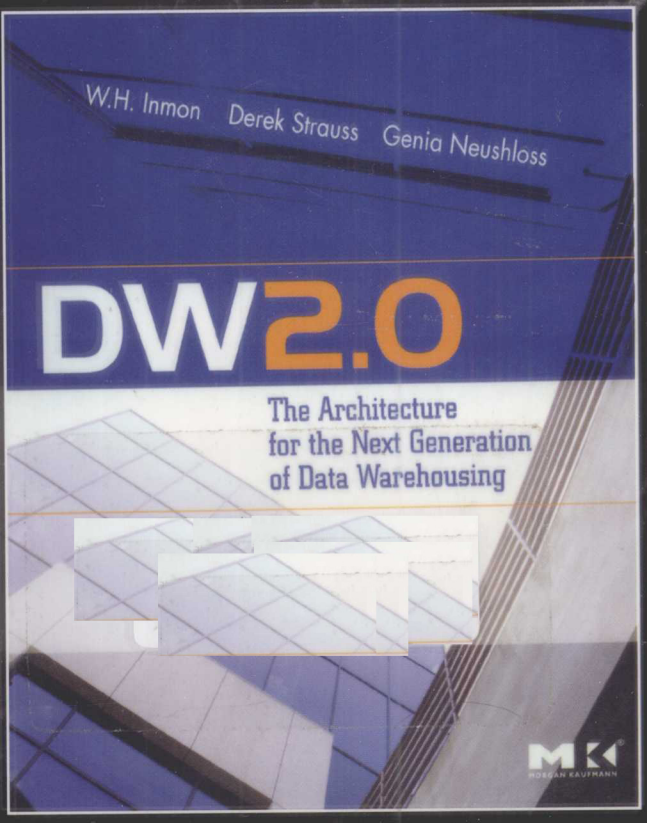
HZ BOOKS
华章教育

计 算 机 科 学 丛 书

ELSEVIER

DW2.0
下一代数据仓库的构架

(美) W. H. Inmon Derek Strauss Genia Neushloss 著 王志海 王建林 付彬 武婷婷 等译



DW2.0
The Architecture for the Next Generation of Data Warehousing

机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

DW2.0
下一代数据仓库的构架

(美) W. H. Inmon Derek Strauss Genia Neushloss 著 王志海 王建林 付彬 武婷婷 等译

DW2.0
The Architecture for the Next Generation of Data Warehousing

机械工业出版社
China Machine Press

本书是数据仓库和商业智能领域的又一部经典著作，讲述了整个生命周期各个环节的具体工作，从业务需求的视角，引导读者全面认识下一代数据仓库系统的构架。本书包含了DW2.0详细的定义和描述，此外，书中对数据仓库的结构、内容及其前景进行了介绍。

本书主要面向数据仓库的业务分析人员、信息构架师、系统开发人员、项目经理、数据仓库技术人员、数据库管理员、数据建模人员、数据管理员等。

W. H. Inmon, Derek Strauss and Genia Neushloss: DW2.0: The Architecture for the Next Generation of Data Warehousing (ISBN 978-0-12-374319-0).

ISBN: 978-981-272-335-2

本书版权登记号：图字：01-2009-2371

图书在版编目（CIP）数据

DW2.0：下一代数据仓库的构架/（美）英蒙（Inmon，W. H.）等著；王志海等译. —北京：机械工业出版社，2010.3

（计算机科学丛书）

书名原文：DW2.0：The Architecture for the Next Generation of Data Warehousing

ISBN 978-7-111-28826-8

Ⅰ. D…　Ⅱ. ①英… ②王…　Ⅲ. 数据库系统　Ⅳ. TP311.13

中国版本图书馆CIP数据核字（2009）第228130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：迟振春

北京诚信伟业印刷有限公司印刷

2010年3月第1版第1次印刷

184mm×260mm · 14.5印张

标准书号：ISBN 978-7-111-28826-8

定价：45.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www. hzbook. com
电子邮件：hzjsj@ hzbook. com
联系电话：（010）88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037

华章教育
华章科技图书出版中心

译者序

在过去二十年中，数据仓库的概念一直在逐步进化，DW2.0是对数据仓库概念最新的理解和描述。自从本书作者Bill Inmon首次给出数据仓库定义之后，该定义就一直被众多研究者和开发者所引用。然而，人们常常陷入什么是数据仓库或什么不是数据仓库这样的混乱或疑惑。在这种情况下，DW2.0尝试对下一代数据仓库进行全方位的定义。与术语“数据仓库”不同，DW2.0有着简明扼要和清晰可辨的含义，本书对其含义进行了详细的论述和准确的定义。

本书是数据仓库和商业智能领域的又一部经典著作，作者Bill Inmon等人在数据仓库领域享有很高的声誉，他们都长期工作在数据仓库系统开发的第一线，将自己多年的经验和感悟融入到了本书的字里行间。本书讲述了整个生命周期各个环节的具体工作，从业务需求的视角，引导读者全面认识下一代数据仓库系统的构架。本书包含了DW2.0详细的定义和描述，所有的内容被分为不同的章节，其中每一个章节都相当于该部分内容的白皮书。此外，书中对数据仓库的结构、内容及其前景进行了介绍。本书主要面向数据仓库的业务分析人员、信息构架师、系统开发人员、项目经理、数据仓库技术人员、数据库管理员、数据建模人员、数据管理员等。

本书的翻译凝结了许多人的智慧。最初，第1章由付彬翻译，第2章由李波翻译，第3章由邵金刚和李亚飞翻译，第4章由冯瑶翻译，第5章由徐闻琦翻译，第6章由王倚丹翻译，第7章与第8章由张森翻译，第9章由刘雪莲翻译，第10章由毛佳敏翻译，第11章由杨磊翻译，第12章由李志尧翻译，第13章由武婷婷翻译，第14章由郑超翻译，第15章由王鑫翻译，第16章与第17章由俞雪娇翻译，第18章由郑超翻译，第19章由邵晓康翻译，第20章、第21章和第22章由武婷婷翻译，第23章由冯瑶翻译。在此基础上，付彬和武婷婷规范了全书的术语，并进行了认真的修订。冯浩、王世强、邵鲁杰、邵进智、孙兴中、贺一航、秦逞、赵飞国、刘礼辉、王辉、张学勇、刘学军、冯岩、杨迪、黄禹钦以及王中锋等参与了本书翻译的讨论。最后，由北京交通大学王志海教授和滨州学院王建林老师审核了全书。

在翻译过程中，我们无一不被Inmon教授等人的睿智和巨大贡献所打动，秉持“形似、意似、神似”的翻译原则，尽最大的努力，希望奉献给广大读者一部真实反映原著风貌的科技书籍。

当然，要译好一本经典著作并不是一件容易的事情，我们的水平还很欠缺，错误之处还望广大读者批评指正。

译　者

2010年1月

前　言

数据仓库已经问世二十多年了，它已成为信息技术基础设施的基本组成部分。数据仓库的出现最初是为了满足对信息而不是对数据的企业需求。数据仓库是一个能够为企业提供整合的、粒度的、历史的数据的结构。

然而，数据仓库存在一个问题，即当前对数据仓库还存在多种不同的解释和实现方式。例如，有联合数据仓库、主动数据仓库、星状模式数据仓库、数据集市数据仓库等。实际上，有多少软硬件供应商，就有多少对数据仓库的诠释和实现方式。

还有一个问题就是，对什么样的结构才是数据仓库适合的，也存在着多种不同的解释和实现方式。而且，每一种实现在构架上都与其他的实现有很大区别。如果走进一个房间，里面联合数据仓库的支持者正在与主动数据仓库的支持者交谈，你也许会听到一些相同的词语，但这些词代表的意思却大相径庭。即使使用相同的词语，你听到的可能也不是有意义的交流。当两个不同背景的人交谈时，即使使用相同的词语，也不能保证他们彼此能够相互理解。

于是，今天的第一代数据仓库就处于这种情况下。

在陷入什么是数据仓库或什么不是数据仓库这样的混乱或疑惑的情况下，出现了DW2.0。DW2.0是对下一代数据仓库的定义。与术语“数据仓库”不同，DW2.0有着简明扼要和清晰可辨的含义。本书对其含义进行了论述和定义。

DW2.0中有很多重要的构架上的特征。这些构架特征代表了DW2.0相对于第一代数据仓库在技术和构架上的进步。在本书中，我们讨论了DW2.0的如下几种重要特性：

- 认识到数据仓库中数据的生命周期。第一代数据仓库仅仅将数据放于磁盘存储器（称之为仓库）中。事实上，数据一旦被置于数据仓库，它就有了自己的生命周期。进入数据仓库后，数据开始老化，数据被访问的可能性也逐渐降低。而数据访问的可能性降低对选择适当的数据管理技术有着深远的含义。另一种现象是，随着数据老化，数据容量会不断增加，并且大多数情况下这种增加是显著的。想要处理访问可能性不断降低的大量数据，就需要一种特定的设计，以免数据仓库的花费巨大，以至于不能有效地使用数据仓库。
- 当既包含结构化数据又包含非结构化数据时，数据仓库是最有效的方法。典型的第一代数据仓库完全由面向事务的结构化数据组成，这些数据仓库提供了大量有用的信息。然而，现代数据仓库应该同时包含结构化数据和非结构化数据。非结构化数据是一些文本数据，包括医疗记录、合同、电子邮件、电子表格以及很多其他的文档。非结构化数据中存在着大量的信息，但如何获取这些信息却着实是一个挑战。对创建同时包括结构化数据和非结构化数据的数据仓库都有哪些要求的具体描述是DW2.0中的一个重要部分。
- 由于多种原因，元数据并没有成为第一代数据仓库的重要组成部分。而在定义第二代数据仓库时，元数据的重要性和作用开始得到认可。在DW2.0中，问题并不是对于元数据的需求。元数据存在于数据库管理系统目录中，存在于业务对象领域中，

存在于 ETL 数据预处理工具中，等等。我们需要的是企业元数据，是从企业级视角理解元数据，需要调节元数据的所有来源并将它们放置在一个能使它们协调工作的环境中。除此之外，在 DW2.0 环境中还需要技术元数据和业务元数据的支持。

- 数据仓库最终建立在一种技术基础之上。数据仓库是围绕业务需求展开的，这通常会反映在数据模型上。随着时间的推移，企业的业务需求会发生变化，但数据仓库的技术基础却不能很容易地改变。这样，就出现了一个问题，即业务需求持续变化，而技术基础却不变。企业中这种不断变化的业务环境与相对稳定的技术环境之间的矛盾会在机构内形成很紧张的局势。在本书的相关部分中，集中讨论了两种解决方案，用于处理数据仓库中这种变化的业务需求和不变的技术基础之间的难题。一种解决方案是采用诸如 Kalido 这样的软件，其为数据仓库提供了一种有延展性的技术基础。另一种解决方案是在数据库定义时，通过设计来分离静态数据和临时数据。这两种方案对数据仓库的技术基础随着业务需求的改变而改变来说有很好的效果。

另外，书中还讨论了其他一些重要的话题。其中一些包括：

- DW2.0 数据仓库基础设施的在线更新。
- ODS 适用于哪里？
- 针对 DW2.0 数据仓库的研究处理过程和统计分析。
- DW2.0 数据仓库环境下的归档处理。
- DW2.0 数据仓库环境下的近线处理。
- 数据集市及 DW2.0。
- 数据仓库中的粒度数据和数据容量。
- 方法论及开发方式。
- DW2.0 的数据模型。

本书的一个重要特色是运用示意图来从整体上描绘 DW2.0 的环境。示意图是经过多次咨询、研讨才确定的，它代表了 DW2.0 中放置在一起的不同组件，是 DW2.0 环境的一个基本构架表现。

此外，书中对数据仓库的结构、内容及其前景进行了介绍。本书适用于业务分析人员、信息构架师、系统开发人员、项目经理、数据仓库技术人员、数据库管理员、数据建模人员、数据管理员等。

关于作者

W. H. Inmon：数据仓库之父。他已编写了 49 本著作，并被译成 9 种语言。Bill 创建了世界上第一个 ETL 软件公司。他在大多数主要的行业期刊上发表了 1000 多篇论文。

除南极洲之外，Bill 在各大洲都组织过研讨会并在各种会议上发言。他拥有九项软件专利。他最新成立的一个公司是 Forest Rim Technology 公司，该公司致力于非结构化数据的存取并将其整合到结构化环境中。每月有超过 1 000 000 人访问 Bill 的网站：inmoncif. com。他在 b-eye-network. com 上的每周通讯已经在业界被广泛阅读，每周有 75 000 个订阅者。

Derek Strauss：Gavroshe 公司的创始人、CEO 和首席顾问。他拥有 28 年 IT 界从业经验和 22 年信息资源管理及商业智能/数据仓库领域的从业经验。

Derek 发起并管理了许多企业项目，他倡导运用商业智能、数据仓库来改善数据质量。Bill Inmon 的 CIF（Corporate Information Factory）理论及 John Zachman 的 EAF（Enterprise Architecture Framework）理论是 Derek 的工作的基石。Derek 同时也是一名专家研讨会主持人，他曾多次在国内及国际的数据仓库会议中演讲。另外，他还是 DW2. 0 认证的构架师和培训师。

Genia Neushloss：Gavroshe 公司的联合创始人和首席顾问。30 多年来，她在保险业、金融业、制造业、采矿业及电信业都拥有相当深厚的管理及技术经验。

Genia 曾举办 JAD/JRP 和系统再造培训课程，是系统再造方法集的编码开发者之一。她拥有 22 年规划、分析、设计和构建数据仓库的专业经验。Genia 多次在欧洲、美国和非洲等与观众见面。另外，她也是 DW2. 0 认证的构架师和培训师。

目 录

出版者的话
译者序
前 言
关于作者

第 1 章 数据仓库简史及第一代数据仓库 …… 1
1.1 数据库管理系统 …… 1
1.2 在线应用 …… 1
1.3 个人电脑和 4GL 技术 …… 2
1.4 蜘蛛网环境 …… 2
1.5 企业角度的演化 …… 3
1.6 数据仓库环境 …… 4
1.7 什么是数据仓库 …… 4
1.8 整合数据——一个痛苦的经历 …… 5
1.9 数据的量 …… 5
1.10 一种不同的开发方法 …… 6
1.11 演变到 DW2.0 环境 …… 6
1.12 数据仓库的商业影响 …… 7
1.13 数据仓库环境的各种组件 …… 7
1.13.1 ETL——抽取/转换/装载 …… 8
1.13.2 ODS——操作数据存储 …… 8
1.13.3 数据集市 …… 8
1.13.4 探索仓库 …… 9
1.14 数据仓库的演变——从企业的角度 …… 9
1.15 关于数据仓库的其他观念 …… 9
1.16 主动数据仓库 …… 10
1.17 联合数据仓库方法 …… 10
1.18 星状模式方法 …… 12
1.19 数据集市数据仓库 …… 12
1.20 建立一个“真正的”数据仓库 …… 13
1.21 总结 …… 14

第 2 章 DW2.0 简介 …… 15
2.1 DW2.0——一种新的范式 …… 15
2.2 DW2.0——从企业的角度 …… 15
2.3 数据的生命周期 …… 17
2.4 设置不同区的原因 …… 19
2.5 元数据 …… 20
2.6 数据访问 …… 21
2.7 结构化数据/非结构化数据 …… 22
2.8 文本分析 …… 24
2.9 “废话” …… 25
2.10 术语问题 …… 25
2.11 特定文本/一般文本 …… 26
2.12 元数据——一个主要组成部分 …… 26
2.13 本地元数据 …… 28
2.14 基础技术 …… 29
2.15 不断变化的业务需求 …… 31
2.16 DW2.0 中的数据流 …… 32
2.17 数据量 …… 32
2.18 实用应用程序 …… 33
2.19 DW2.0 和参照完整性 …… 35
2.20 DW2.0 的报告 …… 35
2.21 总结 …… 36

第 3 章 DW2.0 组成部分——关于不同区 …… 38
3.1 交互区 …… 38
3.2 整合区 …… 41
3.3 近线区 …… 48
3.4 归档区 …… 50
3.5 非结构化处理 …… 56
3.6 企业用户的观点 …… 59
3.7 总结 …… 59

第 4 章 DW2.0 中的元数据 …… 61
4.1 数据和分析的可复用性 …… 61
4.2 DW2.0 中的元数据 …… 61
4.3 主动知识库/被动知识库 …… 63
4.4 主动知识库 …… 64
4.5 企业元数据 …… 64
4.6 元数据和记录系统 …… 65

4.7 分类 …… 66
4.8 内部分类/外部分类 …… 66
4.9 归档区元数据 …… 67
4.10 维护元数据 …… 67
4.11 举例说明如何使用元数据 …… 67
4.12 终端用户的观点 …… 69
4.13 总结 …… 70
第5章 DW2.0技术基础设施的流动性 …… 71
5.1 技术基础设施 …… 71
5.2 快速的业务改变 …… 72
5.3 环状改变 …… 73
5.4 打破循环 …… 73
5.5 缩短IT响应时间 …… 73
5.6 语义暂态、语义常态数据 …… 74
5.7 语义暂态数据 …… 74
5.8 语义稳定的数据 …… 74
5.9 混合语义稳定和不稳定数据 …… 75
5.10 分离语义稳定和不稳定数据 …… 76
5.11 减缓业务的改变 …… 76
5.12 创建数据快照 …… 76
5.13 历史记录 …… 77
5.14 数据划分 …… 77
5.15 终端用户的观点 …… 78
5.16 总结 …… 78
第6章 DW2.0的方法与途径 …… 79
6.1 螺旋式方法——主要特点综述 …… 79
6.2 七流法——总览 …… 82
6.3 企业参考模型流 …… 82
6.4 企业知识协调流 …… 83
6.5 信息工厂开发流 …… 84
6.6 数据归档定位流 …… 84
6.7 数据纠正流（旧称数据清理流） …… 84
6.8 基础设施流 …… 84
6.9 整体信息质量管理流 …… 86
6.10 总结 …… 88
第7章 统计处理和DW2.0 …… 90
7.1 两种类型的处理 …… 90
7.2 使用统计分析 …… 91
7.3 比较的完整性 …… 91
7.4 启发式分析 …… 92
7.5 冻结的数据 …… 93
7.6 探索型处理 …… 93
7.7 分析频率 …… 93
7.8 探索工具 …… 93
7.9 探索型处理数据的来源 …… 95
7.10 更新探索数据 …… 95
7.11 基于项目的数据 …… 95
7.12 数据集市和探索工具 …… 96
7.13 数据回流 …… 97
7.14 在内部使用探索数据 …… 98
7.15 企业分析员的观点 …… 99
7.16 总结 …… 100
第8章 数据模型与DW2.0 …… 101
8.1 智能路线图 …… 101
8.2 数据模型和企业 …… 101
8.3 整合范围 …… 101
8.4 区别粒状型数据和概括型数据 …… 102
8.5 数据模型的层次 …… 102
8.6 数据模型和交互区 …… 104
8.7 企业数据模型 …… 104
8.8 模型转化 …… 105
8.9 数据模型和非结构化数据 …… 105
8.10 企业用户的观点 …… 106
8.11 总结 …… 107
第9章 监视DW2.0环境 …… 108
9.1 监视DW2.0环境 …… 108
9.2 事务监视 …… 108
9.3 数据质量监视 …… 108
9.4 数据仓库监视 …… 108
9.5 事务监视——响应时间 …… 109
9.6 高峰期处理 …… 110
9.7 ETL数据质量监视 …… 110
9.8 数据仓库监视工具 …… 111
9.9 休眠数据 …… 112
9.10 企业用户的观点 …… 112
9.11 总结 …… 113

第 10 章 DW2.0 与安全 …… 114
10.1 保护访问数据 …… 114
10.2 加密技术 …… 114
10.3 缺点 …… 114
10.4 防火墙 …… 115
10.5 使数据脱机 …… 115
10.6 限制性加密 …… 116
10.7 直接转储 …… 116
10.8 数据仓库监视 …… 117
10.9 检测攻击 …… 117
10.10 近线区数据的安全 …… 118
10.11 企业用户的观点 …… 118
10.12 总结 …… 119
第 11 章 时间相关数据 …… 120
11.1 DW2.0 中的所有数据——与时间相关 …… 120
11.2 交互区中的时间相关性 …… 120
11.3 DW2.0 其他部分中的数据相关 …… 121
11.4 整合区中的事务处理 …… 121
11.5 离散数据 …… 121
11.6 连续时间段数据 …… 122
11.7 一个记录序列 …… 123
11.8 非重叠记录集 …… 123
11.9 开始和结束一个记录序列 …… 123
11.10 数据的连续性 …… 124
11.11 时间瓦解数据 …… 124
11.12 归档区中的时间相关变量 …… 125
11.13 企业用户的观点 …… 125
11.14 总结 …… 125
第 12 章 DW2.0 的数据流 …… 127
12.1 贯穿整个构架的数据流 …… 127
12.2 进入交互区 …… 127
12.3 ETL 的角色 …… 128
12.4 进入整合区的数据流 …… 128
12.5 进入近线区的数据流 …… 128
12.6 进入归档区的数据流 …… 129
12.7 下降的数据访问概率 …… 130
12.8 数据的异常流 …… 130
12.9 企业用户的观点 …… 131
12.10 总结 …… 132
第 13 章 ETL 处理与 DW2.0 …… 133
13.1 转换数据状态 …… 133
13.2 ETL 适用范围 …… 133
13.3 应用数据到企业数据的转换 …… 133
13.4 ETL 工作模式 …… 134
13.5 源和目标 …… 134
13.6 ETL 映射 …… 135
13.7 状态转换——实例 …… 135
13.8 更加复杂的转换 …… 136
13.9 ETL 与吞吐量 …… 136
13.10 ETL 与元数据 …… 137
13.11 ETL 与审核记录 …… 138
13.12 ETL 与数据质量 …… 138
13.13 创建 ETL …… 138
13.14 代码创建或参数驱动的 ETL …… 139
13.15 ETL 与丢弃 …… 139
13.16 变化数据的捕获 …… 139
13.17 ELT …… 140
13.18 企业用户的观点 …… 140
13.19 总结 …… 141
第 14 章 DW2.0 与粒度管理器 …… 142
14.1 粒度管理器 …… 142
14.2 提高粒度级别 …… 142
14.3 过滤数据 …… 143
14.4 粒度管理器的功能 …… 144
14.5 本地与第三方粒度管理器的比较 …… 144
14.6 粒度管理器的并行化 …… 144
14.7 作为副产品的元数据 …… 145
14.8 企业用户眼中的粒度管理器 …… 145
14.9 总结 …… 145
第 15 章 DW2.0 和性能 …… 146
15.1 好的性能——DW2.0 的基石 …… 146
15.2 在线响应时间 …… 146
15.3 分析响应时间 …… 147
15.4 数据的流动 …… 147
15.5 队列 …… 147
15.6 启发式处理 …… 148
15.7 分析的生产率和响应时间 …… 149
15.8 索引 …… 149

15.9 移除休眠数据 …… 150
15.10 终端用户培训 …… 150
15.11 监控环境 …… 151
15.12 容量规划 …… 151
15.13 元数据 …… 152
15.14 批处理的并行 …… 152
15.15 事务处理的并行 …… 153
15.16 工作负荷量的管理 …… 153
15.17 数据集市 …… 153
15.18 探索工具 …… 155
15.19 将事务分为不同的类 …… 155
15.20 服务标准协议 …… 155
15.21 保护交互区 …… 156
15.22 数据分割 …… 156
15.23 选择合适的硬件 …… 157
15.24 区分“农民”和“探索者” …… 157
15.25 数据的物理分组 …… 157
15.26 检查自动产生的代码 …… 158
15.27 企业用户的观点 …… 158
15.28 总结 …… 158
第16章 迁移 …… 160
16.1 房屋和城市 …… 160
16.2 在一个完美情况中迁移 …… 160
16.3 完美情况几乎永远不会发生 …… 160
16.4 增量式添加组件 …… 161
16.5 添加归档区 …… 162
16.6 建立企业元数据 …… 163
16.7 建立元数据基础结构 …… 163
16.8 “吞没”源系统 …… 163
16.9 作为缓冲器的ETL …… 164
16.10 迁移到非结构化的环境 …… 164
16.11 企业用户的观点 …… 164
16.12 总结 …… 165
第17章 成本验证和DW2.0 …… 166
17.1 DW2.0的成本值吗 …… 166
17.2 宏观层次的价值验证 …… 166
17.3 微观层次的价值验证 …… 166
17.4 公司B拥有DW2.0 …… 167
17.5 生成新的分析 …… 167
17.6 按步骤执行 …… 168
17.7 总成本是多少 …… 169
17.8 考虑公司B …… 169
17.9 考虑DW2.0的成本 …… 169
17.10 信息的现实情况 …… 170
17.11 DW2.0真正的经济效益 …… 171
17.12 信息的时间价值 …… 171
17.13 整合的价值 …… 171
17.14 历史信息 …… 172
17.15 第一代DW和DW2.0——在经济效益上的比较 …… 172
17.16 企业用户的观点 …… 173
17.17 总结 …… 173
第18章 DW2.0中的数据质量 …… 174
18.1 DW2.0中的数据质量工具集 …… 175
18.2 数据分析工具和逆向工程数据模型 … 175
18.3 数据模型种类 …… 176
18.4 数据分析不一致对自上而下建模的挑战 …… 179
18.5 总结 …… 180
第19章 DW2.0和非结构化数据 …… 182
19.1 DW2.0和非结构化数据 …… 182
19.2 文本读取 …… 182
19.3 在哪里进行文本分析处理 …… 183
19.4 文本整合 …… 183
19.5 简单编辑 …… 183
19.6 无用词 …… 184
19.7 同义词替换 …… 184
19.8 同义词串联 …… 185
19.9 同形异义解析 …… 185
19.10 建立主题 …… 185
19.11 外部术语表/分类法 …… 185
19.12 分词 …… 186
19.13 替换拼写 …… 186
19.14 跨语言的文本 …… 187
19.15 直接搜索 …… 187
19.16 间接搜索 …… 187
19.17 术语 …… 187
19.18 半结构化数据/值=名称数据 …… 188
19.19 准备数据所需的技术 …… 188

19.20 关系数据库 …… 188
19.21 结构化/非结构化连接 …… 189
19.22 企业用户的观点 …… 189
19.23 总结 …… 189
第 20 章 DW2.0 与记录系统 …… 191
20.1 其他记录系统 …… 194
20.2 企业用户的观点 …… 194
20.3 总结 …… 194
第 21 章 多方面的话题 …… 196
21.1 数据集市 …… 196
21.2 数据集市带来的便利 …… 196
21.3 转换数据集市数据 …… 197
21.4 监视 DW2.0 …… 198
21.5 在数据集市间移动数据 …… 198
21.6 不合格数据 …… 199
21.7 用以平衡的条目 …… 199
21.8 重新设置值 …… 200
21.9 数据修正 …… 202
21.10 数据移动的速度 …… 202
21.11 数据仓库工具 …… 203
21.12 总结 …… 206
第 22 章 DW2.0 环境中的处理 …… 207
第 23 章 管理 DW2.0 环境 …… 211
23.1 数据模型 …… 211
23.2 构架管理 …… 211
23.2.1 确定什么时候需要归档区 …… 212
23.2.2 确定是否需要近线区 …… 212
23.3 元数据管理 …… 213
23.4 数据库管理 …… 214
23.5 数据管理 …… 214
23.6 系统和技术管理 …… 215
23.7 DW2.0 环境管理人员的管理 …… 216
23.7.1 优化及优先冲突 …… 217
23.7.2 预算 …… 217
23.7.3 进度表和里程碑的确定 …… 217
23.7.4 资源分配 …… 217
23.7.5 管理咨询人员 …… 217
23.8 总结 …… 218

第 1 章　数据仓库简史及第一代数据仓库

起初，人们仅用一些简单的机制来保存数据。例如，串口卡片、纸带、容量很小的磁芯存储器等。那时，存储器非常昂贵并且容量相当有限。

然而随着磁带的发明和使用，一个崭新的时代也随之来临。使用磁带能够廉价地保存海量数据。并且，磁带对数据的记录格式也没有太大的限制。另外，在磁带中数据不仅可以写入还可以重新写入。因此，与早先的存储方法相比，磁带的使用代表了一个巨大的飞跃。

然而，磁带并不是完美的。由于在磁带中是顺序地访问数据，这样为了访问其中 1% 的数据，常常可能需要物理地访问并读取 100% 的数据。另外对于写数据来说，磁带并不是最稳定的介质。磁带上的氧化物脱落或被划掉，都能导致其无法使用。

磁盘存储代表着数据存储的另一个飞跃。使用磁盘存储，可以直接访问数据，也可以重写。另外，还可以一起访问多个数据。总之，磁盘存储有着各种各样的优点。

1.1　数据库管理系统

磁盘存储产生不久，就随之产生了一种称为“DBMS”（数据库管理系统）的软件。DBMS 软件的产生是为了管理磁盘存储。磁盘存储的管理活动包括：

- 确定数据的合适位置。
- 解决当两个或多个数据单元被映射到同一个物理位置时产生的冲突。
- 允许数据被删除。
- 当无法将一条数据记录存储到一个容量有限的物理空间中时，负责为其寻找合适的物理位置。
- 其他。
- 在磁盘存储的这些优点中，数据的快速定位能力无疑是其中最重要的一个，而正是由 DBMS 完成这一重要的任务。

1.2　在线应用

一旦利用磁盘存储和 DBMS 使数据能够被直接访问后，就很快出现了所谓的在线应用。在线应用使用计算机来实现对数据的快速一致的访问。目前，已有多种商业的在线处理应用，包括 ATM（自动柜员机）、银行出纳处理、投诉处理、航空订票处理、制造控制处理、零售网点的销售处理等。简而言之，在线系统的出现使得各机构进入了能满足顾客日常需求的 20 世纪。在线应用开始变得强大并且普及起来，并且很快成长为交叉应用。

图 1-1 解释了这种信息系统的早期演化。

实际上，在线应用非常受欢迎，增长得很迅速，以至于在短期内就迅速出现了大量的应用。但是这些应用也带来了终端用户的抱怨——“我知道我想要的数据是在某个地方，

只要我能找到它。”这是个实际的情况，公司拥有了一大堆数据，但是查找数据却完全是另外一回事。并且，就算你能找出来，也不能保证你所找到的数据就是正确的。公司的数据正在激增，以至于在任何一个时间点用户都无法保证他们所获得的数据的正确性和完整性。

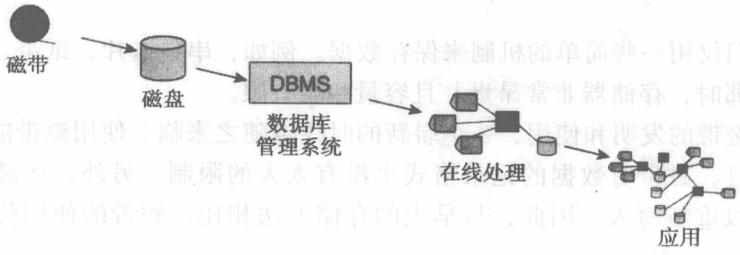


图 1-1 信息系统的早期演化

1.3 个人电脑和 4GL 技术

为了平息终端用户对访问数据的抱怨，两种新的技术应运而生——个人电脑技术和 4GL 技术。

个人电脑技术使得任何人都可以把他/她自己的电脑带进公司，并可以随意地做他/她自己的处理。出现了像电子表格（spreadsheet）这样的个人电脑软件。另外，个人电脑的拥有者可以将他/她的数据存储在自己的电脑上，这样就不再需要集中式的 IT 部门，结果就是——如果用户因为我们不让他们得到自己想要的数据而愤怒，那就给他们好了。

大约在同一时间，另一种技术也出现了，称为 4GL——第 4 代技术。4GL 蕴涵的思想是使编程和系统开发简单到任何人都可以做。这样一来，终端用户就可以摆脱必须从 IT 部门来获取企业数据的束缚。

介于个人电脑技术与 4GL 技术之间的观点是释放终端用户，这样终端用户就可以将命运掌握在自己手中。我们需要给终端用户访问其所需数据的自由，来满足他们对数据的渴望。

个人电脑技术和 4GL 技术很快就在企业中得到应用。

然而，一些没有预料到的事情在这个过程中发生了。当终端用户可以自由地访问数据时，他们发现，除了需要访问这些数据外，想要做出好的决策还有更多事要做。终端用户还发现，即使数据可以被访问，也会存在下列问题：

- 如果数据是不准确的，则没有比这更糟糕的事情了，因为不准确的数据会有很大的误导性。
- 不完整的数据的用处并不是很大。
- 不及时的数据不太符合人们的需要。
- 当同一数据出现多个版本时，依赖于其错误的值会导致糟糕的决定。
- 没有文档的数据的价值值得怀疑。

也只有在终端用户可以访问数据后，他们才能发现数据的所有潜在问题。

1.4 蜘蛛网环境

通常的结果就是一个非常大的混乱，这种混乱有时候可以形象地称为“蜘蛛网”环

境。之所以称为蜘蛛网环境，是因为有如此多的线路通向如此多的地方，这让我们想到了蜘蛛网。

图 1-2 描述了在一个典型的企业 IT 环境中蜘蛛网环境的演变。

图 1-2 在一个典型的企业 IT 环境中蜘蛛网环境的演变

在许多企业环境中，蜘蛛网环境已经发展到了不可想象的复杂程度。为了证实它的复杂度，思考一下如图 1-3 所示的一个企业蜘蛛网环境的真实图表。

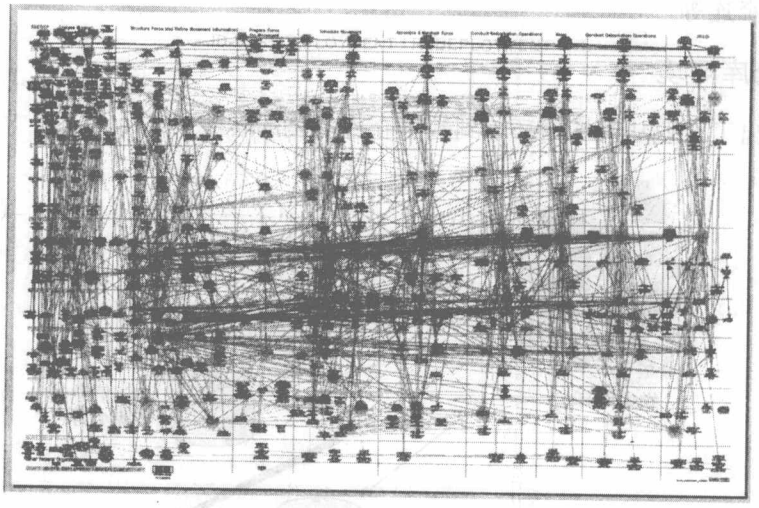


图 1-3 一个真实的蜘蛛网环境

我们看着这个图就觉得恐惧，想一想那些必须要处理如此的环境并试图用它来做一个好的企业级决定的可怜的人吧。令人惊奇的是，任何人都可以把任何事做完，不过很少人能做出好的、及时的决定。

事实上，在目前系统构架备受关注的情况下，蜘蛛网环境对企业来说是一个死胡同，想要使蜘蛛网环境工作是没有希望的事情。

终端用户、IT 专业人员和管理人员的沮丧导致了另一种不同的信息系统构架的发展，这就是以数据仓库为中心的构架。

1.5 企业角度的演化

上述过程是从技术角度出发描述的，还有一个不同的角度——企业角度。从一个企业

人员的角度出发，计算机的发展开始于重复性工作的简单自动化。与人相比，计算机能够以更快的速度、更高的准确率来处理更多的数据。例如，工资单的产生、发票的生成、正在生成的支付过程等工作都是计算机最初进入企业生活的典型应用。

不久后，人们发现计算机还可以用于记录大量的数据，这样就产生了“主文件”。主文件记录了库存、应付款项、应收款项、发货清单等。不久后又产生了联机数据库，利用联机数据库计算机开始进入商业活动的核心。有了联机数据库，航空公司的职员得以解放；有了联机处理，银行出纳员可以担任一些全新的职能；有了联机处理，保险理赔处理比以往任何时候都快。

正是联机处理使得计算机进入了企业网络。换句话说，一旦联机系统被企业人员所应用，那么如果它发生故障，整个业务就要受损并且是立即受损。银行出纳员不能做他的工作，ATM 不能使用，航空订票也会进入人工操作的模式等。

当前，还存在另一个由于计算机进入商业网络而产生的影响，这个影响关系着商业的管理、战略以及决策等方面，即当前企业决策的形成是基于在企业的动静脉等各种网络系统上的数据的。

因此，正在描述的发展过程很难说是一个以技术为中心的过程，它还伴随着一些来自企业的影响和牵连等。

1.6 数据仓库环境

图 1-4 给出了企业从蜘蛛网环境到数据仓库环境的转变。

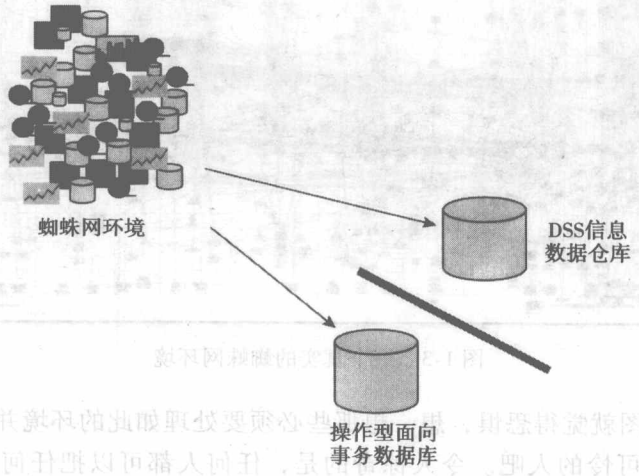


图 1-4 从蜘蛛网环境到数据仓库环境的转变

数据仓库代表了 IT 专业人员思维的重大变化。在数据仓库出现之前，人们认为数据库应该是一种能够满足所有数据需求的东西。但是随着数据仓库的出现，对多种不同种类数据库的需求变得明朗起来。

1.7 什么是数据仓库

数据仓库是信息处理的一个基础。它被定义为：

- 面向对象的。