数据仓库与数据挖掘

原理及应用（第二版）

郑　岩 编著

DATA WAREHOUSE AND DATA MINING

(Second Edition)

清华大学出版社

数据仓库与数据挖掘

原理及应用（第二版）

郑　岩 编著

DATA WAREHOUSE AND DATA MINING

(Second Edition)

清华大学出版社
北京

内容简介

本书全面地介绍数据仓库和数据挖掘的原理及其应用，系统地阐述数据仓库和数据挖掘的主要概念和算法等基础知识，并结合当前各领域的具体应用实例进一步帮助广大读者加深理解，力求学以致用。

全书分为3篇。第1篇介绍数据仓库的发展和演变，主要阐述数据仓库的定义、体系结构、组成、数据模型和ETL过程等，描述数据仓库的设计方法和实现过程，结合实例说明如何构建数据仓库，扼要地介绍数据仓库的应用，如OLAP和OLAM。第2篇介绍数据挖掘的起源和发展，主要阐述数据挖掘和Web挖掘的主要算法，包括聚类、分类、预测和关联分析等，描述如何运用数据挖掘解决实际问题，如客户细分、虚开欺诈识别和WAP日志挖掘等。第3篇阐述数据、信息和知识之间的关系，介绍当前研究热点——语义网和本体的核心技术及其主要应用。

本书可作为计算机及相关专业的研究生和高年级本科生教材，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. —2版. —北京：清华大学出版社，2015
ISBN 978-7-302-37861-7

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(2014)第202718号

责任编辑：刘向威　李　晔
封面设计：何凤霞
责任校对：梁　毅
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：三河市少明印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：24　　**字　　数**：600千字
版　　次：2011年1月第1版　2015年1月第2版　　**印　　次**：2015年1月第1次印刷
印　　数：1～2000
定　　价：45.00元

产品编号：060519-01

前言

数据仓库是将海量数据进行抽取、清洗和转换，并按主题重新组织，可比喻成随时间推移不断丰富的“宝藏”；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，并表示为概念、规则、规律和模式等，可比喻成“淘宝”。随着Internet的迅速普及和广泛应用，每天都会产生大量各种各样的数据，但它们背后到底隐藏着什么，这驱使人们不断探索其中的奥秘。

“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”，让我们在大数据时代不再感到茫然和困惑。未来，数据挖掘将绽放无限的生机和活力，可以辅助、部分代替甚至拓展人的智能和决策，以造福人类。

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，从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技术已经逐步渗透和深入到社会生活的各个方面，并不断催生出新的应用。

本书介绍数据仓库和数据挖掘的原理及其应用；此外，用较多篇幅列举数据仓库和数据挖掘在多个领域的实际应用。

全书分为3篇。第1篇介绍数据仓库的起源和演变，主要阐述数据仓库的定义、体系结构、组成、元数据、数据粒度、数据模型、数据质量及ETL过程，描述数据仓库设计方法和实现过程，结合实际详细阐述如何构建数据仓库及其主要应用，如OLAP和OLAM。第2篇介绍数据挖掘的起源和发展，主要阐述数据挖掘和Web挖掘的主要算法，包括聚类、分类、预测和关联分析等，详细描述数据挖掘的具体应用实例，如客户细分、虚开欺诈识别和WAP日志挖掘等。第3篇阐述数据、信息和知识之间的关系，介绍当前研究热点——语义网和本体的核心技术及方法。

本书作者潜心撰写，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

2014年8月

目录

第一篇 数据仓库

目录

目录

目录

目录

第三篇 语义网和本体

目 录

第一篇

数据仓库

第一章

第1章　数据仓库基础

1.1　概述

人类进入信息时代以来，特别是近些年，数据规模日益扩大，数据呈爆炸式增长。图灵奖获得者吉姆·格雷曾提出一个经验定律，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；天文学家需要从天文望远镜获取的观测数据（其规模可达数千GB）中发现新的遥远天体及其运动规律；医生需要从大量病人的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乃至TB级；其二均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这些海量数据？如何发现其中潜在的知识？这需要一种新的、更为有效的手段对各种数据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；数据挖掘则是利用计算机对海量数据进行快速、有效地分析和处理，从中获取知识，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，引起了知识发现、机器学习和统计分析等领域专家的广泛关注。

1.1.1　演变

数据仓库是建立在传统事务型数据库基础之上，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二十几年的时间，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已经建设不同规模的数据仓库。

传统数据库（普通数据库）和数据仓库的最根本区别在于其侧重点的不同。数据处理分为两大类：事务型处理，又称联机事务处理（Online Transaction Processing，OLTP）；分析型处理，又称联机分析处理（Online Analytical Processing，OLAP）。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，为企业决策提供可靠、有效的科学依据。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适合的,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,建立相对独立的分析处理环境,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,而在数据分析方面,则存在诸多不足,主要体现在缺乏集成性、主体不明确和分析处理效率低等多个方面。

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共享数据,缺乏数据的统一管理和维护。尽管企业内部拥有的数据量庞大,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,处理大量分析型数据的效率得不到保证。

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。且在很多时候,数据仓库是被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。

一般地,数据仓库是在传统数据库的基础上发展起来的,建立在异构业务数据库的基础上。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区别

	数 据 库	数 据 仓 库
内容	与业务相关的数据	与决策相关的数据
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	可能意味着决策延迟

从数据库到数据仓库的演变过程如图 1.1 所示。

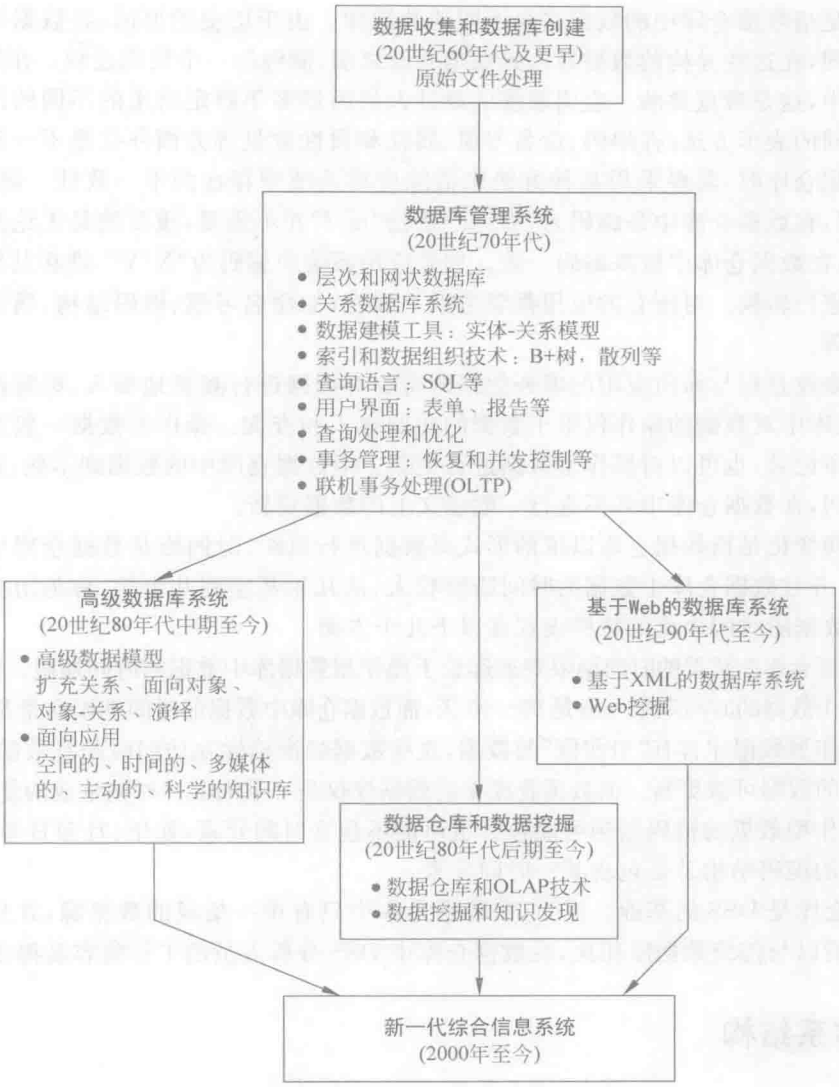

图 1.1　数据库到数据仓库的演变过程

1.1.2　定义

数据仓库的概念最早出现在 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,需经历一个集成过程。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"客户性别"的编码,在数据仓库中是编码为"男/女"还是"m/f"并不重要,重要的是无论是什么原始应用系统,在数据仓库中应该编码一致。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。

不可修改是指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,数据仓库中对数据的操作仅限于数据的初始导入和查询。操作型数据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则不然,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一个维度,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:

- 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。
- 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。
- 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。所以与传统数据库相比,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方式进行数据仓库的数据抽取。外部数据源是指数据来源于企业的外部,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部数据源,同时外部数据源的格式也不统一,如文本、表格和图像等。因此对外部数据源及其数据格式

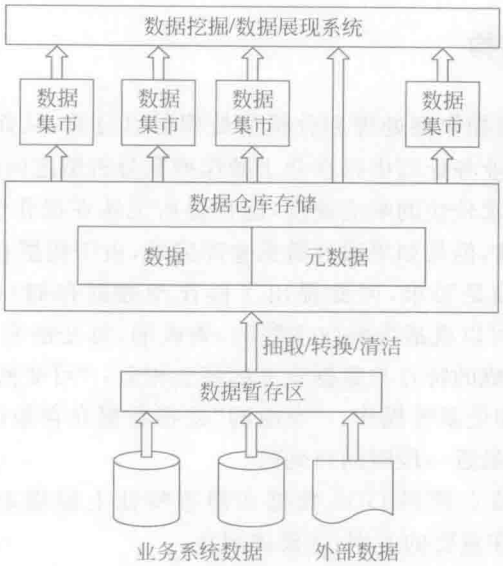

图 1.2　DB-DW 两层体系结构

等都要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。

由于数据仓库中的数据源不统一，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在于：

- 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。
- 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的保证，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。

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的一部分数据，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物理上进行分离，使用户无须到数据仓库的海量数据中查询，只在与本部门有关的数据子集上操作，这样就形成了数据集市(Data Mart)的概念，它是指面向企业中的某个部门(主题)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 DB-DW 两层的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。