数据仓库与数据挖掘

Data Warehousing and Data Mining

武 森 高学东 [德] M.巴斯蒂安 著

冶金工业出版社
http://www.cnmip.com

数据仓库与数据挖掘

武 森 高学东
[德]M.巴斯蒂安 著

北 京
冶 金 工 业 出 版 社
2003

图书在版编目(CIP)数据

数据仓库与数据挖掘/武森等著. —北京:冶金工业出版社,2003.9

ISBN 7-5024-3294-9

Ⅰ.数… Ⅱ.武… Ⅲ.数据库系统 Ⅳ.TP311.13

中国版本图书馆 CIP 数据核字(2003)第 037843 号

出版人 曹胜利(北京沙滩嵩祝院北巷 39 号,邮编 100009)

责任编辑 张 卫(010-64027930,E-mail:bull2820@sina.com) 郭庚辰

美术编辑 李 心 责任校对 刘 倩 责任印制 牛晓波

北京市铁成印刷厂印刷;冶金工业出版社发行;各地新华书店经销

2003 年 9 月第 1 版,2003 年 9 月第 1 次印刷

850mm×1168mm 1/32;12.375 印张;329 千字;379 页;1-4000 册

30.00 元

冶金工业出版社发行部 电话:(010)64044283 传真:(010)64027893

冶金书店 地址:北京东四西大街 46 号(100711) 电话:(010)65289081

(本社图书如有印装质量问题,本社发行部负责退换)

前　言

随着市场竞争的日趋激烈，信息对于企业生存和发展所起的作用越来越重要。由于计算机技术的普遍应用，承载信息的数据随着时间推移而不断积累并与日俱增，使得企业难以从巨量数据中找到真正有用的决策信息，因此企业迫切需要新的技术和方法从巨量数据中提取有价值的信息或知识。数据仓库与数据挖掘是20世纪90年代发展起来的新技术。数据仓库是一种针对大数据集进行数据组织与管理的技术，专门用于支持分析型的数据查询，而数据挖掘则是从大量数据中寻找蕴涵其中但尚未发现的知识。

本书是作者在多年研究与教学工作的基础上完成的，书中全面系统地介绍了数据仓库和数据挖掘领域相关内容的主要思想、基础理论、核心技术与方法。全书共分14章，其中1～8章介绍了数据仓库的基本概念、建模方法、体系结构、主题数据组织与数据预处理技术、OLAP及数据仓库的规划与管理；9～14

章深入讨论了数据挖掘的基本概念、数据准备、数据挖掘技术(包括聚类分析、分类发现和关联规则)及数据挖掘的发展与应用。

本书适用于数据仓库与数据挖掘领域的研究和应用人员,也可作为相关专业研究生和高年级本科生的教学用书或教学参考书。

在本书的写作过程中,王莹参加了第7、8章的写作,龙誉、王文贤分别参加了第12章、第13章的写作,第14章由宫雨写作完成,在此表示感谢!崔焕荣在书稿公式编辑方面做了大量的工作,在此一并表示感谢!在本书的写作过程中,作者参阅了大量文献,在此向文献作者表示感谢!

由于作者水平所限,加之数据仓库与数据挖掘是一个新兴的研究领域,有些专业术语尚未达成共识,书中叙述若有不妥之处,诚恳希望同行和读者批评指正,以便今后改正和完善。

作　者

2002年12月

目　录

1 数据仓库概述

随着市场竞争的日趋激烈，信息对于企业的生存和发展发挥着越来越重要的作用。由于计算机技术的普遍应用，承载信息的数据随着时间的推移而不断增长，并且分布在不同的系统平台上，具有多种存储形式。能否从纷繁复杂、大量沉淀的数据环境中得到有用的决策信息，及时做出正确的分析与决策，已成为企业生存与发展至关重要的环节。自从20世纪70年代提出决策支持的概念以来，人们在决策支持系统(Decision Support System，DSS)理论及应用上做了大量的研究工作，并且在企业决策中发挥了积极的作用。但现有的大部分决策支持系统是基于传统数据库基础之上的，随着企业数据量的不断增加，需要对原有的信息进行提炼和加工，需要为企业领导提供集成化和历史化的数据，需要为企业全局的战略决策和长期趋势分析提供更有效的支持。然而，传统的数据库管理系统因自身的局限性已无法满足决策支持系统对数据的要求。因此，一种适用于决策支持系统的数据组织与管理技术——数据仓库技术(Data Warehouse)应运而生，并逐渐成为支持分析与决策的重要技术。

1.1 数据仓库的产生

20世纪70年代，Micheal.Scott.Morton提出了决策支持的概念，人们力图用现有的数据进行深层次的分析和推理，为决策者提供决策所需的信息。于是，在管理信息系统的基础上，以数据分析和建模定量分析为基础的决策支持系统发展起来了。传统的数据库技术作为数据管理手段，主要用于联机事务处理(On-Line Transaction Process，OLTP)，在这样的数据库中保存的是大量的日常业务数据。它在数据共享、数据与应用程序的独立性、维护数

据的一致性和完整性及数据的安全保密性等方面提供了有效的手段。当它与分析型应用结合时,却出现了许多问题。

首先,决策支持系统为掌握充分的信息,需要访问大量的企业内部数据和外部数据。在一个企业的各部门中,可能存在不同的操作型管理信息系统,如销售管理系统、财务管理系统、物资管理系统等。这些操作型应用系统也可能依赖于不同的数据管理平台,常见的有 Oracle、Sybase、SQL Server、Foxprow、Excel 等。在传统系统中,支持决策的内部数据就来源于这些不同的数据源,它们在数据的定义及组织方式上都可能不同。因此,决策者为获取相应的数据就必须熟悉不同的系统环境和数据定义,花费大量的时间理解数据的具体含义和相互间的关系。这样基于不同的数据源支持决策,不仅对决策者本身的素质要求非常高,而且信息的提取也是非常不方便的。

其次,传统数据库中的大量数据是事务型数据,即该数据是对每一项工作、管理对象的具体的、细节性的描述。在销售业务系统中的数据可能表现为一份合同、一张收款单、一张发票或一张货票等。而企业决策层人员所关心的不可能是这类细节性信息,决策支持系统需要的是综合的、总结性的数据,其特点是涉及的数据量很大,甚至可能涉及到上千条、上万条发票或其他记录,但往往并不需要细节数据或个体数据。决策支持系统的目的在于为决策者提取有用的信息,而决策者不可能也不应该花费时间浏览所有细节数据或个体数据。因此,基于传统的操作型数据库不适合建立分析型应用系统。

再次,事务处理型应用和分析决策型应用对数据库系统的性能要求不同。事务处理型应用的特点是数据存取操作频率高,每日进行成千上万次的输入、修改等记录操作,但每次操作处理的时间短,一般是多个用户分时共同使用系统资源,采用传统数据库系统事务处理型应用运行良好。而在分析决策型应用中,为了获取综合性的、有用的决策信息,应用系统可能需要连续运行几个小时、甚至更多的时间进行必要的计算,大量地占用系统资源。如果

将分析决策型应用与事务处理型应用共同放在同一数据库系统环境中，必然引起系统资源紧张，甚至使事务处理型应用瘫痪。

最后一个问题是在传统数据库中保存和管理的一般是当前数据，即使部分历史数据通过备份或历史数据库等形式保存了下来，却普遍被束之高阁，并没有得到充分的利用。而决策支持系统不仅需要当前的数据信息，而且还要求有大量的历史数据，尤其是需要对历史数据进行分析和比较，找出企业发展变化的趋势。就这一点而言，传统数据库系统也不能满足分析决策型应用的需要。

基于以上的论述可以得出结论，在事务处理型应用环境中直接构建分析决策型应用是不可行的。在这种情况下，面向决策分析型应用而组织和存储数据的数据仓库技术便应运而生。随着市场竞争的日趋激烈，企业经营行为逐渐发生改变，对分析决策的要求越来越高。不论是从效率而言，还是从有效性而言，建立在事务处理系统应用环境中的分析决策系统都无法满足现代企业的需要。为了提高分析决策的效率和有效性，面向分析决策型应用的数据处理及其数据必须与事务处理型应用处理及其数据分离，即必须把分析决策型数据从事务处理系统应用环境中分离开来，建立单独的分析决策型应用环境。数据仓库正是为了构建这种新的分析决策型应用环境而出现的一种数据存储和组织技术，并成为研究数据管理技术的新课题。

1.2 数据仓库的含义

我们现在称之为“数据仓库”的这一技术，最早发轫于20世纪80年代初W.H.Inmon的研究，并存在于其“记录系统”、“本原数据”(Atomic Data)、“决策支持数据库”等研究专题中[1]。数据仓库的概念是W.H.Inmon在其《建立数据仓库》一书中提出的，目前它被认为是解决信息技术(IT)在发展中一方面拥有大量数据，另一方面有用信息却很贫乏这种不正常现象的综合解决方案。W.H.Inmon曾对数据仓库做了这样的描述：“数据仓库是90年