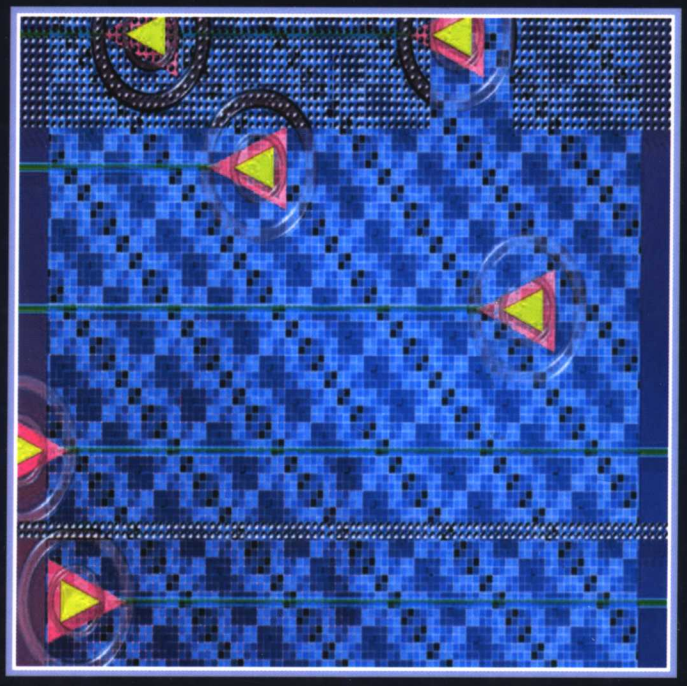
数据库技术丛书

数据仓库设计

Mastering Data Warehouse Design
Relational and Dimensional Techniques



Claudia Imhoff
(美) Nicholas Galemmo 著
Jonathan G. Geiger
于戈 鲍玉斌 王大玲 等译

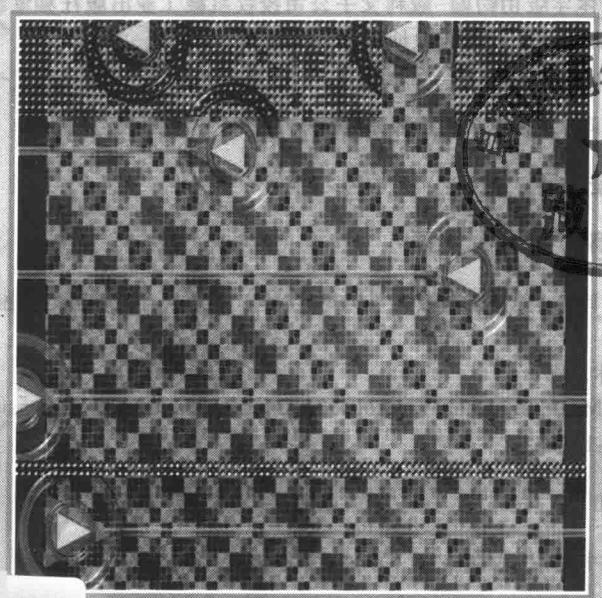
机械工业出版社
China Machine Press

数据库技术丛书

数据仓库设计

Mastering Data Warehouse Design

Relational and Dimensional Techniques



Claudia Imhoff
(美) Nicholas Galemmo 著
Jonathan G. Geiger
于戈 鲍玉斌 王大玲 等译

机械工业出版社
China Machine Press

本书全面论述了设计和建立高效、可持续发展且可扩展的数据仓库的方法，重点论述了建立各种数据模型的方法。主要内容包括业务智能环境和数据模型的概念、数据模型分类、数据模型的开发步骤、各种数据的建模方法、数据仓库的优化与扩展、数据模型的维护、关系型解决方案的部署、多维体系结构与企业信息工厂的比较等。

本书主要面向数据仓库的设计者和构建者以及数据仓库技术研究人员，同时也适合对数据仓库技术和企业信息化建设感兴趣的其他读者阅读。

Claudia Imhoff, Nicholas Galemmo, and Jonathan G. Geiger：Mastering Data Warehouse Design: Relational and Dimensional Techniques (ISBN 0-471-32421-3).

Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书版权登记号：图字：01-2003-6965

图书在版编目（CIP）数据

数据仓库设计 /（美）依默霍夫（Imhoff, C.）等著；于戈等译. -北京：机械工业出版社，2004.12
（数据库技术丛书）
书名原文：Mastering Data Warehouse Design: Relational and Dimensional Techniques
ISBN 7-111-13963-1

Ⅰ. 数… Ⅱ. ① 依… ② 于… Ⅲ. 数据库系统 Ⅳ. TP311.13

中国版本图书馆CIP数据核字（2004）第019616号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）
责任编辑：迟振春
北京中兴印刷有限公司印刷 · 新华书店北京发行所发行
2004 年 12 月第 1 版第 1 次印刷
787mm × 1092mm 1/16 · 20印张
印数：0 001-5 000册
定价：35.00元

译 者 序

我们拿到本书的英文稿时，该书的英文版还没有正式出版。看了该书的书名，首先感到眼前一亮，认为从本书可以详细了解数据仓库的设计方法。再看又感到费解，怎么又是关系与多维技术？通过浏览本书的详细内容后发现，本书确实是一本难得的介绍数据仓库建模方法的好书。

首先，本书的三位作者都具有从事数据仓库建设和咨询的丰富经历。Claudia Imhoff与Bill Inmon合作撰写了*Corporate Information Factory*，提出了企业信息工厂的概念和体系结构，创建了从事分析型CRM及业务智能技术和策略领域的权威咨询公司——Intelligent Solutions公司。Nicholas Galemmo和Jonathan G. Geiger都具有多年的系统开发和咨询经验。

第二，本书成功地将Bill Inmon的基于关系模型的数据仓库设计理念与Ralph Kimball的基于多维模型的数据集市设计理念结合起来，解决了数据仓库设计中令人困惑的局面。长期以来，人们一提起数据仓库的设计，就是多维模型、多维设计，本书给出了很好的答案和选择。

第三，本书是作者长期从事数据仓库建设和咨询工作的经验总结，给出了建立数据仓库的最好的实践方法和建议，即给出了成功建设数据仓库应该采取或避免什么行动。“日历建模”在其他数据仓库书中未见论述，本书作为一章来讨论。有数据仓库经验的技术人员都知道时间属性对于数据仓库至关重要，但是详细讨论日历建模的只有本书。

第四，本书结合两个实例介绍了数据仓库建模过程，并且详细地介绍了各个步骤应该进行的具体工作。数据仓库是一种能够管理企业一定历史时期内的海量数据，并为企业各级决策者提供支持的环境和技术。

全书共分三个部分，第一部分介绍了数据仓库和模型的基本概念，第二部分讨论了模型开发的过程和相关技术，第三部分讨论了数据仓库的操作和管理。本书以两个虚构的企业为实际案例贯穿全书，讨论了企业数据仓库建模和维护的全过程，包括业务建模、数据仓库的三范式建模、数据集市的多维建模、数据仓库的优化、数据仓库的维护、如何在已有的数据集市基础上建立企业数据仓库等。可操作性强、讲解透彻是本书的最大特点。

本书的主要译者从1995年起就开始了数据仓库相关技术的研究，承担了国家863计划中的数据仓库系统开发课题，并且参与了大型企业数据仓库工程的建设，深感建立企业数

据仓库的复杂性。建立企业数据仓库是一个需要投入巨大的人力、物力资源的工程，必须有一个很好的数据仓库建立方法学作为指导。而本书结合实际工程经验，详细地讲解了数据仓库建模中遇到的方方面面的问题，并详细分析了采用各种技术的原因。这些经验和方法对于数据仓库架构师极具参考价值。

本书的翻译、统稿由于戈、王大玲和鲍玉斌共同完成。其中术语表由于戈翻译，第1、11、12章由王大玲翻译，第2、5章由申德荣翻译，第3章由董晓梅翻译，第4章由张天成翻译，第6章由邓庆绪翻译，第7章由于亚新翻译，第8章由孙焕良翻译，第9、10章由赵志滨翻译，第13章由鲍玉斌翻译。

尽管我们具有一定的数据仓库项目研究和开发经验，但是由于本书涉及到许多应用领域的建模问题，所以有些词汇的翻译可能欠准确，译文中难免有不当之处，恳请读者批评指正。如果您有何建议和意见，欢迎发E-mail至：yuge@mail.neu.edu.cn, dlwang@mail.neu.edu.cn, baoyb@mail.neu.edu.cn。

目　录

第三部分 操作和管理

第一部分 基本概念

我们发现，理解一种方法为什么会受到推崇，有助于我们认识其价值，并更好地应用。因此，我们以关于企业信息工厂（Corporate Information Factory，CIF）的介绍作为本部分的开篇。企业信息工厂是一种经过检验的、稳定的体系结构，包括两种用于业务智能（Business Intelligence，BI）的数据存储器，它们分别在业务智能环境中起着特定的作用。

第一种数据存储器是数据仓库，其主要作用是作为数据仓储，存储来自于不同数据源的数据，支持来自第二种数据存储器（即数据集市）的访问。数据仓库作为数据的收集点，其最有效的设计方法是基于实体-联系数据模型和规范化技术，这两种技术分别是Codd和Date在20世纪70年代至90年代针对关系数据库的开创性工作开发的。

数据集市的主要作用是使业务用户很容易地访问高质量的、集成的信息。存在几种类型的数据集市，将在第1章中加以描述。最流行的数据集市是为支持联机分析处理而建立的，其最有效的设计方法是多维数据模型。

介绍了这些概念性主题后，我们将解释关系建模技术的重要性，介绍所需要的不同类型的模型，并且在第2章提出建立关系数据模型的过程。我们还要解释各种数据模型（包括业务的、系统的和技术性的数据模型）之间的关系，以及它们彼此之间如何共享或继承特性，这些数据模型用于为企业构造坚实的基础。

第1章 绪 论

欢迎使用本书，这是第一本详尽描述数据建模技术的书，数据建模技术用于构造多用途的、稳定的、可持续发展的、支持业务智能的数据仓库。本章通过描述BI和数据仓库的目标以及解释这些目标如何符合企业信息工厂（Corporate Information Factory，CIF）的整体结构来介绍数据仓库。本章将讨论数据仓库构造过程的迭代特性，示范数据仓库数据模型的重要性和本书推荐的数据模型格式的合理性。将讨论为什么这种模型格式应该基于关系模型设计技术，说明为什么要求将非冗余性、稳定性和可维护性最大化。还将概述可维护式数据仓库环境的特点。本章最后讨论这种建模方法对最终交付数据集市的影响。通过本章读者将初步了解后续章节中的原理，后续章节会详细描述如何建立数据仓库数据模型。

1.1 业务智能概述

在数据仓库环境中，BI是企业研究其历史行为和运作状况的能力，旨在了解该组织的境地，确定其当前的状态，预测或改变将要发生的情况。20多年来，BI已经走向成熟。下面我们简要地回顾一下过去10年里这一段引人入胜和具有开创性的历史。

读者可能已经熟悉什么是技术采纳曲线。最早采用新技术的那些公司叫做创新者，紧接着的那一批是早期的采纳者，然后是大批的早期采纳者以及大批的晚期采纳者，最后是落伍者。该曲线是一个传统的钟形曲线，开始以指数增长，到了“大批的晚期采纳者”阶段，（由这一新技术产生的）市场增长开始放慢。一种新技术刚出现时，往往难以获得它，昂贵且不完善。随着时间的推移，它的可用性、价格以及特色均改善到几乎只要拥有它就都能从中获利的程度。蜂窝电话就是这方面的一个例子。从前，只有使用该技术的创新者们（医生和律师）佩戴它们，这种电话庞大、沉重且昂贵，服务质量也不稳定，还经常“掉线”。现在，大约60美元即可购买一部蜂窝电话，服务提供商还在开通时额外赠送25美元话费，没有月租金，并且服务相当可靠。

建立数据仓库是另一个技术采纳曲线的例子。事实上，如果你还没有启动第一个数据仓库项目，那么现在恰逢其时。今天的企业主管们期望并且能够经常获得大量好的、及时的信息，他们需要根据这些信息来做出明智的决策，以便带领公司步入下一个10年。然而，

事情并不尽如人意。

恰在10年前，同样是这些主管们批准了主管信息系统（Executive Information System，EIS）的开发以满足他们的需求。启动EIS的创新想法是合理的，因为EIS及时地为主管们提供容易访问的关键性能信息。然而，许多这样的系统并没有达到目标，因为其基础体系结构不能足够快地响应企业日益变化的环境。早期EIS系统的另一个显著缺点是，需要巨大的投入来为这些主管们提供他们所需要的数据。数据获取或抽取、转换与装载（extract，transform and load，ETL）过程是一整套复杂的活动，其唯一目的就是要获得最精确的、集成的数据，并且使企业通过数据仓库或操作型数据仓（Operational Data Store，ODS）可以访问这些数据。

整个过程开始时采用密集型手工劳动，手工编码的“数据抽取器”是唯一手段，它从操作型系统中取出数据，以便业务分析员进行访问。这类似于早期的电话技术，当时接线员不得不穿着旱冰鞋来回奔跑，以手工方式在合适的位置插接线头，从而将你的电话与你的受话方的电话连接起来。

幸运的是，我们已经从那个时期走了出来，数据仓库业界已经开发了大量的工具和技术以支持数据获取过程。现在，技术的进步已使得上述过程的大部分可以自动实现，如同今天电话业的进步一样。同时，与电话技术的发展相类似，这种过程也遗留了一个问题，这个问题即便不是本质的、复杂的，也是困难的。这就是没有任何公司具有同样的数据获取活动，甚至是同样的问题集合。当今大多数在数据仓库建设中给予了大量投入的大企业非常依赖于他们自己的ETL工具来对其BI环境进行设计、构造和维护。

在过去10年间另一个主要的变化是一些工具和建模技术的引进，它们将“易于使用”这一口号带到我们的生活中。由Ralph Kimball等人开发的多维建模概念导致了支持联机分析处理的多维数据集市的广泛应用。

除多维分析外，还有一些成熟的技术也得到了很大的发展，它们可以支持数据挖掘、统计分析以及探索分析的需要。现在，成熟的BI环境要求比星型模式更多的技术。除星型模式外，还需要平面文件、无偏差数据的统计子集、规范化的数据结构等，这些都是数据仓库应该满足的重要数据需求。

当然，我们也不应该低估Internet对数据仓库建设的影响。Internet帮助我们揭开了计算机的神秘性，主管们在他们的日常生活中使用Internet，敲打键盘已不再是小心翼翼的。终端用户工具的厂商意识到了Internet的影响，其中一些人抓住商机设计他们的界面，这种界面模仿了流行的Internet浏览器和搜索引擎中的某些临场感特性。这些工具的成熟性、简易性已经导致了业务分析员和主管们对BI的更为广泛的使用。

前几年发生的另一个重要事件是从技术吸引业务转换到业务需要技术。在BI的早期，

信息技术（Information Technology，IT）界人士认识到了信息技术的价值，并极力向业务群体宣传其优点。可是，在某些场合，IT人员着手建立的数据仓库，只能期待有业务群体使用。今天，功能齐全的决策支持环境的价值已被业务群体广泛认同。例如，一个有效的客户关系管理程序如果没有战略型（对应于集市相关的数据仓库）和战术型（对应于操作型数据仓和操作集市）决策能力，是不能生存的（见图1-1）。

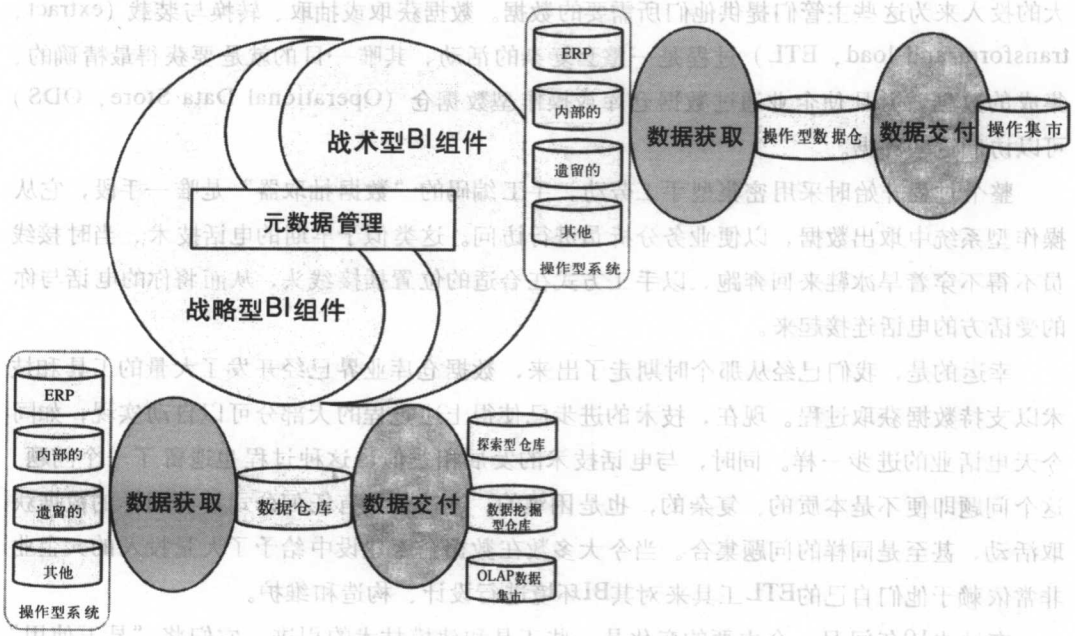


图1-1　一个BI环境的战略型和战术型部分

BI体系结构

过去10年中最有意义的进展之一是引进了支持BI的所有技术需求并被广泛接受的体系结构。该体系结构解决了EIS方法中存在的几个主要缺陷，其中最明显的是，EIS数据结构经常直接由源系统供给（数据），从而导致了一个非常复杂的数据获取环境，该环境需要大量的人机资源来维护。而CIF（如图1-2）（当前在大多数决策支持环境中使用的体系结构）将数据分别装入5种主要的数据库（操作型系统、数据仓库、操作型数据仓、数据集市和操作集市），并且将处理过程合并起来以便高效地将数据从源系统移动到业务用户，从而弥补了EIS的不足。

这些构件可进一步分解为两个分组，每个分组均包含一些构件和过程：

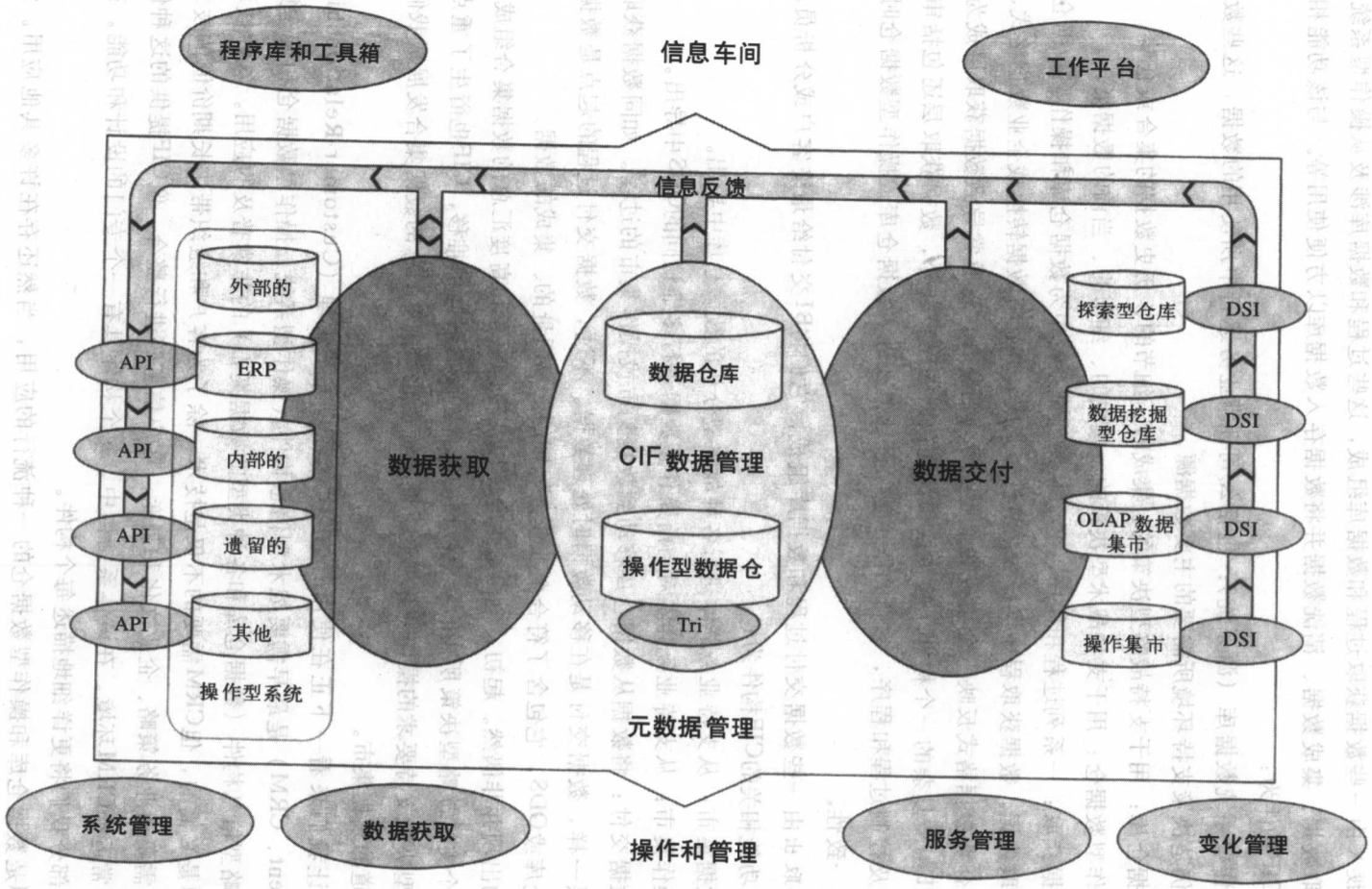


图1-2　企业信息工厂（CIF）

- 数据投入由一些数据获取过程和数据库组成，这些过程和数据库涉及从操作型系统中获取数据、集成数据、清洗数据并将数据存入数据库以方便使用等。与该功能相关的CIF构件为：
  - 操作型系统数据库（源系统）：包括公司日常业务运行中所使用的数据，这些数据仍是决策支持环境所需要的主要数据源。
  - 数据仓库：用于支持战略型决策的、集成的、细节的、历史数据的集合或仓储。
  - 操作型数据仓：用于支持战术型决策的、集成的、细节的、当前的数据集合。
  - 数据获取：一系列过程和程序，它们从操作型系统中为数据仓库和操作型数据仓抽取数据。数据获取程序执行数据清洗、集成，并且将数据转换成企业数据格式，该企业数据格式反映一个集成的企业业务规则集合，这通常导致数据获取层成为CIF中最复杂的一个构件。除转换和清洗数据的程序以外，数据获取层还包括审计及控制过程和程序，它们的作用是保证数据在进入数据仓库或操作型数据仓时的一致性。
- 信息取出由一些数据交付过程和数据库组成，它们将BI交付给最终客户或分析员。与该功能相关的CIF构件为：
  - 数据集市：从支持业务群体进行各种战略型分析的数据仓库中导出。
  - 操作集市：从支持业务群体对当前操作型数据进行多维访问的ODS中导出。
  - 数据交付：将数据从数据仓库移动到数据集市或操作集市的过程。如同数据获取层一样，数据交付是在移动数据时处理数据。不过，数据交付过程的起点是数据仓库或ODS，它包含了符合企业业务规则的、高质量的、集成的数据。

CIF的出现并非偶然。起初，它由数据仓库以及轻度汇总和高度汇总的数据集合组成，它们是一个支持战略型决策所需的历史数据的集合。随着时间的推移，CIF也衍生了重点满足战术型决策支持要求的操作型数据仓。而轻度汇总和高度汇总的数据集合发展成我们现在所知道的数据集市。

下面让我们来看一下正在使用着的CIF。客户关系管理（Customer Relationship Management，CRM）是最早需要战术型信息构件（操作型系统、操作型数据仓和操作集市）和战略型信息构件（数据仓库和各种类型的数据集市）的非常普及的应用。这种技术对于CRM是必需的，但CRM需要的不仅是技术。除了为客户和组织带来长期价值的技术外，它还需要对业务策略、企业文化和制度、客户信息等进行整合。像CIF提供的这种体系结构非常适合CRM环境。在该体系结构中，每个构件具有一个专门的设计和功能。在本章稍后部分我们将更详细地描述每个构件。

CRM是数据仓库和操作型数据仓的一种流行的应用，当然还存在许多其他应用。例

如，企业资源规划（Enterprise Resource Planning，ERP）厂商（如SAP、Oracle以及PeopleSoft）已经接纳了数据仓库，并且增加了能够提供所需功能的工具套件。许多软件厂商正在提供各种“插件程序”，这些插件包括像获利能力分析或关键性能指标（Key Performance Indicator，KPI）分析这样的通用分析应用程序。我们将在本章后面各节中更详细地介绍CIF。

数据仓库技术的发展在帮助公司更好地为其客户服务和增加利润方面一直是至关重要的。它将技术的变化和可持续发展的体系结构相结合。建立这种环境的工具已经发展了很长时间，相当成熟，并且在设计、实现、维护和对关键企业数据进行访问诸方面提供了很多的便利。CIF体系结构获取了这些技术和工具方面的创新，建立了一个包含5种数据仓的环境，每种数据仓都在以正确的时间、正确的地点、正确的形式向业务群体提供正确的信息方面起着关键作用。所以，如果你是建立数据仓库方面的后来者甚至落伍者，请振作起来，它是值得等待的。

1.2 什么是数据仓库

在开始实际描述这种建模技术之前，需要大家对数据仓库的含义、在BI中的作用和目的以及支持数据仓库构建和使用的系统构件等方面达成共识。

1.2.1 数据仓库的作用和用途

正如我们在上一节所讲，整个BI体系结构在过去的10年里已经有了相当的发展，从简单的报表输出和EIS系统，到多维分析统计和数据挖掘，到探索能力以及可定制的分析应用技术的引进，所有这些技术都是健壮和成熟的BI环境的组成部分。图1-3给出了每一个技术进步的一般时间框架。

给定了这些重要的而又显著不同的技术和数据格式要求后，很显然，支持和维护任何BI环境的出发点就一定是建立一个高质量的、可信赖的数据仓库，这些数据仓库具有灵活的、可重用的格式。从一开始，数据仓库就成为BI体系结构的一部分。各种方法学和数据仓库权威对该构件给出了各种名字，如：

- 准备区：数据仓库的一种形式是“后台办公室”（back office）准备区，来自操作型系统的数据首先在这里集合起来。准备区是一个非正式设计和维护的数据分组，其唯一的目的是向多维数据集市输入数据。
- 信息仓库：这是IBM和其他厂商早期使用的数据仓库的名字，它不像准备区定义得那样清楚，在许多场合，它不仅包括历史数据仓储，而且还包括在其定义中的各种

数据集市。

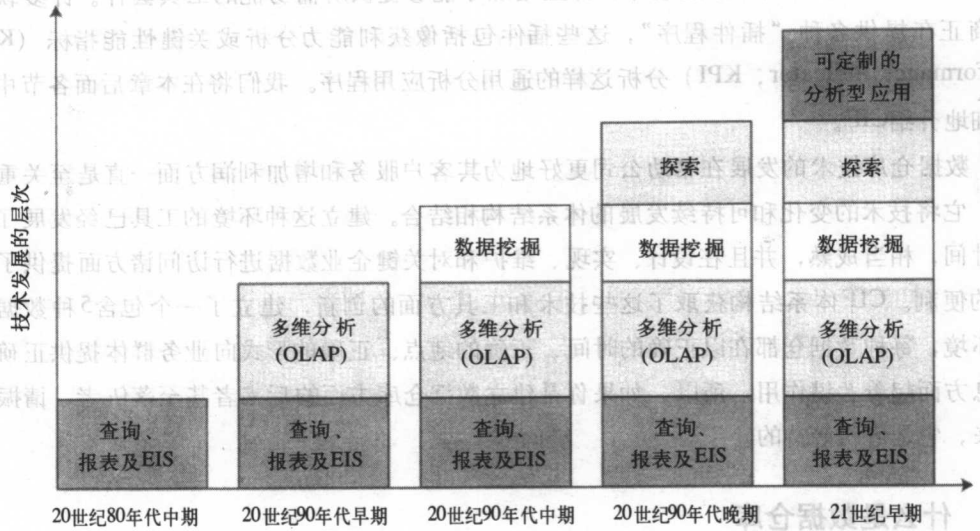


图1-3　BI技术的演进

数据仓库环境必须整合各种技能、功能和技术，因此在设计中必须牢记两个思想。第一，它必须具有一个合适的粒度或细节级别以满足所有的数据集市，即：它必须包括具有最少共同点的细节数据，以便提供聚集的、汇总的集市，以及提供用于事务级探索和挖掘的仓库。

第二，它的设计不能阻碍在数据集市中使用各种技术，该设计必须适应多维集市、统计、挖掘以及探索型仓库。此外，它还必须适应那些正在提供和准备提供的分析型应用，以支持任何新的尖端技术。这样，它必须支持的模式包括星型模式、平面文件、规范化数据的统计型子集以及BI将来需要的模式。给定这些目标后，我们来看看数据仓库如何适应于支持成熟BI环境的复杂体系结构。

1.2.2　企业信息工厂

企业信息工厂（CIF）是一种可对信息进行描述和分类的概念性体系结构，已被广泛接受，它用于操作和管理一个成功且健壮的BI结构。这些信息仓支持三种高层次的组织型处理：

- 业务操作。涉及到每天进行的业务操作，操作型事务处理系统和外部数据在此功能中。这些系统帮助运行业务，通常是高度自动化的。支持这种功能的处理是相当静