数据仓库与数据挖掘

原理及应用

郑岩 编著

清华大学出版社

Data Warehouse
Data Mining

数据仓库与数据挖掘

原理及应用

郑岩 编著

清华大学出版社
北京

内容简介

本书从专业角度全面介绍了数据仓库和数据挖掘的理论、方法、技术及其应用，系统地阐述了数据仓库和数据挖掘的产生、发展和应用及其主要概念、原理和算法，并结合当前数据仓库和数据挖掘中一些新的应用实例进一步加以说明，力求学以致用。

全书分为三篇。第一篇介绍数据仓库的起源和演变过程，阐述数据仓库的定义、体系结构、组成、元数据、数据粒度和数据模型以及 ETL 过程，论述数据仓库设计和实现的方法。结合具体应用详细阐述了如何构建数据仓库及其主要应用，包括 OLAP 和 OLAM 等。第二篇介绍数据挖掘的起源和发展趋势，以及数据挖掘与 Web 挖掘的技术和方法，包括聚类、分类、预测和关联分析等，详细分析了数据挖掘在电信领域的具体应用，如客户细分、重入网识别和 WAP 日志挖掘等。第三篇讨论数据、信息和知识的关系，论述知识表示的主要方法和知识管理的核心技术，介绍当前研究热点——语义网和本体的核心技术和方法，分析了语义网和本体的主要应用。

本书可作为计算机专业研究生或高年级本科生教材，也可以作为计算机研究和开发人员以及相关专业人士的参考资料。

图书在版编目(CIP)数据

数据仓库与数据挖掘原理及应用/郑岩编著. —北京：清华大学出版社，2011.1
ISBN 978-7-302-22819-6

Ⅰ. ①数…　Ⅱ. ①郑…　Ⅲ. ①数据库系统 ②数据采集　Ⅳ. ①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字(2010)第 097105 号

责任编辑：梁　颖　顾　冰
责任校对：时翠兰
责任印制：李红英
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：19.5　**字　数**：464 千字
版　　次：2011 年 1 月第 1 版　　**印　　次**：2011 年 1 月第1 次印刷
印　　数：1～3000
定　　价：32.00 元

产品编号：036980-01

前言

数据仓库是将大量传统数据库数据进行抽取、清洗和转换，并按主题进行重新组织，可比喻为随时间推移不断丰富的“宝藏”；而数据挖掘是从海量数据中发现人们感兴趣的知识，这些知识是隐含的、事先未知的潜在有用信息，挖掘的知识表示形式为概念、规则、规律和模式等，可比喻为“淘宝”。随着 Internet 的迅速普及和广泛应用，每天都产生大量各种各样的信息，但它们背后到底隐藏着什么，这驱使人类不断探索。

工欲善其事必先利其器。在当今信息爆炸的时代，数据挖掘堪比“利器”，让我们面对海量数据时不再感到茫然和不知所措。随着数据仓库的发展和应用，数据挖掘将展现无限的生机和活力，可以辅助、部分代替甚至拓展人的智能和决策，造福人类。

数据经整合汇总为信息，信息经挖掘抽象为知识，知识是智能的基石。因此，信息化到知识化再到智能化将是人类社会发展的必然趋势。数据仓库和数据挖掘正逐步渗透和深入到社会的各个领域，并不断催生新的应用。

本书主要介绍数据仓库和数据挖掘的理论、方法、技术及其应用。此外，用较多篇幅阐述数据仓库和数据挖掘新的应用实例。

全书分为三篇。第一篇介绍数据仓库的起源和演变过程，阐述数据仓库的定义、体系结构、组成、元数据、数据粒度和数据模型以及 ETL 过程，论述数据仓库设计和实现的方法，并结合具体应用详细阐述了如何构建数据仓库及其主要应用，包括 OLAP 和 OLAM 等。第二篇介绍数据挖掘的起源和发展趋势，以及数据挖掘与 Web 挖掘的技术和方法，包括聚类分析、分类、预测和关联分析等，详细分析了数据挖掘在电信领域的具体应用，如客户细分、重入网识别和 WAP 日志挖掘等。第三篇讨论数据、信息和知识的关系，论述知识表示的主要方法和知识管理的核心技术，介绍当前研究热点——语义网和本体的核心技术和方法，分析了语义网和本体的主要应用。

本书编写过程中，参考了许多专家和学者的著作和论文，在此谨向他们表示衷心感谢。

作者潜心撰写历时多年完成，旨在奉献精品以飨广大读者。由于水平有限，不当之处恳请赐教。

作　者

2010 年 8 月

目 录

第一篇 数据仓库

目录

目录

目录

目录

第一篇

数据仓库

第 1 章　数据仓库基础

1.1　引言

进入信息时代以来，特别是近些年，数据库规模日益扩大，数据呈爆炸性增长。图灵奖获得者吉姆·格雷提出了一个经验定律，即网络环境下每 18 个月产生的数据量等于有史以来的数据量之和，仅仅依靠数据库管理系统的查询检索机制和统计分析方法，已经远远不能满足实际需求，面临着“数据爆炸，知识匮乏”的严峻挑战。例如股票经纪人需要从日积月累的大量股票行情变化的历史记录（数据）中发现其规律以预测未来的趋势；天文学家需要从获取的观测数据（其规模可达数千吉字节）中发现新的遥远天体及其运动规律；医生需要从大量病人电子病历中发现某种疾病的起因、症状等。这些数据的共同特点是：其一数据量巨大，一般都是 GB 级乃至 TB 级；其二都以结构化的形式存储在数据库中，包含了大量潜在、有价值的知识，有的已被发现，有的还未被发现。如何有效地管理和利用数据库中的海量数据，以及如何发现其中潜在的知识，需要一种新的、更为有效的手段对各种数据源进行整合并挖掘以发现新知识，更好地发挥这些数据的潜能。因此，数据仓库（Data Warehouse，DW）和数据挖掘（Data Mining，DM）技术应运而生。

数据仓库是一个可更好地支持企业或组织决策，面向主题的、集成的、相对稳定的、随时间不断变化的数据集合；数据挖掘则是使用计算机对大量数据进行快速、有效地分析和处理，从中提取知识，并以一种形式化的、可以理解的方式表达，以便于决策的过程。目前，数据仓库和数据挖掘技术已经成为计算机领域的研究热点之一，引起了数据库、机器学习、统计分析等领域专家的广泛关注。

1.1.1　演变过程

数据仓库是建立在传统事务型数据库的基础之上，为企业决策支持系统（Decision Support System，DSS）及数据挖掘系统提供数据源。到目前为止，国外数据仓库已经发展了十几年的时间，国内虽然起步较晚，但发展较为迅速。目前已有众多的大型公司或企业正在建或计划建设不同规模的数据仓库。

传统数据库（普通数据库）和数据仓库最根本的区别在于其侧重点的不同。数据处理分为事务型处理又称联机事务处理（Online Transaction Processing，OLTP）和分析型处理又称联机分析处理（Online Analytical Processing，OLAP）两大类。事务型处理以传统的数据库为中心进行企业日常的业务处理；分析型处理以数据仓库为中心分析数据背后的关联和规律，为企业的决策提供可靠、有效的依据。事务型处理和分析型处理的分离，划清了数据处理的分析型环境与事务型环境之间的界限。从而由原来以单一数据库为中心的数据环境演变为以数据库为中心的事务处理系统和以数据仓库为基础的分析处理系统。企业的生产环境也从以数据库为中心发展为以数据库和数据仓库为中心。因此，在事务处理环境中直

接构建分析处理应用是不合适的，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，按照决策支持的需要重新组织，建立单独的分析处理环境，数据仓库正是为了构建这种新的分析处理环境而出现的一种数据存储和组织技术。

传统数据库的主要任务是进行事务处理，所关注的是事务处理的及时性、完整性和正确性，在数据分析方面则存在着诸多不足，主要体现在缺乏集成性、主题不明确等多个方面。

1. 缺乏集成性

首先，企业数据库系统与部门条块分割，导致数据分布的分散化与无序化。在一个企业内部，生产、销售和财务等部门往往各自使用一套满足自身工作需要的应用程序。各个部门的应用系统往往不能数据共享，缺乏数据的统一管理和维护。这样企业内部尽管拥有的数据量极大，但各自封闭，构成相互独立的所谓“信息孤岛群”，无法形成统一体。其次，业务数据库缺乏统一的定义与口径，导致数据定义存在歧义。

2. 主题不明确

建立传统数据库的目的是为了满足事务处理的需要，数据库和表的定义与设计完全以此为基础。而对于数据分析而言，这些库和表无疑缺少明确的主题。

3. 分析处理效率低

设计基于传统数据库的应用系统的核心准则是保证事务处理及时而准确。显然，对处理大量分析型数据的效率无法保证。

数据仓库是因为用户需求增加而对某一类数据库应用范围的界定。仅从数据存储容器的角度而言，数据仓库与数据库并没有本质的区别。而且很多时候，数据仓库是作为一个数据库应用系统来看待的。因此，不应该说数据库到数据仓库是技术的进步。

通常，数据仓库是在传统数据库的基础上发展起来的，建立在异构的业务数据库基础上。尽管传统数据库对处理分析型数据存在缺陷，但数据仓库并不是对数据库的彻底抛弃。两者存在诸多差别，如表 1.1 所示。

表 1.1　数据库与数据仓库的比较

	数据库	数据仓库
内容	与业务相关的数据	与决策相关的信息
数据模型	关系、层次结构	关系、多维结构
访问	经常是随机地读、写操作	经常是只读操作
负载	事务处理量大，但每个事务涉及的记录数很少	查询量小，但每次需要查询大量的记录
事务输出	一般很少	可能非常大
停机	可能意味着灾难性错误	可能意味着延迟决策

从数据库到数据仓库演变的具体过程如图 1.1 所示。

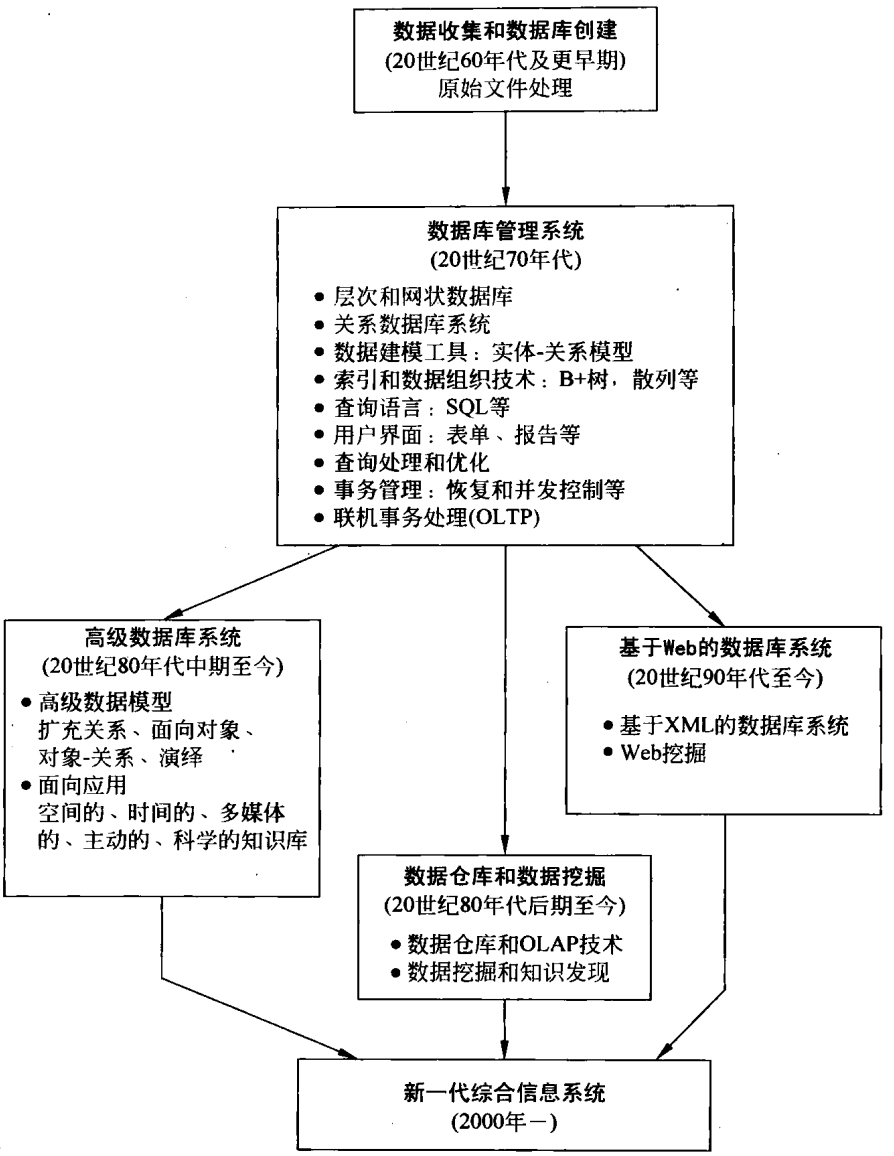

图 1.1　数据库到数据仓库的演变过程

1.1.2　定义

数据仓库的概念最早出现于 20 世纪 80 年代。1993 年，被称为“数据仓库之父”的 William H. Inmon 首次系统地阐述了数据仓库定义，即一个面向主题的、集成的、不可修改的且随时间变化的数据集合，以支持管理人员的决策。

面向主题是相对于传统数据库的面向应用而言。所谓面向应用是指系统实现过程中主要围绕着一些应用或功能，而面向主题则是考虑一个个的问题域，对问题域涉及的数据和分析数据所采用的功能给予同样的重视。数据仓库是面向在数据模型中已定义业务的主要主

题域的，例如在电信领域中典型的主题域包括客户、产品、资源、渠道、服务和竞争等。

集成是指数据仓库中的数据来自不同的数据源。由于历史的原因，各数据源的组织结构往往不同，在这些异构的数据导入到数据仓库之前，必须经过一个集成过程。在数据仓库的所有特点中这是最重要的。应用系统的设计人员历经多年制定出来的不同的设计策略有很多种不同的表示方法，在编码、命名习惯、属性和属性度量等方面往往是不一致的。当数据导入数据仓库时，需要采用某种方法来消除应用系统中存在的不一致性。例如，对“客户性别”编码时，在数据仓库中编码为“男/女”或是 m/f 并不重要，重要的是无论使用什么原始应用系统，在数据仓库中都应该有一致的编码。如果应用系统中编码为 X/Y，则在其导入数据仓库时就应进行转换。对所有的应用都要考虑一致性，如命名习惯、键码结构、属性度量以及数据特点等。

与面向应用的事务数据库需要对数据进行频繁地插入、更新操作不同的是，数据仓库中数据的操作仅限于数据的初始导入和记录查询，而不能修改。数据库处理数据时，一般是一次访问和处理一条记录，也可以对操作型数据进行更新。但数据仓库中的数据通常是一起载入与访问，在数据仓库中并不进行一般意义上的数据更新。

随时间变化是指数据仓库以维的形式对数据进行组织，时间维是数据仓库中很重要的维度之一，并且数据仓库中数据的时间跨度较大，从几年甚至到几十年，称之为历史数据。数据仓库中数据随时间变化的特性表现在以下几个方面：

(1) 数据仓库中数据的时间期限要远远长于操作型数据库中数据的时间期限。操作型数据库中数据的时间期限一般是 60～90 天，而数据仓库中数据的时间期限通常是 5～10 年。

(2) 操作型数据库含有“当前值”的数据，这些数据的准确性在访问时是有效的，同样当前值的数据可被更新。而数据仓库中的数据仅仅是一系列某一时刻生成的复杂快照。

(3) 操作型数据的键码结构可能包含也可能不包含时间元素，如年、月和日等，而数据仓库的键码结构总是包含某一时间元素。

数据仓库是 DSS 的基础。因为在数据仓库中只有单一集成的数据源，并且数据是可访问的。与传统数据库相比，在数据仓库中 DSS 分析人员的工作将容易得多。

1.2 体系结构

1.2.1 两层的体系结构

由数据仓库的定义可知，它是将企业各个业务系统中与分析有关的数据集成在一起，同时数据仓库面向的应用是分析型操作，因此形成了 DB-DW 两层的数据仓库体系结构，如图 1.2 所示。

其中，业务系统作为主要的分析数据来源，其数据格式主要是表的形式。实际中，由于要保证不影响业务系统的正常运行，一般不直接在业务系统中进行数据的查询和抽取，而是采取备份库或者文件传输的形式进行数据仓库的数据抽取。外部数据源是指信息来源于企业的外部，描述企业运营的外部环境与企业经营分析有关的数据，如各个企业的市场份额等，外部数据作为经营分析的补充，对企业经营决策的正确性起着十分重要的作用，因此应保证外部数据的实时性和准确性。外部数据源具有多样性的特点，如年报等都可以作为外

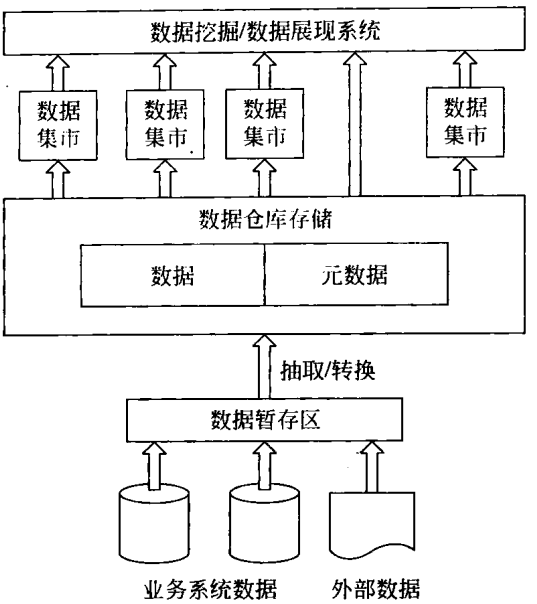

图 1.2　DB-DW 的两层体系结构

部数据源，同时外部数据源的格式也不统一，如文本、数据表格、图像和声音等。因此对外部数据源及其数据格式等都应在数据仓库的元数据中进行记录，同时元数据中还应对外部数据的可信程度有一定评价。

由于数据仓库的数据源不统一，同时源数据的存储形式也不相同，因此有必要在数据进入数据仓库前先将数据存放在一个统一的暂存区中，引入数据暂存区的主要作用如下：

(1) 统一不同数据源的数据格式。将不同数据源中不同的数据格式转换成统一的数据格式，供数据仓库统一处理。

(2) 进行数据的初步检查。在数据进入数据仓库之前，先对数据进行初步检查，鉴于不影响数据仓库的处理时间，这里的检查将仅涉及比较粗略的数据检查，如记录数量、关键字段是否丢失等，对于错误的数据暂不导入数据仓库，这样对进入数据仓库的数据质量有一定保证，但是更复杂的数据清洁工作，如字段格式的统一以及数据内容的清洗这种单一记录级的处理工作则应该在数据抽取时完成。

数据暂存区可以多种存储形式实现，如文件目录或者数据库表的形式。

数据仓库中保存了大量的历史数据，同时数据仓库面向的是整个企业的分析应用，但在实际应用中不同部门的用户可能只使用其中一部分数据，从处理速度和效率的角度出发，可以将这部分数据在逻辑或者物理上进行分离，使用户无需到数据仓库的海量数据中进行查询，只在与本部门有关的数据集合上进行操作，这样就形成了数据集市(data mart)的概念，它是指面向企业的某个部门(主题)而在逻辑上或物理上划分出来的数据仓库的数据子集。将数据仓库按照数据的应用划分为多个数据集市，有利于数据仓库的负载均衡，保证应用的执行效率。同时，由于数据集市具有统一的数据来源——数据仓库，遵循统一的数据模型，保证了各个不同数据集市中数据的统一。

可以看出数据仓库体系结构是一种管道过滤器的结构，数据从数据源进入数据仓库到展示给最终用户，都有一定的关联关系，因此要保证数据仓库中数据处理的合理调度，则需要通过数据仓库的元数据完成。

1.2.2 三层的体系结构

数据仓库的提出使得操作型处理和分析型处理得以分离，从而形成了 DB-DW 两层的体系结构，但是在企业的业务处理中存在介于操作型和分析型之间的需求，需要对短期的历史数据进行分析，同时要求较快的响应速度，这种分析无法在操作型数据库中完成，因为其保存的是数据的瞬态信息，如果通过数据仓库完成，由于数据仓库保存了大量的历史数据，在响应时间上无法满足要求，因此提出了操作型数据存储(Operational Data Store，ODS)的概念，ODS 数据可以概括为面向主题的、集成的、可变的和当前的或接近当前的数据。其中，面向主题和集成的特点与数据仓库的概念相似；“可变的”是指 ODS 数据可以联机改变，包括增加、删除和更新等操作；“当前的”是指数据在存取时刻是最新的；而“接近当前”是指存取的数据是最近一段时间得到的。

面向主题和集成的特点使得 ODS 数据在静态特征上很接近数据仓库的数据，但是 ODS 和数据仓库之间存在重要的差别，主要体现在以下三个方面：

(1) 数据的内容不同。数据仓库中历史数据是指长期保存并可重复查询的数据，既保存细节数据，也保存综合数据。而 ODS 一般只保存细节数据，而且 ODS 数据是可以更新的，即变化的，ODS 中保存的历史数据也是近期的。

(2) 就数据量而言，ODS 保存的数据量要远远小于数据仓库的数据量。

(3) 面向的应用不同。数据仓库用于长期的趋势分析或决策支持，而 ODS 主要是支持企业的全局 OLTP 和即时(up to the second)决策分析应用。

引入 ODS 后，原来 DB-DW 的两层体系结构被扩展为 DB-ODS-DW 的三层体系结构，如图 1.3 所示。

在 DB-ODS-DW 三层体系结构中，ODS 的作用可以概括为：

(1) 为数据仓库提供数据，减少数据仓库数据抽取的复杂性。由 ODS 的定义可知，它具有面向主题和集成两个特点，因此来自业务系统的源数据首先进入 ODS，在进入 ODS 时完成数据清洁和集成的工作，这样再向数据仓库提供的数据就是清洁的和统一的，减轻了数据仓库中数据抽取的工作量。

(2) 即时的 OLAP 分析。由于在业务系统中需要对近期或当前的数据进行分析，如果该任务放在数据仓库中完成，由于数据仓库相应的处理环节较多，同时数据仓库保存了大量的历史数据，如果要完成这种需求势必造成留给数据仓库的数据处理时间减少，所以将这部分任务分配给 ODS，由于 ODS 保存了近期的数据，可以完成用户的即时分析需求。

(3) 全局的 OLTP 操作。由于 ODS 数据的集成性，整合了企业中不同业务系统的数据，同时 ODS 数据是可更新的，因此 ODS 可以提供面向企业全局的 OLTP 操作。