数据库应用系列教材

数据仓库与数据挖掘原理及应用

王丽珍　周丽华　陈红梅　邹力鹃　编著

科学出版社

北　京

内 容 简 介

本书全面深入地介绍了数据仓库、联机分析处理（OLAP）和数据挖掘的基本概念、基本原理和应用技术。全书分成三篇，数据仓库及 OLAP 概念、原理和技术篇的主要内容包括数据仓库的基本概念、体系结构、模型设计、创建和维护，ETL、元数据、数据集市，OLAP 的基本概念、分类、模型设计；数据挖掘技术篇介绍了数据挖掘的基本理论、基本过程、常见模型和算法；工具及实例介绍篇简要介绍了数据仓库产品工具的基本情况，对产品选择和评判进行了一些分析，并较详细地介绍和分析了移动通信业务数据仓库系统。

本书可作为计算机、信息系统等专业的学生学习数据仓库、OLAP 及数据挖掘技术的实用教程，也可供从事数据仓库、数据挖掘研究、设计、开发等工作的科研、工程人员参考。

图书在版编目(CIP)数据

数据仓库与数据挖掘原理及应用/王丽珍等编著. —北京：科学出版社，2005

（数据库应用系列教材）

ISBN 7-03-015657-9

Ⅰ. 数… Ⅱ.王… Ⅲ.①数据库系统-教材②数据采集-教材 Ⅳ.①TP311.13②TP274

中国版本图书馆 CIP 数据核字（2005）第 058632 号

责任编辑：鞠丽娜 韩 洁/责任校对：柏连海

责任印制：吕春珉/封面设计：三函设计

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

印刷

科学出版社发行 各地新华书店经销

*

2005 年 6 月第 一 版 开本：B5（720×1000）

2005 年 6 月第一次印刷 印张：19 3/4

印数：1—4 000 字数：378 000

定价：26.00 元

（如有印装质量问题，我社负责调换<环伟>）

销售部电话 010-62136131 编辑部电话 010-62138978-8002（HI06）

《数据库应用系列教材》编委会

序

近年来，我国高等教育事业飞跃发展，在校学生人数突飞猛进，与此同时，高校教育改革逐渐冲破旧的计划经济模式，新的模式也正在建立。在这种形势下，旧的教材体系已不能适应新的需要，因此迫切需要建立新的教材体系。基于此种情况，我们以计算机相关专业中的数据库系统教材为依托，组织了一套适应不同需求、不同层次、不同目标的数据库系列教材，其组织依据是：

1. 在高等学校中随着老校的调整与改革，新校的不断涌现，过去计划经济的一刀切模式已逐渐改变，各校在培养目标、人才市场定位方面已出现多种模式（如研究型、应用型、开发型等），因此需要有多种不同数据库系统教材以适应不同模式的需求，而现有教材大多只能适应少数模式的需求。

2. 近年来计算机应用飞速发展，计算机与其他专业的交叉应用发展很快，如文科中的数量经济、信息管理、电子商务、财政金融等专业，理工科中的机械、建筑、城市规划、遥感遥测等都急需开设计算机及数据库等相应课程，也需相应的教材，而此方面的合适教材目前较为少见。

3. 随着教学改革的深入，数据库课程自身也需进行改革，它除了需要有主课程外，还需要有若干门配套的辅助性课程与教材，如数据库分析与设计、Web 数据库、数据库应用等课程，以及数据库实验课、实习课以及习题集等配套教材。此外，还需配合使用现代化手段如电子教案及课件等相关音像制品。所有这些教材都需构成一个以数据库主课程为核心的有机组合的系列教材，而此方面的组合教材正是目前所缺少的。

4. 数据库技术本身发展很快，而教材编写相对滞后，同时国内数据库教材又受国外教材影响较大，因此适合国情的本土化教材的建设尤为重要，因此，能编写出既适应目前技术发展水平，又能适应我国经济发展需要的数据库教材是当前之急需。

5. 本系列教材能适应不同模式，不同层次、不同系科（计算机及非计算机专业）的需求，它除追求基本原理的正确性外着重在它的应用性。由于数据库是一门实用性很强的课程，我们希望学生在学了此课程后能在实际应用中发挥作用。

本系列教材正是为适应上面所述的需要而编写的，目前它以计算机及非计算机专业的本科生教材为主，并将逐渐扩充到研究生及大专层次。本系列教材采取开放性组织方式，今后将根据学科发展陆续组织出版数据库领域的优秀图书。

本系列教材的编写人员涉及各个不同层次与专业，有大量实际经验与理论水平，相信这套教材的问世能对数据库教学起一定的促进作用。

《数据库应用系列教材》编委会

2004 年 9 月

前　言

进入信息社会以来，信息技术经历了这样的发展过程：从计算机主机的信息集中处理方式到个人计算机（PC）的信息分布处理形式的转变；从单一的计算机操作系统到计算机互联网络操作的改变；从客户机/服务器（Client/Server）计算体系到多层体系结构计算模式的转变；从单一数据库到大型数据仓库和从局域网到Internet 的改变。现代信息技术的发展和现代科学技术的进步，使人类迈入了新的时期——信息化时代。

信息处理技术的发展，使得各类数据、信息急剧增长，给数据的传输、存储带来了许多新的问题，特别是由于各类不同事务产生大量不同类型的数据，这些数据分别被各个时期建立的许多应用系统所使用。人们希望能够看到所有数据和信息的综合情况，而这些数据和信息有许多不能被统一描述，不能被现有应用系统综合使用。针对这一问题，人们设想专门为业务的统计分析建立一个数据中心，它的数据来自联机的事务处理系统、异构的外部数据源、脱机的历史业务数据等，这个数据中心就叫数据仓库。数据仓库技术的应运而生，成为信息技术领域非常热门的话题之一。

数据仓库技术的提出，建立了一种体系化的数据存储环境，将分析决策所需要的大量数据从传统的操作环境中分离出来，使分散、不一致的操作数据转换成集成、统一的信息。企业内不同单位、不同角色的成员都可以在此单一的环境之下，通过运用其中的数据与信息，发现全新的视野和新的问题，产生用于决策的新分析方法。作为决策支持系统的重要组成部分，数据仓库为决策支持系统提供了分析决策所需的数据；OLAP 的产生进一步增强了决策支持系统快速、一致和交互性的分析能力，它利用存储在数据仓库中的数据完成各种分析操作，并以直观易懂的形式将分析结果展现给决策分析人员；而数据挖掘是从大量数据中提取或“挖掘”知识，从而实现从“数据→信息→知识”的过程，为企业的管理层提供各种层次的决策支持。

本书对数据仓库、OLAP、数据挖掘的原理、技术、工具和应用做了全面深入地介绍和分析，对数据仓库、OLAP 和数据挖掘的发展及应用前景也进行了细致深入地讨论。全书共三篇，分别是数据仓库及 OLAP 概念、原理和技术篇、数据挖掘技术篇和工具及实例介绍篇。内容组织的思路为：基本概念→基本原理→实际应用。

本书在结构的组织上，采用引言→主体内容→小结→习题的结构形式。每章后面的习题可作为课后作业。这些习题或者是短问题，用于测试对内容的掌握；

或者是长问题，需要分析思考甚至查阅资料来完成。在内容的介绍上，除理论联系实际外，还使用了大量的图示及实例，使该书具有较强的可读性和可理解性，因此，凡具有一定数据库基础知识的人，都能学懂本书的内容。

讲授这门课程一般需 70 学时左右，因对问题阐述深入浅出，该书既可作为课堂教学的教材，也可供自学时参考。

本书的写作过程也是笔者学习、研讨、提高的过程。在此过程中，笔者对国内外大量的资料进行了归纳和整理，认真学习了各种数据仓库、OLAP 和数据挖掘工具，对参与及主持开发的两个数据仓库系统进行了全面的分析和总结。

在本书写作过程中，所写内容已经过反复研讨，且有些章节可能由某位老师执笔，而由另一位老师修改，因此难以严格划分每个人之工作量。就执笔而言，其分工如下：第 1、6、7、12、13 章由王丽珍执笔，第 3~5 章由周丽华执笔，第 8~10 章由陈红梅执笔，第 2、11 章由邹力鹍执笔。

本书的写作得到南京大学徐洁磐教授的鼓励和支持，徐教授认真审阅了全书，提出了许多宝贵的修改意见。云南大学研究生夏勇、胥玲芳、秦海林、陈涛、熊芸、陈克平、陈杉等为本书的完成做了大量的辅助工作。另外，本书还得到国家自然科学基金（编号 60463004）和云南省自然科学基金（编号 2002F0013M）的资助，在此一并表示衷心的感谢。

由于笔者水平有限，书中错漏和不妥之处在所难免，恳请读者批评指正。

作　者

2005 年 4 月

目　　录

第一篇　数据仓库及 OLAP 概念、原理和技术篇

第二篇　数据挖掘技术篇

第三篇　工具及实例介绍篇

第一篇　数据仓库及 OLAP 概念、原理和技术篇

第 1 章　数据仓库基本概念

近十几年，随着科学技术飞速的发展，经济和社会都取得了极大的进步，与此同时，在各个领域产生了大量的数据，如人类对太空的探索，银行每天的巨额交易数据。显然在这些数据中蕴藏着丰富的信息，如何处理这些数据得到有益的信息，人们进行了有益的探索。计算机技术的迅速发展使得处理数据成为可能，这就推动了数据库技术的极大发展，但是面对不断增加如潮水般的数据，人们不再满足于数据库的查询功能，提出了深层次问题：能不能从数据中提取信息或者知识为决策服务。就数据库技术而言已经显得无能为力了，这就急需有新的方法来处理这些海量般的数据。在这种情况下，数据库逐步发展到了数据仓库。世界上最早的数据仓库是 NCR 公司为全美、也是全世界最大的连锁超市集团 Wal★Mart 在 1981 年建立的，而最早将数据仓库提升到理论高度进行分析并提出数据仓库这个概念的则是著名学者 W.H.Inmon，他对数据仓库所下的定义是：数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合，用于支持管理决策过程。由此可见，数据仓库是一个综合的解决方案，主要用来帮助企业有关主管部门和业务人员做出更符合业务发展规律的决策。

1.1　从数据库到数据仓库

1.1.1　蜘蛛网问题

在市场经济的激烈竞争中，信息对于企业的生存和发展起着至关重要的作用。企业对信息的需求是多方面的，为了避免企业中各部门或各用户间的冲突和简化用户的数据视图，一种称为“抽取程序”的方法被广泛地应用。比如，市场部人员通常只关心企业的销售、市场策划方面的信息，而不注重企业的研发、生产等其他环节。因此，将销售、市场策划方面的信息抽取出来单独建立部门级的数据库很有必要，这样可以提高数据的访问效率。在部门级数据的基础上可能还要被继续执行抽取程序，以建立个人级的数据库。比如，专门负责制作公司财务报表的数据人员，常常需要从财务部门的数据库系统中抽取数据。又如，部门经理可能经常抽取常用的数据到本地，有针对性的建立个人级数据库就显得尤为重要。

随着数据的逐层抽取，很可能最终导致系统内的数据间形成了错综复杂的网状结构，如图 1.1 所示，人们形象地称其为“蜘蛛网”。一个大型的公司每天进行上万次的数据抽取很普遍。这种演变不是人为制造的，而是自然演变的结果。企业的规模越大,“蜘蛛网”问题就越严重。

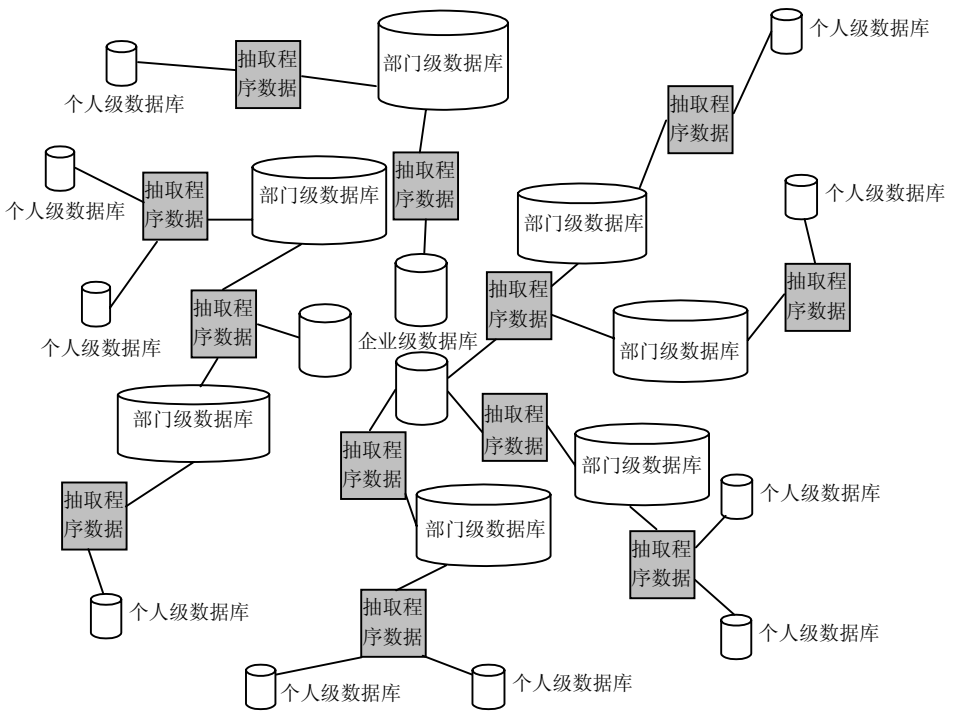

图 1.1　企业中存在的“蜘蛛网”现象

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，因而错综复杂的抽取与访问将产生很多的问题，主要有以下几个方面。

1. 数据分析的结果缺乏可靠性

图 1.2 中展示了某企业的市场部和计划部对项目 I 是否具有市场前景的分析过程和结果。市场部认为“项目 I 的市场前景很好”，而计划部却得到截然相反的结果——“项目 I 没有市场前景”。作为企业的最终决策者，将如何根据这样的结论进行决策呢？

为什么分析同一个企业数据库中的数据，却得到截然相反的结论呢？

首先，两部门可能抽取数据的内容不同。比如，市场部抽取的是项目 I 在大

客户中的应用情况，而计划部抽取的是项目 I 在普通客户中的应用情况。

其次，可能两部门抽取数据的时间不同。如市场部在星期日晚上提取分析所需的数据，而计划部在星期三下午就抽取了数据。有任何理由相信对某一天抽取的数据样本进行分析与对另一天抽取的数据样本进行的分析可能相同吗？当然不能！企业内的数据总是在变的。

再次，引用外部信息的不同。分析项目的发展趋势常常需要引入企业外部的信息，比如报刊信息、国家的政策等。市场部门引用的外部信息来源可能与计划部门不同，而外部信息自然是仁者见仁，智者见智，这也可能是导致最终分析结果不同的原因。

最后，分析程序的差异。市场部门使用的分析程序可能与计划部门不同，分析的内容和指标也可能不同。

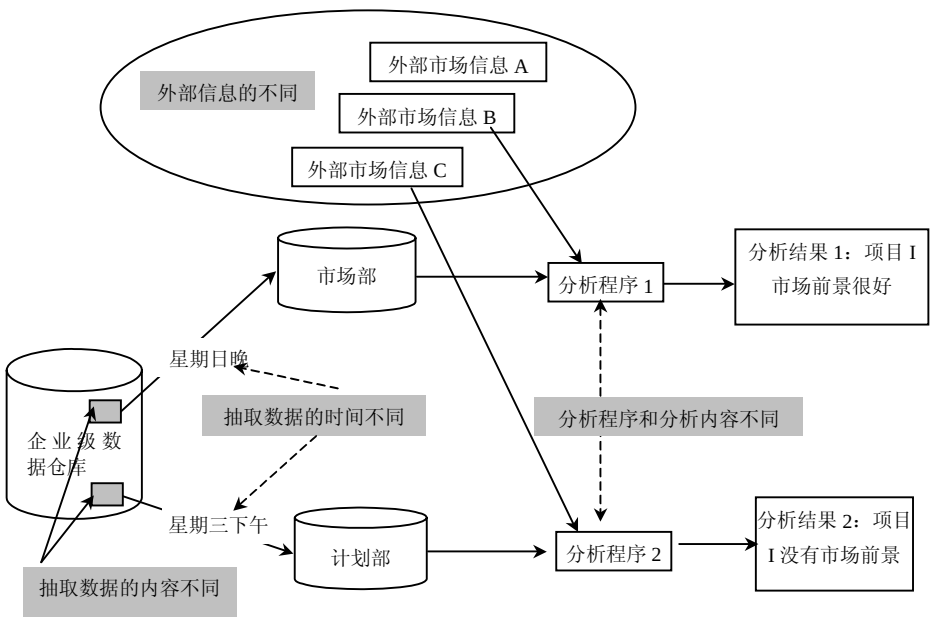

图 1.2　两个分析结果的差异

2. 数据处理的效率很低

数据分析的结果缺乏可靠性并不是蜘蛛网问题中唯一的主要问题。在一个大型企业中，不同级别的数据库可能使用不同类型的数据库系统，对于拥有巨型数据量的企业级数据库可能使用 IBM DB2，而对于部门级和个人级的中小型数据库可能使用 SQL Server。各种数据库的开发工具和开发环境不同，当需要在整个企业范围内查询数据时，数据处理的低效率将是不容忽视的。

如果一个大型企业的决策领导需要一份关于公司整体运营情况的报表，通常

需要动用大量的人力和物力才能达到。首先，定位报表需要的数据，即确定报表涉及的内容分布在哪个数据库的哪个位置，然后调动各个部门的程序员/分析员对应用进行分析、设计和编码。

由于数据分散在各个数据库中，因此需要编写的程序很多。由于企业中使用的数据库类型很多，因此可能需要使用多种技术来实现。可见，面对企业中存在的蜘蛛网现象，为产生一份关于公司整体运营情况的报表，将动用大量的人力、物力和时间才能完成。

如果低效率的过程是一次性的，那么为生成报表花费大量的资源也是可取的。换句话说，如果生成第一份企业报表需要大量资源，生成所有后继报表可以建立在第一份企业报表基础之上，那么不妨为生成第一份报表付出一些代价。但是事实并非如此。

除非事先知道未来的企业报表需求，并且除非这些需求影响到第一张报表的建造，每个新的企业报表总是要花费同前面差不多的代价。

因此，数据处理的低效率是蜘蛛网问题所面临的又一个问题。

3. 难以将数据转化成信息

除了数据处理效率和数据可信度的问题之外，“蜘蛛网”式的结构还难以将数据转化成信息。比如，某电信公司想分析某个大客户今年的情况和过去 3 年有什么不同？大客户的情况可能包括呼叫行为、话费情况、交费情况、咨询问题等。因此要想比较完整地回答这个问题，实际上需要将客户多方面的数据综合成信息。但“蜘蛛网”式的结构中数据缺乏集成性，因此，对综合信息需求的支持确实是不充分的。

另外，每个数据库由于其数据量和业务处理的需求不同，对历史数据的存储时间也不同，因此在蜘蛛网环境中的系统难以提供完整的历史数据。如：记录客户呼叫行为的数据库通常只保留最近 3 个月的呼叫话单，财务数据库可能保留客户今年的交费情况，客户咨询数据库可能只保留客户 2 年内的咨询信息，所以，从这些数据中提取出完整的信息是不可能的。

1.1.2　事务型系统和分析型系统的分离

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，这也正是产生“蜘蛛网”问题的原因之所在。因此，要解决“蜘蛛网”问题，必须将用于事务处理的数据环境和用于分析处理的数据环境分离开。

这样，数据处理被分为事务型处理和分析型处理两大类。事务型处理以传统的数据库为中心进行企业的日常业务处理。比如电信部门的计费数据库用于记录客户的通信消费情况，银行的数据库用于记录客户的账号、密码、存入和支出等一系列业务行为。

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。比如，通过对超市近期数据进行分析可以发现近期畅销的产品，从而为公司的采购部门提供指导信息。又如，对高校大学生就业信息进行分析的结果及结论，可以有效地指导学校制定招生计划和合理设置专业等。

事务型系统的使用人员通常是企业的具体操作人员，处理的数据通常是企业业务的细节信息，其目标是实现企业的业务运营；而分析型系统的使用人员通常是企业的中高层的管理者，或者是从事数据分析的工程师。分析型系统包含的信息往往是企业的宏观信息而非具体的细节，其目的是为企业的决策者提供信息支持。事务型系统和分析型系统的划分如图 1.3 所示。

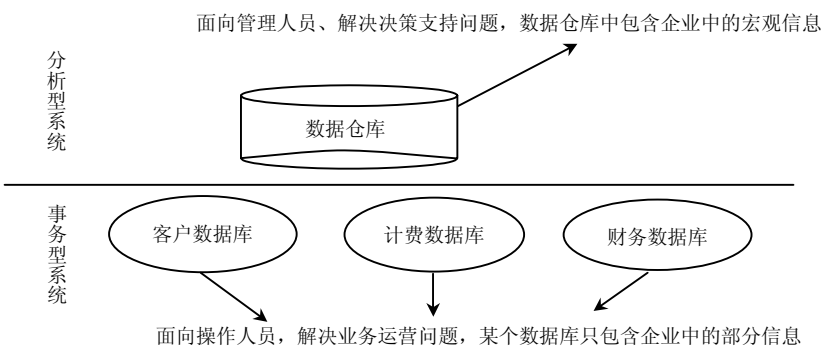

图 1.3　事务型系统和分析型系统的划分

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发展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，也由以数据库为中心的环境发展为以数据库和数据仓库为中心的环境，如图 1.4 所示。

综上所述，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 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。

目前，数据仓库技术正成为企业信息集成和辅助决策应用的关键技术之一。当然，数据仓库的主要驱动力并不是过去的缺点和问题，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。

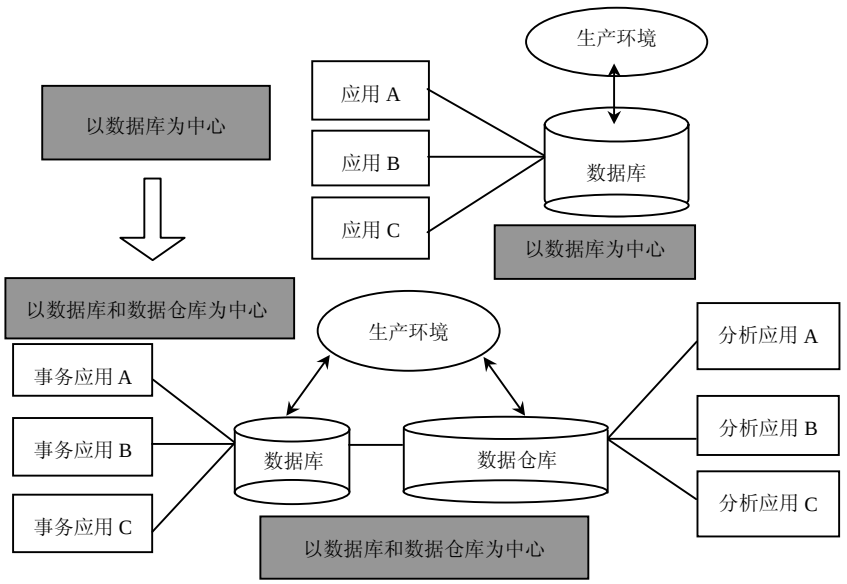

图 1.4　以数据库为中心的环境发展为以数据库和数据仓库为中心的环境的过程示意图

1.2　什么是数据仓库

20 世纪 80 年代中期，“数据仓库”这个名词首次出现在号称“数据仓库之父”W.H.Inmon 的“Building Data Warehouse”一书中，在该书中，W.H.Inmon 把数据仓库定义为“数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合，用于支持管理决策过程”。

对于什么是数据仓库，还有许多不同的定义，如：

“数据仓库是融合方法、技术和工具以在完整的平台上将数据提交给终端用户的一种手段”。

“数据仓库是对分布在企业内部各处的业务数据的整合、加工和分析的过程”。

“数据仓库是一种具有集成性、稳定性和提供决策支持的处理”。

“为查询和分析（不是事务处理）而设计的关系数据库”

在众多的数据仓库定义中，公认的仍然是 W.H.Inmon 的定义。该定义指出了数据仓库面向主题、集成、稳定、随时间变化这 4 个最重要的特征。

1.2.1　面向主题

与传统数据库面向应用进行数据组织的特点相对应，数据仓库中的数据是面

向主题进行组织的。什么是主题呢？首先，主题是一个抽象的概念，是在较高层次上将企业信息系统中的数据综合、归类后进行分析利用的抽象。在逻辑意义上，它是对应企业中某一宏观分析领域所涉及的分析对象，是针对某一决策问题而设置的。面向主题的数据组织方式，就是在较高层次上对分析对象的数据的一个完整、一致的描述，能完整、统一地刻画各个分析对象所涉及的企业的各项数据，以及数据之间的联系。所谓较高层次是相对面向应用的数据组织方式而言的，是指按照主题进行数据组织的方式具有更高的数据抽象级别。

例如在图 1.5 中，我们示例了一个电信企业的情况。该企业基于传统数据库已经建立有计费数据库、财务数据库、客户服务数据库等。其中，计费数据库记录了客户的消费情况，财务数据库记录了客户的缴费情况，客户服务数据库记录了客户的咨询和投诉情况，这些数据库里都有与客户主题相关的数据。如果直接基于传统数据库系统进行决策支持，则需要访问这三个数据库才能获得客户各个侧面的信息，这样将极大的影响系统处理的时间和效率，并且数据之间的不一致性和不同步问题，将极大影响决策的可靠性。

基于以上的原因，数据仓库引入主题之概念，对应某个主题的全部相关数据集中于一个地方，这样决策者可以非常方便地在数据仓库中的一个位置检索包含某个主题的所有数据。

在图 1.5 中，我们选择收益、客户两个主题，则收益主题可以从计费数据库和财务数据库中了解公司各项业务的收入情况；客户主题可以从计费数据库、财务数据库、客户服务数据库中获得客户消费、交费、咨询等全方位的信息。通过这种按主题的数据组织方式，数据仓库极大地方便了数据分析的过程。

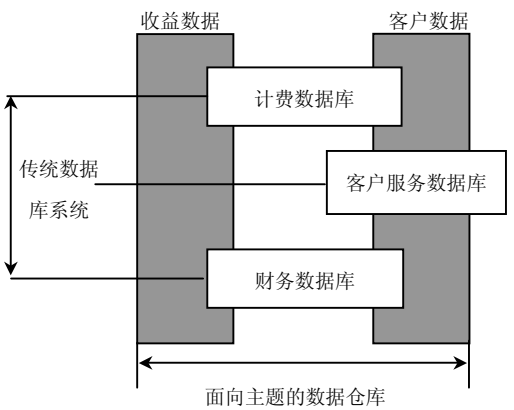

图 1.5　数据仓库面向主题的特性

如图 1.6 所示显示了某电信企业的“客户主题”的数据存储，属于“客户”主题域的数据集合使用相同的公共键码“客户标识”来连接。从图 1.6 中可看到，数据在数据仓库中还是以数据表的形式进行存储，但是，数据的组织方式和建模方法已经同数据库系统有了较大的改变。

1.2.2　集成

数据仓库中存储的数据一般从企业原来已建立的数据库系统中提取出来，但并不是原有数据的简单复制，而是经过统一并综合。这是因为：

1）原有数据库系统记录的是每一项业务处理的流水账，这些数据不适合于

分析处理。在进入数据仓库之前必须经过综合、计算，同时抛弃一些分析处理不需要的数据项，必要时还要增加一些可能涉及的外部数据。

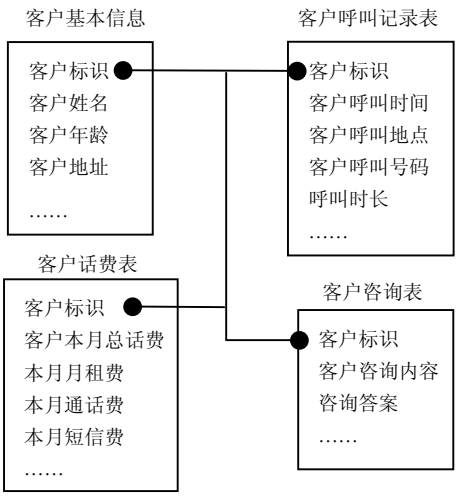

图 1.6　属于相同主题域的数据集合使用相同的公共键码连接

2）数据仓库每一个主题所对应的源数据在源分散数据库中有许多重复或不一致之处，必须将这些数据转换成全局统一的定义，消除不一致和错误之处，以保证数据的质量；显然，对不准确，甚至不正确的数据分析得出的结果将不能用于指导企业做出科学的决策。

事实上，决策支持系统需要集成的数据。全面而正确的数据是有效地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，对源数据的集成是数据仓库建设中最关键，也是最复杂的一步。

1.2.3　稳定性

业务系统一般只需要当前数据，在数据库中一般也只存储短期数据，因此在数据库系统中数据是不稳定的，它记录的是系统中每一个变化的瞬态。

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。没有大量历史数据的支持是难以进行企业的决策分析的，因此DSS 对数据在空间和时间的广度上都有了更高的要求。

在数据仓库中，数据一旦被写入就不再变化了。即数据保存到数据仓库中后，最终用户只能通过分析工具进行查询和分析，而不能修改，即数据仓库的数据对最终用户而言是只读的。由于数据仓库的查询数据量往往很大，并且查询分析的用户多是企业的高层领导，他们是所在领域的专家，但却不一定是计算机专家，所以对数据查询、查询界面的友好和数据的表示提出了更高的要求。

我们在图 1.7 中形象地说明了数据仓库中数据的稳定性，可以看到数据仓库在数据存储方面是分批进行的，定期执行提取过程为数据仓库增加数据，这些数据一旦加入，一般不再从系统中删除。因此我们说数据仓库中数据是稳定的。

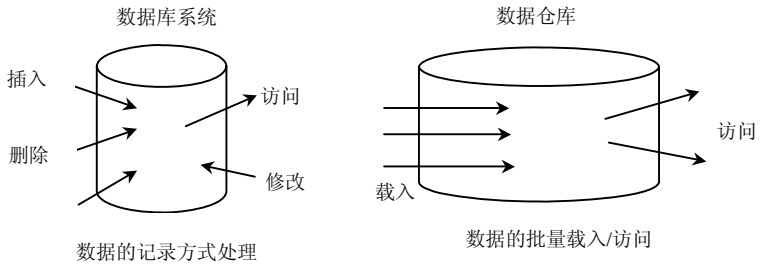

图 1.7　数据仓库的数据稳定性示意

1.2.4　随时间而变化

数据仓库中数据是批量载入的，是稳定的，这使得数据仓库中的数据总是拥有时间维度。从这个角度，数据仓库实际是记录了系统的各个瞬态，并通过将各个瞬态连接起来形成动画，从而在数据分析的时候再现系统运动的全过程。数据批量载入（提取）的周期实际上决定了动画间隔的时间，数据提取的周期短，则动画的速度快，图 1.8 示意了这种特点。

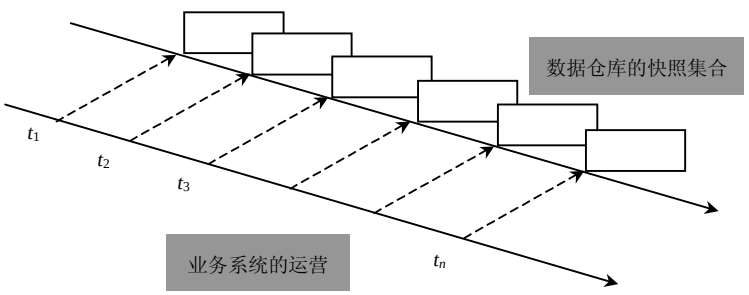

图 1.8　数据仓库数据随时间变化的特点

1.3　数据仓库的体系结构

1.3.1　数据仓库的体系结构

数据仓库的体系结构可以用图 1.9 来表示。由于数据库和数据仓库应用的出发点不同，数据仓库将独立于业务数据库系统，但是数据仓库又同业务数据库系统息息相关。事实上，数据仓库系统 = ETL + 数据存储 + OLAP + 客户端。

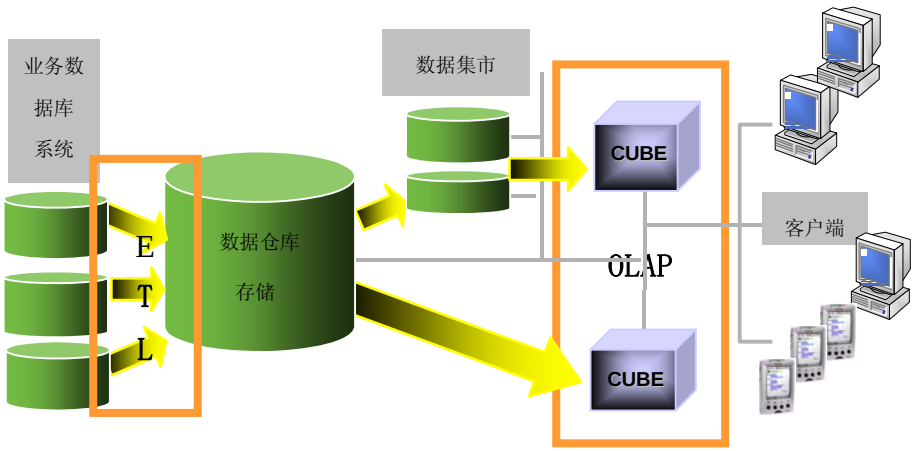

图 1.9　数据仓库的体系结构

1.3.2　数据仓库中的关键名词

下面我们沿着数据的流向详细说明数据在数据仓库中处理的过程，以及一些关键名词。

1. ETL——数据抽取、转换、装载工具

ETL extract/transformation/load 工具就是进行数据的抽取、转换和“净化提炼”处理。所谓数据的“净化提炼”就是对从多个不同业务数据库所抽取的数据，进行数据项名称的统一、位数的统一、编码的统一和形式的统一，消除重复数据。具体来讲，ETL 工具包括：数据抽取（data extract）、数据转换（data transform）、数据清洗（data cleaning）和数据装载（data loading）。

（1）数据抽取（data extract）

从数据仓库的角度来看，并不是业务数据库中的所有数据都是决策支持所必需的。通常，数据仓库按照分析的主题来组织数据，我们只需提取出系统分析必需的那一部分数据。例如，某超市确定以分析客户的购买行为为主题建立数据仓库，则我们只需将同客户购买行为相关的数据提取出来，而超市服务员工的数据就没有必要放进数据仓库。

现有的数据仓库产品几乎都提供各种关系型数据接口，提供提取引擎，从关系型数据中提取数据。

（2）数据清洗（data clean）

由于企业常常为不同的应用对象建立不同的业务数据库，比如一个电信运营公司拥有计费数据库、账务数据库、客服数据库、客户投诉数据库等业务系统，这些业务系统中可能包含重复的信息，比如客服数据库中的部分客户基本信息也在客户投诉数据库中存在，由于不同的数据库可能使用不同数据库公司的产品，

不同的业务系统可能由不同的软件开发商提供，这使得各个业务数据库中的数据可能存在不一致现象。再者，由于数据被冗余地存放在不同的数据库中，如果不同数据库间的数据刷新不是实时的，则可能出现数据不同步的情况。

对于决策支持系统来说，最重要的是决策的准确性，因此确保数据仓库中数据的准确性是极其重要的。从多个业务系统中获取数据时，必须对数据进行必要的清洗，从而得到准确的数据。

所谓“清洗”就是将错误的、不一致的数据在进入数据仓库之前予以更正或删除，以免影响决策支持系统决策的正确性。

（3）数据转换（data transform）

由于业务系统可能使用不同的数据库厂商的产品，比如 IBM DB2，Oracle，Informix，Sybase，NCR Teradata，SQL Server 等，各种数据库产品提供的数据类型可能不同，因此，需要将不同格式的数据转换成统一的数据格式。如时间格式“年/月/日”，“月/日/年”、“日-月-年”的不一致问题等。

（4）数据装载（data load）

数据装载部件负责将数据按照物理数据模型定义的表结构装入数据仓库。这些步骤包括清空数据域、填充空格、有效性检查等。

现在 ETL 工具的功能越来越高级。它具有支持数据的“净化提炼”功能、数据加工功能和自动运行功能（包括处理过程的监控、调度和外部批处理作业的启动等），支持多种数据源，能自动实现数据抽取。

2. 数据仓库存储

数据仓库存储（data repository）就是用于存放数据仓库数据和元数据的存储空间。数据的存储方式主要有 3 种：多维数据库、关系型数据库以及前两种存储方式的结合（在第 6 章 OLAP 的基本概念中将详细讲解）。

数据仓库中存放的数据一部分是从业务系统中提取并经过清洗、转换后的数据，另一部分则是根据 OLAP 分析和数据挖掘的需要，在原始数据的基础上增加的冗余信息，比如，进行大量的预运算，建立多维数据库，以求迅速的展现数据。

数据是对事物的描述，“元数据”就是描述数据的数据，它提供了有关数据的环境，用于构造、维持、管理和使用数据仓库，在数据仓库中尤为重要。

数据仓库中的元数据主要包含两类：管理元数据和用户元数据。

管理元数据是数据仓库设计人员和管理员使用的描述数据仓库的数据信息，用于执行数据仓库开发和管理任务。它包括：

1）数据源信息。

2）转换描述（从业务数据库到数据仓库的映射方法，以及转换数据的算法）。

3）数据仓库中信息的种类、存储位置、存储格式。

4）数据清洗和数据增加的规则。

5）数据映射操作。

6）访问权限、备份历史、存档历史、信息传输历史、数据获取历史、数据访问等。

用户元数据则是帮助用户查询信息、理解结果及了解数据仓库中的数据和组织，它包括：

1）主题区和信息对象类型，包括查询、报表、图像、音频、视频等。

2）Internet 主页。

3）支持数据仓库的其他信息，例如对于信息传输系统包括预约信息、调度信息、传送目标的详细描述、商业查询对象等。

通常，数据仓库将建立专用的元数据库来存放和管理元数据。

3. 数据集市（data market）

数据仓库中存放的是整个企业的信息，并且数据是按照不同主题来组织的。比如市场发展规律的分析主题主要由市场部门的人员使用，我们可以在逻辑上或者物理上将这部分数据分离出来，当市场部门人员需要信息时，不需要到数据仓库的巨量数据中检索，而只需在相应的部门数据上进行分析，因此从效率和处理速度的角度出发，这种划分是合算的。

我们把这种面向企业中的某个部门（主题）而在逻辑上或物理上划分出来的数据仓库中的数据子集称为数据集市。换句话说，数据集市包含了用于特殊目的数据仓库的数据部分。

数据仓库面向整个企业，而数据集市则是面向企业中的某个部门。典型示例是销售部门、库存和发货部门、财务部门和高级管理部门等的数据集市。数据仓库中存放了企业的整体信息，而数据集市只存放了某个主题需要的信息，其目的是减少数据处理量，使信息的利用更快捷、灵活。

4. OLAP

OLAP 是使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。OLAP 的目标是满足决策支持或者满足在多维环境下特定的查询和报表需求，它的技术核心是“维”这个概念。

维是人们观察数据的特定角度。例如，一个企业在考虑产品的销售情况时，通常从时间、地区和产品的不同角度来深入观察产品的销售情况。这里的时间、地区和产品就是维。而这些维的不同组合和所考察的度量指标构成的多维数组则是 OLAP 分析的基础，可形式化表示为（维 1，维 2，……，维 n，度量指标），如地区、时间、产品、销售额。

“维”一般包含着层次关系，这种层次关系有时会相当复杂。通过把一个实

体的多项重要的属性定义为多个维（dimension），使用户能对不同维上的数据进行比较。因此 OLAP 也可以说是多维数据分析工具的集合。

多维分析是指对以多维形式组织起来的数据采取切片（slice）、切块（dice）、钻取（drill-down 和 roll-up）、旋转（pivot）等各种分析动作，以求剖析数据，使用户能从多个角度、多侧面地观察数据库中的数据，从而深入理解包含在数据中的信息。

1）切片和切块是在一部分维上选定值后，关心度量数据在剩余维上的分布。如果剩余的维只有 2 个，则是切片；如果有 3 个，则是切块。

2）钻取是改变维的层次，变换分析的粒度。它包括向上探取（roll up）和向下钻取（drill down）。roll up 是在某一维上将低层次的细节数据概括到高层次的汇总数据，或者减少维数；而 drill down 则相反，它从汇总数据深入到细节数据进行观察或增加新维。

3）旋转是变换维的方向，即在表格中重新安排维的放置（例如行列互换）。

根据数据的组织方式的不同，目前常见的 OLAP 主要有基于多维数据库的 MOLAP 及基于关系数据库的 ROLAP 两种。MOLAP 是以多维的方式组织和存储数据，ROLAP 则利用现有的关系数据库技术来模拟多维数据。在数据仓库应用中，OLAP 应用一般是数据仓库应用的前端工具，同时 OLAP 工具还可以同数据挖掘工具、统计分析工具配合使用，增强决策分析功能。

1.4　数据仓库的数据组织

在对数据仓库的概念、特点及体系结构进行分析后，我们来学习数据仓库的数据组织结构和组织方式。

1.4.1　数据仓库的数据组织结构

一个典型的数据仓库的数据组织如图 1.10 所示。

在数据仓库中，数据一般分成 4 个级别：高度综合级、轻度综合级、当前细节级和早期细节级。

源数据（早期细节级数据）经过综合后，首先进入当前细节级，然后根据应用的需求，通过预运算将数据聚合成轻度综合和高度综合级。由此可见，数据仓库中存储着不同综合级别的数据，一般称之为“数据粒度”。粒度越大，表示细节程度越低，综合程度越高。比如，在电信公司中的电话呼叫数据中记录了每个用户的每次呼叫。进行 OLAP 分析时，常常需要不同层次的数据粒度，因此可以通过预运算将数据综合成每个用户每“天”的通话次数，还可以进一步聚合成每个用户每“月”的通话次数（图 1.10 中右列所示）。

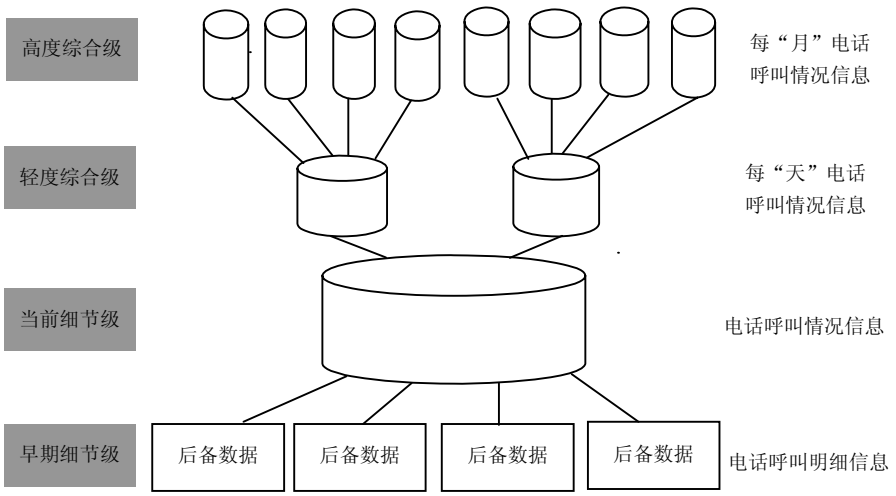

图 1.10　数据仓库的数据组织结构图

在数据仓库中，轻度和高度综合级别的数据一般是由细节数据聚合而来，但需要说明的是轻度和高度是相对的概念，而没有绝对的界限，并且在数据仓库中数据的综合程度常常有很多的级别。

随着时间的推移，系统中的一些细节数据已经“老化”了，很少会被用户使用，此时为了节省系统的存储空间，可以将这些老化的细节数据导出到备份设备上。实际应用中，综合数据也可能被导出系统。比如企业的管理者认为企业的决策只同企业近 15 年来的运营数据有关，则 15 年之前的综合数据也可以导出。对于高度综合的数据，由于其数据量已经很少，所以一般可以不考虑它们的导出问题。

在数据仓库中，处理提取和综合后的数据还包含非常重要的元数据，它描述的是提取和综合后的数据的组织方式，我们在数据仓库的体系结构中已经介绍了元数据。

1.4.2　数据粒度与数据分割

1. 数据粒度

数据粒度是数据仓库中极其重要的概念。粒度可以分为两种形式，一种是对数据仓库中的数据的综合程度高低的一个度量，它既影响数据仓库中数据量的多少，也影响数据仓库中数据的用途。在数据仓库中，多重的数据粒度是必不可少的。由于数据仓库最主要的目的是反映企业整体信息和 DSS 分析，因而绝大多数查询都是基于一定程度的综合数据之上，只有极少数查询涉及到细节。所以，应

该将大粒度数据存储于快速设备（如磁盘）上，而将细节数据定期导出到低速设备（如磁带）上。

粒度的第二种形式是指抽样率，即以一定的抽样率对数据仓库中的数据进行抽样后得到一个样本数据库。这种样本数据库中的粒度不是根据综合程度的不同来划分的，而是由抽样率的高低来划分，抽样粒度不同的样本数据库可以具有相同的数据综合程度。

在数据仓库环境中粒度之所以是一个极其重要的概念，是因为它深深地影响存放在数据仓库中的数据量的大小，同时影响数据仓库所能回答的查询类型，在数据仓库中数据量大小与查询的详细程度之间要做出权衡。

2. 数据的分割

数据的分割是数据仓库中的又一重要概念。所谓数据分割是指将数据分散到各自的物理单元中以便能够独立处理，提高数据处理的效率。数据分割没有固定的标准，分割的方法和粒度应当根据实际情况来确定。分割方法常常可以选择时间、地点、业务领域来划分，也可以是其组合。按照时间进行分割符合数据仓库数据随时间变化的特点，并且分割后数据分布比较均匀，所以是最常用的分割方法。

不过需注意的是：在数据仓库中，围绕分割问题的关键并不是该不该对数据进行分割，而是如何分割。这也是为什么有人说，如果粒度和分割都做得很好的话，则数据仓库设计和实现的几乎所有其他问题都容易解决。但是，假如粒度处理不当，并且分割也没有认真地设计与实现，这将使其他方面的设计难以真正实现。

有关粒度和分割的更为详细的问题将在数据仓库模型设计一章中介绍。

1.4.3　数据仓库的数据组织形式

在数据仓库发展过程中，出现了多种不同的数据组织形式：

1. 简单堆积文件

简单堆积文件就是将每天由业务数据库提取并处理后的数据逐天存储起来，如图 1.11 所示。还有一种形式被称为简单直接文件，它同简单堆积文件非常类似，只是按照一定的时间间隔对业务数据库进行快照并存储，但是时间间隔不一定是每天。

2. 定期综合文件

在定期综合文件这种方式中，数据存储单位被分成日、周、月、季、年等多个级别。首先数据被逐一添加到每天的数据集合中，当一个星期过去了，每天数

据被综合成周数据，依此类推，周数据被综合成月数据……

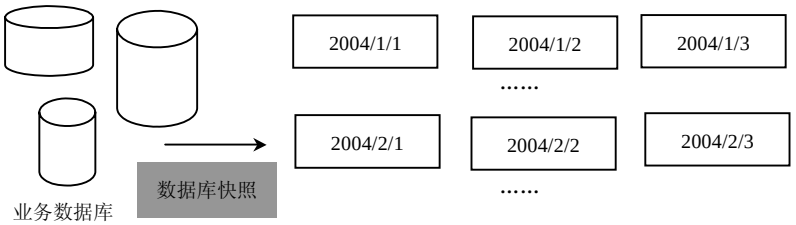

图 1.11　简单堆积数据

定期综合文件的组织方式使得数据量比简单堆积文件方式大大减少，但是由于数据被进行了综合，使得数据的细节在综合中丢失。因此，定期综合文件的形式是牺牲数据的细节信息换取数据量级的减少。

3. 连续文件

定期综合文件其数据量级小时丢失了数据细节，简单堆积文件保留细节但数据量级又很大，是否可以综合两者形式的优点呢？答案是肯定的。

在简单堆积文件中，每天的数据表中有许多雷同的信息，如图 1.12 所示的某商场 2004 年 1 月和 2004 年 2 月的两张采购表，其中“钢笔”和“水杯”在两个表中都出现了。“上海”产的“钢笔”既在 2004/1 购买，又在 2004/2 购买。如果能够用一条记录将两条记录所包含的信息记录下来，则既能保留细节信息,又能大大减少数据量。

2004/1 采购表

商品编号	商品名	商品产地
1	钢笔	上海
2	水杯	昆明
4	帽子	北京

2004/2 采购表

商品编号	商品名	商品产地
1	钢笔	上海
3	毛巾	广州
4	帽子	成都

图 1.12　某商场的两张采购表

图 1.13 中显示了对两张表使用连续文件的形式进行存储的结果。对于两张表中相同的项“钢笔”，只需在时间列上说明购买时间是“2004/1～2004/2”，对于两表不同的表项分别记录。

随着时间的推移，如果又有新的数据表加入，则可以使用连续文件和新的数据表进行类似的处理，以达到“两全其美”的目的。但是我们应当指出：连续文件增加的“时间”列也会为查询带来一定的不便。一个系统某些性能的提高，总