HEWLETT PACKARD

PTR PH

计算机技术详林精选系列

企业数据仓库

规划 建立与实现

〔美〕Eric Sperley 著
陈 武 袁国忠 译

计算机技术译林精选系列

企业数据仓库规划 建立与实现

[美] Eric Sperley 著

陈武 袁国忠 译

人民邮电出版社

计算机技术译林精选系列

企业数据仓库　规划　建立与实现

◆ 著　　[美] Eric Sperley
译　　陈武　袁国忠
责任编辑　俞彬

◆ 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@ pptph.com.cn
网址　http://www.pptph.com.cn
北京汉魂图文设计有限公司制作
北京顺义向阳胶印厂印刷
新华书店总店北京发行所经销

◆ 开本:787×1092　1/16
印张:16.25
字数:395千字　　2000年8月第1版
印数:1－5 000册　　2000年8月北京第1次印刷

著作权合同登记　图字:01－1999－2566号
ISBN 7-115-08601-X/TP·1686

定价:35.00元

前　言

MIS 管理人员和 CIO 的主要目标是使他们的 IT 公司与他们的业务相适应。然而，大多数 MIS 管理人员和 CIO 在技术上训练有素，而在公司策略的把握上则不是那么驾轻就熟。而且，已有的旧系统并不是按照一种有利于集成不同系统中数据以提供新的信息的方式而组织的。因而，通过修改运行系统，并将 IT 与商务有机地结合到一起。提供新信息是具有一定难度的。IT 专家最终会处于这样一种状况：进行改变的需求已知，但是他就是不知道如何选择一种策略或用当前的技术实现这些改变。这种情况就类似于在一个有鲨鱼的水池中——明知有危险，但是不知道它在哪儿以及如何躲避它。

笔者已经编写了二本与这些工作有关的书籍，以帮助读者获得使用数据仓库和开放系统的方法以及使 IT 部门与公司目标相适应所需要的一些知识。许多其他的数据创建方面的书籍已经从较高的层次上介绍了数据仓库的优点和目标，本书将为读者介绍规划、设计、构建和使用数据仓库的详细内容。

多年来，人们一直在各种各样的会议上讨论数据仓库、决策支持系统和执行信息系统。但是仍然有许多专家对数据仓库是什么没有一个清楚的了解。对于那些了解数据仓库基础的人来说，也难以成功地建立数据仓库。已经成功建立了一些项目的数据仓库从业人员发现他们建立的数据仓库不能很好地协同工作。对于其他的情况，用于建立小型数据仓库的技术并不能适用于建立大型数据仓库。最后，他们仍然需要处理在不能很好缩放的旧系统中存在的问题，并处理那些难于集成的数据。

本书作者的目标是介绍一种方法学，以使 IP 专家能够躲过“大鲨鱼”。本书将提供一种描述性策略，以帮助 IT 专家规划、设计和构建企业级的数据仓库。为了理解 IT 社区的当前状况，在第一章中将介绍商务信息技术的发展和历史。在理解了当前的挑战和机会之后，将考查和评价一下数据仓库和运行系统的相对特性。然后在后续章节中介绍一种易于理解的方法学，以便在合理的公司原则和 RAD 技术基础之上建立数据仓库，并作进一步地阐述。

在第二章中，笔者介绍了几种确定公司或 IT 组织的当前地位和发展方向的方法。尽管靠这一本关于数据仓库的书并不能把读者转变成一个公司策略的专家，但是我们可以看一看阐述和理解公司策略的方法，

以及如何选择相应的 IT 策略。作为 IT 专家，我们需要学习的最重要的技术是利用公司经理会谈和应用程序合作开发会议，来了解公司所处情况与经理想要它达到的情况之间的差距。最后，介绍评价建立数据仓库的相关费用的方法。

如果你搭乘一条商用航线上的一架飞机，而飞行员告诉你说他知道如何驾驶飞机，但是不知道要飞向哪儿以及在哪儿着陆，那么你可能会要离开这架飞机。规划对于飞行和数据仓库同样的重要。在第三章中，我们将介绍一些构建数据仓库的方法，以了解构建我们规划的数据仓库需要花费多少，以及需要为这些费用交付什么东西。这将使我们不会像哥伦布一样，出发了但是不知道要去哪儿，到达后不知道自己到了哪儿，并且都是依靠借的钱来做全部事情。

选择具有最大组织影响的数据仓库项目并获得成功是本书第四章的重点。第二章中介绍的 JAD 技术将用来发现对公司具有最大利益的项目，以及该项目的范围。数据仓库的主要目标是为公司知识人员提供信息。由于数据是按照一种有意义的方式组织，并呈现在一种公司环境中，所以它是成功数据仓库的关键。第五章重点介绍数据体系结构和数据构模的原则和指导方针。到第五章结束时，初级数据构模人员应该理解企业和决策支持数据模型的基本组件，而有经验的数据构模人员则将会更好地理解如何将已有的技术扩展到新的领域中。

对于一个数据仓库的成功，理解数据仓库中的数据是非常重要的基础。数据仓库项目失败的一个主要原因是错误地理解了数据仓库中的数据。关于数据仓库中数据的数据叫作“元数据”(metadata)。成功的数据仓库项目使用成功的元数据仓库进行连接。在第六章中我们将介绍构建元数据仓库的意义、理由和方法。

数据仓库项目失败的第二个主要原因是在数据仓库中缺乏高质量的数据。第七章将重点介绍在数据仓库中获得高质量数据的方法。不理解高质量数据的价值，管理层很难投资获得这种数据所需要的资源。因而本章将从一个范例开始，计算数据仓库中数据错误的花费。然后将介绍一种达到这种数据质量的方法。

理解一个研究领域的原理将会使该领域的学生能够应用这些原理解决新问题。在第八章中，我们将学习数据仓库体系结构的原理。这些原理将用来构造一种概念数据体系结构。然后将应用这个概念数据体系结构模型来建立一个逻辑数据仓库。

第九章专门用来帮助理解物理数据仓库。在这一章中，将分析物理数据仓库不同组件的任务、协定和折衷。

软件用于把数据仓库连系到一起并使得其构建成为可能，而第十章的主题就是软件。数据提取、转换和整理软件工具对于数据仓库的构建非常重要。在本章中将介绍这些工具的重要特性。

数据仓库一旦建立之后，必须为数据仓库客户提供适当的工具，以访问该仓库中的数据。第十一章介绍了不同类型的访问工具，并为读者提供了安全地选择适当工具所需要了解的知识。

最后，在第十二章中介绍了数据采集。有几种不同的数据采集方法。在这一章中介绍了数据采集的所有主要方法，以及每一种方法的优点和缺点。

作为一般的声明，本书是为那些有兴趣建立或理解决策支持系统的IT 专家而写的。尤其是 CIO、IT 经理、数据分析员、数据库管理员、设计人员和开发人员将对本书产生兴趣，并觉得它有用。CIO 和 IT 经理将会发现第一至四章特别有用。经理、数据分析员、数据库管理员、设计人员和开发人员将会发现第五到十二章对于数据仓库的实际实现很有帮助，而第一到四章将帮助他们理解其管理的路径。

本书的另一批可能的读者是管理信息系统、商务和计算机科学等专业的学生。笔者花了五年的时间进行教学，发现在本书中可以很容易地包括一些信息，使该书成为一本关于数据仓库方面极好的教科书。在许多章最后都有专门的一节列出了许多练习题，以供正式的和非正式的学生使用。许多章还提出了一些项目，读者既可以把它们当成思考练习题，也可以把它们当成实际的项目来解决。

·作者·

第一章　信息技术简史

1.1 简　介

在本章中我们将调查信息系统组织[1]的当前状态以及如何了解它们，并开始介绍强调策略性企业思考和分析的方法学。用最简单的话说，数据仓库的一个目的就是把公司的信息访问基础从一种非结构化的或发展中的环境改变成一种结构化的或规划良好的环境。这种新环境将提供满足企业需求所需的关键信息。把信息移动到满足企业需求的结构化环境中只是企业数据向开放式灵活体系结构过渡的第一个步骤。数据仓库是在开发满足企业信息需求的结构化环境的过程中的首要步骤之一。

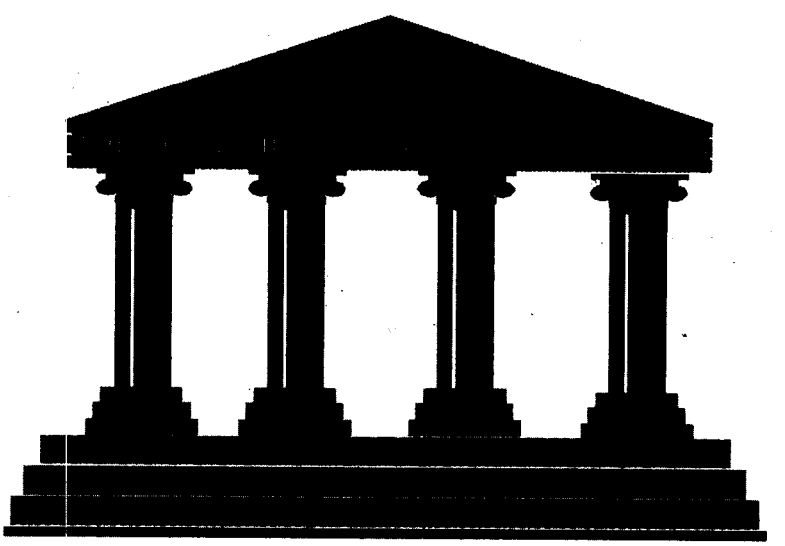

图 1-1　易于识别的古希腊建筑

[1] “组织”（organization）和“企业”（business）在本书中均可以互换使用。许多政府、教育和非赢利性组织都使用数据仓库技术。本书将试图包括所有使用数据仓库技术的情况。

那么什么是“体系结构”（architecture）呢？根据定义，体系结构就是对某些事物结构的研究。处于研究下的体系结构通常具有一些不同的特征。例如，希腊建筑的柱子和屋顶布局就与众不同。大多人看到图 1-1 中所示的结构之后，将会认为它是希腊建筑。而对于数据仓库，其重点将在于信息技术（Information Technology，IT）[1] 部门如何构建事务处理和决策支持系统（DSS）。决策支持与事务处理相结构，就是人们所说的“信息管理”（information management）。研究和设计某种东西结构并参与设计的人叫作“设计师”（architect）。本书的主要目标就是帮助读者成为一名优秀的信息管理和数据仓库设计师。经验丰富的数据仓库设计师将会设计出能以一种高效的方式提供企业信息的决策支持系统。

发展中的或规划不合理的环境的存在，主要是由于在实现最初系统时技术的局限所致。这些早期的系统设计成使用卷到卷的磁带存储和普通文件技术等来满足企业处理需求。尽管现在很容易看清楚这些系统，并奇怪为什么它们会被设计成这样，但是这些老系统中的大多数都使用了当时可用的最好技术。这些系统成功地提高了企业的能力，并获取了利润。最后，很显然它们对于公司的日常运行是如此之重要，以至于公司经理非常关心对这些系统的任何改变或改造是否会导致公司业务的中断。这就是为什么这些系统今天仍然在使用而没有重新开发新系统的原因。这些老系统仍然在通过逐渐地修改发展变化，以添加新的特征，例如 2000 年兼容性。

最初，开发信息技术项目的原因是因为它们与全人工系统相比节省了费用。后来，开发信息技术项目的原因是它们除了能减少运行费用之外，还会给公司提供战略性价值。这种对于信息技术的新思维方式导致了数据仓库的构建，数据仓库为公司提供了战略性和策略性价值，而不仅仅是采用一种新技术。在本章后面，我们将会介绍决策支持系统的特征与运行系统的特征有何不同。

1.2 IT 简 史

世界的历史是一门对事业进行投资以产生更多财富的学科。在农业时代，人们使用金钱来购买土地，并购买或雇佣人来管理和耕种土地。土地上长出庄稼并养育着牲畜，通过卖或交换，它们为财产所有者产生更多的财富。这些财富可以用来获得更多的土地。这些财产的所有权和控制权对于产生利润是非常关键的。由于土地是产生财富的关键，富有的人使用他们的财产在土地上投资，产生更多的财富。

在工业革命之后，人们开始把钱投资到生产物品的工厂中。这些工厂利用制造设备和工人生产物品。生产物品的能力比钱更为重要，因为有些人虽然有钱，但是没有能力制造出能生产出产品（他们可以从产品中获得更多收入）的机器。例如，西班牙人使用他们

[1] 术语“信息技术”（Information Technology，IT）和“信息管理”（Information Management，IM）也可以互换使用。IM 是一个更新的并且更加适当的术语，它实际上就是组织中负责计算机的部门所做的事情。

“找到的”黄金从英国购买船只。在西班牙舰队被摧毁之后，英国仍然有工厂通过生产物品得到收入，但是西班牙再也没有用以获得收入的装备了。在那个时候，这两个王国的相对繁荣戏剧性地对换了一下。

当前，人们和公司把钱投资于硬件、软件和个人，以创建和获得信息。现在使用信息来获得金钱或节省金钱。例如，人们可以使用万维网（World Wide Web）来为某一次旅行查找最低价格的飞机票。在过去，人们只能通过旅行代理得到关于机票价格的信息。而旅行代理可能会出于某些动机，出售不是最低价格的机票。现在人们可以访问信息以节省金钱或找到更精确地满足他们需求的产品。不为其客户提供相关信息的公司面临着失去客户的危险。除了提供更好的服务之外，公司还使用信息来吸引和留住客户。因为资金被投入到信息技术中，期望得到一定利润的收益，所以 IT 项目产生预期投资收益（ROI）是非常关键的。数据仓库成为增加利润的一种工具，用来减少费用、避免将来的花费或通过为公司决策人员提供信息以增加收入。

如图 1-2 所示，与其他专业相比，信息技术的历史相当短暂。其他学科，例如医学、天文学、工程学和建筑学，都具有相当长的时间来发展它们自身的技术。考古学家们相信大脑外科手术最早开始于公元前 2000 年，而内科医生的“希波克拉底誓约”则表明在公元前 400 年就已经有了很好的内科医学实践。另一方面，ENIAC 计算机出于 1946 年。这

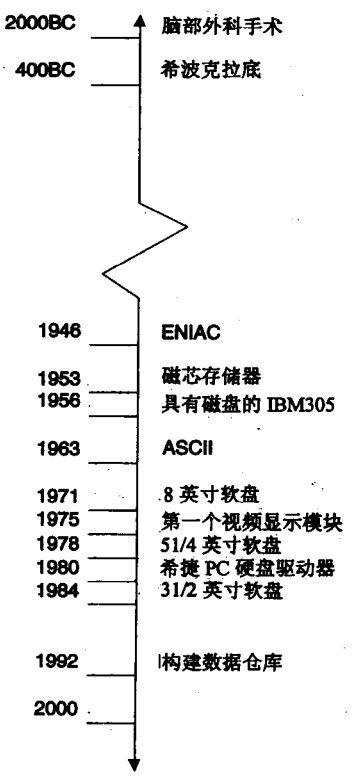

图 1-2 历史时间线

意味着人们还没有太多的时间来开发最好的进程和过程，以开发信息系统。人们所学的内容与当前技术紧密相关，并且随技术的过时而过时。从时间线上可以看到，计算机技术变化非常迅速。这就是信息技术为什么既让人兴奋又让人沮丧的原因。正在建立的新系统要好于任何已经建立的系统，但是明天又将可能建立一个更好的系统。

在信息技术部门的早期历史中，每个新项目的选择和目标都非常明确。一些人工密集的任务被自动化系统所取代，由于减少人工和提高速度而得到的费用节省很“显然”地证明该项目是适当的。原来要用手工过程收集的数据将在一个自动化过程中收集，这个自动化过程将支持公司部门的运行要求。不幸的是，这些项目很少在 IT 小组与公司用户之间进行交互。项目的成功和失败取决于 IT 设计小组对公司理解的程度。在多数情况下，IT 小组都具有进行公司自动化方面的体验，并且交付给公司的最终产品相当成功。在其他一些情况下，IT 小组完成详细的需求分析，并且交付的最终产品也相当成功。但是也有另外一些情况，IT 小组对于公司需求的理解非常有限，他们付出了大量的努力，但是交付给公司客户的最终产品却不太成功。IT 专家不是与将要使用该系统的公司专业人员进行讨论，而是不断地猜测系统将要实现的过程，或与少数的公司专家会谈。如前面所述，项目的目标看起来没有公司用户的帮助也很清楚，因此，必须实现的过程好像也很清楚。

在 20 世纪 60 年代，通常使用过程驱动的方法。因为假定企业所使用的过程就是需要自动化的过程，所以 IT 设计小组分析的起始点就是企业过程。用于支持当前企业过程的数据被标识出来，并设计到适当的系统部分中。尽管企业过程设计到了系统中，但是在 IT 的这个阶段，人们并没有广泛认识到企业可以使用数据的多种不同方法。还需要花费一段时间，才能认识到企业过程将会随企业的发展而发生重大变化，而企业运行所需要的数据也会随时间而改变。例如，银行多年来一直记录存款、提款、收支余额和客户。对于市场的原因，银行帐户的类型和银行帐户的规则发生了改变，但是维护帐户所需要的数据没有改变。在人们思想上发生这种改变之前，过程驱动的方法一直占主导地位。

在 20 世纪 50 年代早期，磁带是唯一广泛可用的大容量存储设备。对于数据处理的本质和任务，磁带有着许多局限。因为磁带仅存储相对较少量的数据，存储在一个磁带上关于特定主题的数据必须限制为单项数据处理任务所必需的数据。从而二个使用部分类似数据元素、部分不同数据元素的不同处理任务通常将需要使用二个不同的磁带“主文件”。因为相同信息将存储在二个不同磁盘主文件上的不同位置中，所以这些“相同”的信息在不同的磁带主文件上时可能会有所不同。例如，一个职员在人力资源主文件中的家庭地址可能与工资主文件中同一职员的家庭地址大相径庭。类似地，在没有制定计划或公司成功所需要的全部信息系统的开发体系结构的情况下，开发出来了许多不同的系统。

通常，新系统在开发时没有考虑已经在其他系统中使用的数据类型和功能。这将会导致这样一种情况，其中不同的运行系统使用不同的数据类型、代码和长度来代表相同的信息。而且，很显然相同的数据在不同的系统中具有不同的意义。例如，看似简单而良好的数据字段地址可以具有许多意义。如货物发送地址、帐单地址和邮寄地址等都是各不相同的。显然，销售系统将具有一个邮寄地址，定购实现系统则具有一个货物发送地址，而帐目支付系统则具有一个帐单地址。

因为读取磁带上信息的物理过程涉及到磁头在磁带上的移动，所以磁带上的信息必

须以顺序的方式读取。因为磁带是一种顺序访问存储设备，所以不可能将磁带 2 上第 5 条记录中的信息与磁带 121 上第 700 条记录中的信息高效地进行组合或比较。这将会使比较一个磁带文件中不同位置处的记录或不同磁带文件上的记录相当困难或非常耗费时间。这种不能随机选择数据记录的缺点，再加上公司中许多不同运行部门中过程自动化所需要的驱动不同，导致了每个运行部门都有它自己的磁带主文件。

如果不同部门的主文件已经合并到一起，并且公司的事务处理转移到了新合并的主文件，则记录使用的磁带长度将会增加，每条记录的处理时间也会相应增加。这意味着不同的公司部门将需要花费更多时间和金钱处理给定数量的记录，才能与公司其他部门共享信息。这时批处理是规范，而不同的批处理将被分配到一个时间窗口中（在其中完成批处理）。记录长度增加导致的处理时间增加可能把批处理挤出分配的窗口。所有这些因素，再加上重新编写计算机程序与合并的磁带格式相匹配的难度，使得大多数 IT 专业人员不愿意选择这种方式。

1.2.1 企业信息筒仓

企业运行部门进行信息系统的开发导致了信息筒仓（silo）或信息井（well）等计算机系统的开发。这些系统可以提供公司特定部门中的详细数据，但是它们不能从公司其他部门中集成信息。因为没有关于公司状况的单一、集成的数据源，所以很难（如果不是不可能的话）通过查询这些筒仓来得到公司的全貌，如图 1-3 中所示。

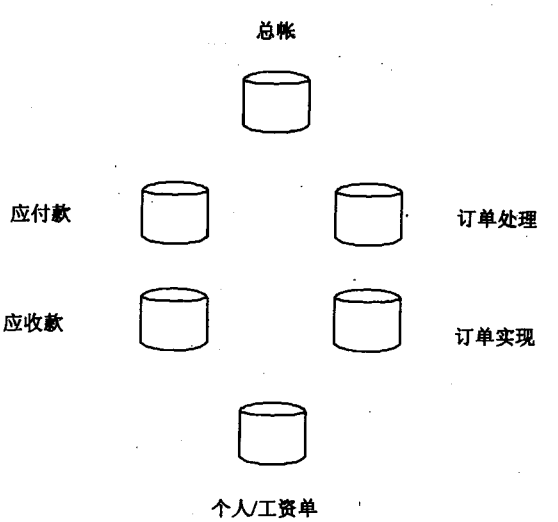

图 1-3 分离的但未必是性质不同的筒仓

作为一个范例，考虑一家电信公司可能需要维护的文件。许多电信公司提供本地电话、长途电话、移动电话、寻呼机和数据等服务。通常，每一种服务有一个运行系统，记录客户、使用率、应付款、实收款和服务条款。每个运行系统在建立时都没有考虑存储在其他运行系统中的数据。这些运行系统无法为一个用户提供一个单一视图，显示他所购买

的那些服务。

为运行部门建立信息系统和数据筒仓不是 IT 部门的唯一功能。大多数公司都组织为筒仓。作为一个范例，考虑一家计算机制造厂商在这个时间的结构。1980 年左右典型的计算机公司都有生产微处理器的部门，生产整机的部门，生产显示器的部门，以及生产其他硬件的部门。通常计算机公司还将开发操作系统和应用程序，并且具有它自己的销售和发货部门。在图 1-4 中描述了这种情况。

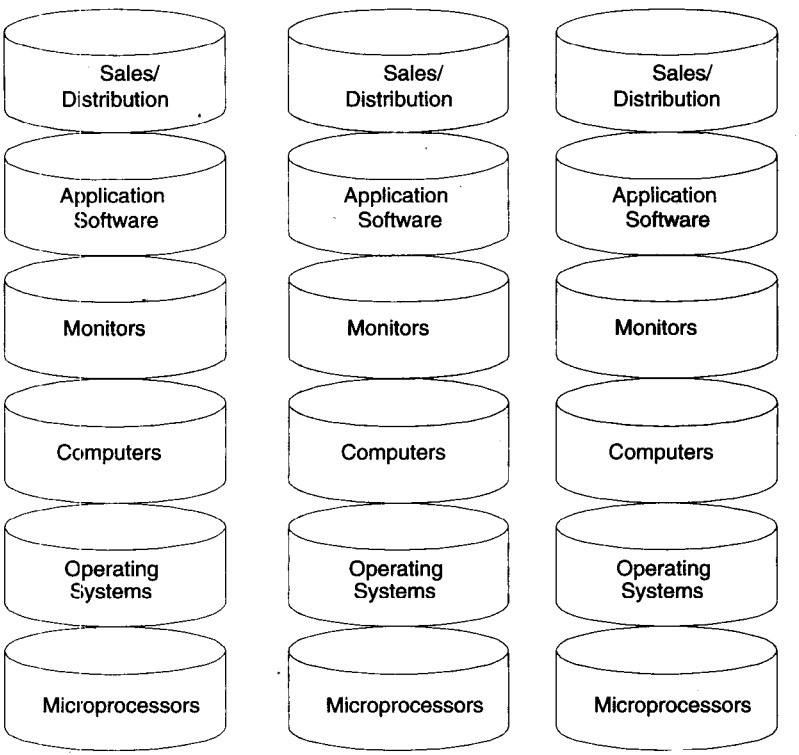

图 1-4 整体的公司运行筒仓

不同的信息筒仓所包含的信息若组织到一起将具有相当大的价值，所以 IT 部门需要编写数据提取程序。这些程序，根据报表的请求而编写，来自公司不同的运行部门。提取程序将从不同的主数据集中读取数据，把数据存储到一种中间数据设备中，并使得数据可作为报表程序的输入。然后报表程序使用其中的数据生成想要的报表。

随着到达 IT 部门的报表请求越来越多，IT 部门需要不断地编写新的提取程序。与需求相比，IT 专业人员总是不足，所以提取和报表程序中的滞后时间将会增长。随着公司前进步伐的增大，滞后时间也随之增加。公司信息客户不断地请求报表的更新和新功能。这将会导致提取和报表程序在其流图中的结构不合理，并且程序也将会变成意大利细面条式的代码。在图 1-5 中描述了一个没有经过规划的决策支持环境。大量的意大利面条式代码提取程序、中间数据存储设备和报表程序，共存于一个未经规划的决策支持环境中，使

得 IT 部门需要处理太多的内容。

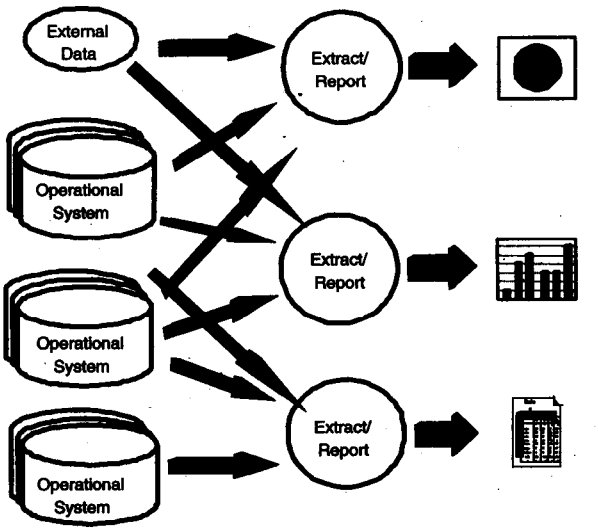

图 1-5　未经规划的决策支持环境

随着 1956 年磁盘引入到 IBM 305 计算机中，直接访问存储设备（direct-access storage devices，DASD）变成现实。记录不再需要按顺序方式访问。只要有足够的空间，就可以很容易地同时处理位于不同硬盘上不同文件中的记录。这种技术允许位于相同或不同文件系统上的记录进行比较、组合，或用于更新其他记录。最初，它较小的存储容量和这种存储技术的高费用限制了它的发展。

1970 年左右，一些研究人员认识到位于许多不同主文件中的数据是相同的或相关的。在有些情况下，数据是或者应该是相同的。在另外一些情况下，不同类型的信息存储在不同的主文件中。还有一些情况下，信息位于不同的主文件中，或者相同类型的不同数据位于不同的主记录中。在最糟的情况下，本应该相同的信息在不同文件中为不同的值，因为这些文件以不同的方式单独维护。

Codd 第一次对不同数据元素可以互相关联这一概念进行了严格的研究。数年来在实际关系数据库产品出现前，它只是关系数据库技术的数学处理。实际上，Codd 最初提出的关系数据库的理论包括一些严格的数学思想，例如域，它们在当今的主流数据库软件产品中仍然没有实现。直到现在，许多大型操作系统仍然使用普通文件或类似的技术。简单地说，在很长一段时间内，理论远远领先于实践。理论和实践都在向前发展，但是理论不必受财政预算和保持公司运行等实际因素的制约。因而，实践赶不上理论的发展。

关系数据库技术许诺改进 IT 专业人员生活的一种方式是通过减少冗余数据的量。公共数据可以放入一个公共表或实体中，而特殊的或重复的数据可以放入相关实体中。许多不同的运行系统可以使用相同的数据以满足不同的运行需求。对于不同运行系统都相同的数据可以维护在一个实体中，所有运行系统都将访问这一个实体以得到必要的信息。关系

数据库技术与直接访问存储设备相结合，使得改革构建 IT 解决方案的方式成为可能。尽管在理论上相当不错，但是在实践中很难实现。运转公司的系统设计在普通文件基础之上。对于大多数公司，把公司的运行系统从旧的普通文件系统移植到关系数据库的任务相当艰巨。

在 20 世纪 70 年代早期，产生了 EIS，即主管信息系统（executive information systems）。这些系统通常由连接到主机计算机的哑终端组成。它们使得公司知识人员可以查看从运行系统中提取出的数据。EIS 提供了把数据组织成信息的方法，但是显然它们没有能力超出运行系统外部，并从全新的报表中提取新数据。如果一名主管想要数据的一个新视图，则他或她需要等待。根据 IBM 主机和兼容主机的一次调查的结果，IT 部门中的平均滞后时间为 13.4 月（Hosier）。

在 1975 年左右，出现了第一批桌面式 PC 和桌面工具。尽管这些早期的 PC 机在能够存储和处理的数据量上非常有限，但是它们让早期“有能力的用户”尝到了没有 IT 部门帮助设计报表也能对数据进行分析的滋味。许多公司信息工作人员发现他们可以访问 IT 部门从运行系统中提取出的数据，或把报表或“绿屏”中的数据输入到一个 PC 数据库中。信息工作人员还可以从外部数据源（例如 Wall Street Journal 或 Business Week）中输入数据。这些数据可以添加到存放有公司数据的数据库中，以便添加到决策过程中。

公司数据的用户现在又步以前 IT 部门的后尘，开发他们自己的未经规划的决策支持系统。不同部门中的人们用他们自己的结构创建自己的数据库。不同用户对数据的定义和字段的计算方式都不同。这些新数据库包含有来自外部源的数据，并且不同的用户选择不同的外部源。因为有些自动化过程并不更新用户驻留在其 PC 机上的数据，所以用户的数据很快就过时了。如果用户在意复制当前数据，他们将需要花费大量的时间来维护他们自己的数据库。如果用户没有更新其本地数据，则可能会从错误数据中得出错误的结论。

公司用户试图工作于一种非结构化老环境中面临的第一个挑战是理解这些数据筒仓之间的关系。如图 1-6 中所示，这些数据筒仓并没有一种标识筒仓之间相同记录的唯一方法。对于不同数据筒仓中看似相同的数据的迥然不同的定义，给公司用户和 IT 社区造成了巨大的混淆。商业专业人员认为 IT 专业人员不对，因为对于相同数量应该产生相同数字的两个不同报表却得出了不同的数字。从而导致这样一种情况，其中公司的决策人员在会议上，对于应该相同的数量报告了不同的数字。

尽管人们的第一个想法可能是 IT 部门不能胜任其工作，或公司同事的沟通不够直接，但是通常情况并不是这样。报表上的数字是使用不同源系统中的不同数据计算得到的。通常，不同源系统中数据的定义方式不同。在其他一些情况下，在不同时期报告公司状况的二个相同报表可能会不一致。通常，这可能是由于生成两个报表期间进行了数据更新而引起的。产生不同报表的另一个原因是外部数据源不同。一个人可能使用了 Wall Street Journal 中的利率数字，而另一个人则可能使用 Rutgers 中的利率数字。因而，导致两个经理的报表中数字不一致的原因有许多。这种情况的结果是公司花费了大量的时间在数据的一致性上，而只花费了很少的时间对数据进行操作。

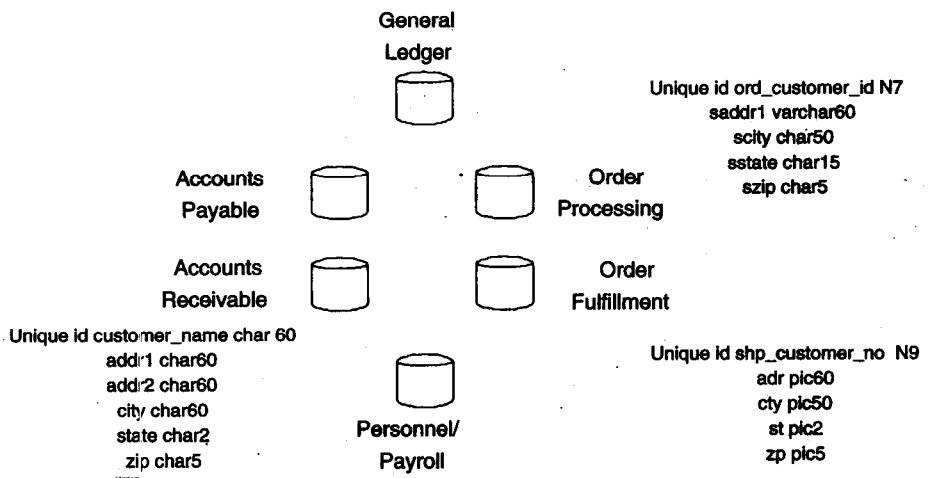

图 1-6 不同字段在非结构化环境中导致混淆

产生混淆的根本原因是决策支持系统的本质未经规划或未经结构化。如果公司提前考虑并规划报表，则可以通过一次提取或一套公共提取，得到产生所有报表所需要的全部数据。这些数据可以存储到一个中央位置。由于大多数报表都是独立开发的，所以不可能规划中央数据位置。

聪明的读者可能会认为关系数据库将解决数据差异或数据混乱的问题。尽管关系数据库技术可以在很大的程度上帮助开发新的运行系统，但是这种技术并不能解决数据筒仓和未规划的系统开发等问题。运行系统仍然在没有考虑其他系统的情况下进行开发，未进行相关设计而开发的系统没有公司的唯一标识符或编码方案，因而关系数据库技术也不能解决这个问题。

总的来说，IT 的情况仍然在变得越来越糟。维护意大利面条式数据提取和报表程序变得越来越困难，但是它变困难的速度非常慢，不足以引起任何人的重视。在高中生物课堂上，人们经常说一个关于青蛙的故事。当把青蛙放入一盘热水中时，它总是跳向安全的地方；但是如果把青蛙放入一盘逐渐加热的冷水中，则有可能将青蛙煮熟。类似地，IT 社区所处的状况也正在逐渐加热，总有一天该社区突然发现太热，难以舒适地继续生活下去。

1.3 数据仓库是什么

许多人都为数据仓库这一思想的出现作出了贡献。尽管早期的作者和思想家写了许多初级的作品，但是大家公认 Devlin 和 Murphy 在 1988 写的一篇文章是第一篇关于数据仓库的论文。1993 年，William H. Inmon 编写了一本具有里程碑意义的书“Building the Data

Warehouse [Inmon]”，在这本书中，他定义数据仓库为“一个面向主题的、集成的、随时间变化的非易失性数据的集合，用于支持管理层的决策过程。”

“面向主题”（subject oriented）意味着数据仓库设计成提供与公司单个过程相关联的数据。作为一个与多数人有关的范例，考虑一所学院或大学存储在不同旧系统中的数据。一所大学可能有应付款、应收款、工资单、助学金和注册系统。注册系统通常保存相当有限的历史数据，并保存相当有限的关于班级中所招收学生以及关于教授各班级的教师的数据。班级注册主题区域将具有关于班级中学生注册和教授该班级教师的历史数据。而注册运行系统将具有学生获得的最终学分，班级注册主题区域则可能与这些细节没有关系。

作为另一个关于数据仓库的面向主题本质的范例，我们回到前面所介绍的通信公司的范例，考虑一家通信公司可能保存的数据。如前所述，许多通信公司提供本地电话、长途电话、移动电话、寻呼机和数据服务。每一种服务都由公司一个部门负责运行，该部门具有自己的运行系统，记录客户、应付款、应收款和服务条款。没有一个数据包含关于客户的所有信息。客户主题区域的构造将为客户提供关于该客户所购买的全部服务的单一视图。这个客户数据仓库将使得通信公司可以根据客户的统计资料和以前的消费趋势，对当前客户推销新的产品。

读者可能已经从未规划的决策支持系统的困难推断出数据仓库的集成性是其最重要的特征。随着多年来不同运行系统的开发，也相应使用了不同的数据类型、大小和编码方法。例如，有些系统把月、日、年编码为 MMDDYY，而另外一些系统则可能使用 DDMMYYYY。再例如，一个系统中的性别字段可能被编码为 m 和 f，而另一个系统中的性别字段则可能被编码为 1 和 0。

当把数据从运行系统移植到数据仓库中时，在把它装入到数据仓库中之前需要对它进行转换。去掉源系统的编码、数据类型和其他“特征”中的不一致性。这时还是检查源系统数据并消除文档数据类型和代码与实际存在的代码之间的不一致性的极佳时机。一般来说，数据仓库的构建将会发现从运行系统中除去不想要“特征”的机会。

数据仓库与运行系统相区别的另一个重要特征是数据仓库的非易失性。当运行系统对存储的数据执行更新、删除和插入操作时，数据仓库装入大量的数据，并且从不会改变它已经加载了的数据中的值。一旦数据加载到数据仓库中之后，它就不能被改变。两个不同的公司用户在不同时间运行相同的查询时，将会得到相同的结果。这样就消除了未规划的提取和报表产生不同结果的情况。

作为数据仓库的最后一个特征是它随时间而变化。尽管运行系统只包含有当前数据，但是数据仓库系统包含有历史数据和上一次数据仓库加载时的当前数据。数据仓库存储的历史时间范围随系统类型的不同而不同，但是 15 个月到 5 年的时间范围是非常常见的。数据仓库中最老的数据通常在它不再适宜存在于数据仓库中时，归档到磁带或光盘上。

运行数据系统和数据仓库信息系统具有大量的相对立的特征，最好将它们放到一起进行对比。表 1-1 中列出了二种系统类型的重要特征。

表 1-1 数据仓库与运行系统比较

数据仓库系统	运行系统
管理层使用	一线工人使用
战略性价值	策略性价值
支持战略定向	支持日常操作
用于事务处理	用于联机分析
面向应用程序	面向主题
仅存储当前数据	存储历史数据
可预测查询模型	不可预测查询模型

数据仓库被建立到无数的主题区域中。这些数据仓库的大小和使用在很大程度上取决于公司的类型以及数据仓库需要提供的公司信息。下面是一些最常见的数据仓库应用程序：

- 风险分析。
- 财务分析。
- 欺诈分析。
- 营销关系。
- 资产管理。
- 呼叫行为分析。

惠普公司有一个叫作 Hewlett-Packard OpenWareHouse 程序，为客户提供了最好的硬件、软件和服务。它的目的是为建立一个数据仓库提供所需要的一切单步采购。一旦决策使用 OpenWarehouse 程序，关系就已经建立好，以交付建立一个成功的数据仓库所需要的全部硬件、软件和服务。从概念上来说，数据仓库可以按照数据流和存储来考虑，如图 1-7 中所示。

另一种结构，“数据市场”（data mart），也与数据仓库非常类似。“食品市场”比“食品超市”具有更好的选择性和可用性，类似地，“数据市场”是一种微型的数据仓库。数据市场通常具有更少的数据，更少的主题区域，以及更少的历史数据。可以把数据市场看成为数据仓库的一个逻辑上或物理上划分的子集。建立数据市场的目的通常是为了服务于特定用户群的需要。与数据仓库类似，数据市场也包含有运行的详细数据和总结性数据。

另外一种结构在某些方面与数据仓库类似，它运行数据商店（store）。尽管与数据仓库类似，但是运行的数据商店（或 ODS）是按照主题来组织的，并且 ODS 只包含有当前或最近的数据。因为 ODS 通常为职员和个人使用，以满足公司和客户的日常需求，所以它不需要数据仓库必须存储那些供战略决策人员使用的历史数据。ODS 的数据设计通常比数据仓库的数据设计更加标准化。