



万水网络与数据库丛书

DATA WAREHOUSING: Building the Corporate Knowledge Base

数据仓库技术

曹增强 王备战 岳晓奎 译
康博创作室 审校



美] TOM HAMMERGREN



中国水利水电出版社

万水网络与数据库丛书

数据仓库技术

[美] Tom Hammergren 著

曹增强 王备战 岳晓奎 译

康博创作室 审校

中国水利水电出版社

内容简介

本书介绍了数据仓库开发人员从接受任务开始到设计开发一个数据仓库直至建成数据仓库的全部过程。重点介绍了开发数据仓库的重要方法——信息打包方法。本书的每一章都提供了生动详实的例子,图文并茂。利用本书介绍的技术,可以更有效、更全面地开发出更易为用户所接受的数据仓库。

本书不仅是建立数据仓库的信息系统专业人员必备的指南,而且对于实际使用数据仓库的用户也有一定的参考价值。

"Original English language edition published by Ventana Communications Group, Inc., P. O. Box 13964, RTP, North Carolina 27709-3964. TEL: 919/544-9404, FAX: 919/544-9472. Copyright ©1997 by Ventana Communications Group. All rights reserved."

本书中文简体字版由中国水利水电出版社出版,未经出版者书面许可,不得以任何方式复制或抄袭本书的任何部分。

版权所有,翻印必究。

图书在版编目(CIP)数据

数据仓库技术/(美)哈默格兰(Hammergren, T.)著;曹增强等译.
—北京:中国水利水电出版社,1998.2

(万水网络与数据库丛书)

ISBN 7-80124-661-6

I. 数… II. ①哈… ②曹… III. 数据库系统, IV. TP311.13

中国版本图书馆CIP数据核字(98)第00866号

书 名	数据仓库技术
作 者	Tom Hammergren
审 校	康博创作室
出版、发行	中国水利水电出版社(北京市三里河路6号 100044) 北京万水电子信息有限公司(北京市车公庄西路20号 100044)
排 版	北京万水电子信息有限公司
印 刷	北京天竺颖华印刷厂
规 格	787×1092毫米 16开本 21.5印张 495千字
版 次	1998年2月第一版 1998年2月北京第一次印刷
印 数	0001—5000
定 价	35.00元

译者序

进入 90 年代后,随着信息技术的飞速发展,信息的存储、管理、使用和维护变得越来越重要,依靠传统的数据库管理方法已很难满足这样的要求。因此数据仓库技术就应运而生并得到了快速发展。本书就是面向那些想了解、使用和推广传授数据仓库的信息系统专业人员及相关人士而编写的。

本书详细介绍了数据仓库的基本知识,以及开发人员从接受任务到设计开发并实施数据仓库的全部过程中所涉及的每一个细节并提供了大量图文并茂的实例。本书重点讲述了开发数据仓库的基本方法——信息打包方法。运用信息打包方法开发人员能够快速交付数据仓库策略和解决方案,而数据仓库的实施又可以帮助用户和企业更好地管理信息资源并使其迅速转化为商业智能。

本书作者已成功地为几家大公司开发建立了数据仓库和决策支持系统,在数据仓库的设计、开发和实施方面积累了丰富的经验。这本书就是作者们为不同类型企业设计、开发决策支持系统和数据仓库经验的结晶。这样使得本书的内容和实例更加生动。本书的推出将为广大信息系统开发人员提供一个很好的参考指南,同时也会成为广大信息系统用户和企业信息管理决策人员的好参谋。

本书翻译人员如下:曹康、曹增强、李东升、王昊、朱琳、李勇、宋勇、李建锋、刘秀英、李增明、刘莉平、许书明、李文博、李娟、吴皓和邓中亮。审校工作由康博创作室组织人员集体完成,感谢他们在百忙中抽出宝贵时间认真审校了全书的译稿,并提出了一些很有价值的修改意见。

译者

1997 年 8 月于北京

致 谢

感谢是表示礼貌的最优雅的形式。

——Jacques Maritain

写一本书远不是听起来那么容易,它包含许许多多人的支持。虽然书的封面上印着我的名字,但这实际上是许多人的心血。当我考虑需要给那些在这项工作中给予我支持和帮助的人表示谢意时,我想到了很多很多人。

也许,我最感激的是我的妻子, Kim, 以及我可爱的孩子, Brent 和 Kristen。他们为我创造了一个很好的环境,从而使我能够成功地完成这本书的写作——一个我与他们共同分享、同时也使我们大家牺牲了很多东西的成就。

另外,还有 Waterside Productions 的 Carole McClendon 与她的出色职员,她们给予了很大的支持;还有 Thomson 国际出版公司的 Jim DeWolf, Trudy Neuhaus, Jerry Olsen 及支持他们的同事; Sandy Emerson, Jose Cartagena 及其在 Sybase 出版公司的同事; Alan Rottenberg, Lynne Angus, Mickey Gill, Jean Peirre 以及在 Cognos 的一些幕后工作人员;还有 Sybase 公司的 Cathy Murray, 给予了很大的支持并评审了本书,其中 Thom King 提出了很有价值的修改意见,使本书更有意义。作者还要特别感谢 Jo-Ann Campbell, 这本书的图几乎由他一个人完成。如果没有这些人的共同努力,就不可能有这本书。

也不应忘记对我的一生产生很大影响的人,包括我的父亲和母亲, Betty 和 Gordon Hamnergren。他们牺牲了很多时间,陪伴我度过难关,取得成功。如果没有他们长期的教诲和支持,我将不能成为现在的我,还有我的姐姐, Jane 和 Beth, 也一直和我在一起,不断鼓励我。对他们我有说不完的感谢——我非常感谢。另外,我妻子的家人也经常给本书提出修改意见并持续地支持我。当一个人有这么多的支持和信任时,很容易完成要做的工作。

另外,公司的同事和领导也传授给我许多东西,这其中也包括 Don Leonardo 和 Eric Schurr, 是他们很早说服我加入当时还不太稳定的软件行业,然后进一步教导永远不要自满,要不断进取,还包括那些使我真正懂得质量含义的用户,特别是当我在 Cognos 工作期间开发 PowerPlay 和 Impromptu 时与我合作的 Proctor & Gamble 公司的朋友。这其中包括当我在 Cognos 工作早期及智能事务部(以前称 Desktop 部)刚诞生时与我共同分担失败的痛苦的那些朋友,有 Jeff Papows, Ron Zambonini, Mike Potter, Alan Rottenberg, Jim Sinclair, Joe Smarkala, Graham MacIntosh, Robin McNeill, Rob Rose, Glen Rassmussen, Colin McAlpin, Rick Soderstrom, Ron Nordien, Mike Green, Mickey Gill, Sue Hardeman, Barb Paradis 以及 PowerPlay 和 Impromptu 项目组的每一位成员。

最后,我还要感谢读者购买此书。希望它能帮助读者取得成功并有助于完成学业。同时也希望将来能与读者分享一些新的想法、建议、批评。本人地址是 hamnergren @ objx.com., 也可通过我们的 World Wide Web 站点获取有价值的支持信息。网址: <http://www.objx.com>。

目 录

译者序

致谢

序言	1
0.1 未来将会带给我们共同的知识库	1
0.2 为什么写这本书?	3
0.3 本书的使用对象	3
0.4 未来的数据仓库:自适应数据结构	4
0.5 本书的起源	7
0.6 数据仓库什么时候将成为现实	7
第一章 数据仓库的市场爆炸	9
1.1 什么是数据仓库?	11
1.2 操作系统与数据仓库系统	12
1.3 为什么需要数据仓库系统?	15
1.4 数据仓库处理	17
1.5 数据仓库报表示例	20
1.6 小结	25
第二章 影响数据仓库项目成功的因素	26
2.1 取得管理层的信任	26
2.2 从管理项目开始:面向主题的数据仓库	27
2.3 清楚地交流实现希望	30
2.4 任命一个面向用户的项目经理	31
2.5 采用成熟的方法	32
2.6 设计必须注重查询,而不是注重事务	33
2.7 只加载所需数据	35
2.8 定义合适的数据源:元数据映射	36
2.9 明确定义唯一的主题	38
2.10 加强对所有面向决策的数据体的使用和参考	39
2.11 小结	40
第三章 数据仓库结构设计	41

3.1	定义结构	42
3.2	开发结构	42
3.3	评估当前的结构	48
3.4	小结	56
第四章	制订结构蓝图	58
4.1	结构需求	58
4.2	定义单个结构	60
4.3	数据结构需求	68
4.4	数据仓库蓝图	74
4.5	小结	76
第五章	项目生命周期及管理	79
5.1	结构和系统计划	81
5.2	分析和设计	84
5.3	实施	88
5.4	使用	90
5.5	小结	92
第六章	组织项目组	93
6.1	知识的产生	94
6.2	项目管理组	94
6.3	项目资源库	98
6.4	项目工作流程	103
6.5	小结	104
第七章	数据采集:信息用途分析	105
7.1	商务数据的多维特性	105
7.2	采集需求	109
7.3	维数示例	120
7.4	信息包的进一步讨论	122
7.5	小结	123
第八章	建立数据模型	125
8.1	星形图设计	126
8.2	数据仓库中的实体	127
8.3	访问信息包	129

8.4	数据模型中实体的图形表示	132
8.5	将信息包图转换成星形图	134
8.6	扩展星形图:雪花图	140
8.7	统一的实体定义	141
8.8	附加信息包的定义	144
8.9	小结	146
第九章	数据库设计	148
9.1	精炼层次	148
9.2	建立物理数据模型	151
9.3	物理数据仓库细观	163
9.4	小结	168
第十章	数据提取和净化	169
10.1	管理公司数据资产	169
10.2	提取规格	173
10.3	加载数据	178
10.4	用复制代理优化提取	181
10.5	数据发行	184
10.6	提取处理标准	184
10.7	小结	187
第十一章	发行和访问数据	189
11.1	访问工具:现时付款或以后付款	189
11.2	应用程序框架	194
11.3	发送货物	204
11.4	小结	211
第十二章	避免生产障碍	213
12.1	数据集成	213
12.2	需要不断调试	216
12.3	安全性	229
12.4	小结	229
结束语	走向未来	230
	软件销售商群体进展	230
	开发人员进展	233

数据管理的进展	236
小结	237
附录 A 信息包样本	239
A.1 自动化:产品缺陷与质量分析	239
A.2 商标管理:促销分析	241
A.3 咨询服务:费用分析	242
A.4 咨询服务:使用分析	244
A.5 金融方面:帐目分析	245
A.6 金融服务:信贷分析	247
A.7 保健:服务使用	248
A.8 人力资源:雇员停工期和调动	250
A.9 人力资源:工资津贴分析	251
A.10 生产:循环时间	253
A.11 生产:库存事务处理分析	254
A.12 生产:劳动时间分析	256
A.13 生产:定期发贷分析	257
A.14 生产:产品成本分析	259
A.15 生产:供应能力	260
A.16 产品管理:市场分析	262
A.17 销售:顾客人口统计分析	264
A.18 销售:商品供应线管理分析	265
A.19 销售:销售分析	267
A.20 销售:软件产品销售分析	268
A.21 销售:电信产品销售分析	270
A.22 服务中心:寻听分析	271
A.23 小结	273
附录 B 一个面向主题的数据仓库实例	274
B.1 结构	274
B.2 小结	291
附录 C 开发数据仓库的指南	292
C.1 成功的关键因素	292
C.2 工作分析和技巧	292
C.3 信息打包方法参考图	295
C.4 小结	301
附录 D 数据仓库表格的实例	302
术语表	322

序 言

——未来将带来什么？

过去已不重要,现在亦不重要,人们要面对的是未来。因为人们将不再拥有过去,对现在人们也不应考虑太多,而未来则是无所不能的。

——Oscar Wilde

在 2000 年的某个时候……,汤姆已升任企业首席战略家,企业的前景、方向、生产计划等,这些都属于他的管辖范围。他的许多同事都很敬畏他。他们公司信息系统的前任领导是如何取得如此大权力。然而一旦进行深入调研,可以很明白地显示出他是如何取得这么大成绩,如何管理这么多事情。

汤姆:计算机,我们正打算对那些选择我们产品和我们竞争者的产品的用户进行统计。

计算机:汤姆,键控指示器认为:那些年收入在 \$ 75,000 到 \$ 150,000 之间的人中选择你们产品的人越来越多。这些人一般拥有自己的住房,已经结婚,居住在郊区,常在超级市场等买回他们所需的所有东西,我们的对手已赢得年收入在 \$ 50,000 到 \$ 75,000 之间的用户的支持。这些消费者的趋向还没有固定,他们的职业易于改变,单身,无家庭负担。他们一般住在大城市中,有自己的特殊购物习惯,往往在专卖店购买东西。而对那些年收入小于 \$ 50,000 的消费者其购买趋势还不明朗,这些人往往根据自己的收入租一套公寓安家,一般很少随意购买你们或竞争对手的产品。

汤姆:他们什么时候购买我们的产品? 什么原因会促使他们这样做?

计算机:根据消费者数据代理(Consumer Data Agency)的研究,刺激销售能很敏感地影响销售额,其中像中奖销售、打折优惠等会吸引这些人。我们及竞争对手都有这方面的例子,你要看看宣传电视片、宣传印刷品及准备的宣传品吗?

汤姆:给我看看电视片及我们公司准备的宣传资料,如果增长率最大的公司不是我们公司的话,给我们介绍这个公司。

汤姆:(看完宣传资料后)这些促销的成本是多少? 这种促销活动的最佳时机是什么时候? 这种促销活动会吸引多少新用户? 这种促销活动将取得收效并会吸引潜在的用户吗? 给我看看用户是如何通过有线电视或在未来商店对待这些宣传活动的。

当汤姆进入一个虚拟现实的未来商店以及如何向消费者展示这些宣传活动时对话继续进行。

0.1 未来将会带给我们共同的知识库

在成功的狂热之中,我们所有人对所谓的数据仓库系统开始推波逐浪。当技术市场正经

历数据仓库的急速发展时,许多人都问:“数据仓库是什么?”技术市场的目的是为用户提供一个一致的共同信息源,一个共同的知识库。

在数据仓库方面存在的问题是,许多人能解释清什么是数据仓库,但很少有人知道怎样去建造数据仓库,下面是这个问题存在的许多原因中的一部分。

- 许多人不明白操作和分析系统之间的区别。
- 设计数据库系统的技术和工具不支持数据仓库的许多很重要的方面,包括多维数据、聚集和概要层数据。
- 如何去设计、实施和管理数据仓库的知识只能从那些有建造数据仓库经验的人那儿获得。

即使存在这些固有的问题,数据仓库市场的增长仍很快速。增长的原因是因为用于某个事务的操作方面的共同计算机中有二十多年的数据。行政部门的人员正在了解为什么他们把这些资料保存了这么多年。谁是我们的用户?用户的主要行为模式是什么?他们的购买趋势是什么?他们将如何受市场影响?商品的质量和用户满意程度之间的关系是什么?销售人员是如何较好地预测市场?如何尽快扭转销售情况?什么原因使用户长期购买我们的产品?我们与什么人共同占有80%的市场?等等此类问题。

信息系统专业人员需要利用当前系统中已有的数据,来考虑容易但不可预测的事查询和问题。然而业界当前缺乏一种获取所需数据并将其转换成建造和交付数据仓库的统一方法。

这本书就是用来解决以上提到的问题的。本书将解释从信息系统专业人员接受领导的指令开始到开发一个数据仓库直至建成数据仓库的整个过程。这本书为读者提供设计、建模及交付数据仓库方面的技术。这些技术联系在一起就形成了一套完整的方法,称为信息打包方法,可以让信息系统专业人员去为特定公司开发、交付数据仓库。

0.1.1 信息打包方法

信息打包方法贯穿于整个过程,开始是从用户那儿收集事务需求,然后分析并将它们模型化,生成数据仓库,最后管理和维护这个数据仓库系统。通过开始定位事务需求——用户的要求——本书所讨论的方法可以让信息系统专业人员用一种比以前更有效的方式提交“事务智能”。

数据仓库的设想不是刚刚才有。然而,它是从以前的面向用户的支持系统如信息中心、决策支持系统、领导信息系统等发展而来的。数据仓库仍处于发展的初期。

根据以上的建议,未来的数据仓库中会增加许多新的改进,这些改进包括改进数据质量,丰富内容,改进可使用性,以及在数据仓库操作过程中整体质量的提高。为达到这一目的,必须通过新技术寻找一种更好的方法来传输数据。在商业方面,如何通过信息系统提高企业效益的工作刚刚开始。数据仓库将在这方面起带头作用,因为数据仓库提供了一种具有竞争力的资本——以智能事务为特征的信息。

0.2 为什么写这本书?

在信息系统和技术方面如果要完成前面所讲的全部改变需要花很多的时间。我们大多数人不可能等待软件公司来完成像 Xanadu 这样的系统,往往要自己开发。最终的信息访问系统需要许多企业去建立正确的、技术超前和更稳定的、适应性更广的结构。数据仓库将驻留在这一新结构中。对我们中的某些人,数据仓库的革命已经来临;另一些人则仍将继续努力以满足需要。无论你属于哪一部分,有一点是很清楚的:数据仓库将使我们不得不再仔细看一看我们管理所有数据资源的方法。

本书是提高用户交付基于可更改结构的数据仓库的解决方案的能力的系列丛书中的第一本。在这本书中,我们主要介绍信息打包方法。这些技术将允许读者快速提交数据仓库策略和解决方案。这些技术将立即帮助读者及其企业更好地管理信息资源,并使管理信息转换为商业智能。

本书在讨论信息打包技术时,有几个重要的概念。下面是其中的一些概念:

- **结构**——用于建立企业数据仓库的合适的基础。
- **进程**——在企业中成立项目组,并为公司内的数据仓库操作建立一个可重复的过程。
- **内容**——当传输可更改的和可重新使用的数据结构时,为用户输出合适的信息。
- **访问**——使公司内的每个人在线了解数据仓库内的信息。

这方面的其他一些书还包含如何改进数据仓库处理技术方面的内容。这方面的热门领域包含有诸如 Internet 及多媒体(图像、声音、视频等)这样的技术。每本书都是根据个人经验及与建立数据仓库者多次讨论所写。我们希望这些书有助于读者在企业中开发更好的信息资源,扩展用户访问这些信息的能力。

我们写这本书的一个目的是开发和发展信息打包方法,使读者开发的数据仓库在一个具有更高适应性的结构中,这将有助于在未来吸收更多的用户信息需求。

0.3 本书的使用对象

本书是面向那些需要传授和理解如何建立数据仓库的信息系统专业人员。其中包括项目经理、设计师、数据库管理人员、数据建模人员、分析人员、开发人员等。

实际从事数据仓库开发的信息系统专业人员将从本书受益非浅。本书各章提供的技术包括开发数据仓库的首次充分集成的生命周期开发进程信息系统专业——信息打包方法。利用这些技术的人员将发现开发的数据仓库更有效、更全面,也更易被用户接受。这些技术的作用是提供更易维护的对象集合,而这些对象管理和维护那些用以提供商业智能的合适的数据。因此,本书适合于数据仓库项目组的所有成员。

0.4 未来的数据仓库:自适应数据结构

随着像信息高速公路这样的技术进步(Internet 及它的全球网),信息系统正变得像电话一样普通。当一个公司成立的时候,它就需要购买这种需求——计算资源已开始成为一种需求。这些系统提供给我们的生产力是巨大的,电子结算的概念已开始成为商业交易中的合法货币。

甚至我们的孩子也正成为计算机的狂热爱好者。我们像儿女这么大的时候所懂得的自动化方面的知识远比不上他们。这些意味着什么?我们正在学习如何更好地使用新技术吗?我们正在改进以前在信息系统方面的不足吗?

要使我们的信息系统能完成企业的主要任务,还需做很多工作。将来大部分改进工作将面向工程应用,但我们认为最重要的是数据结构的概念和先进性。

未来的整个数据结构提供一个推动商业活动的真实数据的全面定义。这种结构将把用于日常事务处理的操作系统,维持公司运转的能力,以及在我们的权限范围内用于分析重要决策的历史数据和外部数据联系在一起,能使我们在以后进一步赢得更大的生意。这种共享数据的概念将很快被那些已经开始进行数据仓库信息活动的人接受。

0.4.1 共享数据

共享数据概念的一个前提是,一个数据项只能在生成该数据的最佳位置,由人或数据源采集一次。这些数据源称为记录系统,或正式数据源。这种记录系统主要负责定时维护数据以保证其正确性,并最大限度地被其他系统所使用。

总的来讲,这一方法避免了当数据不是由一个记录系统维护时所带来的问题。这种格式的数据将使它面对整个企业而不仅仅是面对一个特定的组织。也许共享数据最重要的作用只有当在整个企业中对所有人都能获得所需数据时才能被认识到。

对整个企业的影响是商业决策将会是真正的信息决策——也就是说,商业决策是基于一种企业知识库作出的。现在,在大多数环境中,同样的资料在不同的计划中往往以不同的形式保存而不能共享。往往输入一个系统后再输入另一个系统,这些数据以它自己的形式存在,根本无法与原始数据源联系起来。结果使提供的信息不准确或不完整,商业决策缺乏实际有用的信息而完全凭经理的直觉。有了共享数据,则只有一个源,数据将是完整的,做出的决策及推荐意见将是符合实际的——来源于企业知识库这个唯一的信息源。

这一概念使得在共享数据环境中更易控制信息质量,因为只有一个记录系统负责维护特定的数据,管理质量也能保证,增强了标准化程度。而现在的环境很难确定源到底是什么,由谁来负责维护数据。

这一概念对企业也提供了较高的成本效益,要保存这些信息需要很少的系统,也只需很少的人维护和录入数据。净效应如下:

- 有更多的时间分析这些数据。
- 从数据中产生的信息具有更高的置信度。

- 在寻找、匹配及确认数据的准确性时浪费的时间减少了。

0.4.2 实现统一的数据源

信息系统从来都不太可能从物理上来支持一个万能的数据库概念。虽然许多软件公司否认这一点。实际上每一个公司都使用许多种(有时是10种以上)不同的数据库销售商去管理公司的数据财产。因此,需要支持多数据源的策略。

用双重策略来讨论这一问题往往比较容易。双重数据库策略是一个从数据获取(操作)及数据访问(决策支持)环境产生标准的可共享的数据的概念。数据获取系统及处理用于输入商业事务处理及参考数据。这些数据随后要经过加工、格式化,然后送入数据仓库。在数据数据仓库里面,数据按主题范围存储,从这些范围可以用一个特别的方法或其他预先优化的格式直接访问数据。

这种双重数据库策略要求数据存贮于数据获取和数据访问环境中。这种数据获取环境允许更新数据,而数据访问环境固定这些数据,只允许读取。对这种策略有很明显的原因。下面是两个主要原因。

- **数据一致性** 大多数企业的决策支持数据必须稳定且完整。根据现有系统设计和实施的实际情况,在报表和决策支持的数据库中并不总是保证数据的一致性。如果对这种数据库执行报表和决策支持功能,则在同一天不同时刻就会获得不同的结果。对报表和决策支持来讲,只有提供一个独立的、一致的数据库才能解决这一问题。双重数据库策略只有在用一个数据库进行事务处理管理而其他数据库用来进行决策支持时才能保证较好的数据一致性。
- **性能** 当前的技术不允许在一个数据库中同时有效地更新和读取数据。这样,双重数据库当用一个数据库去执行更新活动,一个用于决策支持读取活动时提供较好的性能。通过生成两个独立的数据库,可以对每一数据库单独调整从而提供优化的性能。在更新频繁的事务数据库上的索引数可以最少,而对只读活动频繁的决策支持数据库上的索引数可最大化。这方案中的每一种在给定环境都能提供各自的一致性以及最优的性能。

因为处理需求似乎与技术进步同步发展,人们可以看到当前技术性能的局限性在以后还会持续较长的时间。如果这些局限性最终被解决,性能将不再是驱动结构发展的因素。

在多数据源中心,信息仓库具有获取和访问信息的双重功能。中间数据仓库可看成是提供详细数据的词典——或者像有些人提到的,是关于数据的数据。这种信息包含一个数据源清单及其相关的标准。

0.4.3 信息打包

当开始数据仓库处理时,先要清理数据。数据经过转换送入信息包。这些信息包在公司间是经过严格标准化的。每一公司都有用户、产品、地理位置、报表图及其他共用的主题。每

一主题的核心数据对类似的事务经过标准化处理。然而,对整个信息包来讲还需另外一些附加数据。生成这些数据后作为重要的及对核心数据起补充作用的数据保存下来。

随着数据仓库技术的发展及其在企业中的全面实施,人们将开发通用模板以帮助企业管理信息包。这一主题范围的标准化也许是数据仓库对信息系统回报的一个具体例子。当人们为报表而进行标准化时,也许就能认识到有些通用的主题存在相同的方法。数据仓库使选择方法更稳健,操作系统则扩充获取和维护的方法。

0.4.4 小结

计算机软件工业的焦点已开始集中于几个关键领域,以改进用户数据仓库解决方案。正如读者从开始进入数据仓库市场的解决方案中所看到的那样,其中包括如下一些先进技术。

- **逻辑应用程序划分** 许多销售商正开始将前端查询报表工具分为描述引擎,计算及表示引擎。虽然从工程前景来看,他们已做了不少工作,但用户现在刚开始在决策支持代理商和应用程序服务器那儿看到这种物理分割。这些改进允许用户工作站正确地描述这些数据,而同时可让一个功能更强大的、潜在的并行处理机去获取并正确地筛选所需数据。
- **应用程序划分之间的开放联接性** 销售商已经认识到一个所谓的能满足各方面需要的工具实际上并不能真正满足用户的需求。因此,那些允许多个应用程序更好地相互集成的工具已开始出现。出版对管理者提供标准信息库的简要资料可使不同对象集成,而不管这些对象是执行信息系统报表、决策支持系统报表、电子表格、计划或新闻传送。复合文档的重要性对数据仓库来说正在增强,正如它在办公自动化方面一样。
- **数据和逻辑的对象标支持** 正如读者将在本书中所看到的,在数据仓库中按照其相关的加载与访问模式进行管理的全局定义实体的概念是一种正在发展的技术,它将极大地提高部署数据仓库的速度和效率。虽然在面向对象数据库或语言中没有采用这一技术,但在再利用原理和提高生产效率方面这一技术用得很多。另一方面,数据和代码已紧密地连结在一起。
- **工业标准对象** 这些对象当前包含了数据仓库最薄弱的区域,尽管在有些团体中这是正在发展的或固定的。例如,顾客包装产品工业有一个标准对象集,如市场与产品编码标准。另一方面,在卫生系统,正在发展针对病人记录的数据结构的明确定义。那些主要依赖数据交换的工业领域很明显在这一方面将起带头作用。
- **与更先进、技术更丰富的应用程序销售商的共同合作** 应用程序商及更好地使用应用程序销售商数据店的能力现在已发展到许多致力于数据转换销售商正在销售用于某一应用程序转换模块。这一趋势将随着诸如 PeopleSoft 和 SAP 这样的应用程序的继续发展而更加普及。

0.5 本书的起源

这本书是在总结我们过去几年为几个大公司开发数据仓库和决策支持系统的经验即我们称之为信息打包技术的基础上,在许多同事的推动下写出的。这一方法是多年为不同类型企业设计、开发决策支持系统及数据仓库的经验结晶。

也许我们最大的感受来自于在 Cognos 公司工作时为顾客包装产品业的一个世界领先的公司,即 Proctor & Gamble 公司进行 PowerPlay 及 Impromptu 系统的开发。当在 Cognos 及其战略合作伙伴 P&G 工作时,我们学到许多交付高质量决策支持数据库及数据仓库所需要的策略及技术;寻找用户需求。像 P&G 这样的公司是数据消费大户。当他们消化和整理这种数据后,这些数据就会变成一种市场武器和公司的财富,它确保公司在市场上能推出高质量、有竞争力的产品,确实满足消费者的需求。

在 Cognos 公司的工作经历也使我们接触到了其他公司,包括 Dun & Bradstreet 及它们的 SmartStream 系列产品;Equifax 公司及其人口统计信息服务系统;AT&T 公司及其销售管理程序。所有这些经历使得我们能够建立一套易于理解的、用于开发成功的数据仓库的方法。这些经历既有成功的,也有失败的,都使我们学到不少东西,使得我们更好地安排本书所描述的技术。

在未来几年,我们将进一步改进这种方法。然而,我们感到读完这本书后,读者会深刻理解如何成功地从数据仓库传输用户所需信息。一个公司知识库提供信息——或更好地表示为知识——关于如何去做生意及用什么去改进它的知识。不管是基于质量、利润率还是其他准则,公司的知识库将是每一个公司的战略武器,可以提高生产率和竞争力。这就是我们选择“数据仓库:建立公司知识库”这样的题目的原因。传输和维护这一知识库是数据仓库工作方面的主要工作。

0.6 数据仓库什么时候将成为现实

有一些问题几乎无法回答,其中之一是本序言所述目标什么时候能达到,但我们希望本书中的内容能够激发读者的创造性,使你更接近这一目标。建立一个支持用户的基本环境需要很长时间和大量经费。然而,经过读者的努力和共同感受,可以为用户开发一个小型 Xanadu 系统,给他们提供信息以更好地完成他们的工作,同时也使你的公司更富有知识,更具竞争力。

同时,也可以看到那些等待最终解决方案的人。到获得最终解决方案时,你的公司也许已经赢得那些等待的人——或者他们已经消失!无论选取那种方式,我们认为不要等待技术来到你身边。自己创造未来,自己作出解决方案。不要相信充斥技术工业的市场宣传,对技术工业提出挑战,使它给你真正需要的东西。不久,未来将会成为现实——也许你就是本书开始时的对话中的一员,或者是下面对话中的人物。

TOM(或者插入你的名字):计算机,请将这个修改过的全息图分发给项目小组成员,每

人可以通过自己的全息名分析它。如果每项工作都完成后,我们下周生产就会增长,并从我们的竞争对手那儿夺得市场。

四周后在一次工作午餐上:

董事长:TOM 在这儿又做了一次。他及其计算机助手制定了一套促销方案,从而最大程度地利用了我们的促销投资,在以前不太成功的基础上获得了 20%的市场占有率,这是我们及我们竞争对手都梦寐以求的。TOM,请讲讲你们是怎样取得这么大成功的。

TOM:好,所有这些都开始于 4 年前,那时我……