SHUJU CANGKU YU SHUJU WAJUE SHIWU

21世纪高职高专规划教材
高等职业教育规划教材编委会专家审定

数据仓库与数据挖掘实务

主　编　谷　斌
副主编　耿科明　张　昶
靳艳峰　赵宝柱

北京邮电大学出版社
www.buptpress.com

21 世纪高职高专规划教材
高等职业教育规划教材编委会专家审定

数据仓库与数据挖掘实务

主　编　谷　斌
副主编　耿科明　张　昶
　　靳艳峰　赵宝柱

北京邮电大学出版社
www. buptpress. com

内 容 简 介

本书力求通过浅显易懂的语言和贴近生活的案例，深入浅出地介绍数据仓库与数据挖掘技术的概念和相关理论。本书内容覆盖数据仓库的概念、结构、设计、使用、维护、优化方法，以 SQL Server 分析服务器为例介绍了数据仓库的具体构建和使用方法。在数据挖掘部分，本书从数据挖掘的基础工作和流程开始，对常见的模型和方法做了全面介绍，并利用 Clementine 工具介绍了如何通过工具实施真实的数据挖掘过程。

本书适合作为高职高专类院校电子商务、信息管理、数据库营销等专业教材，也可作为数据分析方向培训教材。

图书在版编目(CIP)数据

数据仓库与数据挖掘实务 / 谷斌主编. -- 北京 : 北京邮电大学出版社，2014.8
ISBN 978-7-5635-4050-1

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

中国版本图书馆 CIP 数据核字(2014)第 153516 号

书　　名：数据仓库与数据挖掘实务
著作责任者：谷　斌　主编
责 任 编 辑：刘　颖
出 版 发 行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：
开　　本：787 mm×1 092mm　1/16
印　　张：13
字　　数：338 千字
版　　次：2014 年 8 月第 1 版　2014 年 8 月第 1 次印刷

ISBN 978-7-5635-4050-1　　定　价：28.00 元

前　言

随着互联网经济的迅猛发展，人类掌握的数据越来越多，如何使用数据以创造价值成为商业领域关注的热点。除了数据挖掘应用成熟的零售、金融和电信等行业外，依托数据和数据分析开展的数据库营销也为更多的行业所接受。在此背景下，市场营销、电子商务等专业对数据分析相关知识和能力提出了越来越明确的需求。

数据仓库与数据挖掘课程在高校中一直作为高年级本科生或研究生的专业课程。相关教材主要围绕数据仓库、数据挖掘的概念、设计、算法等方面展开，重点侧重知识教授。但是对于高职高专类院校，这种教材编写思路无法适应基本教学。一方面，职业技术学院的学生逻辑思维能力较弱，对需要大量计算的理论推导和模型推演难以理解，更愿意学习操作性强的技能类课程。另一方面，企业数据分析的应用仍主要集中在传统分析应用上，如分类、聚类、关联规则等方向，研究型、开发型工作较少。企业更迫切需要将多维分析和数据挖掘应用到决策和营销一线，更需要能够利用各类工具解决实际岗位问题的员工。在这种情况下，在高职高专类院校中开设本课程就必须对课程重新定位，不能直接沿用本科或研究生教材。

根据多年教学实践和企业数据分析工作的经验，本书在原来各类教材的基础上，对数据仓库和数据挖掘两部分教学内容进行了仔细筛选，对操作技能做了重点倾斜。在具体的工具软件方面，本教材结合市场上较为流行的工具，简要介绍了数据分析工作开展的基本思路和方法。

数据仓库部分主要强调了数据仓库与数据库的差异性，在数据仓库管理、建设、优化等几个方面做出了侧重介绍。结合 SQL Server 分析服务器介绍数据仓库的构建、实施和 OLAP 分析操作。数据挖掘部分侧重介绍了传统分类、聚类、关联规则挖掘，同时介绍了互联网和与电子商务紧密相关的 Web 挖掘和文本挖掘。

全书分为 8 章。第 1 章综合介绍数据仓库与数据挖掘的基本情况。第 2 章主要介绍数据仓库的生命周期和基本体系结构。第 3 章为数据仓库的设计，从概念模型、逻辑模型和物理模型三个层面介绍数据仓库的设计方法。第 4 章为如何使用数据仓库，包括 OLAP 分析、元数据、数据仓库的管理和维护以及数据仓库的优化方法等内容。第 5 章为数据预处理，将数据仓库的数据导入过程和数据挖掘的预处理过程合并在一章介绍。第 6 章介绍数据挖掘的难点和知识表示等数据挖掘的基础知识。第 7 章是本书的重点章节，介绍了数据挖掘中主要的几种挖掘方法。第 8 章简单介绍了大数据的概念。

参与本书编写的还有耿科明、张昶、靳艳峰、赵宝柱等几位老师。

由于作者水平有限，欢迎对书中不足给予指正。

对参考文献中列出的以及未列出的所有文献作者表示由衷的感谢。

作　者

目　录

第1章　数据仓库与数据挖掘概述

1.1　数据库与数据仓库

1.1.1　数据的层次性

随着互联网的迅速发展和“大数据时代”的到来，数据已经成为人们生活和工作中不可缺的组成部分。无论哪个行业、哪个地区，想要在这场浩大的变革中获取优势地位，制胜的关键在于如何对数据掌握、理解和使用。

对于数据的理解，可以从生活中的实例开始。39 是一个用来描述对象的数值，比 38 大，比 40 小，但是这个数值的具体含义只有在提供了额外的描述后才能更好地理解。补充上℃这个描述，39 变为 39℃，对这个数据的理解就映射为对温度的感受。当明确为体温 39℃后，对这个值的理解就更丰富到“发烧”、“生病”了。对于同一个数值理解的巨大差异体现在数字背后，这些就是数据内涵和抽象。

按照以彼得·德鲁克博士(Peter F. Drucker)和斯威比博士为代表的知识管理理论来看，我们已经生活在知识经济和知识管理的环境当中。每时每刻，身边都充满了各种各样的数据。只有将这些杂乱无章的数据，转换为信息和知识，才能帮助我们做出合理的、科学的选择。可见知识是从数据到智慧划分为不同层次的。

图 1-1　知识层次

从图中可以看出来，数据、信息、知识(可扩充智慧层次)构成了知识层次结构，代表了人类认识世界的不同层次。

第一个层次是数据。数据作为最基础的层次提供了对现实世界的理性描述。在众多的定义中从多样的角度定义出数据是对客观事物的数量、属性、位置及其相互关系进行抽象表示，以适合在这个领域中用人工或自然的方式进行保存、传递和处理。这个层次是我们认识世界的基础和最直接的手段。

第二个层次是信息。“信息”是现在出现频率很高的一个词，由于很难给出基础科学层次上的信息定义。系统科学界曾下决心暂时不把信息作为系统学的基本概念，留待条件成熟后再作弥补。到目前为止，围绕信息定义所出现的流行说法已不下百种。以下是一些比较典型、比较有代表性的说法。1948 年信息论的创始人香农在题为《通信的数学理论》的论文中指出：“信息是用来消除随机不定性的东西。”1950 年数学家、控制论的奠基人诺伯特·维纳认为，信息是人们在适应客观世界，并使这种适应被客观世界感受的过程中与客观世界进行交换的内容的名称。1963 年 Weaver，Bar-Hillel，Carnap，Popper 等人提出信息论研究应当从香农信息发展到语义信息。语义不仅与所用的语法和语句结构有关，而且与信宿对于所用符号的主观感知有关，是一种主观信息。

作为第二个层次，信息与数据紧密相关。虽然信息表现为各种各样的数据，但是其所蕴含的内在意义是单纯的数据无法提供的，比如 39℃不简单代表了数值大小，还作为温度的度量描述了冷暖差异。所以如果用一句话来分辨什么是数据，什么是信息的话，那就是“数据是信息的载体，信息是数据的内涵”。如果用公式来表达信息与数据的关系的话，可以描述为“信息＝数据＋处理＋时间”。也就是说，信息是具有一定时效性的、有逻辑的、经过加工处理的、对决策有价值的数据流。

第三个层次是知识。知识之所以在数据与信息之上，是因为它反映了客观世界的规律性，与决策相关。一般认为这些知识的经典定义都有其价值和意义，信息虽然给出了数据中一些有一定意义的内涵，但是在时间效用失效后其价值开始衰减，只有通过人们的参与对信息以归纳、演绎、比较等手段进行挖掘，使其有价值的部分沉淀下来，并与已存在的人类知识体系相结合，这部分有价值的信息才会转变成知识。例如，北京 7 月 1 日，气温为 30℃；12 月 1 日气温为 3℃。这些信息一般会在时效性消失后，变得没有价值，但当人们对这些信息进行归纳和对比就会发现北京每年的 7 月气温会比较高，12 月气温比较低，于是总结出一年有春、夏、秋、冬四个季节，有价值的信息沉淀并结构化后就形成了知识。知识作为对信息的抽象和提炼，是人类改造客观世界的重要指导。

除了以上三个层次外，如果再进一步划分的话，还可以划分出智慧层次。智慧是人类解决问题的一种能力，是人类特有的能力。智慧的产生需要基于知识的应用。一般认为，智慧是人类基于已有的知识，针对物质世界运动过程中产生的问题，根据获得的信息进行分析、对比、演绎，找出解决方案的能力。这种能力运用的结果是将信息的有价值部分挖掘出来并使之成为已有知识架构的一部分。

本书在数据、信息、知识的框架下，介绍商业领域中影响力越来越大的数据仓库和数据挖掘技术，以及相关技术在数据库营销、电子商务等方向的应用。

1.1.2 数据仓库出现的原因

随着 20 世纪 90 年代后期互联网的兴起与飞速发展，人类进入了信息爆炸的时代。企业中大量的信息和数据，需要用科学的方法去整理，需要从不同视角对企业经营的各方面信息精

确分析、准确判断。面临激烈的竞争环境，及时做出正确决策是企业生存与发展的重要环节。企业利润的降低使得很多企业必须从粗放经营转变到集约经营，经营决策需要快速、尽可能多的定量分析，而不是似是而非的定性分析。而随着 ERP、CRM 等信息系统的广泛应用以及互联网的蓬勃发展，企业数据量激增，企业需要获得更高层次的数据分析，而数据库越来越难以满足这种需求。

关系型数据库管理系统(RDBMS)作为目前最重要的数据库应用，在企业经营的方方面面都起到了极其重要的作用，承担着重要责任，同时在日常生活中，如 QQ、淘宝、各大银行的业务系统等，绝大部分重要应用也都基于关系型数据库系统。但是正如上面说到的，数据库系统在经历了几十年的发展后正面对着越来越多的挑战。有些问题可以通过新技术来加以解决，如并行、分布、NoSQL(非关系型数据库系统)等。但是就数据库管理系统整体而言，面对越来越复杂的决策需求和综合性快速分析需求显得力不从心。

从整体上看，数据库管理系统在新形势下表现出的问题主要有以下几个方面。

1. 数据量增长迅速，处理复杂问题的性能下降明显

数据库系统的性能与其承载的数据量紧密相关，且两者之间并不是线性关系。当数据量的增长达到一个量级后，数据库系统的性能会迅速下降。一般来说，数据库性能与系统架构和硬件性能有紧密关系，如磁盘、网络、内存、CPU 等。要解决其中任何一个问题都需要不菲的、持之以恒的投入。特别是在这个数据不断膨胀的时代，企业数据量从过去的 MB 到 GB 再到 TB，增长到现在的 PB 级数据规模。虽然近十来年分布式、内存数据库等新技术应用越来越多，但是数据增长的速度相比硬件投入和系统优化带来的性能提升要快得多。数据库性能与需求之间的矛盾越来越突出。特别是在当今，电子商务如此发达，越来越多的商业应用集中于对客户的分析和特征模式识别，再加上企业管理中科学决策和数据分析方面的广泛需求，传统数据库系统已经难以满足复杂查询要求，需要一种主要针对分析应用的高性能数据管理工具。

2. 存在信息孤岛现象，异构环境的数据转换和共享困难

现在企业的各项经营管理活动已经无法离开各类信息系统的支持，对数据库系统的依赖程度越来越高。但在企业实际运行中，业务数据库系统的条块与部门分割，导致数据分布的分散化与无序化。在一个企业内部，供应、生产、销售、财务等部门往往各自使用着一套满足自身工作需要的应用程序。建立、使用、维护着本部门的业务数据库系统。另外，各部门的应用程序与业务数据库系统在规划、建立时，缺少通盘的考虑，有些应用系统更是由行政管理部门指定使用的，企业自己没有选择的权力，也就没有统筹考虑的可能。这样，企业内部尽管拥有的数据量极大，但却各成体系、封闭存在。构成相互独立的所谓“信息孤岛”，无法形成一个统一的整体。

有些企业认识到了这种问题的存在，试图以 IT 部门为主。连接企业内的各个信息孤岛，以改变这种状况完成数据共享。但企业的 IT 部门并不从事一线的业务工作，大多缺乏足够的业务知识，且受限于开发商的技术保密等原因，往往无法真正打破信息孤岛之间的藩篱。同时由于业务数据库缺乏统一的定义与规划，导致数据定义存在差异。由于各应用程序源自不同的开发商，所使用的数据库系统依托的平台和种类各异、结构不同，变量的定义也缺乏统一的规范和标准，往往出现变量类型和名称完全相同的字段，出现在不同应用系统的数据库中具有完全不同的含义，又或者具有相同含义的字段，在不同应用系统的数据库中变量类型和名称完全不同。这些问题都严重限制了企业整合内部数据的努力，制约了以内部数据库系统存储的数据开展综合决策分析的尝试。

3. 数据主要面向事务处理，缺少对决策和数据分析的支撑

数据库核心工作是完成事务处理，为了保证由于任何原因造成的异常不会影响数据库的安全和数据的一致性，数据库管理系统设计了各种复杂的机制来解决这些问题。典型的如并发控制问题，为了解决可能存在的并发冲突，数据库管理系统通过两段锁协议实现事务的可串行化。但是在处理分析问题时，这些机制并没有太多的用处，反而会影响分析的效率。

因此，传统数据库在当前数据量增长迅速、经营管理中决策支持、数据分析要求越来越高的背景下，越来越力不从心，无法担当作为大规模数据综合分析平台的重任，管理决策任务需要有一种新的理论、技术和工具来提供支持，这就是数据仓库。

1.1.3 数据仓库的概念

前面提到数据仓库相对于数据库更适合于企业管理决策，具有更高的效率，更能针对企业的分析要求。具体到数据仓库的定义则有多种，如 Jiawei Han(韩家炜)在数据挖掘概念与技术中提到的“数据仓库是一种数据的长期存储，这些数据来自多数据源，是有组织的，以便支持管理决策。这些数据在一种一致的模式下存放，并且通常是汇总的。数据仓库提供一些数据分析能力，称作 OLAP(联机分析处理)”。还有一些厂商认为数据仓库是一种信息系统，能给一个组织或机构提供商务智能以支持管理决策的制订。在各类定义中比较典型，引用最广泛的则是 W. H. Inmon 在 1992 年出版的 *Building the Data Warehouse* 中提出的“数据仓库是面向主题的、集成的、随时间变化的、非易失的数据集合，用于支持管理层的决策过程(A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions)”。

在 Inmon 的定义中数据仓库有四个关键特性和一个最终应用落脚点，分别是“面向主题的”、“集成的”、“随时间变化的”、“非易失的”和面向管理的决策问题，这四个特点和主要应用方向代表了目前对数据仓库的主流共识。

1. 面向主题

“面向主题”是数据仓库中数据组织的最基本原则。传统数据库系统围绕着企业的功能应用组织和设计，目的是完成具体的业务，具体业务操作和过程一般总会有具体对应的关系或属性。而在主题方面，不同公司会有较大差异、商业公司(零售企业)的主要应用可能是销售管理、客户管理、进货渠道管理、仓储管理等；而对于商业保险公司来说，可能主要围绕保单处理、投保人管理(客户管理)、保险政策、保险代理人管理(销售团队管理)等方面；生产型企业则围绕其他的核心功能组织数据库和各类业务系统。数据仓库以支持管理层的决策为目的，围绕着某些具体的分析主题而组织。

数据仓库中的所谓“主题”，是一个逻辑概念。在信息管理的层次上，主题就是从管理的角度出发，对数据进行综合分析而抽取出的需要作进一步分析的对象。数据仓库的构造过程，首先就是确定主题的过程。数据仓库的设计者必须明确该数据仓库所支持的决策内容，即数据仓库的用途，并将决策内容归纳为若干个具体的、易于利用数据组织加以分析的主题。

主题的抽取，必须体现出独立性和明确性的特点。即主题要有独立的内涵，各个主题之间要有明确的界限，不应有依存关系。要保证与主题相关的所有数据都能得到正确的组织，避免数据的缺失与冗余。

在数据仓库内部数据组织的层次上，主题体现为若干数据集合，每个数据集合内的数据，

各自描述一个共同的对象的某方面的特征。这些数据组合起来，共同形成对该对象的较为完整、一致、准确的描述，这一被描述的对象就是“主题”。

在构造数据仓库的过程中，确定了主题之后，就应对业务数据库的内容加以组织归类。需指出的是，业务数据库的内容和主题之间并不体现出一一对应的关系，数据也不是从数据库直接复制到数据仓库中的。有两点在划分数据仓库主题时需要明确注意：

(1) 数据库中数据的多重归属问题。由于主题之间存在一定的逻辑联系，有些业务数据库中的某些属性，可能对多个主题的分析有用。例如，销售数据库中的“销售金额”属性，在“销售”、“人力资源”等主题的分析中，都要用到。但这种多重性的实质，是同一数据的多次使用，而不是同一数据的重复物理存储。

(2) 不是所有数据库中的数据都需要导入数据仓库，并不是业务数据库中的所有内容都对主题分析有用。有些内容完全是为了便于进行业务处理而产生的，与任何主题都无关，在导入数据仓库时，应当舍弃。

2. 数据的集成性

数据仓库中数据的集成性，是指在构建数据仓库的过程中，多个外部数据源内格式不同、定义各异的数据，按既定的策略经过抽取、清洗、转换等一系列处理，最终构成一个有机的整体。要再次强调的是，这个整合过程绝不是简单地将数据从业务数据库复制到数据仓库中。和基于传统数据库的业务处理程序不同，数据仓库的数据并不直接取自业务的处理过程，而是在对业务数据库的内容进行处理后得到的。传统业务处理程序的侧重点在于迅速、正确地处理所有业务，记录业务内容和处理结果，而不是对决策提供支持。数据仓库直接使用传统业务处理程序的处理结果，这样就节省了业务处理的开销，可将精力完全集中在数据分析上。

如图 1-2 所示，数据仓库从业务数据库中获取数据后，并不直接将其导入，而是进行一系列的预处理工作，即对数据进行筛选、清洗和转换、综合等工作(ETL)，以解决数据中存在的以下问题：

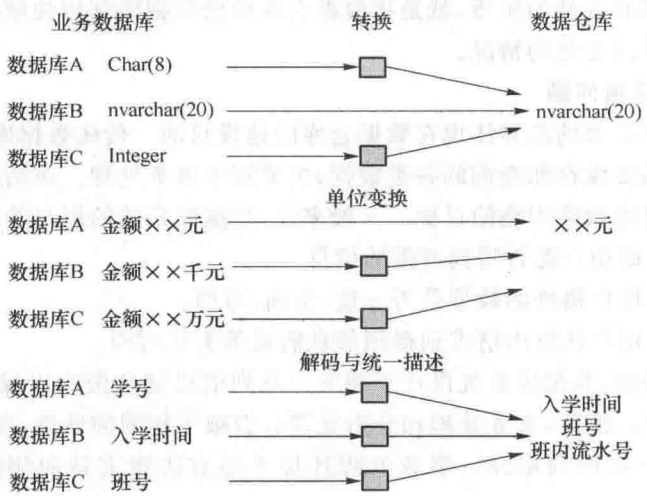

图 1-2　数据的集成

(1) 数据格式的差异。不同的业务系统，所依据的数据库系统可能是不同的，而且即使是基于同一种数据库系统，同一属性在不同应用中的定义也可能是不同的。以“雇员 ID”字段为例，在有些系统中定义为 char(8)，而在有些系统中则定义为 integer。

(2) 计量单位问题。不仅在基于不同数据库系统的业务系统中,对同一属性的计量单位可能不同,就是在同一业务数据库的不同表中,对同一属性的计量单位也可能缺乏统一,存在差异。例如,"单价金额",在有些表中,以"元"为单位,而在有些表中,则以"万元"为单位。

(3) 数据编码的处理。在业务系统中,为了便于存储许多属性定义了各种编码。一方面,这些编码形式不一,存在很大差异,必须统一解码或转换后才能存储在数据仓库中;另一方面为了便于后续分析,也应当将复杂的编码解码为不同的字段。例如,性别属性使用 0 代表未知,1 代表男性,2 代表女性;学生类别使用 0 代表正常学生,1 代表留级学生,2 代表休学;将学号由 1 个 10 位整数字段分解为入学年份、班级、小学号 3 个字段,等等。

(4) 字段的统一。在集成业务数据库系统数据时还需要解决一词多义和多词一义的问题,要在数据仓库中定义统一的字段意义。

3. 数据的非易失性

数据按照业务要求在操作型数据库系统产生、更新、删除和查询。但是数据仓库则体现出一种不同数据的特性。数据被装载(load)到数据仓库后,被打上一个时间戳。数据仓库中的这个数据代表了在某一时刻业务数据库中对应数据项的描述,可以称之为数据快照。虽然随着时间的流逝,在实际业务中这个数据字段可能早已发生变化,但是在数据仓库中,该数据仍代表在这个时间戳时刻,该数据项的值,不会随着后续装载进来的新数据而发生变化。这样不断导入业务数据,在数据仓库中会保留下来该数据项的历史变化记录,为后续分析和决策提供了依据。

可以看出来,因为不需要更新已经导入的数据,并且也很少有必要进行删除操作,所以数据在数据仓库中是稳定的,也就是非易失的,即数据一旦导入数据仓库就很少发生变化。

4. 数据是随时间变化的

数据的时变性,是指数据仓库的内容随时间的变化而不断得到增补、更新。正如上面谈到非易失性时说的,数据仓库对导入其中的数据给定一个时间戳,使之成为一个描述特定时刻特征的数据快照。数据时变性的实质,就是指数据仓库中的数据能利用快照数据,形成历史数据的轨迹描述业务随时间变化的情况。

5. 面向管理的决策问题

与数据库系统的最大的差异体现在数据仓库的建设目的。传统数据库系统是为了处理企业在业务操作中所需要保存和查询的各类数据,主要用于事务处理。在结构上、设计上和应用方式上都体现出数据库始终围绕的目标。一般来说,数据库系统的用户关心如下问题:

(1) 可访问性。即用户能否得到所需的信息。

(2) 完整性。即用户得到的数据是否一致、全面、可信。

(3) 及时性。即用户从发出请求到得到信息需要等多长时间。

为了解决这些问题,数据库系统设计了相关一系列的机制和策略以保证数据库系统能够很好地应对各项要求。但是,事务处理和分析处理有着极不相同的性质,直接使用事务处理环境来支持决策存在一定的局限性。事务处理环境不适宜决策支持应用的原因主要有以下几种:

(1) 事务处理和分析处理的性能不同。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;在分析处理环境中,用户的行为特点完全不同,某个决策支持应用程序可能需要连续运行数个小时,从而消耗大量的系统资源。因此,将事务处理和分析处理这两种性能差异很大的应用放在同一个环境中运行是不合适的。

(2) 数据集成问题。决策支持应用需要全面、正确的数据。全面、正确的数据是实现有效分析和决策的前提，相关数据收集得越完整，得到的分析结果就越准确，决策就越可靠。然而由于企业内部的事务处理应用比较分散，导致企业业务数据分散，形成信息孤岛困境。数据集成可以使企业能够拥有全面、正确的数据。

(3) 历史数据问题。事务处理一般只针对当前数据，因此，在数据库中一般只存储短期数据，即使有一些历史数据被保存下来，也往往没有得到充分利用。但对于决策分析而言，历史数据是相当重要的。没有对历史数据的详细分析就很难把握企业的发展趋势，绝大多数分析方法都是建立在大量的历史数据基础之上的。同时，决策支持系统在空间和时间的广度上对数据提出了更高的要求，事务处理环境难以满足这些要求。

(4) 数据综合问题。在事务处理系统中积累了大量的细节数据，如何利用这些数据进行决策分析，一般需要在决策分析前对它们进行不同程度的综合。然而，在事务处理系统中，这种综合往往认为是一种冗余而被限制。例如，在设计数据库时，一般要遵循数据库规范化理论，其主要目的在于尽量避免数据的冗余，保证数据的一致性。

因此，要提高分析和决策的效率和有效性，就必须把分析型数据从事务处理环境中提取出来，按照决策支持处理的需要进行重新组织，建立单独的分析处理环境。也就是说，分析型处理及其数据必须与操作型处理及其数据相分离。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。

1.1.4　数据仓库与数据库的差异

从数据仓库的定义和特点可以看出数据仓库(Data Warehouse，DW)与数据库(Data Base，DB)的主要差异。在这里用一张表作一个简单的对比，使得大家能对这两个概念有更清晰的认识。

表 1-1　数据库与数据仓库的差异

	数据库(DB)	数据仓库(DW)
数据的粒度	细节性数据	以综合数据为核心的多粒度设计
时效性	实时数据	以历史性数据为主
数据能否更新	可更新	不更新
操作需求的明确性	事先明确具体业务需求	需求事先不明确
应用目标	事务应用	分析应用
一次操作数据量	小	大

需要解释的有如下项目：

(1) 数据粒度的差异。粒度是在数据仓库建设过程中比较重要的一个方面。所谓粒度是指确定数据仓库中数据单位的细节和汇总程度描述。一般情况下，根据数据粒度划分标准，可以将数据仓库中的数据划分为：详细数据、轻度综合、高度综合三级。粒度的基本原则是细化程度越高，数据量越大，粒度越小；细化程度越低，数据量越小，粒度越大。由于数据仓库主要针对决策分析问题，多考虑宏观层面，故此，一般来说，DW 中的数据粒度偏向综合数据。但是

为了避免大量损失具体细节数据，数据仓库的粒度设计，通常采取多级别的粒度设计方案来兼顾各类问题并获得可接受的性能。相反由于数据库系统面向具体的业务，必须要保存所有业务需要用到的所有细节数据。

(2) 操作需求的明确性差异。基于数据库系统的各类信息系统在开发时需要作较为全面的需求调研和分析，要对整个业务流程中所有可能出现的要求做出合理设计。因此，数据库系统的结构和使用方法是精心设计好的、优化过的。相反，让一个企业经理或决策者仔细描述他有可能遇到的所有决策需求是基本不可能完成的任务。正是由于数据仓库面向的决策问题很难在具体问题出现前做好分析和设计，所以设计数据仓库时要能灵活应对各类突然出现的新决策需求。

(3) 一次操作的数据量差异。在目前的商业应用中，数据库系统的一次查询的数据量越来越大，有时甚至可以达到 TB 级别。这也是现在各类并行、分布式系统，甚至是大数据系统出现的一个原因。但是由于决策分析一方面需要多年的历史数据来开展，另一方面操作也较为复杂。所以一般来看，面向业务的事务处理相比面向决策支持的分析处理要简单得多，数据量也要小。

1.1.5 数据仓库的商业应用

一些企业较早认识到数据仓库的价值，投资购买数据仓库。这种投资使它们在本行业竞争中处于主动地位。传统营销理念下，企业围绕产品这个中心开展各类商务活动，开发了一个新产品后通过各类营销活动希望大家都来买，而新一代的商业模式则侧重于客户的需求，以客户为中心，以需求定制产品。这个转变代表了营销从传统的 4P 向 4C 的转变。有了数据仓库后，企业可以通过大量的、各方各面的数据分析客户是谁，他喜欢什么样的产品和服务，应该如何提供更好的产品和服务给他，并以此创造更多利润。

沃尔玛(Walmart)以营业额计算为全球最大的零售公司，同时也是世界上雇员最多的企业，连续三年在美国《财富》杂志世界 500 强企业中居首。其经营法则的核心是控制成本，保证在竞争对手面前具备最低的价格。严谨的采购态度、完善的发货系统和先进的存货管理是促成沃尔玛做到成本最低、价格最便宜的关键因素。几年来他们的数据仓库规模从 6TB 增加到现在的 100TB。利用数据仓库，他们通过网上供货商随时补充货源，实现对库存商品更有效的控制，达到最小库存量。

数据仓库在电信等行业发挥着巨大作用。当电信行业出现竞争时，就会出现客户从甲公司跳到乙公司的现象，这种现象会使电信公司浪费巨额资金。有了数据仓库，就能预测客户流失，知道谁可能跳到竞争对手那里去。如果企业能够在客户流失之前采取适当的措施，就能够在一定程度上减少客户流失，从而给企业带来巨大的收益。这种效益十分明显，它可以帮助企业在一到两年的时间内收回在数据仓库方面的投资。

数据仓库在银行的应用也非常普遍。现在，发达国家的大型商业银行，特别是美国的许多大银行都建立了自己的数据仓库系统，其中存储的客户信息量可以用千亿字节和万亿字节来计算。数据仓库可以有效地帮助银行从这些海量客户数据中开展分析过程。例如，从中找出现有客户潜在的消费行为，分析客户信用卡的使用情况和信用卡犯罪的可能性，银行从特定客户得到赢利的模式，比较不同类型客户的赢利情况，客户使用各种金融产品的频率和爱好，分