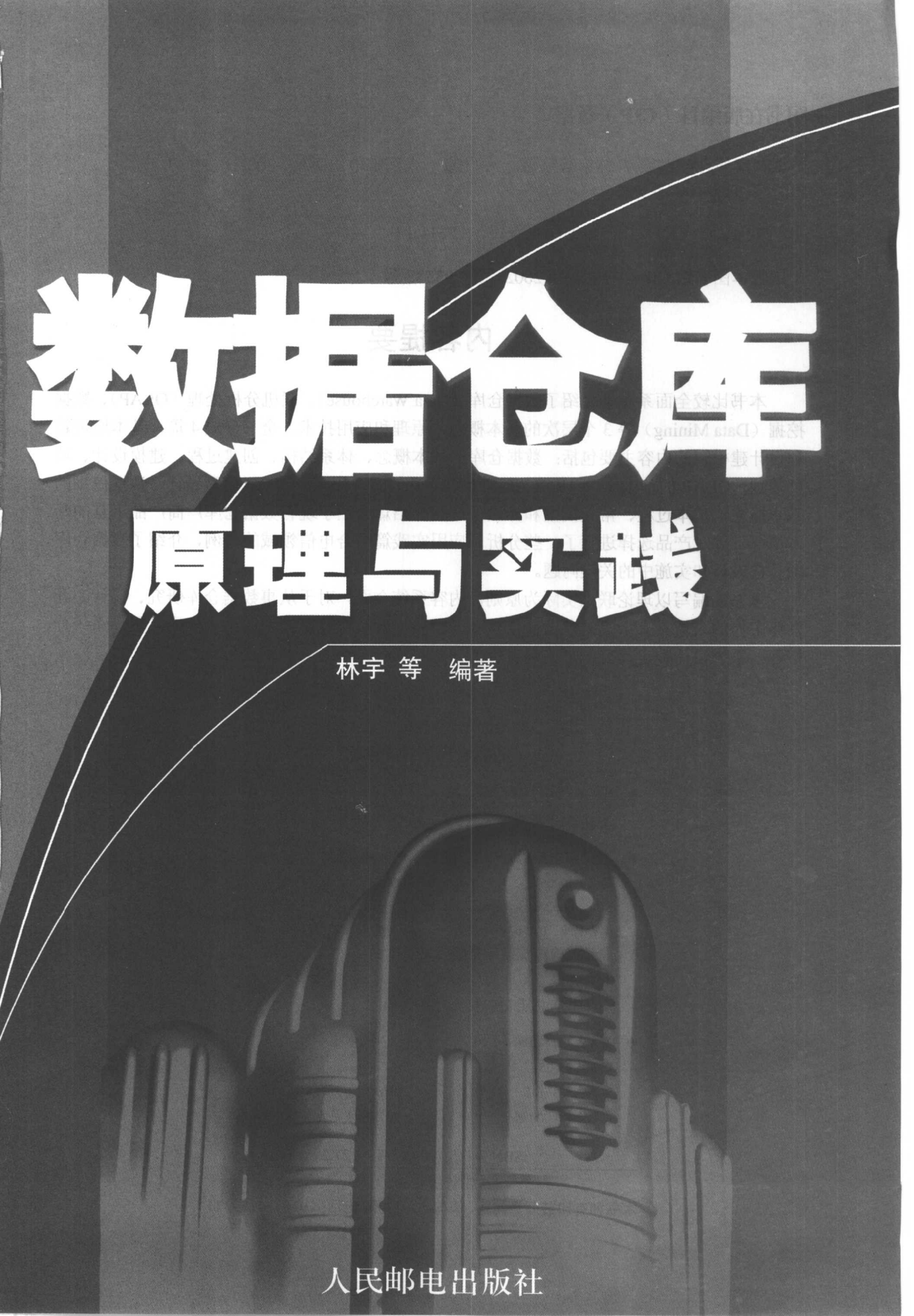
数据仓库
原理与实践

林宇 等 编著

人民邮电出版社
POSTS & TELECOMMUNICATIONS PRESS

数据仓库
原理与实践

林宇 等 编著



人民邮电出版社

图书在版编目（CIP）数据

数据仓库原理与实践 / 林宇等编著. —北京：人民邮电出版社，2003.1
ISBN 7-115-10044-6

Ⅰ. 数… Ⅱ. 林… Ⅲ. 数据库系统Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字（2002）第 093538 号

内容提要

本书比较全面系统地介绍了数据仓库（Data Warehouse）、联机分析处理（OLAP）、数据挖掘（Data Mining）等 3 个层次的基本概念、原理和应用技术。全书分成 4 篇，基本原理篇和设计建模篇的内容主要包括：数据仓库的基本概念、体系结构、创建过程、建模设计、项目规划，OLAP 的基本概念、ROLAP 和 MOLAP 的实现原理、OLAP 模型设计，数据挖掘的基本概念、基本过程、常见模型和算法。产品介绍篇介绍了现有数据仓库厂商产品工具的基本情况，并对产品选择进行了一些分析。应用实践篇结合电信领域的实例，介绍了数据仓库项目在设计和实施中的关键问题。

本书的编写以理论联系实际为原则，内容系统全面，对于从事数据仓库研究、设计、开发等工作的人员具有宝贵的参考价值，对于需要了解数据仓库技术的系统集成人员、系统分析师、系统设计师也具有一定的参考价值。

数据仓库原理与实践

◆ 编　著　林 宇 等
责任编辑　张立科

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号
邮编　100061　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
读者热线　010-67132692
北京汉魂图文设计有限公司制作
北京密云春雷印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：787×1092　1/16
印张：23.25
字数：563 千字　　2003 年 1 月第 1 版
印数：1-4 000 册　　2003 年 1 月北京第 1 次印刷

ISBN 7-115-10044-6/TP・2729

定价：42.00 元

前 言

随着计算机应用和网络计算的发展，“计算”正向两个不同的方向拓展：一是广度计算，二是深度计算。广度计算是把计算机的应用范围尽量扩大，同时实现广泛的数据交流。深度计算是人们对以往计算机的简单数据操作。目前对深度操作已提出了更高的要求，希望计算机能够更多地参与数据分析与制定决策的工作。传统的数据库技术是单一的数据库资源，它适合操作型事务处理，但分析型处理（或信息型处理）能力较弱。数据仓库的出现，将操作型环境和分析型环境进行了分离，划清了数据处理的分析型环境与操作型环境之间的界限，从而由原来的以单一数据库为中心的数据环境发展为以数据仓库为中心的一种新的体系化环境。

数据仓库技术以改进后的数据库技术作为存储数据和管理资源的基本手段，以统计分析技术作为分析数据和提取信息的有效方法，通过人工智能、神经网络、知识推理等数据挖掘方法来发现数据背后隐藏的规律，从而实现从“数据→信息→知识”的过程，为企业的管理阶层提供各种层次的决策支持。

本书从数据仓库、联机分析（OLAP）、数据挖掘等3个层次对数据仓库系统的关键技术进行了深入细致的介绍和分析。

全书共分成4篇，分别是基本原理篇、设计建模篇、产品工具篇、应用实践篇。

基本原理篇从第1章至第4章。第1章“数据仓库概述”介绍从数据库到数据仓库的演变过程，着重说明了要解决“蜘蛛网”问题，必须在体系结构上加以变革，将操作型环境和分析型环境分离。第2章“数据仓库基本原理”主要介绍与数据仓库相关的基本概念和数据处理的流程，着重说明了数据仓库（DataWarehouse）中数据清洗、数据转化、元数据、数据集市等一系列重要的概念，比较了数据仓库和数据库的特点和关键技术，详细介绍了数据仓库面向主题、数据集成、数据稳定、数据随时间变化的根本特征。第3章“OLAP的基本原理”主要介绍OLAP技术的概念和展现方式。第4章“数据挖掘的基础”主要介绍与数据挖掘相关的概念，本章首先给出数据挖掘的形式化定义，然后介绍数据挖掘的描述型挖掘算法和预测性挖掘算法，并列举了几个数据挖掘在网络中的实际应用。

设计建模篇从第5章至第10章，它在基本原理篇的基础上进一步介绍了数据仓库模型设计、OLAP模型设计、数据挖掘模型、数据仓库的规划以及数据仓库的维护等问题。第5章“企业模型设计”和第6章“数据仓库的模型设计”用于介绍数据仓库的建模知识。第7章“OLAP建模方法”对OLAP建模方法进行了比较细致的介绍。第8章“数据仓库规划和开发方法”介绍了数据仓库的投资分析和两种软件开发方法学。第9章“数据挖掘的实施过程”介绍了数据挖掘的基本方法。第10章“数据仓库的建立和维护”从业务系统、数据仓库、OLAP、数据挖掘、数据展现等方面完整地说明了数据仓库建立的全过程，并介绍了在数据仓库维护阶段设计人员和维护人员需要了解的技术问题。

产品介绍篇包括了第11章和第12章。在第11章中，简单介绍了现有的著名的数据仓库厂商及其主要产品，并选择了有代表性的产品进行说明。在第12章中，从技术和需求两方面对产品选择问题作了一些分析。

应用实践篇从第13章至第16章。本篇结合电信领域的特点，通过一个电信领域的应用

实例来说明数据仓库的具体实施方案。第 13 章“项目的需求和目标分析”主要说明数据仓库在电信领域的应用规划。第 14 章“系统结构和模型设计”介绍了系统结构设计、数据仓库模型设计、OLAP 模型设计三部分。第 15 章“数据装载设计和数据挖掘”介绍了数据载入、建立多维数据库、数据挖掘相关的设计和实现问题。第 16 章“界面设计和项目总结”说明了界面设计的相关问题，并对项目的总体工作流程、数据流程、模块构成进行了总结。

本书的写作过程也是笔者们学习、探讨、实践的过程。在此过程中，笔者阅读了大量的国内外著作、论文，考察了数据仓库各大厂商的产品特征和性能，对大量的资料进行了归纳和整理。

在本书的编写过程中，得到了程时端、卢美莲、金跃辉、龚向阳、阙喜戎等教授的息心指导。参加本书研讨和写作的有林宇、郭凌云、王令成、伦勇、高波、翟朝阳、胡颖峰、李劲松、徐自军、王重钢、乌海涛、王茜、白刚、范锐、彭泳、柴平宣、乐辉华等。另外，高波为本书的筹备和组织做了大量工作。在此对他们的辛勤工作表示深深的谢意！

由于数据仓库技术是一个新的技术，加之笔者水平有限，书中错误在所难免，希望各位专家和广大读者给予批评指正，多提宝贵意见。

编者

2002.11

目　录

第一篇　基本原理篇

第二篇 设计建模篇

第三篇 产品介绍篇

第四篇 应用实践篇

第一篇　基本原理篇

本篇主要对数据仓库、OLAP 分析、数据挖掘的基本原理进行了详细介绍，以求读者能够对基本概念有比较明确的认识。

本篇共 4 章。第 1 章“数据仓库概述”介绍了数据库到数据仓库的演变过程，重点说明了“蜘蛛网”问题的产生原因以及随之而来的种种问题。要解决这些问题就必须在体系结构上加以变革，将操作型环境和分析环境分离，使企业由原先以数据库为中心的生产环境过渡到以数据仓库为中心的生产环境。

第 2 章“数据仓库基本原理”主要介绍数据仓库相关的基本概念和数据处理的流程。重点说明了数据仓库（DataWarehouse）中数据清洗、数据转化、元数据、数据市场等一系列重要的概念。我们还比较数据仓库和数据库的特点和关键技术，详细阐明数据仓库面向主题、数据集成、数据稳定、数据随时间变化的根本特征。接着，介绍数据仓库的数据组织结构和组织方式，如何进行数据追加，以及数据颗粒度和数据分割的概念。

第 3 章“OLAP 的基本原理”主要介绍 OLAP 技术的概念和展现方式。首先介绍了 OLAP 的基本概念：变量、维、维层次和维分类、事实、多维数据立方体。然后说明了 OLAP 分析的基本动作：数据钻取、数据聚合、数据切片、数据切块、数据旋转等。接着介绍了 OLAP 丰富多彩的展现方式，比较了 OLAP 和 OLTP 处理的差异。2 最后介绍了以多维数据库为基础的 MOLAP 的实现原理和以关系型数据库为基础的 ROLAP 的实现原理，详细比较了 ROLAP 和 MOLAP 技术的优缺点。

第 4 章“数据挖掘的基础”主要介绍了与数据挖掘相关的概念。本章首先给出了数据挖掘的形式化定义，并根据形式化定义简要说明了数据挖掘的过程。然后介绍了数据挖掘的算法分类，说明了描述型挖掘算法，它包括关联分析、序列分析、分类分析、聚类分析。接着介绍了预测性挖掘算法，它包括决策树算法、神经网络算法等。最后讲述了数据挖掘的体系结构，并列举了几个数据挖掘在网络中的实际应用。

第 1 章　数据仓库概述

本章介绍了从数据库到数据仓库的演变过程，着重说明了“蜘蛛网”问题的产生原因以及随之而来的种种问题。要解决这些问题就必须在体系结构上加以变革，将操作型环境和分析环境分离，使企业由原先以数据库为中心的生产环境过渡到以数据仓库为中心的生产环境。本章还介绍了以数据仓库为基础的商业智能系统的功能和构成以及商业智能系统中的核心技术。最后还简要介绍了数据仓库技术的应用前景。

1.1　数据库到数据仓库的演变

1.1.1　蜘蛛网问题

随着数据库技术的广泛运用，企业的运营环境逐渐转化成以数据库为中心的运营环境。企业对数据的需求是多方面的，除了在企业中建立企业级的数据库外，常常还要建立部门级数据库。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还要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，很可能会形成如图 1-1 所示的“蜘蛛网“现象，使数据的抽取和访问显得错综复杂。一个大型的公司每天进行上万次的数据抽取很普通。这种演变不是人为制造的，而是自然演变的结果，如果不再体系结构上进行调整，”蜘蛛网“问题将越来越严重。

错综复杂的抽取与访问将产生很多的问题，诸如数据分析的结果缺乏可靠性、数据处理的效率很低、难于将数据转化成信息。

1．数据分析的结果缺乏可靠性

图 1-2 中展示了某电信公司的市场部和计划部对业务 A 是否具有市场前景的分析过程和结果。市场部认为“业务 A 的市场前景很好”，而计划部却得到截然相反的结果——“业务 A 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？

两个分析的数据都来自于企业数据库，但是结论却不同，下面通过分析两个过程的差异来寻求原因。

首先，市场部门和计划部门从企业数据库中抽取的数据可能不同，比如，市场部抽取的是在大客户中对业务 A 的使用情况，而计划部抽取的是在普通客户中对业务 A 的使用情况。两者分析数据的内容存在差异。

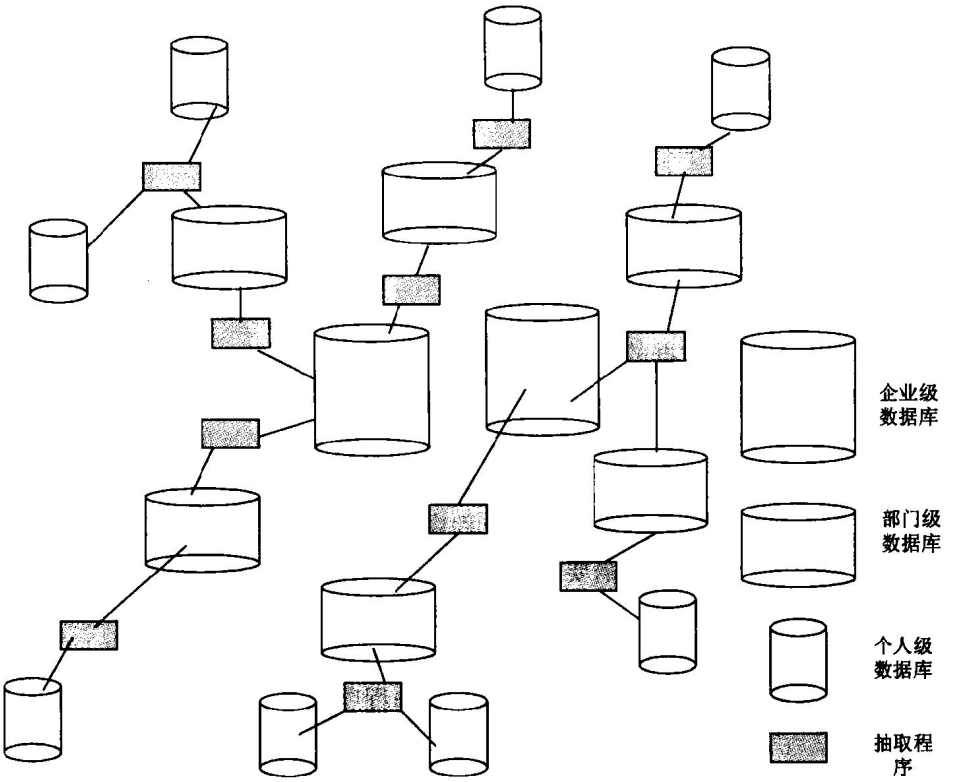


图 1-1　企业中存在的“蜘蛛网”现象

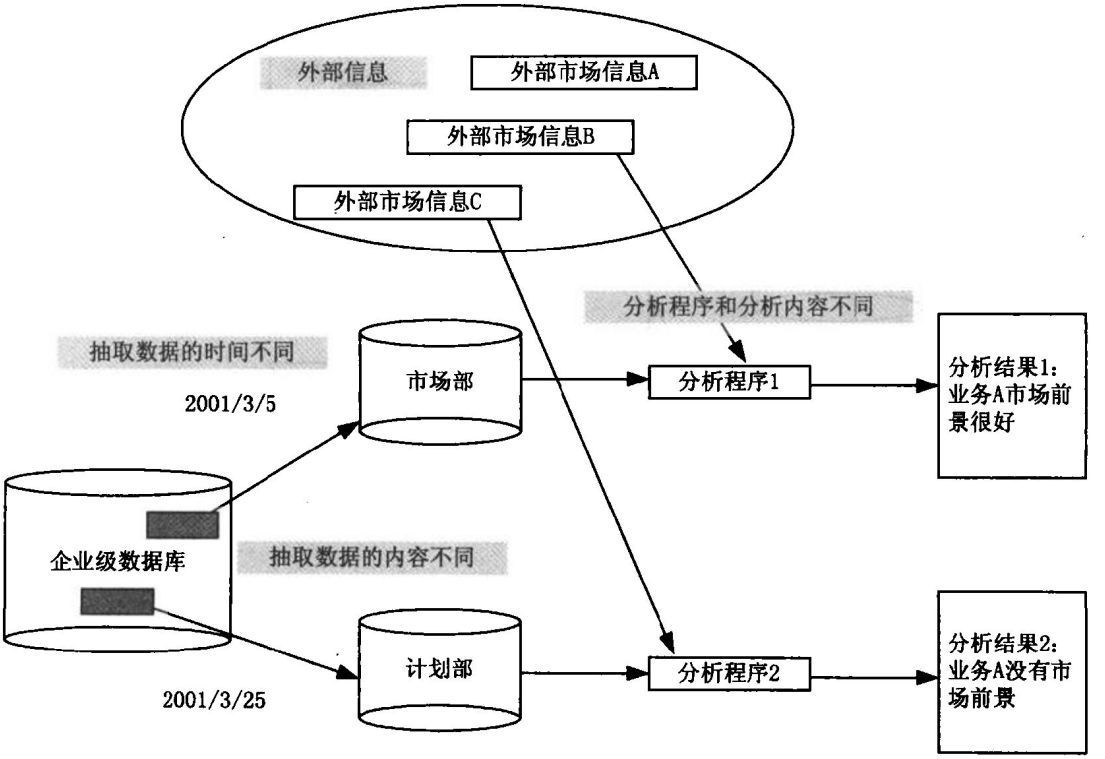


图 1-2　两个分析过程的差异

其次，市场部分析的时间是 2001 年 3 月 5 日，而计划部分析的时间是 2001 年 3 月 25 日，两个分析过程从企业数据库中抽取数据的时间不同，企业数据库中的内容已经发生了变化，这很可能导致分析的结果不同。

再次，分析业务的发展趋势常常需要引入企业外部的信息，比如客户的满意程度、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，市场部门使用的分析程序可能与计划部不同，分析的内容和指标也可能不同。通过上面的分析，我们可以看出导致两个分析过程出现截然相反的结论的根本原因是数据的来源不一致，对于不同来源的数据的分析结果显然是不一致的。

2．数据处理的效率很低

在错综复杂的体系结构中，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，对于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，抽取程序应用的技术不同，因而难以集成。

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，需要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。由于在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。程序的重用性很差，因为决策者明天想看的内容很可能与今天不同。可见，动用大量的人力、物力和时间才能完成的报表不仅时效性很差，数据处理的效率也很低。报表形成的过程如图 1-3 所示。

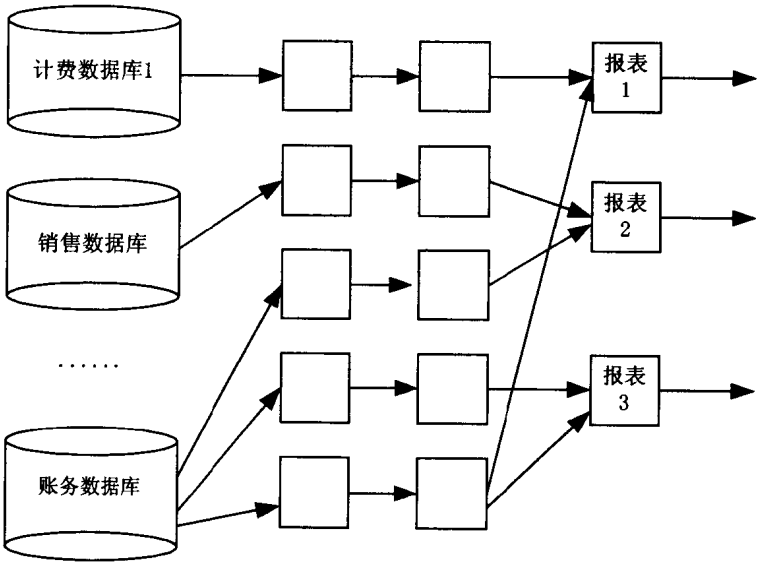


图 1-3　报表形成过程示意图

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为信息。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客户的呼叫行为、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。

在实际的数据库系统中，记录客户呼叫行为的数据库通常只保留客户最近 3 个月的呼叫话单，账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，如图 1-4 所示。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，因此以现有的数据库系统难以提供完整的历史数据。鉴于这样的原因，用户根本不可能从这些数据中提取出完整的信息。

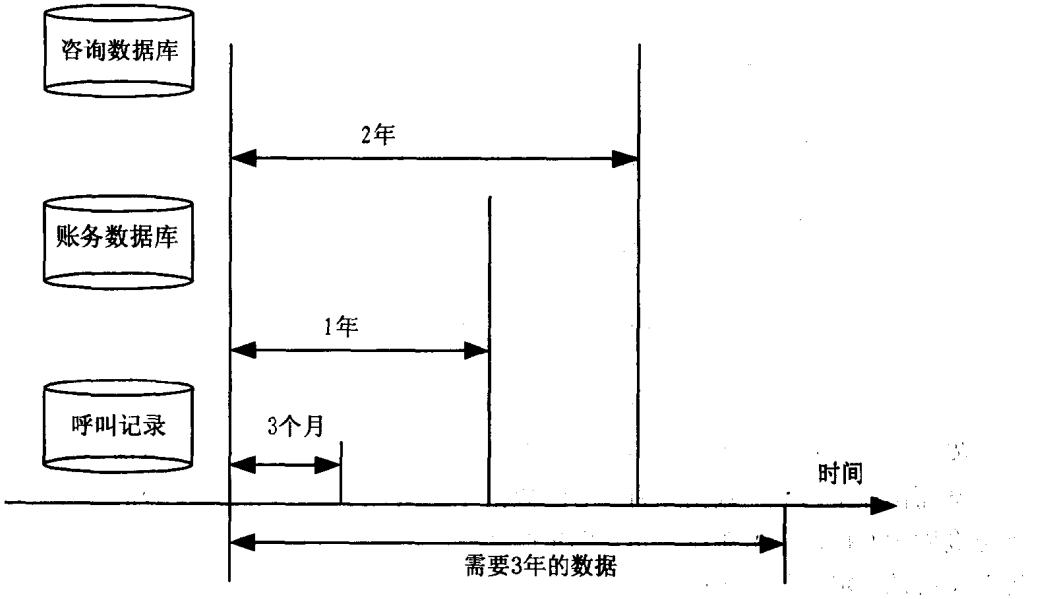


图 1-4　现有数据库系统难以提供完整的历史数据

1.1.2　操作型系统和分析型系统的分离

随着人们应用需求的发展，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数据分析的数据环境分离开。

这样，数据处理被分为操作型处理和分析型处理（或信息型处理）两大类。操作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，对于一个大型的连锁超市，如果能够将各个营业点不同时期的营业情况以非常直观的方式展现给管理人员，则管理人员可以根据这些分析结果决定是否需要撤销营业情况极差的营业点，而在客户流量特别大的超市附近增设营业网点。

操作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支持信息。操作型系统和分析型系统的划分如图 1-5 所示。