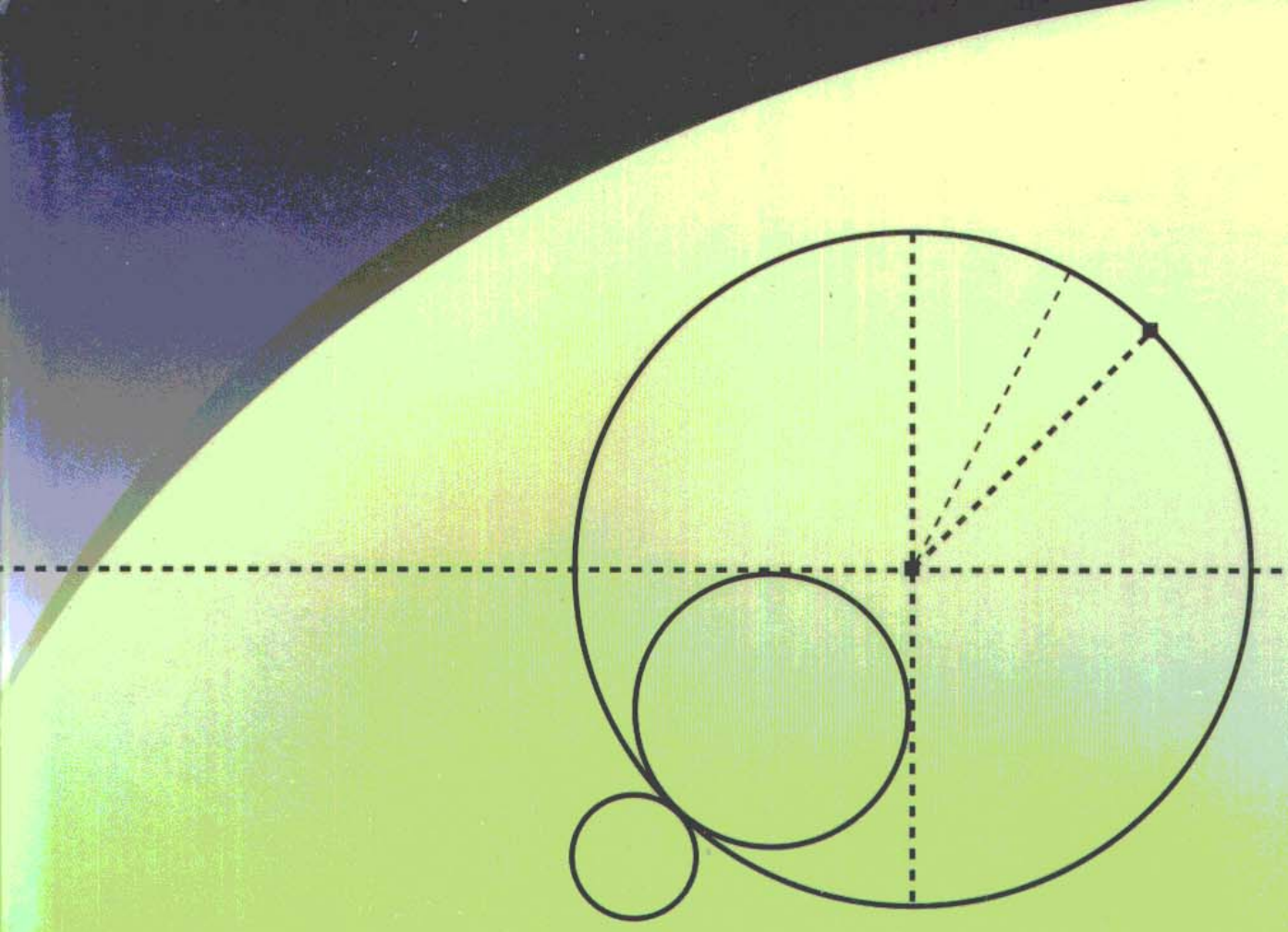

DECISION SUPPORT IN THE DATA WAREHOUSE

数据仓库中的决策支持

〔美〕 Paul Gray Hugh J.Watson

陈朔鹰 周志逵 赵丰年 等译

北京理工大学出版社

数据仓库中的决策支持

[美]Paul Gray Hugh J. Watson 著

陈朔鹰 周志逵 赵丰年 等译

北京理工大学出版社

内 容 简 介

本书由综述、创建数据仓库、数据仓库环境中的决策支持和用于决策支持的数据仓库产品四部分组成。第一部分简述了数据仓库和决策支持涉及的主要问题；第二部分解释了数据仓库的技术术语，并提供了建立和维护用于决策支持的数据仓库的框架；第三部分描述了使用数据仓库的决策支持应用程序的建立过程；第四部分介绍了数据仓库领域中的一些主要厂商及其典型产品。本书可作为从事数据仓库决策支持开发研究人员的参考书，也可作为大学本科生或研究生的选修课教材。

图书在版编目（CIP）数据

数据仓库中的决策支持/（美）格瑞（Gray,P.）等著；陈朔鹰等译.—北京：北京理工大学出版社，2001.1

ISBN 7-81045-766-7

Ⅰ. 数… Ⅱ. ①格… ②陈… Ⅲ. 数据库系统－决策支持系统 Ⅳ. TP311.131

中国版本图书馆 CIP 数据核字（2000）第 78222 号

北京市版权局著作权合同登记号图字：01-1999-0963 号

责任印制：毋长新 责任校对：陈玉梅

北京理工大学出版社出版发行
（北京市海淀区中关村南大街 5 号）
邮政编码 100081 电话（010）68912824
各地新华书店经销
北京房山先锋印刷厂印刷
*
787 毫米×1092 毫米 16 开本 13.5 印张 300 千字
2001 年 1 月第 1 版 2001 年 1 月第 1 次印刷
印数：1—4000 册 定价：20.00 元

前 言

数据仓库是本世纪90年代最热门的发展领域之一。由于需要向决策者提供干净、一致、相关的数据，越来越多的组织正在建立数据仓库。这些昂贵的、耗时的工作需要花费数百万美元和许多人年的努力。这些组织也得到了很好的回报，获得了许多收益，例如高质量的决策，把IS(Information Systems department 信息系统部门)的职员从频繁的数据请求中、以及对客户购买行为的深刻洞察中解脱出来。

建立数据仓库是一项具有挑战性的工作，因为它涉及到许多方面。例如，必须获得项目投资。这些必须的资金是来自 IS，还是业务部门，或是两者一起出？必须解决定义数据仓库数据需求等开发问题。还有许多技术决策，如使用什么样的服务器，是否需要并行处理等。而且，还存在行政方面的问题，例如：业务部门对失去某些数据控制权的态度。

出于对数据仓库的兴趣，人们每年都要举行多次会议并撰写大量的书籍和文章。这些书籍和文章同时也提供了有关数据仓库的丰富信息。然而，问题在于：要获得关于数据仓库全貌所需的信息无法只从一个数据来源中获取，而解决这一问题也正是我们编写此书的目的。从我们的观点来看，重要的是理解如何创建和维护数据仓库，如何开发使用数据仓库、数据仓库产品和工具、公司的数据仓库经验的范例。

本书内容

为满足以上目标，本书分为四个部分：

1. 综述
2. 创建数据仓库
3. 数据仓库环境中的决策支持
4. 用于决策支持的数据仓库产品

第一部分，由第一章——数据仓库：新的决策支持环境组成。它简要描述了数据仓库和决策支持涉及的主要问题。

第二部分解释了数据仓库的技术术语，并提供了建立和维护用于决策支持的数据仓库的框架。这部分说明了数据仓库的建立和操作，并展示了所需使用的正确的数据仓库工具。它集中说明了诸如如何提取、清洗、处理和加载原数据，如何维持关系型数据仓库和多维度数据仓库和数据集等细节。第二部分由第二章数据仓库基础和第三章建立和维护数据仓库组成。

第三部分描述了使用数据仓库的决策支持应用程序的建立过程。它描述了用于数据仓库的决策支持工具的使用，其中包括从查询、联机分析处理的响应到数据挖掘和数据库营销。它也描述了作为决策支持集成交互界面的万维网技术的作用。它逐渐显示了如何在数据仓库环境下建立决策支持软件。最后，这部分包含了实用信息：会出现什么错误，对目前现实情况的调查、可预期的未来发展等。这部分的三章分别是决策支持工具（第四章）、创建决策支持应用程序（第五章）、创建数据仓库的最后思考（第六章）。

第四部分由七章组成，介绍了数据仓库领域中的一些主要厂商及其典型产品。这些销售商是：

- NCR 公司
- Oracle 公司
- Red Brick Systems 公司
- Platinum technology 公司
- Brio Technology 公司
- Comshare 公司
- Seagate Software 公司

每个销售商都描述了他们的数据仓库、决策支持产品和服务，举出了成功使用他们产品的实例。通过这些章节，读者能够比较这些厂商的产品，并决定哪一个最适合自己的需求。这些章节是经过仔细编辑的，以便能够对性能和结果进行真实的描述。章节内容的责任由资料的提供者各自负责。

总之，本书是第一本从决策支持观点出发的有关数据仓库内容的书。另外，本书的独特之处在于：提供了描述主要厂商的章节，这使读者能够同时比较各种方案。

数据仓库的作用

同所有的创新一样，数据仓库正在形成其生命周期。目前它还处于快速成长期，但是当越来越多的公司成功地开发了数据仓库时，它将很快进入成熟期。然而，对于数据仓库的基本需要没有改变，因为决策支持总是需要数据。基于目前的各种趋势（例如，竞争的增加、产品生命周期的缩短、同客户维持更紧密关系的需要），可以认为将来数据仓库会对企业更为重要。为了满足这些需求，数据仓库将变得更大、更复杂，并且包含实时访问数据。同时成为主角的还有技术的进步，例如：多媒体数据库、面向对象技术、维护和使用大型数据库的能力。我们认为数据仓库将继续成为 IS 领域中的重要部分。

Paul Gray
Hugh J. Watson

数据仓库：新的决策支持环境

1.1 简介

计算机用于数据处理并为决策提供信息。早在本世纪 70 年代，人们就认识到了：为进行决策支持将特殊的经过预处理的数据放在不同的平台上具有明显的益处，这种方法使用户可以轻松访问所需要的数据，同时改善了系统的响应时间，并增强了数据的完整性和安全性。决策支持系统是第一个使用此方法的应用系统，终端用户计算的出现使其他许多应用系统（如行政管理信息系统）得益于经过特殊准备和储备的数据。

在本世纪 90 年代，许多组织发展了数据仓库技术以便向最终用户提供决策支持数据。然而，在早期企图为这种需求服务的各种努力之间却有很大的不同。一种是使用特殊的软件帮助从数据仓库（或数据市场）中提取、清理、清除以及加载数据。依靠这种应用软件，许多服务器软件可以用于作数据存贮：多维数据库、Lotus Notes 服务器和基于网络的服务器都加入了关系数据库的行列。增强的数据访问工具使终端用户更容易存取、分析、显示信息，例如，可以不必编写 SQL 查询。

随着技术的进步，商业压力和机遇促使经营者的兴趣转移到数据仓库上。今天如此混乱而高速的环境使决策的周期变得更短，因此需要依靠信息技术支持决策。许多组织越来越将注意力集中到客户身上，他们已经认识到，包含大量客户信息的数据库能提供产生商业优势的信息。另外，在许多组织中，用户及时获得决策支持信息的惟一办法是自己进行分析，因为组织规模的减小降低了雇佣信息系统人员为他们工作的可能性。

1.2 数据仓库

数据仓库的作用

数据仓库的基本作用是为决策支持提供数据。在一般情况下，使用数据仓库的应用软件（如行政管理信息系统）已经存在了一段时间，不过数据仓库通过改善和扩大数据的范围、准确度和易访问性，为这些软件增加一些新的活力。它们也产生和方便了其它的应用程序，

例如：基于关系数据库或多维数据库技术和数据挖掘的联机分析处理 OLAP（On-Line Analytic Processing）。

DSS 和 EIS 传统

决策支持系统（DSS）这个概念是本世纪 70 年代在帮助管理者进行决策的信息系统中提出的。这些系统中使用模型和数据解决管理中存在的问题，这些问题的复杂程度可以从使用简单的电子表格进行预算到使用整体程序来进行优化选择。最初的想法是管理者能够自己创造和操作这个系统，后来证明这个设想是错误的，因为大多数管理者不具备所需要的经验、技巧和时间。因而，从本世纪 80 年代开始，许多组织和厂商提供了简化的信息系统，称为行政管理信息系统（EIS）。

EIS 中基本的假设是管理者想得到关于他们公司和外界的常规信息，这些信息包括在组织的生产和处理的历史中，以及对将来的预测。办法是使管理者们能够借助于他的 EIS 立即知道组织中正在发生的事件。这些 EIS 包括公司的财务信息、生产历史、目前的状况、计划、人员、以及诸如竞争者、电子邮件等外部信息。最初的 EIS 没有 DSS 的分析能力。正如 Rocket 和 Delong 所说：老资格的管理者使用 EIS 来发现问题；职员使用 DSS 学习并提供代替品。

尽管 EIS 和 DSS 十分有用，但是它们经常缺少一个强大的数据库部件。总会出现这样的情况：为一个目的收集的信息不能直接用于其它目的。特别是，大多数组织的信息收集主要是维护现在的单一事务和客户的信息。然而，管理决策要求考虑过去和将来的情况，而不仅仅是现在的情况。结果，大多数 DSS 和 EIS 的开发者们不得不创建他们自己的数据库，而这种创建数据库的工作却不是他们所长，同时要进行这项工作，也是费力费时的。

DSS 和 EIS 软件厂商迅速地捕捉到由数据仓库技术带来的机会。DSS 和 EIS 变成了“旧新闻”，用它们来激发兴趣和销售越来越困难。进而，通用程序语言，比如：Visual Basic 和 Power Builder，开始为那些想开发应用程序（如与 EIS 相关的应用软件）的公司提供低成本的、有吸引力、容易使用的解决办法。通过内部发展、包括在产品中增加新的功能、形成合作联盟（例如：与多维数据库提供商合作）等办法重新调整产品供应，厂商可以提供“新的新闻”，以提供他们所强调的 OLAP 和数据挖掘等性能。

数据库传统

数据库的开发商一直明白：他们的软件需要具备事务处理和分析处理的能力。然而，他们的主要发展集中于不断扩大的事务数据库，而以信息数据库为代价。即使操作数据和分析数据分离，随着不同的需求和不同的用户团体，这个过程也会发生。

20 世纪 80 年代是发展迅速的时代，在此期间，前十年引进的关系模型发展到了顶峰。用于在线事务处理的关系数据库市场是巨大的，而客户/服务器计算机体系的到来增加了这种需求，这样就没有必要转移到其他处理。

正如大部分新技术方案一样，一旦每个人拥有这项技术，数据库就变成稳定的替换者，而不是成长的市场。此时，数据库开发商开始注意使用他们所获得知识的新方向。他们偶然发现不但应该存储现在的数据，而且应该存储过去的数据，他们也知道在系统中的数据并不总是准确的和一致的。可能更重要的是，他们发现了当复杂的查询应用于现存的数据库时，所需的反应时间太长，并且阻碍了关系系统以至于它们不能执行想要的处理功能。

了解了这些不同点之后，就创建出专门用于分析的新数据库。由于这些数据库存储大量的数据，远远超出以往普通数据库的容量，而且数据保存和使用的时间较长，因此这些数据库被称作数据仓库。

数据仓库用户

一个数据仓库服务于两类用户：

- 直接的用户
- 应用程序所有者

直接用户是具有影响力的用户，如市场分析家、财政计划者，他们需要得到数据以进行他们的工作。通过使用简单的软件，如 Excel 或 Access，他们在数据仓库中直接存取数据。一旦他们有了数据，他们将进行进一步处理。这些用户需要较好地知道数据仓库中什么数据是有用的、数据怎样存储以及怎样存取它。

一个应用程序所有者的责任是为大量的用户创建应用程序。例如，在许多组织中，许多通过 EIS 提供的数据是来自于数据仓库的。对于 EIS 用户，数据来源是透明的，无论它是一个数据仓库，还是一组不同的文件。除非数据有问题或响应时间太慢，他们不会注意，也不会对如何准备和存储数据担忧；对于数据仓库，他们不需要了解太多。然而，应用程序所有者必须清楚有关数据仓库的大量信息，因为它可能是应用程序中数据的主要来源。

什么是数据仓库？

数据仓库通常是一个专用的数据库系统，它独立于这个组织中的联机事务处理系统（OLTP）。在以下方面它不同于操作型系统：

- 它比事务系统跨越更长的时间
- 它包括多个被处理过的数据库（对数据库进行处理是为了使数据仓库的数据具有统一定义）
- 它被优化以便回答来自直接用户和应用软件的查询

数据仓库至少有三种类型：

- 提供数据并支持整个企业的传统数据仓库
- 数据市场，它是微型的数据仓库，用来支持特殊的商业单位或部门
- 操作型数据仓库，它把数据仓库技术应用到事务系统中

定义

定义数据仓库有许多方法。Inmon（1992 年）把数据仓库定义为：

- 面向对象的
- 集成的
- 不同时间的
- 非易失性的

以支持管理决策处理的数据集成。

Inmon 被许多人誉为数据仓库之父。这个定义非常有用，因为它给出了数据仓库可测量的属性。在后面，我们将详细讨论每一个属性。

Inmon 的数据仓库定义作了两个隐式的假设：

1. 数据仓库与操作型系统在物理上是分开的。
2. 数据仓库同时包含用于管理的聚集数据和事务数据（原子数据），而与用于在线事务处理的数据库相分离。

其他一些采用 DSS 观点的定义也是有趣的。

Imhoff （1995 年）定义数据仓库为：

- 为支持 DSS 功能而设计的一组集成的、面向对象的数据库，每一个数据单元与一些时间段相关。
- 经过优化并用于决策支持的一组数据库。

Red Brick 系统的奠基人 Ralph Kimball 将数据仓库称为人们能够存取数据的地方（1996 年）。

Oracle 公司的 Corey 和 Abbey 在书上将数据仓库定义为：

- 直接从操作型系统和一些外部数据源获得的公司信息的集合，它的特定目的是支持商业决策，而不是商业运作。

1995 年 Babcock 在《计算机世界》中叙述到：

- 数据仓库是以简化形式从操作型系统汇总或聚集的数据集成，面向最终用户的数据存取和报告工具使用户可以获得决策支持所需的数据。

Babcock 继续补充说数据仓库是：

- 信息的，而非操作型的
- 分析和决策是面向支持的，而非面向事务处理的
- 通常是客户/服务器模式，而非传统的基于主机模式

数据仓库的特性

表 1.1 汇总了数据仓库的特性。这些特性在以后的章节中会详细讨论。

表 1.1 数据仓库的特性

特性	描述
面向对象的	数据是通过用户如何查阅进行组织的
集成的	清除术语和信息冲突中的矛盾；也就是，数据清理
非易失的	只读数据；数据不能被用户更新
时间级数	数据是时间级数，而不是现行状态
汇总的	适当的时候，操作数据被聚集成可以决策的状态
较大的	保持时间级数隐含着维持了更多的数据
非正规化的	数据可能是冗余的
元数据	用于用户和数据仓库管理员的数据
输入	操作数据加上所需要的外部数据

面向对象

在一个数据仓库中，数据是围绕着企业主要对象（如销售）组织的，而不是个别事务。也就是，这种组织是通过企业涉及的主题领域，而不是以软件应用为基础。产生这种差异的原因是应用软件围绕过程和功能设计，每一个过程和功能有它需要的特定数据，但是许多数据元素只涉及到功能的一部分。这些操作数据要求与应用程序的立即需求相关，并且以现在

的商业制度为基础。另一方面，数据仓库包含面向决策支持的数据，这些数据跨越一段时间，并允许更复杂的关系。

数据集成

在数据仓库中，信息应该是：

- 干净的
- 确定的
- 适当聚集的

通过整理，相同信息仅能用一种方法查阅。不幸的是，原有系统中许多方法查阅出相同的信息。这样，在两个系统中，可以用男女区分性别，在第三个系统中用 0、1 表示。类似的现象有，一些系统使用两个阿拉伯数字表示年份（产生 2000 年问题），而其它系统使用 4 位数字。

当数据进入数据仓库时，它被聚集起来，并用一种方法来查阅，对可衡量的属性而言，它含有相同的格式和单位。这样，在数据仓库中，数据符合一种单一的、全球通用格式，即使来源是不同的。

同样，数据必须是正确的。差错将破坏数据库，除非有外部干预，否则错误就会永远停留在那里。类似地，有些信息可以省去。需要特定步骤确保数据仓库中的数据经过校验。在数据仓库中的一些数据被聚集起来，必须重视并保证数据聚集是正确的。如果这些操作被正确执行，用户可以集中精力使用数据，而不必关注其可信度和稳定性。

非易失性环境

在一个操作环境中，当进行更新（插入、删除、改变）等操作时，它们以记录为基础有规律地执行。在这方面，数据仓库不能更新。但是，在用户存取之后，数据被载入数据仓库，这种方法会使技术环境更简单。事实上，操作环境和数据仓库之间没有什么关系，除非操作数据被复制。

当新的数据载入数据仓库时，它被过滤和转化，仅存储决策支持需要的数据。另外，使用一些计算生成操作型数据中没有发现的汇总数据。例如，数据仓库可能包括每周的数据，这些数据从每天的销售额中累计生成，是操作型数据中不存在的汇总数据。这种汇总改善了数据仓库用户使用时的响应时间。

时间级数

在一个操作型环境中，决策是在线做出的（例如：通过电话查询得到客户的信用情况），因此，数据在存取时必须是准确的，这对于一个数据仓库来说不成问题。数据在某些时候是准确的，但不必立刻准确。特别是，当数据被装入数据仓库时，它必须是完全准确的。

通常，数据仓库所包含的数据有 5 年到 10 年的年限，而另一方面，操作型数据维持 60 天到 90 天。

在数据仓库中，组织起来的数据要一直包含相关的时间单位（例如：天或周），因此，正确记录的数据仓库数据是不能被用户更新的。

数据仓库结构

数据仓库包含5类数据：

- 当前的详细数据
- 旧的详细数据
- 轻度汇总的数据
- 高度汇总的数据
- 元数据

这些数据没有必要存储在同一种媒质中。然而，利用软件可以存取各种媒质中的数据，需要注意的是：数据仓库设计中决定性的关键因素是所包括的细节数量，也就是数据的“尺寸”。在组织的事务处理系统中保存的原始数据通常是最详细的。

当前的详细数据

当前的详细数据反映最近的事件。如果数据以最低标准的尺寸存储，数据可能变得非常多。通常要求这种信息能够快速存取，因此，一般保存在磁盘上。它通常是经过简单清理和安装的当前事务数据的复制品。然而，并不是事务系统中的所有领域都能移到数据仓库中。值得注意的是，尽管认为是当前数据，但是，这个数据只有在从事务系统中移出时才是最新的。

许多决策支持问题涉及到要直接使用从事务详细记录中获得的数据，例如，今天我们卖了多少东西？各周硬件销售的走向如何？收入怎么样？在回答一些合法查询时，在数据仓库中之所以要复制详细数据是因为，在相当长的一段时间内，一些合法的查询有可能要涉及到细节，此时对细节的查询就不会干扰到事务系统或导致事务系统被终止。

旧的详细数据

当详细数据到达某一期限，数据从磁盘移到海量存储介质时，大多数数据仓库有规则描述。尽管详细形式的数据是可以恢复的，但是因为介质必须伴随一个较慢的工作过程，所以存取时间有些长。然而，这个信息是相同的。

轻度汇总的数据

许多决策支持应用软件以汇总事务数据为基础。经验表明，通过预测请求的标准次数并据此汇总数据，将缩短数据仓库的响应时间并提高使用效率。

从设计者的角度看，存在两个决策要求：

1. 选择要汇总的属性，
2. 选择要汇总的时间单位。

这些问题涉及到权衡问题，如果不需要重复执行计算，就必须有更多的存储空间。很明显，经常被查询的属性或属性集合应该被汇总，然而那些几乎不被查询的属性或属性集合不应该被汇总。一旦属性被选中，下一个问题就是如何经常汇总各个特殊属性。例如，我们每天汇总销售额吗？每周汇总销售额吗？每月汇总销售额吗？保存结果吗？为了汇总，必须确定选中属性的时间单位，应该按用户要求加速回答。

高度汇总的数据

一些信息，特别是高级别管理者要求的信息，形式上必须精简并且容易存取，该信息通常包括反复咨询的信息，这种信息不局限于保持汇总的事务数据。它具有长时间保持汇总数据能力，以便建立对趋势的认识。通过存储高度汇总的数据，信息的响应时间也将缩短。

元数据

元数据被定义为关于数据的数据。它保存关于数据仓库的信息而不是数据仓库要提供的信息。对于数据仓库的工作人员和用户来说，元数据作为要素是不可见的，每组人员要求不同的信息。对于数据仓库的工作人员来说，元数据包括：

- 数据仓库中目录的内容。这个目录指出数据仓库的地点，通过查询发现正确的信息时将使用该目录作为索引。
- 将操作数据映射为数据仓库形式时的指导。当数据载入数据仓库时，数据必须是标准格式，并且遵循数据仓库的规则。这就是说数据必须是干净的。指导介绍了如何转化每一套特殊数据以保证其格式的正确。例如，如果美国电话电报公司在一个数据集中存为 AT&T，在另一个数据集中存为 ATT，要确保数据仓库使用一种格式记录它们。
- 用于汇总的规则。对于数据仓库的用户来说，元数据包括：
  * 用于描述数据的商业术语，
  * 与商业术语对应的用来存储数据的技术术语，
  * 当生成元数据时，数据、规则的来源。

数据形式

事务系统中正常数据的概念不能应用于数据仓库，这在关系数据库中很普遍。事务系统认为消除数据冗余是值得的，以至于一个事务过程的所有数据都在同一个地点。这个观点就是把数据区域组织成一组容易使用且有意义的表格。人们普遍认为存储空间很昂贵，不应该浪费。

数据仓库中的观点是，组织数据以使它是有用的，并且在使用时能够快速重新查到，数据冗余绝对正确。

数据流

如图 1.1 所示，几乎所有数据都从操作环境进入数据仓库。数据被清理并移入数据仓库，数据会继续停留在数据仓库中，除非有下面三个操作之一：

1. 数据被清除，
2. 同其它信息在一起的数据被汇总，
3. 数据被归档。

数据仓库的老化过程处理把当前数据加入到旧的详细数据之中，汇总过程以详细数据为基础。

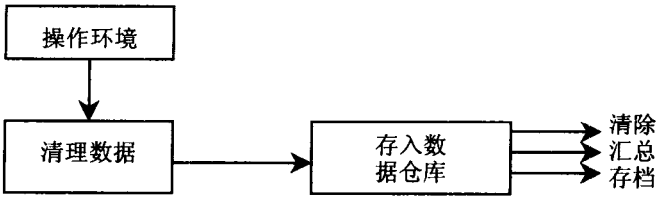

图 1.1 数据流

数据仓库体系结构

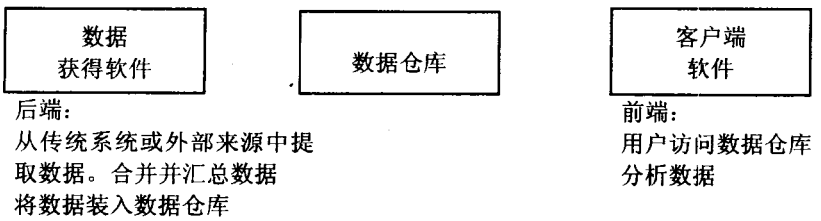

图 1.2 数据仓库的三层体系结构

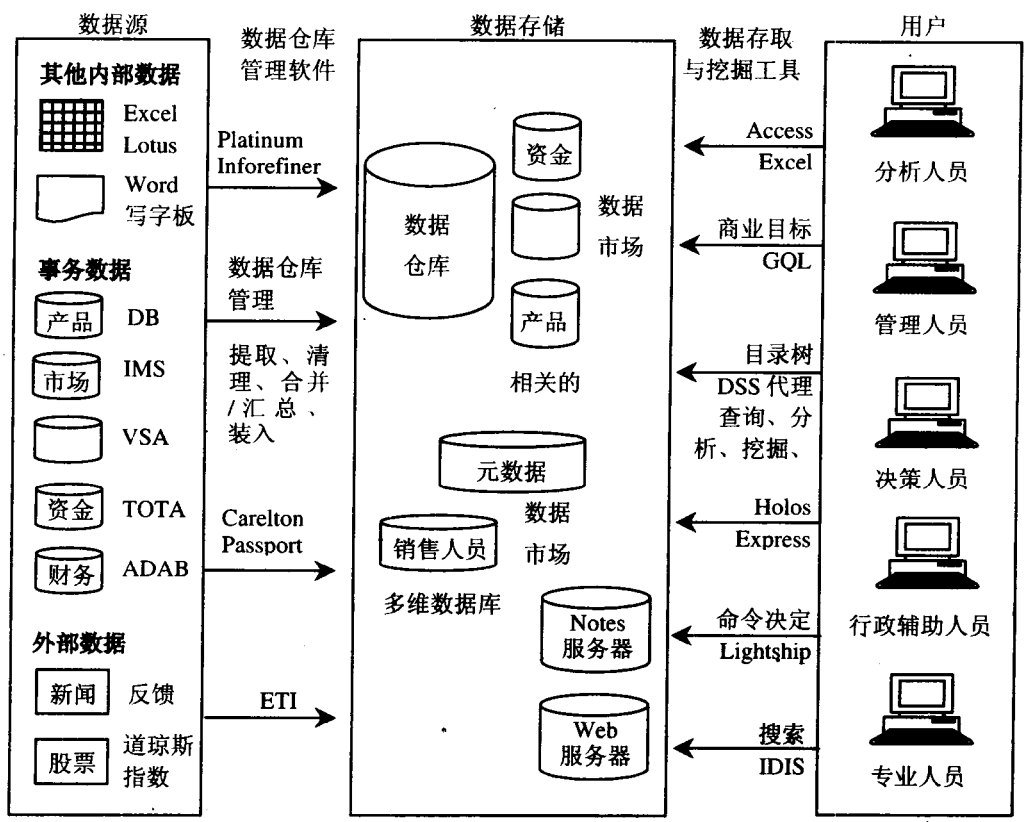

图 1.3 实际的数据仓库体系结构

数据仓库体系结构一般有三个部分（见图 1.2）。

1. 数据仓库本身包含数据和相关软件
2. 数据获得软件（后端），该软件从传统系统和外部来源获取数据，合并并汇总数据，把它装入数据仓库。
3. 允许用户存取和分析数据仓库中数据的客户（前端）软件。

图 1.3 扩展了图 1.2，展示了许多可能的与数据仓库相关的数据源、软件、用户。数据源（左边所示）包括内部的信息、外部的信息和事务信息。数据仓库管理软件用来获得数据仓库中的信息。数据库房(Data Store)不仅包括数据仓库而且还包括多种数据市场和服务器。数据存取和挖掘工具为用户提供通道和数据分析能力。

数据仓库也有其他的体系结构。例如：在两层环境中，数据源和客户在同一台计算机上，而数据仓库和数据获取软件在另一台计算机上。单层体系结构中，三个功能都在同一台物理机器上，有时在数据量有限且用户的数量很小时会使用这种结构。

为什么使用独立的数据仓库？

在整个这一章，都假设数据仓库和联机事务处理系统是独立的，实际上这种独立是存在的，有四种原因：

1. 性能。存取要求的高峰和低谷降低了 OLTP 系统的性能，OLTP 系统中的数据是面向操作的结果而不是决策支持的结果。这样，甚至看上去很简单的查询都对 OLAP 系统产生了很大干扰。
2. 数据存取。许多组织经常维护多种支持不同 OLTP 功能的数据库，数据仓库作为所有企业数据的集成，结合了所有数据来源和外部数据源，决策支持应用软件通过这些多种来源使用数据。典型的数据仓库用户不会注意数据存储在哪里，他们需要访问的数据，而不管 OLTP 系统中是否有。
3. 数据格式。数据仓库中的数据包括汇总数据和基于时间的数据，它们不在 OLTP 系统中保存。因为数据仓库中的数据是集成的，所以信息要以单一的标准格式保存。
4. 数据质量。数据仓库中的数据是干净、可确定、适合聚集的。数据经过转换以确保各数据项的惟一值存储在数据仓库中。数据仓库的基本概念是它包含正式确认的数据。这样，当人们遇到或使用数据仓库时，他们只要花费时间理解数据的意思，而不用再讨论正确的数据值是什么。数据仓库提供惟一正确的数值。

事实数据表和维表

为在联机分析过程（1.3 节所讨论）中进行检索，数据仓库中的数据要进行优化。因此，数据既可以组织为多维数据库（称为 MOLAP）也可以组织为关系数据库（称为 ROLAP）。在这两种情况中，目的是加速数据检索，以便快速地答复一位分析员，例如：分析员的查询涉及 9 月份在 Michigan 州打了折扣的销售产品数量 * 。在这部分，我们讨论使用特殊形式的关系型数据库以便它能处理多维。

要使关系数据库具有多维性，要特别介绍两种表格：事实数据表格和维表。

* 包括五个方面：产品、地区、价格、客户类型和时间

- 事实数据表包含大量的事实数据，如果所有的数据被存放在单一事实数据表中，结果是一个非常大的表。
- 维表包含事实数据表的指针,它显示在哪里能发现信息。每一维提供一个单独的表。

事实数据表又长又细；维表则短、小、宽。在一个查询中，系统首先存取一维或更多维的维表（例如 5 个维表），然后存取事实数据表，这种安排称为星型结构（star structure）或模式（schema）。星型结构允许数据表示一个“虚拟”的而不是物理的数据超立方体，因为维表中的指针允许用不同于查询的方法进行组织。图 1.4 说明了一个传统的星型模式(star schema)。

在星型结构中，每一维被它自己的表格描述。事实数据被安排在单一大事实数据表中，事实数据表被各维关键字组成的联合关键字索引。这样，图 1.4 中，包括了关键字 store、product 和 period，同时还包括了销售货物时的 dollars、units 和 price。

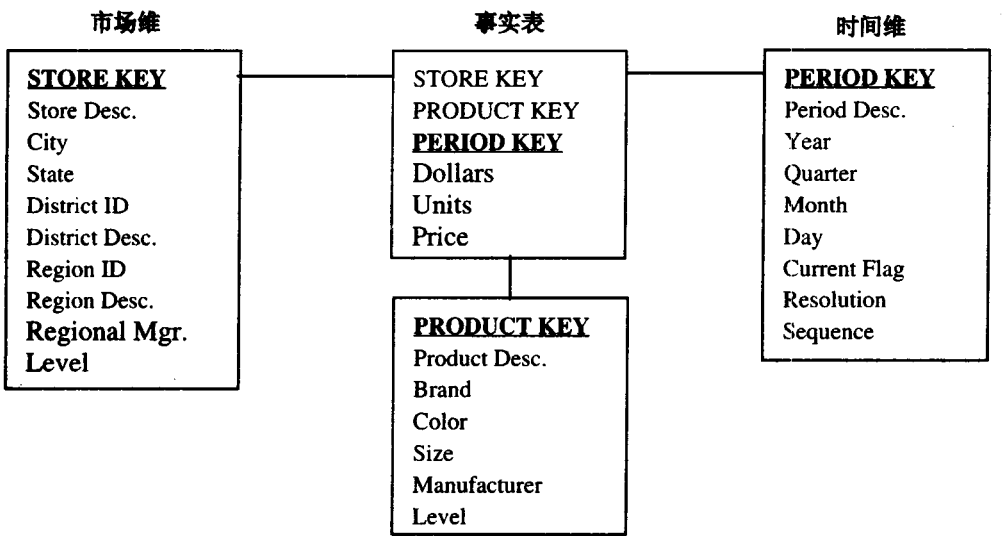

图 1.4 星型结构

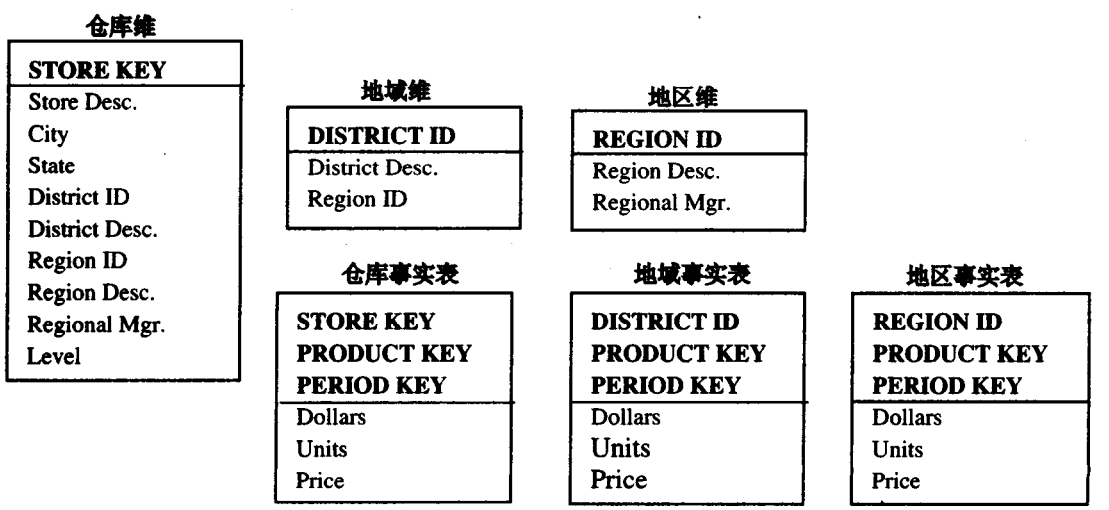

图 1.5 雪花结构

图 1.5 所示，雪花结构(snowflake schema)，是星型结构的一种替代。当数据库的种类很多而且维数很大时，使用这种结构。在一个维表中，雪花结构创建属性表格，对于超维，这种结构允许更快的数据恢复。

在图 1.5 雪花结构的简单说明中，使用了三个事实数据表而不是一个事实数据表。每一个表（仓库、地域、地区）有其相关的维表，也都含有产品和周期维。值得注意的是，如果地域表被查询，不需要处理地区和仓库的事实数据表。

使用数据仓库：汇总和索引

像任何软件产品一样，只有有规律地使用数据仓库时它才有价值。数据仓库的经验表明：

- 汇总的程度越高，数据使用得越多。
- 汇总的数据越多，检索越快。

这样，数据仓库按惯例创建和存储经常存取的数据汇总。预计算不仅加速了汇总数据的查询响应，而且加快了对其它查询的响应，因为它们减少了数据仓库忙的时间。

值得注意的是，高级别汇总数据能够被索引并可重新构造以方便使用。然而，低级别汇总下的数据过于庞大以至于不能索引和改变结构。在组织中增加数据，为外部数据源汇总数据，例如公司的 SEC 文档和竞争者股价，也被数据仓库保存。

费用和大小

数据仓库十分昂贵，数百万美元的费用很普通，一项研究报告表明平均费用为 220 万美元（Watson and Haley，1996）。

因为它们为企业而设计以便每个人有能力访问公共数据集，所以随着时间的变化，它们的体积在增加，典型的存储空间从 50GB 到 100GB 以上。因为它们的容量大，一些公司使用并行计算加速数据索引。尽管并行计算机的价格有所降低，但仍然不菲，而且，它们需要熟悉并行计算机的程序员及有更多程序员的组织支持。

数据市场

数据仓库的高费用限制了它们在大公司中的使用，许多公司选择的是建立称为数据市场(data mart)的低花费、小规模数据仓库。数据市场是为关键的商业单元（SUB）和部门而设计的小数据仓库，它是入门数据仓库和提供学习机会的一种途径。它也用来提供概念样板。数据市场存在的主要问题是它们随着部门的不同而不同。因此，如果在上述事实之后要开发一个综合数据仓库，那么集成它们将很困难。

可以使用两个方法来克服集成中遇到的问题：

1. 一些公司从单独的数据市场开始，但是有计划地在以后集成它们，这是一种循序渐进的方法，并最终以拥有一个企业范围的系统为目的。
2. 其它厂商以分配给个体单位分布式数据市场的形式建立起一个完整数据仓库。优点是数据市场较小、更容易针对局部需要；缺点是获得企业范围内的解决办法将更困难。

数据市场导致增强的信息孤岛，它给许多组织带来麻烦。为避免这个问题，一些公

司首先创建一个数据仓库，用它加装数据市场。这种方法使用户具有数据的企业视角，但是也带有数据市场的性能特征。

数据仓库的厂商起初发展应用广泛的数据仓库，然而，许多组织不愿意在没有小项目经验的情况下就承诺大项目所要求的资源。为了响应这个市场需求，许多厂商现在提供数据市场策略和解决办法，也为成熟的数据仓库提供策略和解决办法。

产业规模

数据仓库是一个主要产业。虽然估计有所不同，但是很明显，财富杂志全球前 500 家中一半多的企业正在开发数据仓库项目或正在计划之中。表 1.2 列出了数据仓库和决策支持软件的主要厂商（按英文字母顺序排序）。一些厂商提供两个方面的产品。

表 1.2 主要厂商的列表

[illegible]	[illegible]
Arbor	Andyne
Carleton	Ardor
IBM	Brio
Information	Business Objects
NCR	Cognos
Oracle/IRI	Comshare
Red Brick	Pilot
Sybase	Planning Sciences
	Platinum
	SAS
	Seagate

市场

在咨询公司中，进行市场预测是一项很麻烦的工作。下面是在 1996 年所进行的典型估价：在 1997 年 Bulter Group(U.K.)估价数据仓库有 90 亿的交易额。国际数据公司（IDC）则说在 1995 年美国有 57%的市场份额。

IDC 公司在 1996 年进行了名为“数据仓库的金融冲击”研究。结果，以拥有数据仓库的 62 家公司为基础，他们得到了表 1.3 所示的这些公司的结果。

表 1.3 数据仓库的金融冲击

3 年来平均返回的赢利（ROI）	401%
25% 有 ROI	>600%
中间资金回收	1.67 年
中间 ROI	167%
平均回收时间	2.3 年
平均赢利	220 万美元

世界范围内的市场

以 1996 年 IDC 公司的白皮书为基础，世界范围内数据仓库市场年收入预测显示在表 1.4 和图 1.6 中。像预测国家债务将减少一样，大部分数据仓库年收入的增长是在未来几年。