软件工程师丛书

数据仓库技术与实现

彭木根　编著
雨人科技　策划

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

95

软件工程师丛书

数据仓库技术与实现

彭木根　编著

雨人科技　策划

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 提 要

数据仓库作为近些年来发展迅速的一种新兴技术，它把收集到的数据转变成有意义的可用在分析和报表等应用程序中的信息。并且通过多步进程执行处理和分析，这些进程包括收集数据、净化数据和存储数据等。本书首先详细介绍了数据仓库技术的理论和实现方法，然后详细阐述数据仓库的解决方案。并且通过实例阐述了如何创建、管理和维护数据仓库。

全书内容翔实，示例丰富，结构合理，语言简洁，图文并茂。作为一本数据仓库技术的专著，结合实际系统地讲解了数据仓库技术的理论知识。在说明当前常用数据仓库解决方案的基础上，全面分析了微软和 SAS 两种数据仓库解决方案的具体操作过程。

作为有效解决数据仓库技术的最佳参考资料，本书主要面向数据仓库和数据库的系统管理人员，以及从事数据仓库系统应用开发的专业人员。对于从事数据仓库技术理论研究的人员，本书提供了研究数据仓库的理论和方法。本书可作为 MIS、计算机科学，以及商务等专业的参考书和数据仓库用户及系统管理员的必备手册。

图书在版编目（CIP）数据

数据仓库技术与实现/彭木根编著. —北京：电子工业出版社，2002.6
(软件工程师丛书)
ISBN 7-5053-7622-5

Ⅰ.数... Ⅱ.彭... Ⅲ.数据库管理系统 Ⅳ.TP311.13

中国版本图书馆 CIP 数据核字（2002）第 032480 号

责任编辑：寇国华
印　　刷：北京市天竺颖华印刷厂
出版发行：电子工业出版社　http://www.phei.com.cn
　　北京市海淀区万寿路 173 信箱　邮编　100036
经　　销：各地新华书店
开　　本：787×1092　1/16　印张：27.75　字数：654 千字
版　　次：2002 年 6 月第 1 版　2002 年 6 月第 1 次印刷
印　　数：5000 册　　定价：42.00 元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系。联系电话：（010）68279077

出 版 说 明

随着我国加入 WTO，现代化建设也将以前所未有的步伐向前迈进。我们面临更大的挑战，也面临更多的机遇。一个不争的事实是计算机的应用普及将更加深入，将需要数量更多、水平更高的软件工程师。

我国的软件工程师队伍已有了长足的发展，软件开发水平已有了长足的进步。作为中国人，我们期盼的是中国软件业走自主创新之路，在世界上的地位越来越高。作为出版工作者，为发展我国的软件事业尽最大努力，是我们义不容辞的责任，这正是我们于 1999 年底推出《软件工程师》丛书的初衷。

目前这套丛书已出版了 40 多种。从市场销售和读者反馈的情况看，这套丛书已经得到了读者的首肯和厚爱，这也是对我们下一步工作的激励。

可以说，计算机应用系统的多样化、规模化和复杂化对软件工程师提出了更高的要求，同时也为软件工程师提供了更多的施展个人才华的机会。

针对这种形势，我们正在扩充《软件工程师》丛书的选题范围，进一步界定这套丛书的特色，并把丛书按如下类型整合。

一是开发类，通过大量实例说明如何使用各种流行的高级语言、工具类软件开发不同的应用系统，说明开发思想、开发过程、难点及其解决方案。为了适应我国软件工程师开发综合软件系统的需求，我们把包含编程功能在内的高级应用软件的开发应用也纳入到丛书中。

二是技巧类，通过大量实例说明在不同应用系统开发过程中，有关缩短开发周期、提高开发质量、解决开发中的疑难问题的各种技巧。

三是技术类，介绍软件开发的有关理论和技术，以及在实践中的应用，如系统分析与系统设计、软件测试和系统安全等。

四是手册类，即每个软件工程师必备的案头书。

在新的一年开始之际，这套丛书从内容、开本、印刷及装帧等方面都将以全新的面貌与广大读者见面，目的在于使其更受读者的欢迎，每本书能容纳更多的信息。

我们以为软件工程师提供图书信息服务为宗旨，坚持以图书质量为生命。我们希望《软件工程师》丛书能对读者有所帮助，希望读者提出更多的宝贵建议和意见，包括工作中遇到的技术难点、疑点和问题。希望更多的作者加入我们的专家行列，推介自己的实践经验和累累硕果。我们的网址是 www.phei.com.cn，请和我们联系。

为了我国软件业的更加美好的明天，让我们共同努力。

電子工業出版社

前　言

当今信息技术正在从数据处理向数据使用方向转变，为了解决企业中普遍存在的“数据监狱”和“数据贫穷”现象，为企业决策分析人员方便、迅速地提供更准确、高质的信息，已成为企业当前迫切需要解决的问题，因此数据仓库技术应运而生。数据仓库(Data Workhouse)技术是近年来出现并发展迅速的一种技术，它可以充分利用数据仓库中有已存储的信息帮助决策者进行决策。数据仓库概念起源于20世纪80年代中期，而后又经过被誉为“数据仓库之父”的PrismSolution公司副总裁W.H.Inmon加以定义与发展。W.H.Inmon对其定义为：“数据仓库是支持管理决策过程的、面向主题的、集成的、随时间而变的、持久的数据集合。”数据仓库的信息源具有分布和异构的特点。数据仓库管理系统根据企业的原始操作数据和来自外部的数据汇集和整理成数据仓库，为企业提供完整、及时、准确和明了的商业决策支持信息。

本书是根据作者多年的对数据仓库技术的理论研究以及开发实践经验编写而成的。之所以编写这本书，是因为现在许多企业越来越依赖于从信息系统中收集信息，企业的各种客户也希望能够在许可的情况下访问企业的有关数据。为了管理好数据，保持数据的一致性，以及从企业的角度分析数据，必须有精通数据仓库技术的人才。但是由于数据仓库技术的本身特点，所以造成这方面的人才急缺，这方面的权威资料极少。读者迫切需要一本理论和实际结合紧密的书籍出现，本书因而应运而生。

本书共分三大部分，其中第一部分“基础篇”主要介绍了数据仓库的基础知识，对数据仓库技术的研究热点进行了深入剖析；第二部分“工具篇”介绍了现在常用的数据仓库工具和数据仓库的实现过程；第三部分“实例篇”主要讲述了具体的数据仓库解决方案，通过实例阐述了如何进行数据仓库的创建，管理和维护。

参加本书编写工作的还有花淑琴、梁普选、甘加强、刘强、马林、李剑、周保成、王成刚、徐涛、白俊、赵冰宇和康博峰。本书总策划姜云峰和作者对本书的所有参与者表示感谢，是他们的勤奋工作使这本书得以尽快与读者见面。

数据仓库技术是一个正在研究的热点，限于作者的水平和时间，很难将其做一个全面而且深入的阐述，本书可能会有谬误之处，恳请读者批评指正。

我们的电子邮件地址是：cityantique@sina.com。

作　者

目　录

第一部分　基础篇

第二部分 工具篇

第三部分 实例篇

第一部分　基础篇

第1章

从数据库到数据仓库

在激烈的市场竞争中，信息对于企业的生存和发展起着至关重要的作用。表达信息的数据随着时间和业务的发展而不断膨胀，因而有人惊叹道：当今的时代是信息爆炸的时代。同时数据分布在不同的系统平台上，具有多种存储形式，作为领导和决策者如何从这样复杂的数据环境中得到有用的决策数据呢？随着分布式结构的成熟，数据库技术的提高和数据处理技术的发展，数据仓库(Data Warehouse，DW)和决策支持系统(Decision Support System，DSS)应运而生。本章中将主要介绍数据仓库产生的历史背景、定义、数据库体系化环境和数据仓库中的数据组织结构，以及数据仓库中所应用的方法论。同时，本章还将阐述数据仓库与传统数据库的区别，并从学术界公认的信息系统发展的3个阶段入手，强调数据仓库技术在未来的信息领域中的特殊地位及其光明前景。

1.1 数据仓库的由来

在市场经济的激烈竞争中，企业必须把业务经营同市场需求联系起来，在此基础上做出科学、正确的决策，以求生存。为此，企业纷纷建立起了自己的数据库系统，由计算机管理代替手工操作，以此来收集、存储、管理业务操作数据，改善办公环境，提高操作人员的工作效率。比如，人们在日常生活中经常会遇到这样的情况：超市的经营者希望将经常被同时购买的商品放在一起，以增加销售；保险公司想知道购买保险的客户一般具有哪些特征；医学研究人员希望从已有的成千上万份病历中找出患某种疾病的病人的共同特征，从而为治愈这种疾病提供一些帮助……对于以上问题，现有信息管理系统中的数据分析工具很难给出答案。即传统的数据库应用系统并不能很好地支持决策，因为它是面向业务操作设计的，无论是查询、统计，还是生成报表，其处理方式都是对指定的数据进行简单的数字处理。虽然能简化具体操作人员的劳动强度，但不能对这些数据所包含的内在信息进行提取，所以对于企业的中高层领导来说并没有相应的决策支持系统。企业需要新的技术来弥补原有数据库系统的不足，需要把已经广泛收集到的数据集成到数据仓库中，以从业务数据中提取有用的信息，帮助他们在业务管理和发展上做出即时正确的判断。数据仓库技术应运而生，成为信息技术领域非常热门的话题之一。

数据仓库是计算机和数据应用发展到一定阶段的必然产物。如今，信息处理部门的工作重点已不在于简单的数据收集。随着企业计算机应用的不断深入，企业已经积累了大量的生产业务数据，企业内的各级人员都希望能够快速、交互并方便有效地从这些大量杂乱无章的数据中获取有意义的信息，决策者希望能够利用现有数据指导企业决策和发掘企业的竞争优势。由此我们可以看到数据仓库的目的是为了建立一种体系化的数据存储环境，将分析决策所需的大量数据从传统的操作环境中分离出来，使分散、不一致的操作数据转换成集成、统一的信息。企业内不同单位、不同角色的成员都可以在此单一的环境之下，通过运用其中的数据与信息，发现全新的视野和新的问题、新的分析与想法，进而发展出制度化的决策系统，并获取更多经营效益。我们不妨再深入地想一想，要实现这个目的，必须获得大量的历史数据和汇总数据。而且与联机事务处理(OLTP)系统不同，用户不需要修改数据，而使用的大量随机查询是基于联机分析处理(OLAP)的应用。用户对这类应用越来越迫切的需求推动了数据仓库技术的发展。

数据仓库同时也是适应决策支持系统的需要而产生的，所以采用的软件产品应该能支