计 算 机 科 学 丛 书

数据仓库

Building the Data Warehouse

Second Edition

(美) W. H. Inmon 著

王志海 等译

机械工业出版社
China Machine Press

WILEY

计算机科学丛书

数据仓库

(美) W. H. Inmon 著

王志海 等译

黄厚宽 田盛丰 审

机械工业出版社
China Machine Press

本书论述在设计和建造数据仓库中涉及的所有主要问题，论述分析型环境(决策支持系统环境)以及在这种环境中的数据构造。主要内容包括数据仓库的设计与建造步骤，传统系统到数据仓库的迁移，数据仓库的数据粒度、数据分割、元数据管理、外部数据与非结构化数据，分布式数据仓库、高级管理人员信息系统和数据仓库的设计评审等。

本书主要是面向数据仓库的设计、开发和管理人员，以及构造和使用现代信息系统的人员，也适于信息处理方面的高校师生和从事传统数据库系统技术工作的人阅读。

本书版权登记号：图字：01-2000-1170

图书在版编目(CIP)数据

数据仓库／(美)因曼(Inmon,W.H.)著；王志海等译. －北京：机械工业出版社，2000.5
(计算机科学丛书)
书名原文：Building the Data Warehouse,Second Edition.
ISBN 7-111-07889-6
Ⅰ.数… Ⅱ.①因… ②王… Ⅲ.数据库系统－基本知识 Ⅳ.TP311.13

中国版本图书馆CIP数据核字(2000)第15806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：李伯民 李新阳
北京昌平第二印刷厂印刷·新华书店北京发行所发行
2000年5月第1版·2000年6月第2次印刷
787mm×1092mm 1/16·15印张
印数：5 001-8 000册
定价：25.00元

译 者 序

计算机网络与数据库技术的迅速发展和广泛应用，使得企业管理进入一个崭新的时代。广大基层管理人员摆脱了繁重的制表业务和数据处理工作，管理工作得到进一步规范化，许多业务得到了联机事务处理信息系统的支持。然而，面对当今竞争日趋激烈与瞬息万变的市场经济，各级管理人员迫切需要面对不同层次的大量信息迅速作出抉择。这就要求各级管理人员能够从大量复杂的业务数据中获取各自权限内的决策信息，及时把握市场变化的脉搏，作出正确有效的判断和抉择。特别是随着数据库系统的逐日运行，数据的堆积将越来越庞大，这种需求就比以往任何时候都更加迫切。从各级决策者的角度来看，数据处理的重点应该从传统的业务过程扩展到对业务数据的联机分析处理，并从中得到面向各种管理主题的统计信息和决策支持信息。

数据仓库就是针对解决上述问题所产生的一种技术方案，是基于大规模数据库的决策支持系统环境的核心。正如本书作者数据仓库之父W. H. Inmon 所定义的，数据仓库是一个面向主题的、集成的、不可更新的且随时间不断变化的数据集合，用来支持管理人员的决策。本书详尽地讲述了数据仓库系统的基本概念与基本原理。主要包括数据仓库的结构、数据分割与粒度划分、数据仓库模型、数据仓库中的数据访问方式、数据仓库的各种组织技术、分布式数据仓库、数据仓库与管理人员之间的关系、设计复查以及数据仓库的开发方法等。本书面向数据仓库的开发者、管理者、设计者、数据管理员、数据库管理员以及在现代数据处理环境中的相关人员。对于计算机专业的本科生和研究生也有重要的参考价值。

我们研究小组的博士生导师和教授组织博士研究生和硕士研究生对数据仓库和数据库知识发现进行了长期的研究与讨论，在此过程中翻译了一些有关数据仓库的著作。出版社希望将这本数据仓库的最经典著作的中文版推荐给国内读者。为此，我们对该书的译稿重新进行了修正与审校。下面是本书各章的主要翻译者：前言由王志海和王琨翻译，第1章由王琨翻译，第2章由王继奎翻译，第3章由董隽、王志海、刘犇、林友芳翻译，第4章和第5章由高思宇翻译，第6章由王春花翻译，第7章由王琨翻译，第8章由王继奎翻译，第9章和第10章由林友芳翻译，其余部分由刘犇、王志海、王继奎、林友芳等翻译，参加本书翻译与讨论工作的还有宁云晖、李晓武、蔺永华、范星艳等。全书由王志海负责统一定稿，由黄厚宽教授审定前言、第1章、第2章、第4章至第7章，其余部分由田盛丰教授审定。

由于译者水平有限，以及一些术语的翻译目前尚缺乏规范，错误之处望广大读者批评指正。

译 者

北方交通大学计算机科学技术系

2000年3月

审、译者简介

王志海　博士，副教授，1985年毕业于郑州大学计算机科学系，获理学学士学位，1987年毕业于哈尔滨船舶工程学院计算机与信息科学系，获工学硕士学位，1998年毕业于合肥工业大学计算机与信息学院，获博士学位。中国机器学习学会理事，曾任澳洲MONASH大学计算机科学与软件工程学院客座研究员。主要完成了国家教委博士学科点专项科研基金项目“从大规模数据库中自动提取领域知识的算法与实现研究”，主持了安徽省自然科学基金项目“基于超媒体面向智能CAD概念设计的知识获取系统”等项目。目前，作为第一主要研究者正在进行国家自然科学基金资助项目“基于粗糙集合理论的概念结构模型研究”、主持铁道部科技专项经费资助项目“大规模异构数据库的知识发现方法及其可视化技术研究”的研究工作，发表论文三十余篇。

黄厚宽　教授，博士生导师。1940年9月生，1963年毕业于北京大学数力系六年制数学专业，1966年哈尔滨军事工程学院应用数学研究生毕业。1970～1980年参加我国首次洲际火箭发射落点水声测量系统研制，主持总体数学模型论证计算及专用计算机系统软件编制，获中央军委嘉奖及原国防科工委重大科技成果三等奖。1983～1985年先后在美国亚拉巴马大学和佛罗里达大学信息研究中心任访问教授。十多年来主持完成多个专家系统与工具及计算机应用系统，进行机器学习、专家系统、分布式人工智能的研究。共获省部级科技进步奖6项，已发表论文80多篇，指导硕士与博士研究生50多人，俄罗斯高级访问学者1人。现任中国计算机学会人工智能与模式识别专委会副主任兼秘书长等。

田盛丰　教授，1944年11月生，1967年毕业于哈尔滨军事工程学院电子工程系，1968～1977年在七级部五院五零四研究所任实习研究员、1977年至今在北方交通大学计算机系任教。其中1982～1984年在美国纽约州立大学石溪分校作访问学者，主要研究人工智能；1997年在英国伦敦大学Royal Holloway学院计算机科学系合作研究人工智能项目。曾主持和参加了多项科研项目，包括国家自然科学基金项目“隧道工程预测专家系统”、“工程建设中知识系统的应用研究”、“断裂地质构造遥感图象判释专家系统”，教委博士点基金项目“隧道岩溶预测专家系统”，部委级项目“国防交通铁路工程保障指挥决策专家系统的改进与应用”等。发表论著2部及论文50多篇。

前　言

数据库及其理论已经出现好长时间了。早期数据库主要是一些单独的数据库，应用于已知信息处理领域的各种目的---从事务处理到批处理和分析处理。大多数情况下，早期的数据库系统主要集中于操作型的日常事务处理。近年来，数据库较为高级的思想已经产生，一方面是为了满足操作型数据处理的需求，另一方面是为了满足信息型或分析型数据处理的需求。从某种程度上讲，数据库的这种新颖思想是随着个人计算机(PC)技术、第四代程序设计语言(4GL)技术以及最终用户的推动而出现的。

下面的诸多因素导致了操作型数据库和信息型数据库的分离：

- 满足操作型需求的数据从物理上不同于满足信息型或分析型需求的数据。
- 支持操作型处理的技术从根本上不同于支持信息型或分析型需求的技术。
- 操作型数据的用户范围不同于信息型或分析型数据所支持的用户范围。
- 操作型环境的处理特点与信息型环境的处理特点从根本上是不同的。

由于这些原因(以及很多其他原因)，当今建造数据库系统的方式是将操作型处理及数据同信息型或分析型处理及数据分离。

本书论述分析型或决策支持系统(DSS)环境以及在这种环境中的数据构造，主要讨论数据仓库(或信息仓库)及其相关问题。数据仓库是处于信息型DSS处理的核心。

本书所讨论的问题是面向管理者和开发者的。就大多数章节而言，本书是关于数据仓库的问题与技术的，在某些适当的地方也将讨论一些技术层次上的问题。本书旨在作为数据仓库的设计者和开发者的一本指导性读物。

什么是分析型、信息型处理呢？分析型或信息型处理是针对制定决策过程中管理方面的需求而进行的处理。分析型处理是浏览大量数据以找出其中的趋势，这就是所谓的DSS处理。当DSS分析员进行分析型处理时，不是只查看一个或两个数据记录，而是要访问很多记录，不像在操作型处理中的那种情况。

另外，DSS分析人员极少更新数据。在操作型系统中，数据以单个记录的方式频繁地更新，而在分析型处理中需要不断访问记录以及收集记录内容进行分析，但是各个记录的内容极少更改或根本不更改。

在分析型处理中，响应时间的需求与传统操作型处理相比要宽松得多。分析的响应时间可从30分钟到24小时，而对操作型处理而言这样的响应时间范围将会是一种巨大的灾难。

作为分析型环境的计算机网络比作为操作型环境的计算机网络的规模要小得多。通常情况下，分析型网络的用户远比操作型网络的用户要少。

与作为分析型环境的技术不同，操作型环境的技术必须关注其本身的数据、事务锁定、数据争用和死锁，等等。

因此，在操作型环境和分析型环境之间存在着很多重要区别。本书是关于分析型的DSS环境的，着重阐述下列问题：

- 数据粒度。

- 数据分割。
- 元数据。
- 数据可信度的需求。
- DSS数据的集成。
- DSS数据的时基。
- 识别DSS的数据源——记录系统。
- 迁移和方法。

本书是针对数据仓库的开发者、管理者、设计者、数据管理员、数据库管理员以及在现代数据处理环境中构造系统的其他人员的。也适用于信息处理方面的大学生们。

本书是与数据仓库有关的系列从书的第一本，此套从书的下一本是《USING THE DATA WAREHOUSE》，该书着重阐述了在已经建造好的数据仓库中会遇到的问题。此外，还介绍了一种更大的体系结构的概念和一种操作型数据储仓(ODS)的思想。操作型数据储仓是与数据仓库相似的一种体系结构，不同之处在于操作型数据储仓只适用于操作型系统，而不适用于信息型系统。此套从书的第三本书是《BUILDING THE OPERATIONAL DATA STORE》，书中阐述什么是操作型数据储仓与如何建造操作型数据储仓。

有很多人直接或间接地为本书的完成做出了贡献。下面的名单仅列出了他们当中的一些人：

- Sue Osterfelt,Nations Bank
- Claudia Imhoff,Intelligent Solutions
- John Zachman,Zachman International
- Jim Kerr,independent consultant
- Ed Young,Prism Solutions
- Jim Ashbrook,Prism Solutions
- Cynthia Schmidt,Prism Solutions
- Peter LaPorte,Tandem Computers
- Edie Conklin,independent consultant
- George Coleman,Prism Solutions
- Jeanne Friedman,Logica
- Cheryl Estep,Chevron Corporation
- Kevin Gould,Sybase
- Chuck Kelley,Pine Cone Systems
- George Comeaux,Bank of Boston
- J.D.Welch,Prism Solutions
- Arnie Barnett,Barnett Data Systems

目　录

第1章　决策支持系统的发展

信息系统领域是一个“不成熟”的领域。“不成熟”这个词通常具有消极的含义，因而公开使用这个词不得不多加小心。但是从历史的观点来看的确如此。如果我们将信息处理的历史与其他技术领域的历史进行比较的话，就没有争议了。我们知道古埃及的象形文字主要是当时的帐房先生用来表示所欠法老谷子的多少。当漫步在罗马市区，我们就置身于两千多年前土木工程师所设计的街道与建筑物之间。同样，许多其他的领域也可追溯到远古时代。

因为信息处理领域只是从60年代初期才出现的，所以，历史地来看，信息处理领域是不成熟的。

信息处理领域的年轻性表现之一就是其倾向于面面俱到。有这样一种说法，如果细节都正确了，那么我们就可以坐享其成。这就好象是说，若我们知道如何铺水泥、如何钻孔、如何安装螺母与螺栓，就不必操心桥梁的外型与用途了。如此态度会驱使一个成熟的土木工程师发疯的。

数据仓库的历史是伴随某种发展过程开始的，在此发展过程中，业界中人士所考虑的是投入更大的力量。更大规模的体系结构正在被勾勒出来——在这种体系结构中数据仓库处于中心地位。最好从一种广阔的视角去观察这个体系结构，而不是从某种细节去认识。

1.1　演化

有趣的是，决策支持系统(DSS)处理是一个漫长而复杂的演化进程的结果，而且它仍在继续演化。DSS处理的起源可以追溯到计算机发展的初期。

图1-1表明了从20世纪60年代初期直到1980年的DSS处理的演化进程。在60年代初期，创建运行于主文件上的单个应用是计算领域的主要工作。这些应用的特点表现在报表和程序，常用的是COBOL语言。穿孔卡是当时常用的介质。主文件存放在磁带文件上。磁带适合于廉价地存放大容量数据，但缺点是需要顺序地访问。事实上，我们常说，在磁带文件的一次操作中，100%的记录都要被访问到，但是只有5%或更少的记录是真正需要的。此外，访问整条磁带的文件可能要花去20~30分钟时间，这取决于文件上是什么数据及当前正在做什么处理。

大约在60年代中期，主文件和磁带的使用量迅速膨胀。很快，处处都是主文件。随着主文件数量的增长，出现大量冗余数据。主文件的迅速增长和数据的巨大冗余引出了一些严重问题：

- 需要在更新数据时保持数据的一致性。
- 程序维护的复杂性。
- 开发新程序的复杂性。
- 支持所有主文件需要的硬件数量。

简言之，属于介质本身固有缺陷的主文件的问题成为发展的障碍。如果仍然只用磁带作为存储数据的唯一介质，那么难以想象现在的信息处理领域会是什么样子。

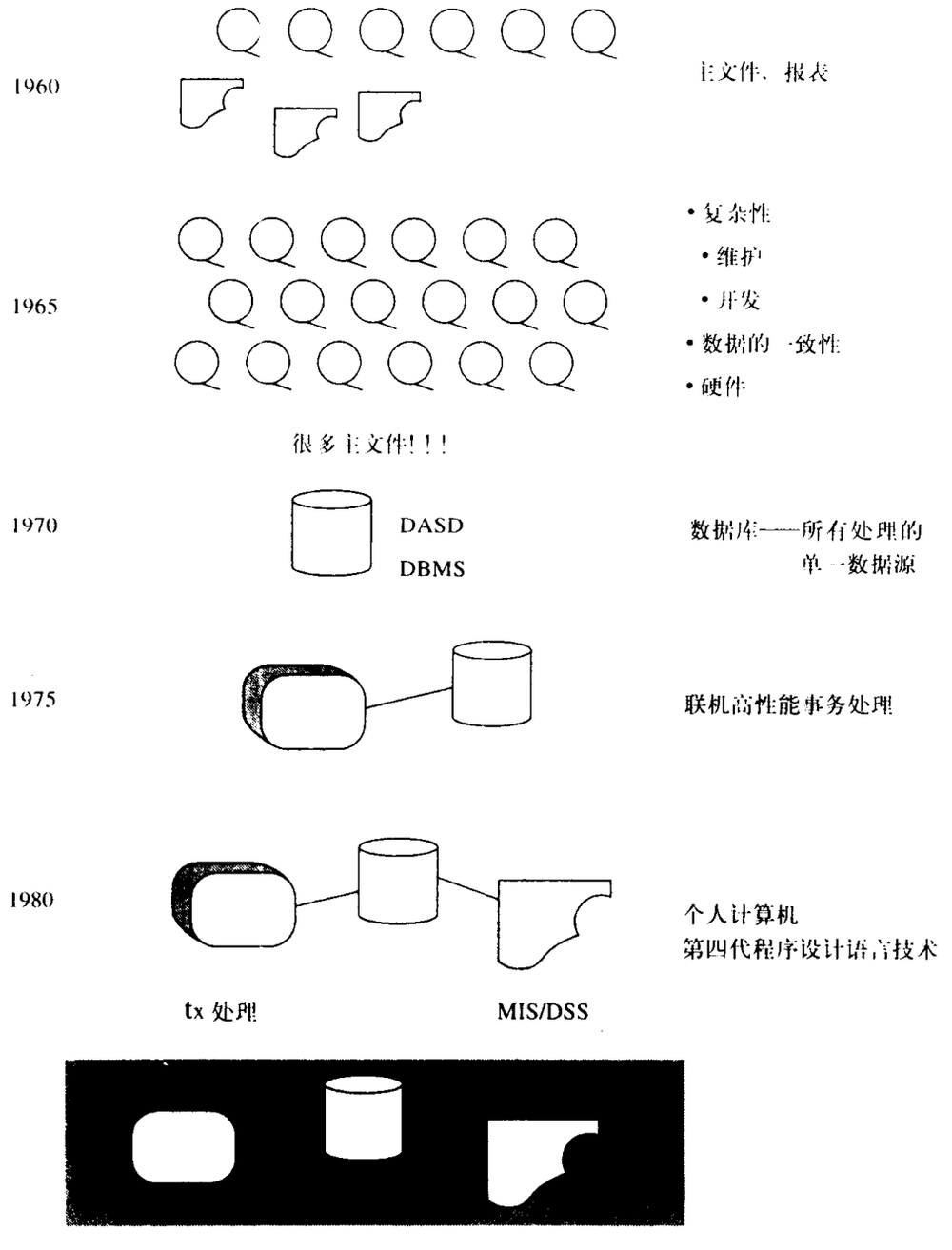

图1-1 体系化环境的早期演化阶段

如果除了磁带文件以外没有别的东西可以存储大量数据，那么世界上将永远不会有大型、快速的预定系统，ATM系统，以及其他系统。而事实上，在除磁带文件之外的种种介质上存储和管理数据的能力，为采用不同的处理方式和更强有力的处理类型开辟了道路，从而把技术人员和商务人员前所未有地聚集到一起。

1.2 直接存取存储设备的产生

到了1970年，一种存储和访问数据的新技术出现了。这就是20世纪70年代见到的磁盘存

储，或者称之为直接存取存储设备(DASD)。磁盘存储从根本上不同于磁带存储，因为DASD上的数据能够直接存取。DASD就不需要经过第1条记录、第2条记录……，第n条记录，才能得到第n+1条记录。一旦知道了第n+1条记录的地址，就可以轻而易举地直接访问它。进而，找到第n+1条记录需要的时间比起扫描磁带的时间少得多。事实上，在DASD上定位记录的时间是以毫秒(ms)来计量的。

随DASD而来的是称之为数据库管理系统(DBMS)的一种新型系统软件。DBMS的目的是使程序员在DASD上方便地存储和访问数据。另外，DBMS关心的是在DASD上存储、索引数据等任务。随着DASD和DBMS的出现，解决主文件系统问题的一种技术解决方案应运而生。“数据库”的思想就是DBMS的产物。纵观主文件系统所导致的混乱以及主文件系统累积的大量冗余数据，就不会奇怪为什么把数据库定义为——所有处理工作的单一数据源。

但这一领域的发展并未在1970年停止。到70年代中期，联机事务处理开始取代数据库。通过终端和合适的软件、技术人员发现更快速地访问数据是可能的——这就开辟了一种全新的视野。采用高性能联机事务处理，计算机可用来完成以前无法完成的工作。当今，计算机可用于建立预定系统、银行柜员系统、工业控制系统，等等。如果仍然滞留在磁带文件系统时代，那么今天我们认为理所当然的大多数系统就不可能存在了。

1.3 个人计算机/第四代编程语言技术

到了80年代，一些更新颖的技术开始涌现出来，比如个人计算机(PC)和第四代编程语言(4GL)。最终用户开始扮演一种以前无法想象的角色——直接控制数据和系统，这超出了对传统数据处理人员的界定。随着PC与4GL技术的发展，诞生了一种新思想，即除了高性能联机事务处理之外，对数据可以做更多的处理。管理信息系统(MIS)——(早期被如此称呼)也可能实现了。MIS如今称为DSS，是用来产生管理决策的处理过程。以前，数据和技术不能一并用来导出详细的操作型决策。一种新的思想体系开始出现，即一个单一的数据库既能用作操作型的高性能事务处理，同时又用作DSS分析处理。图1-1表明了这种单一数据库的范例。

1.4 进入抽取程序

大型联机高性能事务处理问世后不久，就开始出现一种称为“抽取”处理的程序(见图1-2)，这种程序并不损害已有系统。

抽取程序是所有程序中最简单的程序。它搜索整个文件或数据库，使用某些标准选择合乎限制的数据，并把数据传到其他文件或数据库中。

抽取程序很快就流行起来，并渗透到信息处理环境中。至少有两个理由可以用来解释它为什么受到欢迎：

- 因为用抽取程序能将数据从高性能联机事务处理方式中转移出来，所以在需要总体分析数据时就与联机事务处理性能不发生冲突。
- 当用抽取程序将数据从操作型事务处理范围内移出时，数据的控制方式就发生了转变。最终用户一旦开始控制数据，他(她)就最终“拥有”了这些数据。

由于这些原因(以及其他众多原因)，抽取处理很快就无处不在。到了90年代已有了很多抽取程序，如图1-3所示。

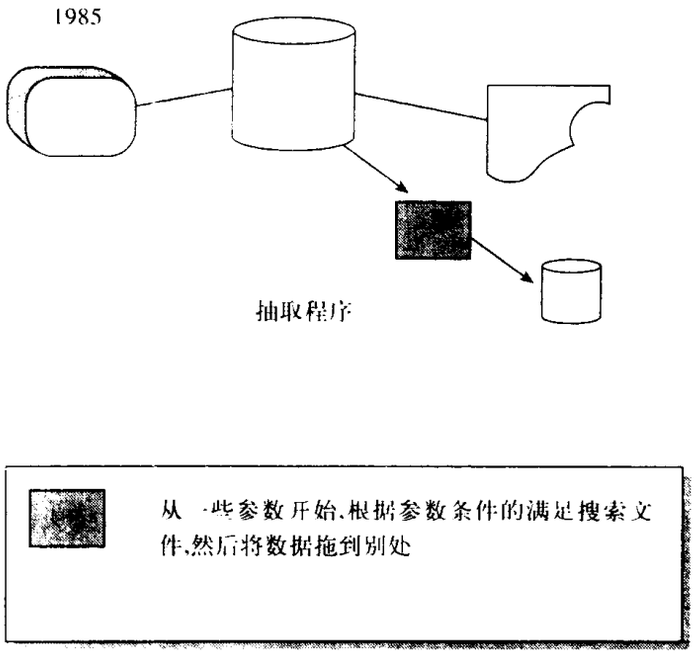

图1-2 抽取处理的特性

1.5 蜘蛛网

图1-3显示抽取处理的蜘蛛网开始形成。起初只是抽取，随后是抽取之上的抽取，接着是在此基础上的再次抽取，如此等等。对于一个大公司，每天进行多达45 000次的抽取不是没有听说过的。

贯穿于公司或组织的这种抽取处理模式很常见，以致得到一个专有名称。这种由失控的抽取过程产生的结构被称为“自然演化体系结构”——当一个组织以放任自流的态度处理整个硬、软件体系结构时，就会发生这种情况。组织越庞大，越成熟，自然演化体系结构问题就变得越严重。

从总体上看，抽取程序形成了蜘蛛网，这正是自然演化(或“传统系统”)体系结构的另一

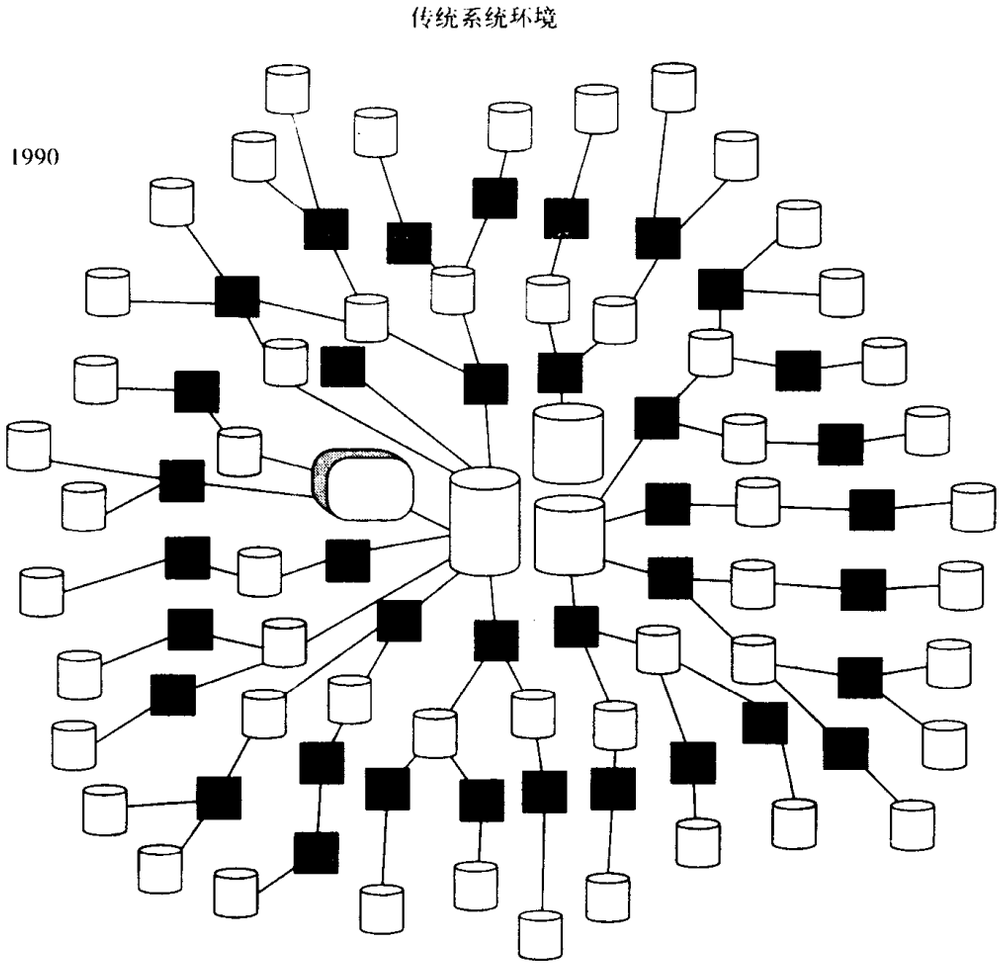

图1-3 抽取处理广泛采用必然是件好事情

个名称。

1.6 自然演化体系结构的问题

与自然演化体系结构相关联的困难到底是什么呢？问题很多，主要有：

- 数据可信性。
- 生产率。
- 数据转化为信息的不可行性。

1.6.1 数据缺乏可信性

以上问题之首是数据缺乏可信性，如图1-4所示。两个部门向管理者呈送报表，一个部门说业绩下降了15%，另一个部门说业绩上升了10%。两个部门的结论不但不吻合，而且相去甚远。另外，两个部门的工作也很难协调。除非十分细致地编制了文档，否则对任何应用目的而言，协调是不可能的。

当管理者收到这两张报表时，他们不知如何是好。管理者面临着根据政策和个人意志做决定的状况。这是在自然演化体系结构中可信性危机的一个实例。

这种危机很广泛存在，而且是可以预想得到的，为什么？有五个理由可以解释危机的可预测性(见图1-4)，它们是：

- 数据无时基。
- 数据算法上的差异。
- 抽取的多层次。
- 外部数据问题。
- 无起始公共数据源。

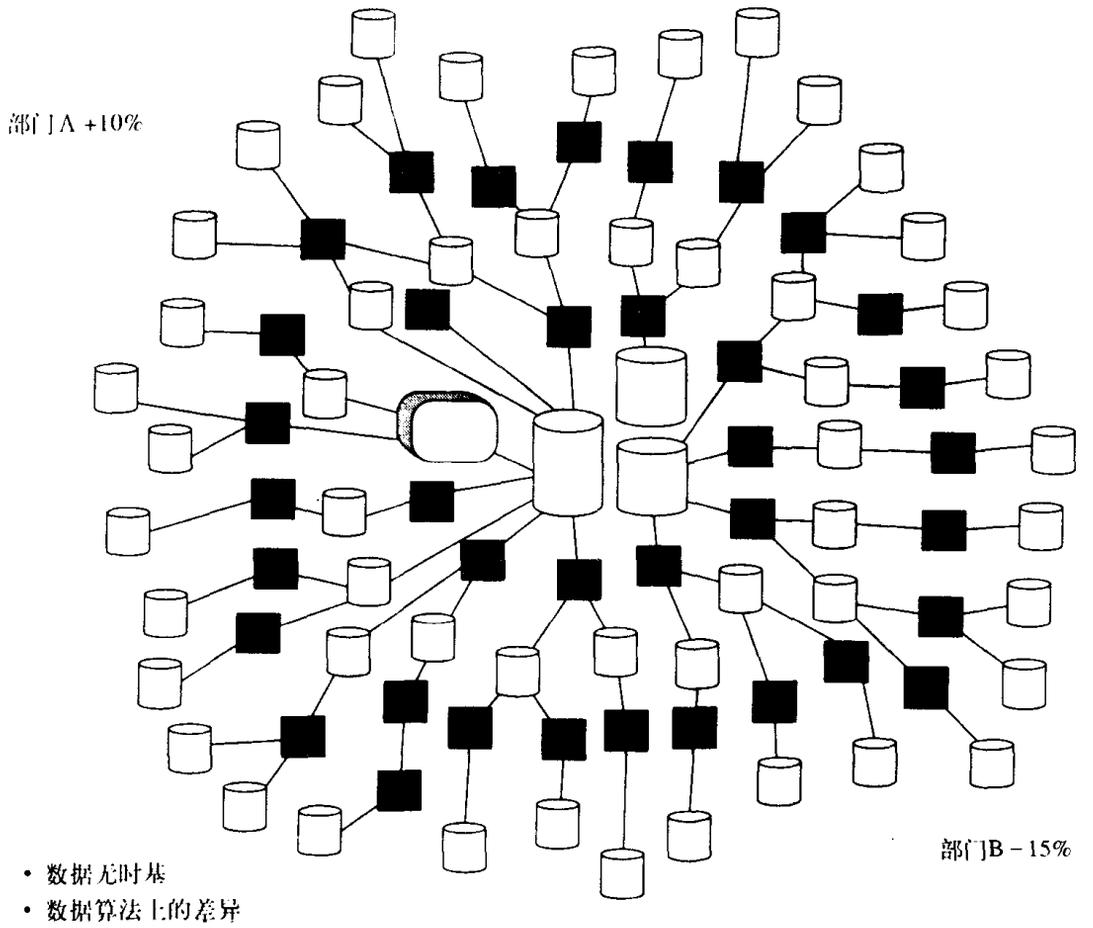

图1-4 在自然演化体系结构中缺乏数据可信性

图1-5显示一个部门在星期日晚上提取分析所需的数据，而另一个进行分析的部门在星期一下午就抽取了数据。有任何理由相信对某一天抽取的数据样本进行的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！公司内的数据总是在变的。任何在不同时刻抽取出来用于分析的数据集之间只是大致相同。

在自然演化体系结构中，数据可信性危机具有可预见性的第二个理由是算法上的差异。

比如，一个部门选择所有的老帐号作分析。而另一个部门选择所有大帐号作分析。在有老帐号的顾客和有大帐号的顾客之间存在必要的相关性吗？可能没有。那么分析结果大相径庭就没有什么可大惊小怪的了。

可信性危机可预见性的第三个理由是前两个理由的扩展。每次新的抽取结束，因为时间和算法上的差异，抽取结果就可能出现差异。对一个公司而言，从数据进入公司系统到决策者准备好分析所采用的数据，经过八层或九层抽取不是罕见的。

缺乏可信性的第四个理由是由外部数据引起的问题。利用当今在PC层次上的技术很容易从外部数据源取得数据。在图1-5所示的例子中，一个分析人员从《华尔街日报》取得数据放入分析流中，而另一个分析人员从《商业周刊》中取得数据。分析人员在取得数据之时所做的第一件事就是从大量外部数据中抽出所需要的部分。数据一旦进入PC，就不再属于《华尔街日报》了，而简单地变成了可能出自于任何数据源的普通数据。

并且，从《华尔街日报》取得数据的分析人员对从《商业周刊》中取得的数据是一无所知的，反之亦然。这就不足为怪，外部数据导致自然演化体系结构中的数据缺乏可信性。

导致数据缺乏可信性的最后一个因素是通常没有一个公共的起始数据源。部门A的分析工

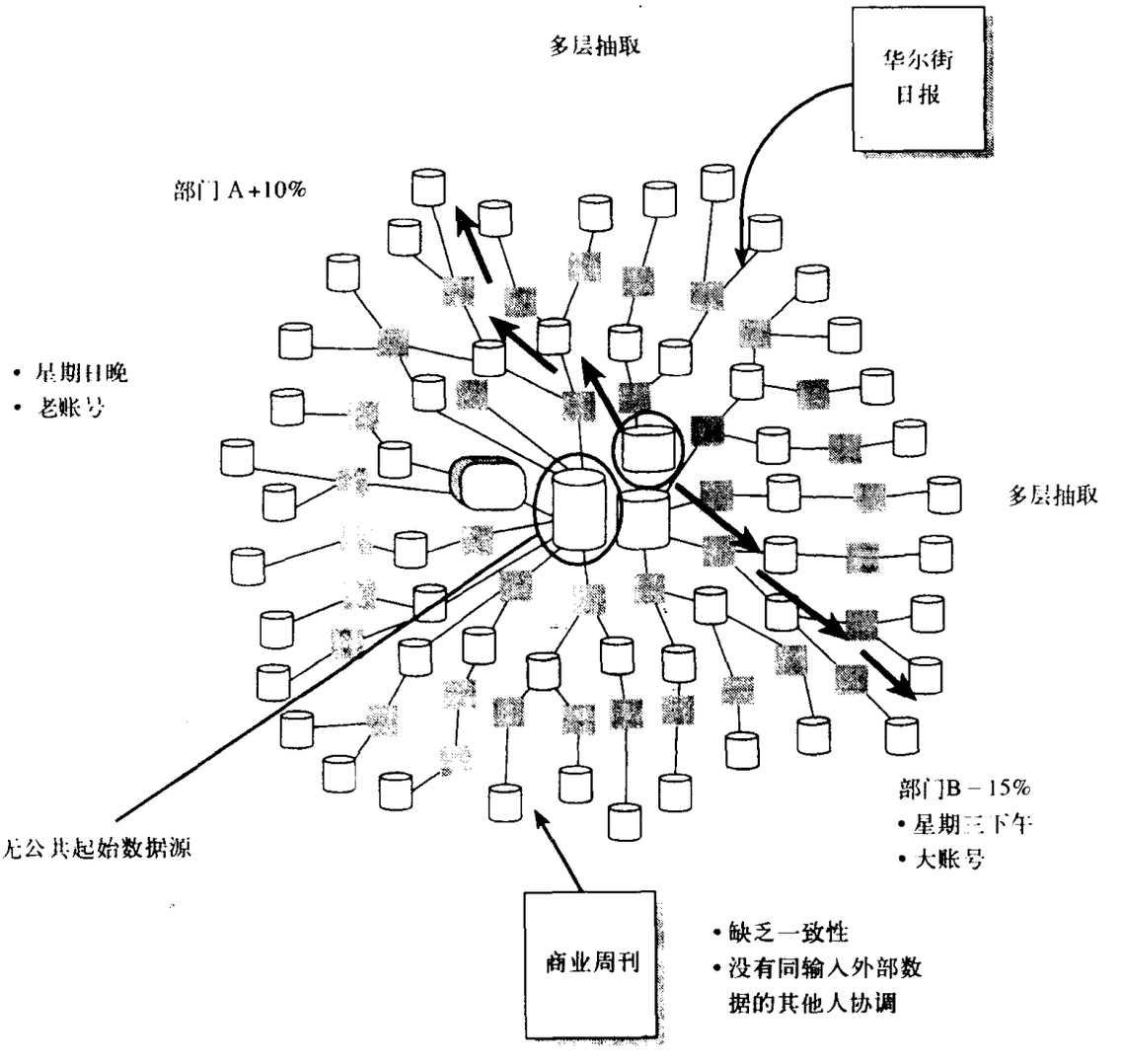

图1-5 自然演化体系结构中可信性危机可预见性的原因