王 珊 等 编著

数据仓库技术与联机分析处理

数据库丛书

科学出版社

数据库丛书

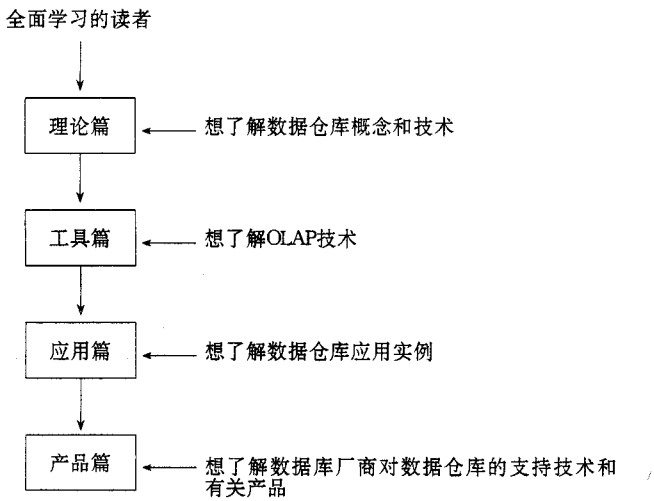
数据仓库技术与联机分析处理

王 珊 等编著

科学出版社

1998

中注意做到既使全书各章是一个相互联系的整体，同时又使各篇章能自成一体。读者可以选择其中的某些篇章来阅读。例如，想了解 OLAP 技术的读者可跳过理论篇直接阅读工具篇(如下图所示)。



阅读本书指南

本书的写作过程也是我组织研究所师生们学习、研讨和实践的过程。两年多来，我们阅读了大量国外的著作、论文，考察和使用了 DBMS 和 OLAP 的有关产品，进行了数十次研讨，写出了数十篇报告和论文，在此基础上进一步充实、整理成为本书。

参加本书研讨和写作的有：王珊、陈红、楼文武、李纪华、周胜、罗立、刘方、张孝、王秋月、李鹏、刘爽、周勇等。特别应指出的是，楼文武在资料整理、内容调整和书稿录入等方面付出了辛勤劳动。

本书收入了中国银行广东省分行建立和应用数据仓库的实例。对国内广大用户来说，他们成功的经验比起国外的应用实例更为亲切，容易借鉴。在此，我向撰写该实例的温巩先生、罗亦文女士、温家明先生等表示衷心感谢。

中国科学院研究生院邵佩英教授审阅了全稿并提出了许多有益的意见，在此向她致以衷心的谢意。

我们在编写本书的过程中，尽可能做到深入浅出，力求概念正确，理论联系实际。由于数据仓库作为一个新的领域，发展非常迅速，加之我们水平有限，故书中一定存在许多不足之处，希望同行和广大读者提出批评和建议。

王　珊

1997 年金秋　于中国人民大学数据与知识工程研究所

内 容 简 介

数据仓库技术(Data Warehousing)和联机分析处理(On-Line Analytical Processing,简记为OLAP)是信息领域中近年来迅速兴起的计算机技术。

本书全面而系统地介绍了数据仓库技术和基于数据仓库的OLAP应用技术,主要内容包括数据仓库的基本概念、创建技术和方法、数据仓库的体系结构以及投资回报分析(理论篇),数据分析工具、数据分析模型、OLAP的基本概念、多维数据库、OLAP的实现技术,以及数据挖掘技术等(工具篇)。本书还在应用篇中给出了数据仓库的若干应用实例,特别是我国自己的应用例子。最后在产品篇中介绍了著名的数据库厂商Informix,Oracle,Sybase关于数据仓库的解决方案和相关产品。

本书是学习、掌握和运用数据仓库技术的综合指南,是从事数据库和数据仓库的研究和开发者、设计开发人员,以及需要了解数据仓库实际技术的系统集成人员、系统设计师和有关专业人员的良师益友,也可作为大学高年级学生或研究生相关课程的教材和参考书。

图书在版编目(CIP)数据

数据仓库技术与联机分析处理/王珊 等编著.-北京:科学出版社,1998.5
(数据库丛书)
ISBN 7-03-006412-7

Ⅰ.数… Ⅱ.王… Ⅲ.数据库系统-信息处理 Ⅳ.TP 311.13

中国版本图书馆CIP数据核字(98)第02360号

科学出版社出版
北京东黄城根北街16号
邮政编码:100717
中国科学院印刷厂印刷
新华书店北京发行所发行 各地新华书店经售
*
1998年6月第 一 版 开本:787×1092 1/16
1998年6月第一次印刷 印张:16
印数:1—3 100 字数:356 000
定价:30.00 元

《数据库丛书》是我国数据库专家学者团结协作、合力撰写的一套系列著作。它比较全面地反映了国际数据库技术的丰富内容与最新发展，和我国数据库科技工作者多年来的主要研究成果，具有较高的理论水平和学术价值。

数据库是计算机科学技术中发展最快的领域之一，也是应用最广的技术之一。是计算机信息系统与应用系统的构成基础。相信《数据库丛书》的编撰出版，必将有益于推动我国数据库技术的研究与发展，促进我国数据库技术的普及与提高，加快数据库应用的推广与深入，为我国社会经济信息化作出贡献。

张效祥

九六年六月

《数据库丛书》编委会

序

数据库是计算机领域发展最快的学科之一，因为它既是一门非常实用的技术，也是一门涉及面广、研究范围宽的学科。因此，它吸引了理论研究、系统研制和应用开发等不同方面众多的学者、专家和技术人才致力于其研究和实践。

数据库系统所管理、存储的数据是各个部门宝贵的信息资源。在信息化时代来临、Internet高速发展的今天，信息资源的经济价值和社会价值越来越明显。建设以数据库为核心的信息系统和应用系统，对于提高企业的效益、改善部门的管理、改进人们的生活均具有实实在在的意义。正因为数据库技术与经济、社会的发展和信息化建设有着密切的关系，这门学科才获得了巨大的源动力和深厚的应用基础。

数据库系统已从第一代网状、层次数据库系统发展到第二代关系数据库系统和第三代以面向对象为主要特征的数据库系统。数据库技术与网络通信技术、面向对象技术、并行计算技术、多媒体技术、人工智能技术等互相渗透，互相结合，成为当前数据库技术发展的主要特征。它使数据库领域中新的技术内容层出不穷，新的学科分支不断涌现，形成了新一代数据库系统的大家族。与传统的数据库相比，当今数据库的整体概念、技术内容、应用领域，甚至某些原理都有了重大的发展和变化。

面对如此丰富的学术内容和技术方法，如此广阔的研究方向和应用领域，从事数据库研究、开发和应用的科技人员，攻读数据库方向的研究生都迫切希望有一套丛书能系统而全面地介绍数据库学科的多个分支和相关领域。

《数据库丛书》的编写宗旨是把当前数据库学科各个分支的最新学术成果介绍给读者，以促进国内的学术研究；同时，又介绍数据库技术的发展过程，各分支之间的内在联系及在数据库大家族中的位置，以促进数据库和计算机科学的其他领域技术的结合。

本丛书由各分册组成，包括《数据库进展》、《分布式数据库》、《分布式数据库管理系统实现技术》、《并行关系数据库管理系统引论》、《数据仓库技术与联机分析处理》等。本丛书的每一分册涉及数据库学科的一个或几个分支。其中《数据库进展》则与其他分册有所不同，是本丛书的总纲、指南和补充，是给本丛书穿针引线、铺垫基础，从而使丛书成为一个各部分既相互独立又相互联系的整体。

《数据库丛书》是开放的，故丛书的分册将随着数据库学科的发展而不断补充。

本丛书各分册的主编和作者，多是长期从事数据库各分支领域研究工作的专家、学者。他们学术造诣高深，实践经验丰富，书中许多内容是他们长期研究的成果。本丛书不仅反映了国际数据库技术的最新成果和发展方向，也展示了我国数据库工作者的学术成果和研究深度，具有较高的理论水平和学术价值。它的出版是我国数据库学术界的一件大喜事。我向本丛书的所有作者和编委的辛勤工作表示崇高的敬意。

萨师煊

1998 年 1 月

前　言

近年来，国外信息技术领域中悄然兴起并日益成熟的数据仓库技术引起了国内同仁的广泛重视。1996 年 7 月 15 日，我们在《计算机世界报》上发表了一组有关数据仓库的文章，引起了学术界、企业界很大反响和浓厚的兴趣。许多同仁来电话、来信，提出了许多问题和要求，希望能结合他们目前的业务，应用数据仓库技术建立决策支持系统，以充分利用现有的数据资源，提取管理决策所需要的信息。在这样的大环境下，本书应运而生。

根据国外数据仓库技术的发展和国内对数据仓库技术的需求，本书全面而系统地阐述了数据仓库的基本概念和方法，介绍了数据仓库作为决策支持系统(DSS)的一种有效而可行的体系化解决方案应包括的三个方面的技术内容：数据仓库技术(Data Warehousing，简记为 DW)、联机分析处理技术(On-Line Analytical Processing，简记为 OLAP)、数据挖掘技术(Data Mining，简记为 DM)，以使读者对该领域的理论、技术和应用情况有一个全面的了解。

数据仓库是一种解决问题的方案，而不是可以买到的现成产品。数据仓库以传统的数据库技术作为存储数据和管理资源的基本手段，以统计分析技术作为分析数据和提取信息的有效方法，以人工智能技术作为挖掘知识和发现规律的科学途径。因此，它是诸多学科相互结合、综合应用的技术。本书从数据仓库系统的架构中展示这些技术的作用和相互之间的关系。

本书共十二章，分为理论篇、工具篇、应用篇和产品篇四个部分。

理论篇包括第一至四章，介绍了：数据仓库产生的背景及定义，数据仓库的特征，数据仓库中数据的组织，以及数据库体系化环境；创建数据仓库的方法、模型和步骤；操作数据存储(ODS)的概念、特点，以及 ODS 和数据库、数据仓库的关系，创建 ODS 的两条技术路线等；数据仓库的投资回报定量分析和定性分析，其中列举了 IDC 随机调查的 62 个数据仓库中 3 个有代表性的例子。

工具篇包括第五至七章，论述了数据仓库系统中数据仓库、数据仓库管理系统和数据仓库工具三部分的作用，简要介绍了数据仓库中查询工具、OLAP 分析工具和数据挖掘工具等三类主要的工具(只有通过高效的工具，数据仓库才能真正发挥出数据宝库的作用)，介绍了 OLAP 的基本概念、OLAP 产品的 12 条准则和基于多维数据库和关系数据库的两种 OLAP 实现技术，以及数据挖掘的概念、分析方法及其有关技术。

应用篇包括第八、九章，概述了数据仓库的应用实例，这些例子是经过精心挑选的，其中有中国银行广东省分行数据仓库的应用和实践。

产品篇包括第十至十二章，介绍了：Informix 公司对数据仓库的支持技术及前端 OLAP产品 Metacube；Oracle 公司提供的对数据仓库的解决方案及 OLAP 工具 Express 的功能和产品构成；Sybase 公司所支持的数据仓库的多层体系结构及其 Sybase IQ 数据库服务器，数据仓库设计工具 Warehouse Architect。

为尽可能面向各个层次及不同部门的数据库理论和应用的工作者，我们在编写过程

目　录

理　论　篇

工 具 篇

应 用 篇

产 品 篇

理论篇

第一章　从数据库到数据仓库

随着计算机技术的飞速发展和企业界不断提出新的需求，数据仓库技术应运而生。传统的数据库技术是以单一的数据资源，即数据库为中心，进行从事务处理、批处理到决策分析等各种类型的数据处理工作。然而，不同类型的数据处理有着其不同的处理特点，以单一的数据组织方式进行组织的数据库并不能反映这种差异，满足不了数据处理多样化的要求。近年来，随着计算机应用，特别是数据库应用的广泛普及，人们对数据处理的这种多层次特点有了更清晰的认识。总结起来，当前的数据处理可以大致地划分为两大类：操作型处理和分析型处理（或信息型处理）。操作型处理也叫事务处理，是指对数据库联机的日常操作，通常是对一个或一组记录的查询和修改，主要是为企业的特定应用服务的，人们关心的是响应时间，数据的安全性和完整性。分析型处理则用于管理人员的决策分析。例如，DSS，EIS 和多维分析等，经常要访问大量的历史数据。两者之间的巨大差异使得操作型处理和分析型处理的分离成为必然。这种分离，划清了数据处理的分析型环境与操作型环境之间的界限，从而由原来的以单一数据库为中心的数据环境发展为一种新环境：体系化环境。

1.1　从数据库到数据仓库

数据库系统作为数据管理手段，主要用于事务处理。在这些数据库中已经保存了大量的日常业务数据。传统的 DSS 一般是直接建立在这种事务处理环境上的。数据库技术一直力图使自己能胜任从事务处理、批处理到分析处理的各种类型的信息处理任务。尽管数据库在事务处理方面的应用获得了巨大的成功，但它对分析处理的支持一直不能令人满意，尤其是当以业务处理为主的联机事务处理（OLTP）应用与以分析处理为主的 DSS 应用共存于同一个数据库系统中时，这两种类型的处理发生了明显的冲突。人们逐渐认识到，事务处理和分析处理具有极不相同的性质，直接使用事务处理环境来支持 DSS 是行不通的。

具体来说，事务处理环境不适宜 DSS 应用的原因概括起来主要有以下五条：

（1）事务处理和分析处理的性能特性不同。

在事务处理环境中，用户的行为特点是数据的存取操作频率高而每次操作处理的时间短，因此，系统可以允许多个用户按分时方式使用系统资源，同时保持较短的响应时间，OLTP 是这种环境下的典型应用。

在分析处理环境中，用户的行为模式与此完全不同，某个 DSS 应用程序可能需要连续运行几个小时，从而消耗大量的系统资源。将具有如此不同处理性能的两种应用放在同一个环境中运行显然是不适当的。

（2）数据集成问题。

DSS需要集成的数据。全面而正确的数据是有效的分析和决策的首要前提，相关数据收集得越完整，得到的结果就越可靠。因此，DSS不仅需要整个企业内部各部门的相关数据，还需要企业外部、竞争对手等处的相关数据。

事务处理的目的在于使业务处理自动化，一般只需要与本部门业务有关的当前数据。而对整个企业范围内的集成应用考虑很少。当前绝大部分企业内数据的真正状况是分散而非集成的。造成这种分散的原因有多种，主要有事务处理应用分散、“蜘蛛网”问题、数据不一致问题、外部数据和非结构化数据。

上述问题是事务处理环境所固有的，尽管每个单独的事务处理应用可能是高效的，能产生丰富的细节数据，但这些数据却不能成为一个统一的整体。对于需要集成数据的DSS应用来说，必须自己在应用程序中对这些纷杂的数据进行集成。可是，数据集成是一项十分繁杂的工作，都交给应用程序完成会大大增加程序员的负担。并且，每做一次分析，都要进行一次这样的集成，将会导致极低的处理效率。DSS对数据集成的迫切需要可能是数据仓库技术出现的最重要动因。

①事务处理应用的分散。

当前企业内部各事务处理应用间实际上几乎都是独立的，之所以出现这种现象有多种原因。有的原因是设计方面的，例如，系统设计人员为了减少系统开发费用和加快开发进度，总是采用简单而“有效”的设计方案，这种“有效”仅指对解决当前面临的问题有效，而不能保证对以后新出现的问题继续有效。有的原因是经济方面的，当经费有限时，企业总是考虑先对关键的业务活动建立应用系统，然后再逐步建立其他业务的信息处理系统。还有的原因是历史、地理方面的，例如，某个大公司由分散在各地的多个子公司组成，企业的兼并，等等。

由于这种事务处理应用分散状况的存在，DSS应用需要对分散在多个事务处理应用中的相关数据进行集成，以向分析人员提供统一的数据视图。

②“蜘蛛网”问题。

DSS应用中为了避免与其他用户的冲突和简化用户的数据视图，一种称作“抽取程序”的方法目前被广泛地应用，用户利用抽取程序从文件或数据库中查找有用的数据，然后这些数据被提取出来放入其他文件或数据库中供用户使用。这些经抽取得到的新文件或数据库又被某些用户再进行抽取，这种不加控制的连续抽取最终导致系统内的数据间形成了错综复杂的网状结构，人们形象地称为“蜘蛛网”。企业的规模越大，“蜘蛛网”问题就越严重。

虽然网上的任意两个节点的数据可能归根结底是从一个原始库中抽取出来的，但其数据没有统一的时间基准，抽取算法各不相同，抽取级别也不相同，并且可能参考不同的外部数据。因而对同一问题的分析，不同节点却会产生不同甚至截然相反的结果。这当然使决策者无从下手。

③数据不一致问题。

前述的应用分散和“蜘蛛网”等多个问题，导致了多个应用间的数据不一致。这些数据不一致的形式是多种多样的：

同一字段在不同应用中具有不同的数据类型。例如，字段Sex在A应用中的值为

“M/F”，在B应用中的值为“0/1”，在C应用中又为“Male/Female”。

同一字段在不同应用中具有不同的名字。例如，A应用中的字段balance在B应用中名称为bal，在C应用中又变成了currbal。

同名字段，不同含义。例如，字段weight在A应用中表示人的体重，在B应用中表示汽车的重量，等等。

为了将这些不一致的数据集成起来，必须对它们进行转换后才能供分析之用。数据的不一致是多种多样的，对每种情况都必须专门处理，因此，这是一项很繁重的工作。

④外部数据和非结构化数据。

在决策中经常用到外部数据，这部分数据不是由事务处理系统产生的，而是来自于其他外部数据源。例如，权威性刊物发布的统计数据、业界的技术报告、市场比较和分析报告、股票行情等，这些数据通常都是非结构化数据。在事务处理系统中，由于没有对外部数据进行统一管理，用到这些数据的DSS应用必须自行集成。

（3）数据动态集成问题。

由于每次分析都进行数据集成的开销太大，一些应用仅在开始对所需数据进行了集成，以后就一直以这部分集成的数据作为分析的基础，不再与数据源发生联系，我们称这种方式的集成为静态集成。静态集成的最大缺点在于，如果在数据集成后数据源中数据发生了改变，这些变化将不能反映给决策者，导致决策者使用的是过时的数据。对于决策者来说，虽然并不要求随时准确地探知系统内的任何数据变化，但也不希望他所分析的是几个月以前的情况。因此，集成数据必须以一定的周期（例如24小时）进行刷新，我们称其为动态集成。显然，事务处理系统不具备动态集成的能力。

（4）历史数据问题。

事务处理一般只需要当前数据，在数据库中一般也只存储短期数据，且不同数据的保存期限也不一样，即使有一些历史数据保存下来了，也被束之高阁，未得到充分利用。但对于决策分析而言，历史数据是相当重要的，许多分析方法必须以大量的历史数据为依托。没有对历史数据的详细分析，是难以把握企业的发展趋势的。

通过（2）、（3）、（4）所述可见，DSS对数据在空间和时间的广度上都有了更高的要求，而事务处理环境难以满足这些要求。

（5）数据的综合问题。

在事务处理系统中积累了大量的细节数据，一般而言，DSS并不对这些细节数据进行分析。这主要有两个原因，一是细节数据数量太大，会严重影响分析的效率；二是太多的细节数据不利于分析人员将注意力集中于有用的信息上。因此，在分析前，往往需要对细节数据进行不同程度的综合。而事务处理系统不具备这种综合能力，根据规范化理论，这种综合还往往因为是一种数据冗余而加以限制。

以上这些问题表明，在事务型环境中直接构建分析型应用是一种失败的尝试。数据仓库本质上是对这些存在问题的回答。但是数据仓库的主要驱动力并不是过去的缺点，而是市场商业经营行为的改变，市场竞争要求捕获和分析事务级的业务数据。建立在事务处理环境上的分析系统无法达到这一要求。要提高分析和决策的效率和有效性，分析型处理及其数据必须与操作型处理及其数据相分离。必须把分析型数据从事务处理环境中提取出来，按照DSS处理的需要进行重新组织，建立单独的分析处理环境，数据仓库正