“贴近用户”的思路,回答用户关心的核心问题,不仅介绍主数据管理的产生背景、概念、模型和技术等理论知识,同时涵盖主数据管理项目的实施方法和过程、主数据管理

的产品和应用案例,使读者对主数据管理项目从底层技术知识到上层应用实践都能有系统的理解。同时,本书有针对性地对行业主流厂家的主数据管理产品进行了全面介绍,让读者能够更加深入地了解行业主流产品与趋势。书中案例都是近几年国内相关行业的领先企业的优秀实践,对其他企业的主数据管理和数据治理工作具有很高的参考价值。另外,书中对大数据、云计算、人工智能和区块链等新兴技术与主数据管理的结合应用也进行了探讨及趋势分析。

全书通过对主数据管理的背景、概念、模型、技术、实施、产品、案例、发展等内容的全面介绍,为读者揭开主数据管理这一新兴概念的神秘面纱,为读者进行数据治理、主数据管理实践提供重要参考。全书分为4篇,共14章。第一篇数据治理概念,包括第1~3章,其中第1章介绍数据治理的背景、意义和核心内容,并且引入数据管理的成熟度模型,使用户可以根据自评表得到成熟度评估和治理建议;第2章讨论主数据和主数据管理的概念和意义,为读者揭示主数据管理的必要性;第3章讨论主数据驱动的数据治理,系统地介绍治理框架、治理过程和数据治理工具。第二篇数据治理实施,包括第4~8章,其中第4章介绍主数据治理项目的准备工作;第5章讨论主数据体系规划方法;第6章说明主数据治理项目的具体实施步骤;第7章介绍主数据项目的运维和管理;第8章介绍目前国内主流的主数据管理解决方案和产品,并分析国内主数据管理的先进案例。第三篇数据治理技术,包括第9~13章,其中第9章介绍数据架构和模型的相关技术知识;第10章讨论数据集成技术及其企业应用;第11章介绍数据质量管理的定义、评估框架以及数据质量战略;第12章讨论数据生命周期管理的概念、内容和体系架构;第13章介绍数据安全管理和数据隐私保护。第四篇数据治理前景,包括第14章,主要展望主数据与大数据、云服务、人工智能和区块链应用的发展趋势。

本书既可补充从事信息化建设的IT部门人员的专业知识,更能为组织管理者提供信息化知识储备和工作思路,助力IT架构的组织优化。本书也面向咨询公司的顾问和实施人员,不仅针对主数据管理项目,而且对处理各类信息系统项目中可能出现的数据问题都具有一定参考价值。本书还可以作为企业管理软件开发人员的自学参考书,以及相关专业在校师生开阔视野、理论联系实践的参考书。

在本书的编写过程中参考和引用了国内外很多书籍和网站的相关内容,部分图片素材和个别实例的初始原型也来源于网络,部分互联网相关资源无法一一列举出处,在此向其作者一并予以感谢。众所周知,一本书难免出现不足和疏漏之处,恳请广大读者将意见和建议反馈给我们,以便在后续版本中不断改进和完善。有关数据治理的更多信息,可关注北京三维天地科技有限公司微信公众号。

编者

2019年1月

目录

CONTENTS

第一篇 数据治理概念

第 1 章 数据治理概述	3
1.1 数据治理背景	3
1.2 数据资产和数据管理	5
1.2.1 数据资产的概念和重要性	5
1.2.2 数据资产的构成	7
1.2.3 数据管理的内容、现状和问题	11
1.3 数据治理的目标和挑战	14
1.3.1 数据治理的概念	14
1.3.2 数据治理的目标	15
1.3.3 数据治理的挑战	16
1.4 数据治理的核心内容	17
1.4.1 数据治理的内容	17
1.4.2 数据治理的基本过程	20
1.4.3 数据治理的重点	22
1.5 数据治理的评估——成熟度模型	22
1.5.1 数据管理的成熟度模型	22
1.5.2 您的企业需要数据治理吗	24
第 2 章 主数据和主数据管理	28
2.1 主数据的概念	28
2.1.1 主数据的定义	28
2.1.2 主数据的特征	29
2.1.3 主数据的范围	30
2.2 主数据管理的概念	31
2.2.1 主数据管理的定义	31

2.2.2	主数据管理体系	32
2.2.3	主数据管理系统的功能	33
2.3	主数据管理的意义	34
2.3.1	主数据管理的必要性	35
2.3.2	主数据管理的意义	37
第3章	主数据驱动的数据治理	39
3.1	数据治理框架	39
3.1.1	国际标准化组织	40
3.1.2	国际数据管理协会	41
3.1.3	国际数据治理研究所	42
3.1.4	IBM 数据治理委员会	43
3.1.5	中国电子工业标准化技术协会信息技术服务分会	43
3.1.6	现有数据治理框架的局限	45
3.2	主数据驱动的数据治理框架	46
3.2.1	治理思路和治理目标	46
3.2.2	治理框架	48
3.2.3	技术架构	50
3.3	主数据驱动的数据治理过程	51
3.3.1	过程框架	51
3.3.2	架构阶段	51
3.3.3	治理阶段	52
3.3.4	任务、角色、分工、职责	53
3.4	数据治理工具和系统选型	54
3.4.1	软件公司的行业实践	55
3.4.2	产品特性	56
3.4.3	软件公司的实力	56
3.4.4	软件公司的实施	56
3.4.5	软件的价格	56

第二篇 数据治理实施

第4章	主数据项目的准备	59
4.1	主数据项目实施的主要风险	60
4.1.1	组织风险	60
4.1.2	数据风险	62

4.1.3	集成风险	66
4.1.4	其他风险	67
4.2	数据治理管理组织	69
4.2.1	项目组织	69
4.2.2	人员配置	72
4.2.3	管控角色	73
4.2.4	管控流程	75
4.2.5	绩效考核	75
4.3	数据管理规范体系	76
4.3.1	主数据管理规范	76
4.3.2	主数据应用标准	77
第5章	主数据体系规划方法	82
5.1	主数据体系规划的任务和步骤	82
5.1.1	主数据体系规划的任务	82
5.1.2	主数据体系规划的步骤	83
5.2	主数据体系评估方法论	84
5.2.1	主数据管理成熟度模型	84
5.2.2	主数据管理成熟度模型的评价指标	88
5.2.3	主数据管理成熟度评估方法	90
5.3	现状调研与需求分析	93
5.3.1	现状调研	93
5.3.2	现状评估与差距分析	97
5.3.3	需求分析	99
5.4	主数据识别分析方法	104
5.4.1	多因素分析方法	104
5.4.2	主数据类型识别分析	105
5.4.3	主数据元属性识别分析	106
5.5	主数据体系规划设计	107
5.6	主数据体系架构设计	109
5.6.1	主数据管控体系	110
5.6.2	主数据标准体系	111
5.6.3	主数据质量体系	112
5.6.4	主数据安全体系	113
5.7	主数据管理实施规划	114
5.7.1	实施策略	114

5.7.2	实施计划	115
5.7.3	投资预算	116
第6章	主数据项目实施步骤	117
6.1	实施方法概述	117
6.1.1	传统软件开发项目的实施方法	118
6.1.2	主数据项目的实施方法	119
6.2	项目实施阶段的主要任务	121
6.2.1	第一阶段：体系规划阶段	121
6.2.2	第二阶段：平台实施阶段	124
6.3	各主要阶段的任务分工	126
6.3.1	项目启动与需求调研阶段	126
6.3.2	体系规划与架构设计阶段	126
6.3.3	标准建立及主数据平台设计阶段	127
6.3.4	客户化设计、开发、测试、数据清洗阶段	128
6.3.5	系统上线启用阶段	129
6.3.6	系统运维与持续优化阶段	130
6.4	数据准备	130
6.4.1	数据准备方案制订	130
6.4.2	数据采集	131
6.4.3	数据清洗	131
6.4.4	数据导入	133
6.5	人员培训	135
6.6	程序设计	136
6.6.1	程序设计的基本要求	136
6.6.2	程序设计方法	137
6.6.3	产品定制开发	138
6.7	系统集成	139
6.7.1	系统集成架构	140
6.7.2	集成流程	140
6.7.3	系统集成技术	141
6.8	系统测试	142
6.9	系统试运行及上线	146
6.9.1	系统试运行	146
6.9.2	系统切换	147
6.10	系统评价	149

6.11	项目管理	152
第 7 章	主数据项目的运维和管理	159
7.1	主数据运维管理体系	160
7.1.1	主数据运维管理组织	160
7.1.2	主数据运维管理流程	161
7.2	主数据运维管理内容	162
7.2.1	主数据模型运维管理	162
7.2.2	主数据 workflow 运维管理	163
7.2.3	主数据生命周期运维管理	164
7.2.4	主数据质量运维管理	165
7.2.5	平台基础服务运维管理	166
7.2.6	主数据存储运维管理	167
7.2.7	数据库系统运维服务	168
7.2.8	主数据安全运维管理	169
7.2.9	基于云服务的运维管理	170
7.3	主数据运维应急响应措施	173
7.4	对外部供应商的运维要求	174
第 8 章	典型主数据管理产品及实施案例	176
8.1	主数据管理系统模式的分类	176
8.1.1	基于 ETL 工具的主数据应用	176
8.1.2	基于 SOA 的主数据管理平台	178
8.2	典型产品和解决方案及其对比	179
8.2.1	SunwayWorld 的主数据全生命周期管理平台	179
8.2.2	SAP 的 MDM 解决方案	183
8.2.3	IBM 的 MDM 解决方案	186
8.2.4	Oracle 的 MDM 解决方案	188
8.2.5	Informatica MDM 解决方案	190
8.2.6	产品对比	191
8.3	先进企业的主数据管理现状	192
8.4	主数据典型应用案例介绍	194
8.4.1	石油行业应用举例——某大型石油总公司的主数据管理	194
8.4.2	煤炭行业应用举例——某大型能源集团公司的主数据管理	196
8.4.3	有色金属行业应用举例——某大型有色金属公司的主数据管理	198
8.4.4	建筑行业应用举例——某大型建筑股份有限公司的主数据管理	200

8.4.5	航空航天行业应用举例——某航天建设集团有限公司的主数据管理	203
8.4.6	基建行业应用举例——某工程建设有限责任公司的主数据管理	205
8.4.7	电器行业应用举例——某大型电器集团的主数据管理	207
8.4.8	机械制造行业应用举例——某大型饲料机械集团的主数据管理	209
8.4.9	水泥行业应用举例——某水泥控股有限公司的主数据管理	212
8.4.10	交通运输行业应用举例——某交通投资建设有限公司的主数据管理	215
8.4.11	政府部门主数据应用举例——某省经信委项目的主数据管理	217

第三篇 数据治理技术

第9章	数据架构和模型	223
9.1	数据架构	223
9.1.1	数据架构规划	224
9.1.2	数据架构设计	228
9.2	数据模型	230
9.2.1	数据模型的定义	230
9.2.2	数据模型的类型	231
9.2.3	数据的物理特征	232
9.2.4	元数据模型	233
9.2.5	主数据模型	235
9.2.6	信息链和信息生命周期	238
9.2.7	数据谱系和影响分析	238
第10章	数据集成	240
10.1	企业应用集成	240
10.1.1	企业应用集成的概念	241
10.1.2	企业应用集成的分类	243
10.1.3	企业应用集成的方法	245
10.1.4	企业服务总线	248
10.1.5	微服务架构	250
10.2	数据集成交换服务	254
10.2.1	制定数据集成交换规范和架构	254
10.2.2	搭建数据交换平台	255
10.2.3	实现数据交换管理	257

10.3	构建数据服务体系	258
10.4	形成数据资产全局视图	258
第 11 章	数据质量管理	262
11.1	数据质量的定义	262
11.1.1	数据质量	262
11.1.2	数据质量维度	263
11.1.3	数据质量评估	265
11.1.4	数据剖析	266
11.1.5	数据质量问题和数据管理问题	267
11.1.6	合理性检查	267
11.1.7	数据质量阈值	268
11.1.8	过程控制	269
11.1.9	联机数据质量的检测和监控	269
11.2	数据质量评估框架	270
11.2.1	数据质量评估框架的背景	270
11.2.2	数据质量评估框架的范围	271
11.2.3	数据质量评估框架的质量维度	273
11.2.4	数据质量期望	274
11.3	数据质量评估测量类型	275
11.3.1	数据模型的一致性	275
11.3.2	数据内容的有效性	276
11.3.3	评估数据内容的一致性	277
11.4	数据评估方案	279
11.4.1	数据初步评估	279
11.4.2	数据质量改进评估	287
11.4.3	数据质量持续改进	288
11.5	数据质量战略	291
11.5.1	数据质量战略的概念	291
11.5.2	数据战略和数据质量战略	292
11.5.3	把数据作为资产	294
11.5.4	监控数据质量	295
第 12 章	主数据全生命周期管理	297
12.1	主数据全生命周期管理及意义	297
12.2	主数据全生命周期管理内容	299

12.2.1	数据申请	299
12.2.2	数据审核	301
12.2.3	数据变更	302
12.2.4	数据集成和数据分发	303
12.2.5	数据查询	306
12.2.6	数据归档	308
12.3	数据清洗管理	308
12.3.1	数据清洗的内容	310
12.3.2	数据清洗的一般过程	312
12.3.3	数据清洗的工具	313
12.4	建立主数据全生命周期管理体系	314
12.4.1	概述	314
12.4.2	建立信息架构	314
12.4.3	发现数据对象	314
12.4.4	分类数据对象和定义服务水平	314
12.4.5	建立测试数据管理策略	315
12.4.6	归档数据	315
第 13 章	数据安全 管理	316
13.1	数据安全的意义和作用	316
13.1.1	数据安全的概念	316
13.1.2	数据安全的意义和作用	316
13.2	数据安全的关键内容	317
13.2.1	数据存储安全	317
13.2.2	数据传输安全	317
13.2.3	数据使用安全	318
13.3	数据隐私保护	320
13.3.1	数据隐私保护的意义和作用	320
13.3.2	数据隐私保护面临的问题和挑战	321
13.3.3	数据隐私保护技术	321

第四篇 数据治理前景

第 14 章	主数据管理应用前景展望	325
14.1	主数据管理应用市场发展趋势	326
14.2	大数据时代的主数据管理	327

14.2.1	大数据的定义和特征	327
14.2.2	大数据时代企业管理的新模式	329
14.2.3	主数据管理在大数据分析中的作用	331
14.2.4	大数据对主数据管理的挑战	333
14.3	基于云服务的主数据管理	334
14.3.1	云服务的定义和发展现状	334
14.3.2	主数据管理的云服务模式	338
14.3.3	主数据管理云服务平台的技术基础	339
14.3.4	云服务对企业主数据管理的影响	341
14.4	面向人工智能的主数据管理	342
14.4.1	人工智能的定义及应用领域	342
14.4.2	人工智能在企业中的实践	345
14.4.3	主数据管理与人工智能的关系	347
14.4.4	主数据管理在人工智能中的作用	348
14.5	区块链技术与主数据管理	351
14.5.1	区块链的定义及特征	351
14.5.2	区块链技术的应用领域	352
14.5.3	区块链技术在主数据管理中的应用	354
14.6	主数据管理——企业发展的坚实根基	356