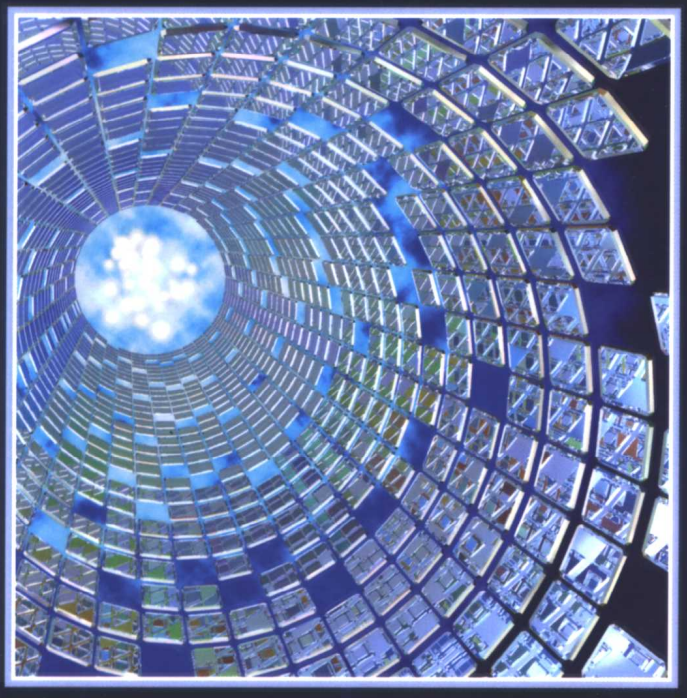
数据库技术丛书

元数据仓储的构建与管理

Building and Managing the Meta Data Repository

A Full Lifecycle Guide



(美) David Marco 著

张铭 李钦 等译

附赠CD-ROM

机械工业出版社
China Machine Press

数据库技术丛书

元数据仓储的构建与管理

Building and Managing the Meta Data Repository

A Full Lifecycle Guide

(美) David Marco 著

张铭 李钦 等译

机械工业出版社
China Machine Press

本书从元数据的基本概念入手，详细介绍了元数据的理论内容、市场前景，并对元数据开发项目的完整生命周期进行了细致阐述，从评估工具、制定计划到配置人员和设计体系结构等各个方面，不一而足。整本书脉络清晰，深入浅出，无论是对元数据理论的研究者还是对具体实施元数据项目的工程人员都不无裨益。

David Marco: Building and Managing the Meta Data Repository: A Full Lifecycle Guide (ISBN: 0-471-35523-2).

Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书版权登记号：图字：01-2002-4819

图书在版编目（CIP）数据

元数据仓储的构建与管理 /（美）麦考（Marco, D.）著；张铭等译. -北京：机械工业出版社，2004.5

（数据库技术丛书）

书名原文：Building and Managing the Meta Data Repository: A Full Lifecycle Guide

ISBN 7-111-14109-1

Ⅰ.元… Ⅱ.①麦… ②张… Ⅲ.元数据-研究 Ⅳ.G201

中国版本图书馆CIP数据核字（2004）第018059号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）

责任编辑：李云静

北京昌平奔腾印刷厂印刷·新华书店北京发行所发行

2004年5月第1版第1次印刷

787mm × 1092mm 1/16 · 17印张

印数：0 001 - 4 000册

定价：35.00元（附光盘）

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

译者序

随着信息技术的发展和应用的日渐普及，人们对信息系统的要求不断提高，对数据的认识也不断深入。从早期文件系统的应用，发展到完善的数据库技术，直到复杂而庞大的数据仓库和数据集市的出现。此时，如何对这些信息系统和数据集合进行管理和访问已经成为数据的生产者和使用者面临的最突出问题。数据的生产者需要有效的数据管理和维护方法，而使用者则要求信息系统易于使用、数据易于理解。在这种情况下，元数据成为信息资源有效管理和应用的重要手段。顺理成章地，元数据也就成为当前信息技术领域中研究的热点之一。

然而，目前的元数据发展现状与早期的软件方法学类似，丰富的理论研究掩饰着缺少实践经验的尴尬。作为企业实施元数据项目时，技术人员缺少完整而实用的工程指导，为了解决这一问题本书应运而生。顾名思义，本书并没有涵盖元数据理论的方方面面（这也不是其目的所在），本书的目的是为实施元数据仓储项目的人员从各个环节上提供实际的指导。本书的作者有着多年为企业构造IT系统的经验，书中列出了作者从业以来经历并解决的许多实际问题，这些问题可能是读者目前正待解决的，也可能是以后将要遇到的。在译者看来，这些实例正是本书的精华之处。

本书从元数据的基本概念入手，详细介绍了元数据的理论内容、市场前景，并对元数据开发项目的完整生命周期进行了细致阐述，从评估工具、制定计划到配置人员和设计体系结构等各个方面，不一而足。整本书脉络清晰，深入浅出，无论是对元数据理论的研究者还是对具体实施元数据项目的工程人员都不无裨益。

全书内容主要包括11章和3个附录，其中11章又分为两大部分。第一部分是一些基础的介绍性内容，第1章和第2章完整地讲述了元数据的意义以及一些基本概念，第3章则从总体上介绍了当前主要的元数据标准。第二部分则根据元数据仓储项目生命周期的各个阶段，分别在第4章到第9章介绍了元数据工具的评估，元数据仓储项目的组织和人员配置，如何制定元数据项目计划，元数据体系结构的设计，如何通过元数据来提高数据质量以及元模型的构造。此外，第10章还详细列出了开发周期的各个阶段应该交付的产品以及取得的阶段性成果。最后，第11章展望了元数据的未来发展趋势。

为了便于读者在实际项目中对上述知识加以应用，本书的附录A和附录B分别列出了通用的元数据工具评估表和元数据项目方案。而附录C则列出了建立元模型时所需要的详细DDL代码，读者在建立自己的元模型时可以参考这些代码。

张铭、李钦组织并参加了本书的翻译和审校工作，参与本书部分翻译工作的还有薛明、吴先荣、丁婷、包小源、李丽、陈佳、韩近强等。由于本书涉及的是比较前沿的主题，许多术语

还没有形成固定的译法，因此翻译起来有一定的难度。尽管译者经过反复讨论和推敲，但由于水平所限，有些中文词语可能难以精确地表达原文的意思，译文中也可能存在一些不当之处，希望读者不吝赐教。

译 者

2003年冬于北京大学

书　评

“David Marco的著作提供了许多实用观点，这些观点是他本人作为顾问从实际经验和教训中得来的。一旦机构开始理解元数据对于企业成本效率管理的重要性，则本书对于初学者、管理员和IT专家都将成为无价之宝。即使读者所在的机构还没有准备开始制订企业范围的元数据战略，本书也将提供必要的概念，使得他们能够估计其目前的计划能提供什么帮助，并能为将来成为智能化的企业而铺平道路。”

Katherine Hammer
Evolutionary Technologies International（ETI）公司总裁和CEO（首席执行官）
元数据联盟主席
《Workplace Warrior》作者

“这是第一本讨论数据仓库中的元数据问题的著作，非常引人入胜。尽管‘元数据’容易让人误以为是知识技术，David Marco的著作还是写得易于接受、实践性强并可以立即被用于实践。本书是IT专业人士的优秀资源。”

Steve Murchie
微软公司产品经理

“元数据泰斗David Marco在这本新著中再一次证明了他在本领域的造诣。本书从重要的原则和实际解决方案的角度讨论了元数据领域的关键问题。对于那些希望了解元数据仓储的战略重要性和实现方法的人士来说，本书是必读教材。”

Charlie Chang
Informix软件公司高级副总裁

“如果把元数据看成是数据仓库的‘粘合剂’的话，这本书就是数据仓库工程管理者组织其工程项目的关键要素。就像优秀的元数据一样，本书中的信息正确、综合性强、易于理解。本书应该是数据仓库工程开发者的必读之物。”

Wayne Eckerson
Data Warehousing Institute教育和研究主管

“元数据是数据仓库的关键成功因素之一。元数据的实现对于很多机构来说是很头疼的问题，因为这些机构不知道实现元数据的正确方向。David Marco的这本著作指明了这个方向，并可以作为实现元数据的蓝图。”

Sid Adelman
Sid Adelman & Associates公司总裁

“元数据管理是通向未来成功电子商务的关键。David Marco的著作中介绍了许多实践经验。每一位考虑元数据实现策略的人，不管其元数据是为数据仓库工程、商业智能还是电子商务服务的，都应该把本书放在他们的案头。”

G. Allen Houpt
Computer Associates International公司知识管理、业务经理

序　言

计算机出现的早期，数据传输和存储是使用穿孔卡片和纸带，然后是磁盘和随机访问介质。不久数据库出现了，联机应用程序紧随其后。接下来又出现了爬虫Web环境（spider web environment），这导致了数据仓库的出现。在数据仓库基础之上又产生了数据集市、操作型数据存储和信息发掘仓库（exploration warehouse，或称探索型数据仓库）。

每种形式的信息处理都会导致另一种更复杂形式的产生。这些处理形式逐渐发展成被称为“企业信息源”（corporate information factory）的框架。

但是不同信息处理形式之间的整合是很难实现的。每种处理形式都有自己的目标和技术，其中大多数都是自身特有的。很难在不同的信息处理形式之间形成并维护某种统一形式。

企业级整合的惟一希望在于元数据。但元数据本身也是一个令人困惑的主题，因为其形式也是多种多样的。无论如何，企业中的每种处理形式都有自己的元数据形式。而磁带机的元数据完全不同于近线存储的元数据，也不同于数据集市的元数据。此外，将数据仓库连接到ODS（operational data store，操作型数据存储）所需要的元数据也不同于ETL（extraction，transformation and load；抽取、转换和加载）中的元数据。

我们需要的是一点点次序和组织。如果要在整个企业中实现集成和协调，起点毫无疑问应该是元数据。

但是，试图完全理解元数据，就好比是与章鱼较劲，要在水下屏住呼吸。仅仅是因为问题有如此之多的方面，就会使取得进展十分困难。遭到“灭顶之灾”的可能性很大。

David Marco的书称得上是在克服这种困难方面所作努力的一个里程碑。从概念到实际，David对元数据的很多方面都有很好的解释。当前，对元数据的论述往往是从特定工具的角度出发，片面地讨论元数据的一到两个方面。David的与众不同之处在于：从元数据的一个方面到另一个方面，直至方方面面，他都乐于面对。

要想了解当代的元数据观点，应该听听David Marco的见解。

——W. H. Inmon

Pine Cone Systems公司首席技术官

前　言

20世纪50年代和20世纪60年代人们开始构造计算机系统，从那时起人们就意识到需要用“各种材料”（bunch of stuff，即知识）来构造、使用和维护这些系统，但却不清楚如何将计算机系统的知识与了解市场和行业所必需的“其他材料”集成起来。幸运的是，经过一段时间后人们发现，信息系统需要的是关于正在使用的业务数据的数据。换句话说，需要的是元数据（meta data）。

当我们讨论元数据时，实际上是在讨论“知识”——关于系统、业务、竞争、客户、产品和市场的知识。在当今的时代，这些知识可以提供足以决定业务成败的竞争优势。这个时代对企业的要求远胜于从前，企业必须比其竞争对手更出色，才能更好地生存和发展。元数据可以提供非常显著的竞争优势，但是人们必须完全理解并懂得如何有效地应用元数据。

本书的组织

当作者购买关于信息技术（或其他任何主题）的书时总是从几个主要方面来挑选，但最主要还是寻找一本能够与作者个人产生共鸣、寓教于乐的书。作者还喜欢购买那些既有理论基础，又有可靠的实践经验的书。作者特意寻找的是那些只能通过实践获得的信息——如果一本书给出的例子提供了有用的教训（即使只有一个）或是防止项目中可能会犯的错误，那么这本书就算得上物有所值。在写作本书时，作者尽量记住自己最关心的内容，为读者提供关于元数据的可靠基础知识（而不假设读者已有这方面的知识），并根据自己从事顾问的多年经验来提供实用经验。

本书的第一部分除了提供理解元数据的基础以外，还讨论了元数据为机构带来的特殊价值，即元数据如何帮助企业提高收入或降低开支。对于那些向行政管理层推销元数据概念的人来说，这些优点是非常有用的。第一部分还讨论了影响元数据行业的一些主要趋势，例如愈演愈烈的标准之争和可扩展标记语言（XML）的出现。元数据无可争辩地成为信息技术中变化最快的领域之一。元数据还是（尽可能地）理解纷至沓来的变化的关键因素，从而使构造的仓储的灵活性足以应对这些变化。

第二部分主要关注元数据仓储的实现，详细介绍如何建立合理的体系结构、组建仓储团队、构造元数据模型以及选择必要的元数据工具。这一部分还详细讨论了如何使用元数据来保证数据仓库和数据集市中数据的质量，以及从仓储和决策支持系统（DSS）中生成有用信息。

众所周知，真相往往比虚构的内容还奇怪，现实生活也常常比虚构的戏剧还有趣。本书第一部分和第二部分中提到的一些“战争故事”可以使读者体会到，决策支持和元数据仓储项目通常比小说还要奇怪和有趣。其中的许多故事使读者沉浸在娱乐之中，但所有的故事都是为了指导读者应该做什么以及避免做什么。

本书的读者对象

只要对元数据仓储合理运用，并且每个人都理解元数据能做什么而不能做什么，元数据仓储就可以为企业带来巨大价值。当然，“每个人”这个词的范围十分广泛，但通过阅读本书或其中一部分，下列人员最有可能从中获益：

- **业务用户** 元数据仓储可以显著提高决策支持系统和操作型系统中保存的信息的价值，因为它提供了信息技术（IT）系统和业务用户之间的语义链接。当业务用户理解了如何有效地使用元数据时，他们对决策支持信息的精确性和完整性才更有信心，也才更有可能依靠决策支持信息做出战略性业务决策。
- **IT经理** IT经理可以利用元数据仓储为其支持的业务单元带来大量价值，并确保数据仓库中信息的质量，从而协助业务用户和行政管理人员根据精确及时的信息做出可靠的决策。此外，元数据仓储可以提高IT开发人员的劳动效率，并降低部门的开发成本。
- **开发人员** 开发人员需要了解实现元数据仓储项目的关键任务。这些任务包括物理元数据建模、项目方案制订、程序设计、元数据工具评估尺度、元数据访问技术和高级技术体系结构设计等。
- **项目发起者** 项目的发起者需要理解元数据如何使企业获益，这样他们才能向行政管理人员推销元数据概念。低估元数据仓储项目的作用是项目失败的主要原因，而发起者需要对元数据及其潜在的投资回报率（ROI）有清晰的认识，只有这样才能确保项目在初期能通过审批并保证项目实施过程中资金和人员的充分投入。没有这样的认识，发起者就不可能成为元数据的坚定倡导者。

关于Web站点

本书的相关Web站点是www.wiley.com/compbooks/marco。这个免费的Web站点上有一些指向元数据集成和访问工具生产商的链接，还有一些其他的元数据相关特性。此外还欢迎本书的所有读者在www.EWSolutions.com/newsletter.asp上登记免费订阅《Real-World Decision Support》（RWDS）。RWDS是一份电子通讯，致力于提供丰富的、生产商中立的实际解决方案，以帮助解决决策支持系统和元数据仓储实现中遇到的困难。

目　录

第二部分　元数据仓储的实施

第三部分 附 录

第一部分 基础铺垫

第1章 介绍元数据及其投资回报

在决定构造元数据仓储之前，需要全面地考虑什么是元数据以及什么不是元数据，还要考虑元数据仓储究竟能为所在机构带来多大价值。本章简单地回顾元数据的历史，然后很快会探讨为什么需要元数据以及合理使用元数据的企业是如何借此提高自身竞争力的。

1.1 起步阶段

信息技术（IT）仍然处于幼年期，同时，就如幼儿一样，其成长之快令人不可思议。全世界用于IT方面的支出在1999年预计为22 000亿美元，到2002年以前预计将会攀升到33 000亿美元。如果与过去相比，这种增长就更为显著了。第一台通用计算机出现于20世纪40年代末期，而就在短短的20年之前，我们还在用穿孔卡片来编程。（许多人都还对卡片弄乱后重新排序的噩梦记忆犹新！）

如今，工业发展的步伐已经放缓。计算机几乎改变了人类生活的每一方面，即便如此，计算机的发展也还只是处在一个“呀呀学语”的时期。

信息技术的啼声初探

现在的IT系统已经足以满足公司的日常业务交易。如果这些业务是静态的，那么这已经足够了。但众所周知，业务根本不会是静态的。业务会随着社会、技术、政治以及行业因素的变化而不断变化。企业由IT系统控制，这些系统必须进行相应的改变，否则企业就无法对大量而多变的市场因素做出反应。

但是，现在的这些计算机系统都是不可变化的。事实上，现在构造的这些系统只不过是一些“数据孤岛”，对这些系统做出改变的难度也几乎等同于移动一座岛屿。即使对于当今最完备的系统也是这样。至于为什么会发生这种情况是很容易理解的。回想20世纪70年代末和80年代初，数据存储的成本很高，IT开发人员的成本则相对低廉，所以“能干”的程序员们就决定尽可能地节省存储空间。这种做法不仅使得IT系统难以维护，并且还有可能导致将来出现问题。在节省存储空间的尝试中，最有名的例子就是使用2位数字来表示年/日期字段。当初采取这种做法时，没有人想到这些IT系统会一直用到新千年的到来。所有人都抱着这样的想法：20年以后这些系统都会被全新的系统替换。多么幼稚的错误！构造更好的全新系统远比人们预计的要复杂得多。

刚刚所提到的那个例子正是臭名昭著的千年虫（Y2K）问题，几乎所有人都对这个问题有

所耳闻。Y2K问题清楚地表明，现有的系统无法轻易地适应变化。这个问题也使人们意识到，必须对数据有更深入的理解，对系统有更好的控制，才能够适应不断变化的业务需求。幸运的是，这一行业开始变得成熟而理智。元数据提供了满足这些需求的答案，并当之无愧地受到了整个行业的关注。

1.2 定义元数据

关于元数据最简单的定义是描述数据的数据（data about data）。这一定义并不能完整地涵盖元数据的所有领域。在第2章中提供了元数据的详细定义，但现在还是先用下面这个简短的定义：

> 元数据是指与业务和技术过程及企业使用数据有关的所有物理数据以及包含知识的信息。

现在再对这个定义进行一些扩展。

> 元数据是指来自企业内外的所有（软件和其他介质中含有的）物理数据和（员工和各种媒介中含有的）知识，包括物理数据的格式、技术和业务过程、数据的规则和约束以及企业所使用数据的结构。

谈论元数据实际上是在谈论知识（knowledge），包括系统、业务和市场的知识。另一方面，讨论元数据仓储也是在讨论存储元数据所使用的物理数据库表，这些元数据被传递给业务用户和技术用户（参见图1-1）。尽管元数据启动的物理实现需要许多操作，但元数据仓储是物理实现的“中枢”。

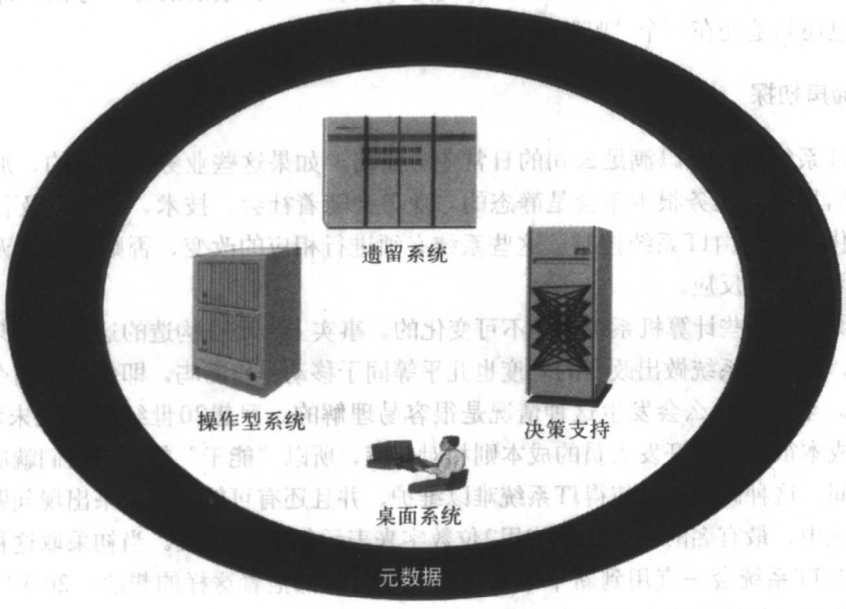
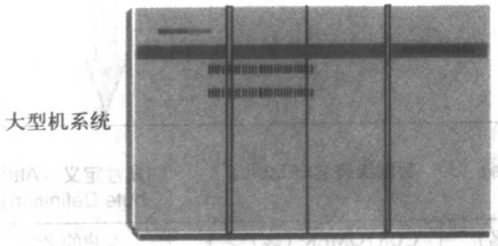


图1-1　元数据交互

1.3 元数据——开端

许多人以为元数据和元数据仓储是全新的概念，但实际上这些概念的起源都可以追溯到20世纪70年代。出现的第一代商业元数据仓储系统被称作数据字典（data dictionary）。这些数据字典更关注对数据（data）而不是知识（knowledge）的研究。数据字典提供了数据相关信息的集中仓储，包括定义、关系、起源、作用域、用法和格式，其目的在于帮助数据库管理员（DBA）制定计划、控制和评估集合、存储以及数据的使用。例如，早期的数据字典主要用于定义需求、对企业数据建模、生成数据定义以及支持数据库。

现在需要面对的挑战是区分元数据仓储与数据字典的不同。尽管元数据仓储执行数据字典的所有功能，但其作用范围要大得多（参见图1-2）。



数据字典信息 属性名（Attribute Name）	数据库表名（Table）	属性定义（Attribute Definition）
Customer_Name（客户名）	CUSTOMER（客户表）	40字节的字母数字串
Social_Security_Number（社会保险号）	CUSTOMER（客户表）	必须是9字节的数字值，中间不允许有破折号
Customer_State（客户所在州）	CUSTOMER（客户表）	必须是2字节的字母数字串，而且应该是合法的州名缩写。例如，IL（伊利诺伊州）、NY（纽约州）、CA（加利福尼亚州）等

图1-2 20世纪70年代：以数据字典的名义出现的仓储系统

1.3.1 元数据的商业性演化

20世纪70年代引入的计算机辅助软件工程（CASE，Computer Aided Software Engineering）工具就是提供了元数据服务的第一代商业工具之一。

CASE工具不仅极大地推动了数据库和软件应用程序的设计进程，也将关于所管理数据的数据保存下来。用户很快就开始要求CASE工具生产商能够提供接口，以便将元数据与不同CASE工具链接在一起。而生产商并不愿意提供这样的接口，因为他们认为自己的工具能够提供必需