

高等学校信息管理示范教材

数据仓库与数据挖掘技术 原理及应用

■ 姚家奕 编著

<http://www.phei.com.cn>



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

高等学校信息管理示范教材

数据仓库与数据挖掘技术 原理及应用

姚家奕 编著

电子工业出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书从逻辑层次上分为导论、原理、技术和实践四大部分,第1章和第2章是导论部分,首先介绍了数据仓库与数据挖掘的基本概念,然后从应用层面介绍了数据仓库与数据挖掘技术在多个热点行业的最新应用情况。第3章至第9章是原理部分,系统介绍了数据仓库、OLAP和数据挖掘技术的基本原理,以及关联规则分析算法、聚类分析算法、分类分析算法和序列模式分析算法。第10章至第14章是技术部分,以微软SQL Server 2000为数据管理平台,系统介绍了OLAP分析功能、多维数据集设计、维度和指标的建立、MDX语言的应用、多维数据集的优化、数据挖掘和管理技术。第15章是实践部分,主要介绍了数据仓库系统的开发方法,以一个实际的数据仓库系统开发项目为背景,详细介绍了该系统的体系结构设计和模型设计。

本书既可作为高等院校硕士研究生和本科生的教材和参考书,也可作为程序设计人员的参考书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

数据仓库与数据挖掘技术原理及应用 / 姚家奕编著. —北京: 电子工业出版社, 2009.8

高等学校信息管理示范教材

ISBN 978-7-121-09398-2

I. 数… II. 姚… III. ①数据库系统—高等学校—教材 ②数据采集—高等学校—教材 IV. TP311.13 TP274

中国版本图书馆CIP数据核字(2009)第135214号

策划编辑: 刘宪兰

责任编辑: 李光昊

印 刷: 北京东光印刷厂

装 订: 三河市鹏成印业有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路173信箱 邮编 100036

开 本: 787×1092 1/16 印张: 28 字数: 645千字

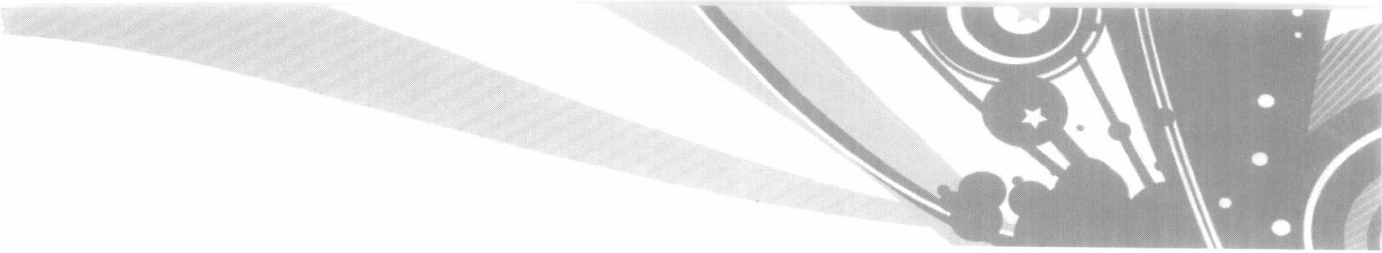
印 次: 2009年8月第1次印刷

定 价: 43.00元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zlts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。



前 言

回顾数据仓库和数据挖掘的发展历程,我们发现,这两个概念和理论原本是从不同的层次和角度提出,并且是相互独立,各自发展的。数据仓库是为提高分析和决策的效率和有效性,按照决策支持的需要对相关数据进行重新组织、建立单独的分析处理环境而出现的一种数据存储和组织技术。数据挖掘从技术的角度来讲,就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。从商业应用角度看,数据挖掘是对商业数据中的大量业务数据进行抽取、转化、分析和模式化处理,从中提取辅助商业决策的关键信息,即从数据中自动发现相关商业模式。

经过数十年的发展和进化,数据仓库和数据挖掘逐渐融合在一起,形成一种新的模式,即“DW(数据仓库)+OLAP(联机分析处理)+DM(数据挖掘)”,如今,无论是从应用层次,还是从理论和技术层次,我们已经不能把二者分开,数据仓库成为数据挖掘的基础和数据保障,数据挖掘也不再神秘,高不可攀,已经从专家学者们独享的象牙塔中走出,把它的应用扩展到各个领域。

笔者长期从事数据仓库和数据挖掘的理论教学和科研项目研究,本书的编写就是在这些教学积累和科研成果基础上完成的。在本书的编写中坚持理论、技术和实践充分融和的原则,贯穿整书每一章节的基本思想是以解决实际问题为目标,书中的理论公式和技术方法都力图用一个具体的示例加以解释和说明,最后以一个实际开发的数据仓库项目为案例,将系统实际开发中不同环节和阶段遇到的问题给出具体的解决方法,不仅加深了对知识的理解,更增强了数据仓库系统实际开发能力。同时本书还配备了相应的实验教材。

本书从逻辑层次上分为导论、原理、技术和实践四大部分,第1章和第2章是导论部分,首先介绍了数据仓库与数据挖掘的基本概念,然后从应用层面介绍了数据仓库与数据挖掘技术在多个热点行业的最新应用情况。第3章至第9章是原理部分,系统介绍了数据仓库、OLAP和数据挖掘技术的基本原理,以及关联规则分析算法、聚类分析算法、分类分析算法和序列模式分析算法。第10至第14章是技术部分,以微软SQL Server 2000为数据管理平台,系统介绍了OLAP分析功能,多维数据集设计、维度和指标的建立、MDX语言的应用、多维数据集的优化、数据挖掘和管理技术。第15章是实践部分,主要介绍了数据仓库系统的开发方法,以一个实际的数据仓库系统开发项目为背景,详细介绍了该系统的体系结构设计和模型设计。

本书既可作为高等院校硕士研究生和本科生的教材和参考书,也可作为程序设计人员的参考书。

本书由姚家奕主编。在本书的编写过程中得到了张润彤教授、关中良教授的大力支持和帮助，并为本书提出了很多建设性的意见，在此一并感谢！同时还要感谢的是为本书的资料收集和文档编辑投入大量时间的研究生王一清、孔淑慧、石丹丹、葛洵洵和徐进华。最后应该感谢的是为本书的出版给予巨大支持的电子工业出版社。

本书难免会有许多不足之处，恳请专家、同行和读者提出宝贵意见。

编者
2009年5月

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396；(010) 88258888

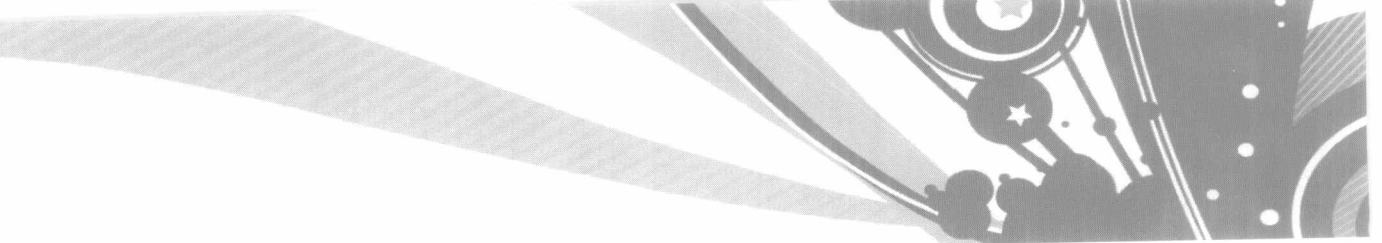
传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036



目 录

第 1 章 数据仓库与数据挖掘概述	(1)
1.1 从数据库到数据仓库	(2)
1.1.1 数据库遇到的困境	(2)
1.1.2 操作型系统和分析型系统的分离	(7)
1.1.3 数据仓库的产生	(9)
1.1.4 传统数据库与数据仓库的区别	(12)
1.2 数据挖掘	(13)
1.2.1 数据挖掘的产生	(13)
1.2.2 数据挖掘的发展历程	(15)
1.2.3 数据挖掘与数据仓库的关系	(17)
1.3 关联学科和技术介绍	(18)
1.3.1 统计学	(18)
1.3.2 人工智能技术与机器学习	(19)
1.3.3 商业智能	(20)
1.3.4 OLAP (Online Analytical Process, 联机分析处理)	(20)
1.4 数据仓库产品介绍	(21)
1.4.1 Business Objects	(21)
1.4.2 Oracle	(22)
1.4.3 IBM	(22)
1.4.4 Sybase	(23)
1.4.5 Informix	(24)
1.4.6 NCR	(24)
1.4.7 SAS	(25)
1.4.8 CA	(25)
本章小结	(26)
本章习题	(26)
第 2 章 数据仓库与数据挖掘的应用和发展	(27)
2.1 金融行业的应用	(28)
2.1.1 银行	(28)
2.1.2 证券	(30)
2.1.3 保险	(33)

2.2	通信与安全行业的应用	(36)
2.2.1	电信	(36)
2.2.2	信息安全	(39)
2.3	生产制造与零售行业	(40)
2.3.1	生产制造	(41)
2.3.2	零售	(42)
2.4	医疗与生物医学行业	(44)
2.4.1	医疗	(44)
2.4.2	生物医学	(46)
2.5	其他行业	(48)
2.5.1	公安	(48)
2.5.2	税务	(50)
2.5.3	竞技运动	(50)
2.6	数据仓库与数据挖掘技术的发展趋势	(52)
2.6.1	数据仓库的发展趋势	(52)
2.6.2	数据挖掘技术的发展趋势	(56)
	本章小结	(61)
	本章习题	(61)
第3章	数据仓库的基本原理	(62)
3.1	数据仓库的体系结构	(63)
3.1.1	数据仓库体系的三个层次	(63)
3.1.2	数据仓库体系结构的基本特点	(65)
3.1.3	数据仓库体系结构的计算模式	(66)
3.2	数据仓库的基本概念	(66)
3.2.1	数据仓库中的数据	(66)
3.2.2	数据仓库处理过程中的关键名词	(69)
3.2.3	数据集市 (Data Mart)	(70)
3.3	数据仓库的特点	(75)
3.3.1	面向主题	(76)
3.3.2	数据的集成性	(77)
3.3.3	数据的非易失性	(77)
3.3.4	数据的时变性	(78)
3.4	数据仓库的数据组织	(79)
3.4.1	数据仓库的数据组织结构	(79)
3.4.2	数据的颗粒度	(80)
3.4.3	数据的分割	(82)
3.4.4	数据仓库的数据组织形式	(83)

3.4.5 数据追加技术.....	(83)
3.4.6 数据仓库中的数据清理.....	(85)
3.5 数据仓库的数据管理.....	(86)
3.5.1 元数据的管理.....	(86)
3.5.2 外部数据与非结构数据的管理.....	(87)
本章小结.....	(89)
本章习题.....	(89)
第4章 OLAP 的基本原理.....	(90)
4.1 OLAP 的体系结构.....	(91)
4.2 OLAP 中的基本概念.....	(93)
4.2.1 OLAP 设计的基本术语.....	(93)
4.2.2 OLAP 服务管理的基本术语.....	(98)
4.2.3 OLAP 与数据仓库、OLTP 的关系.....	(102)
4.3 OLAP 的基本特征与功能.....	(104)
4.3.1 OLAP 的基本特征.....	(104)
4.3.2 OLAP 的基本功能.....	(105)
4.4 OLAP 的分类.....	(107)
4.4.1 MOLAP.....	(107)
4.4.2 ROLAP.....	(109)
4.4.3 HOLAP (Hybrid OLAP).....	(111)
4.4.4 MOLAP 与 ROLAP 的比较.....	(111)
4.5 OLAP 的展现.....	(114)
4.5.1 OLAP 的展现方式.....	(114)
4.5.2 OLAP 的展现方法.....	(115)
本章小结.....	(117)
本章习题.....	(117)
第5章 数据挖掘的基本原理.....	(118)
5.1 数据挖掘的概念.....	(119)
5.1.1 数据挖掘的形式化定义.....	(121)
5.1.2 数据挖掘的技术定义.....	(122)
5.1.3 数据挖掘的商业定义.....	(123)
5.1.4 数据挖掘与 OLAP 的关系.....	(124)
5.2 数据挖掘的体系结构.....	(126)
5.3 数据挖掘的基本功能.....	(126)
5.3.1 概念描述.....	(127)
5.3.2 信息摘要.....	(127)

5.3.3	信息抽取	(128)
5.3.4	元数据挖掘	(128)
5.4	数据挖掘的对象	(128)
5.5	数据挖掘的步骤与过程模型	(130)
5.5.1	数据挖掘的步骤	(131)
5.5.2	数据挖掘的过程模型	(132)
5.6	数据挖掘的分类	(138)
5.6.1	关联分析 (Association)	(139)
5.6.2	聚类分析 (Clustering)	(140)
5.6.3	分类分析 (Classification)	(141)
5.6.4	序列分析及时间序列 (Sequence Analysis and Time Sequence)	(142)
5.6.5	其他分析	(143)
	本章小结	(143)
	本章习题	(144)
第 6 章	关联规则分析算法原理与应用	(145)
6.1	关联规则的典型应用	(146)
6.2	关联规则挖掘算法的基本原理	(148)
6.2.1	关联规则算法的基本概念	(148)
6.2.2	挖掘频繁项集的经典算法——Apriori 算法	(150)
6.2.3	生成关联规则	(156)
6.2.4	预测	(157)
6.3	关联规则挖掘算法的使用	(158)
6.3.1	关联规则算法的参数	(158)
6.3.2	DMX 查询	(159)
6.3.3	模型内容	(161)
6.3.4	解释模型	(162)
6.4	关联规则挖掘的优化算法	(162)
6.4.1	AprioriTid 算法	(162)
6.4.2	AprioriHybrid 算法	(163)
6.4.3	基于粗糙集的关联规则算法	(164)
6.4.4	具有自适应能力的动态递增的关联规则算法	(165)
6.4.5	关联规则的增量式更新算法	(166)
6.4.6	多层关联规则发现算法	(168)
6.4.7	约束性关联规则发现算法	(169)
	本章小结	(172)
	本章习题	(172)

第 7 章 聚类分析算法原理与应用	(173)
7.1 聚类分析的典型应用	(174)
7.2 聚类算法的基本原理	(175)
7.2.1 聚类分析的基本概念	(177)
7.2.2 聚类分析的基本方法	(181)
7.3 聚类分析算法的使用	(184)
7.3.1 聚类分析算法的参数	(184)
7.3.2 聚类模型的使用	(186)
7.4 聚类分析方法的优化算法	(190)
7.4.1 聚类分析的基本优化算法	(190)
7.4.2 面向流数据和孤立点挖掘的新型聚类算法	(195)
本章小结	(199)
本章习题	(199)
第 8 章 分类分析算法原理与应用	(200)
8.1 分类分析算法的典型应用	(201)
8.2 分类分析算法的基本原理	(201)
8.2.1 分类分析算法的基本概念	(202)
8.2.2 决策树基本算法介绍	(203)
8.3 基于信息论 (Information Theory) 的分类分析算法	(206)
8.3.1 概念与定义	(207)
8.3.2 ID3 分类算法	(208)
8.3.3 C4.5 分类算法	(209)
8.4 分类与回归树算法	(212)
8.4.1 构建决策树	(213)
8.4.2 决策树修剪 (Pruning)	(215)
8.4.3 决策树评估 (Estimate)	(218)
本章小结	(219)
本章习题	(220)
第 9 章 序列模式分析算法原理与应用	(221)
9.1 序列模式分析的典型应用	(222)
9.2 序列模式分析的基本原理	(224)
9.2.1 序列模式分析的基本概念	(224)
9.2.2 序列模式的发现步骤	(226)
9.3 序列模式分析典型算法的使用	(228)
9.4 序列模式分析的新算法	(235)
9.4.1 基于 Apriori 的候选码生成——测试的方法	(235)

9.4.2 基于垂直格式的候选码生成——测试的方法	(236)
9.4.3 模式增长方法	(238)
本章小结	(239)
本章习题	(239)
第 10 章 Microsoft SQL Server 2000 数据仓库基本操作	(240)
10.1 Analysis Manager 的配置	(241)
10.1.1 注册服务器	(241)
10.1.2 创建数据库	(241)
10.2 数据源的管理	(241)
10.2.1 指定 ODBC 数据源	(242)
10.2.2 指定 SQL Server 数据源	(244)
10.3 多维数据集和维度的创建	(245)
10.3.1 创建维度	(245)
10.3.2 创建多维数据集	(252)
10.4 管理与使用权限的设置	(255)
10.4.1 系统管理员的安全性控制	(255)
10.4.2 数据库角色定义与管理	(256)
10.4.3 多维数据集角色的管理	(261)
10.5 数据库的存档与恢复	(265)
10.5.1 数据库的存档	(265)
10.5.2 数据库的恢复	(265)
10.6 DTS 在数据仓库中的应用	(267)
10.6.1 DTS 概述	(267)
10.6.2 数据导入/导出工具	(267)
10.6.3 DTS 中的数据转换	(271)
本章小结	(272)
本章习题	(273)
第 11 章 Microsoft SQL Server 2000 OLAP 的基本设计	(274)
11.1 多维数据集的建立	(275)
11.1.1 度量值的添加	(275)
11.1.2 时间维度的建立	(275)
11.1.3 雪花模型维度的建立	(277)
11.1.4 星型模型维度的建立	(278)
11.1.5 父子维度的建立	(278)
11.1.6 完成多维数据集的创建	(279)
11.2 多维数据集的编辑与管理	(279)

11.2.1	维度的编辑	(280)
11.2.2	多维数据集的编辑	(282)
11.3	多维数据集的设计存储和处理	(284)
11.4	多维数据集分析模式的应用	(286)
11.4.1	直接使用“Analysis Manager”进行数据浏览以及 OLAP 的实施	(286)
11.4.2	使用 Excel 作为前端分析工具	(289)
11.4.3	使用 OLAP 的 Web 动态数据透视	(292)
	本章小结	(298)
	本章习题	(298)
第 12 章	Microsoft SQL Server 2000 OLAP 的高级设计	(299)
12.1	计算成员的建立与应用	(300)
12.1.1	度量值成员的导出与应用	(300)
12.1.2	维度成员的导出与应用	(302)
12.2	计算单元的应用	(304)
12.2.1	建立计算单元	(304)
12.2.2	编辑计算单元	(308)
12.3	“对策”的建立与应用	(308)
12.4	“命名集”的建立与应用	(313)
12.4.1	建立命名集	(313)
12.4.2	命名集在 MDX 中的应用	(314)
12.5	成员属性与虚拟维度	(314)
12.5.1	成员属性的建立与应用	(315)
12.5.2	虚拟维度的建立与应用	(316)
12.6	多维数据集的分区与合并	(318)
12.6.1	建立多维数据集分区	(319)
12.6.2	编辑分区与筛选条件设置	(322)
12.6.3	分区的合并	(323)
12.7	虚拟多维数据集的建立与应用	(324)
12.8	钻取选项的设置	(328)
12.8.1	钻取的基本概念	(328)
12.8.2	启用多维数据集的钻取功能	(328)
12.8.3	给角色提供钻取权限	(329)
	本章小结	(331)
	本章习题	(331)
第 13 章	Microsoft SQL Server 2000 MDX 技术	(332)
13.1	MDX 概述	(333)

13.1.1	MDX 语句的基本概念和组成元素	(333)
13.1.2	MDX 语句与 SQL 语句的比较	(334)
13.2	MDX 基础	(335)
13.2.1	MDX 语句的基本结构	(335)
13.2.2	成员、元组和集合	(337)
13.2.3	轴维度和切片器维度	(341)
13.2.4	建立多维数据集上下文	(343)
13.3	高级 MDX	(343)
13.3.1	指定单元格属性	(343)
13.3.2	生成 MDX 中的命名集	(349)
13.3.3	生成 MDX 中的计算成员	(351)
13.3.4	生成 MDX 中的计算单元	(352)
13.4	Microsoft SQL Server 2000 MDX 示例应用程序的使用	(354)
13.5	Analysis Services 中的 MDX 函数	(356)
13.5.1	成员函数	(356)
13.5.2	集合函数	(364)
13.5.3	维度函数	(372)
13.5.4	级别函数	(374)
13.5.5	数值函数	(376)
	本章小结	(381)
	本章习题	(381)
第 14 章	Microsoft SQL Server 2000 数据挖掘	(382)
14.1	Microsoft SQL Server 2000 中的数据挖掘模型	(383)
14.1.1	建立关系数据挖掘模型	(383)
14.1.2	建立 OLAP 数据挖掘模型	(394)
14.1.3	挖掘模型角色管理	(401)
14.2	Microsoft SQL Server 2005 中数据挖掘的改进	(402)
14.2.1	新增加的算法	(403)
14.2.2	易于使用的数据挖掘工具	(404)
14.2.3	简单而强大 API	(404)
14.2.4	与同类 BI 技术的集成	(404)
	本章小结	(405)
	本章习题	(405)
第 15 章	数据仓库系统开发方法、项目管理及实例分析	(407)
15.1	螺旋式开发方法	(408)
15.2	数据仓库项目开发管理	(410)

15.3	数据仓库开发应避免的问题	(414)
15.4	数据仓库系统分析与设计实例	(417)
15.4.1	aCRM 系统背景介绍	(417)
15.4.2	aCRM 系统目标和需求分析	(418)
15.4.3	aCRM 系统体系结构设计	(423)
15.4.4	aCRM 系统模型设计	(424)
	本章小结	(432)
	本章习题	(433)

第 1 章

数据仓库与数据挖掘概述

本章引言

本章首先介绍了从数据库到数据仓库的演变过程，着重说明了传统数据库所面临的问题和局限性。要解决这些问题就必须在体系结构上加以变革，将操作型环境和分析型环境分离，使企业由原先以数据库为中心的生产环境过渡到以数据仓库为中心的生产环境。还介绍了数据挖掘的概念及其发展过程，并分析了数据挖掘与数据仓库的关系。描述了数据仓库和数据挖掘技术与相关学科和相关技术之间的关系，本章最后介绍了目前市场上数据仓库的主流产品。

本章重点

- 数据仓库产生的原因
- 数据仓库基本概念
- 数据仓库与数据库的区别
- 数据挖掘基本概念
- 数据挖掘与数据仓库的关系
- 数据仓库与数据挖掘的学科定位
- 对八大主流数据仓库产品的产品特点以及主要工具的了解



1.1 从数据库到数据仓库

信息技术的不断推广应用,将企业带入了一个信息爆炸的时代。每天都有如潮水般的信息出现在管理者的面前,等待管理者去处理、去使用。这些管理信息的处理类型主要有事务型处理和分析型处理两大类。事务型处理,也就是通常所说的业务操作处理。这种操作处理主要是对管理信息进行日常的操作,对信息进行查询和修改,目的是满足组织特定的日常管理需要。在这类处理中,管理者关心的是信息能否得到快速的处理,信息的安全性能否得到保证,信息的完整性是否会遭到破坏。分析型处理则是指对信息做进一步的分析,为管理人员的决策提供支持。这种类型的信息处理在现代企业中应用越来越广泛,越来越引起管理人员的重视。管理信息的分析型处理,必须访问大量的历史数据才能完成,而不像事务型处理那样,只对当前的信息感兴趣。因此,在分析型处理中,产生了与操作型处理所采用的传统数据库有很大差异的数据环境要求。

传统数据库在操作型处理中获得了较大的成功,但是对管理人员的决策分析要求却无法实现。为实现管理人员的决策分析需要,在数据库基础上产生了能够满足决策分析所需要的数据环境——数据仓库(DW, Data Warehouse)。

1.1.1 数据库遇到的困境

传统数据库处理日常事务十分理想,但是要基于事务处理的数据库来帮助决策分析,就产生了很大的困难。其原因主要是传统数据库的处理方式和决策分析中的数据需求不相称,导致传统数据库无法支持决策分析活动。这些不相称性主要体现在决策处理中的系统响应问题、决策数据需求问题、决策数据操作问题和“蜘蛛网”问题。

1. 决策处理的系统响应问题

在传统的业务处理系统中,用户对系统和数据库的要求是数据存取频率要高,操作时间要快。由于用户对数据进行操作时间短暂,使系统能够在多用户的情况下,也可保持较快的系统响应时间。

在决策分析处理中,用户对系统和数据的要求则发生了很大的变化。在决策分析中,有的决策问题请求,可能导致系统长达数小时的运行,有的决策分析问题的解决则需要遍历数据库中大部分数据。这就必定消耗大量的系统资源,而这些是事务联机处理系统所无法承担的。

2. 决策数据需求问题

在进行决策分析时,需要有全面、正确的集成数据,这些集成数据不仅包含企业内部各部门的有关数据,而且还要包含企业外部的、甚至竞争对手的相关数据。但是在传统数据库中,只存储了本部门的事务处理数据,而没有与决策问题有关的集成数据,更没有企业外部数据。如果将数据的集成问题交给决策分析程序解决,将大大增加决策分析系统的负担,使原先执行时间冗长的系统运行时间进一步延长,用户将更加难以接受。