Microsoft SQL Server 2000
数据仓库设计和使用指南

康博创作室　主编

清 华 大 学 出 版 社

需要全本请访问 www.ertongbook.com

(京)新登字 158 号

内 容 提 要

数据仓库是一种把收集的数据转变成有意义的信息技术。微软公司在 Microsoft SQL Server 2000 中提供了完整的数据仓库解决方案。本书对 Microsoft SQL Server 2000 的数据仓库解决方案进行详细的阐述。本书共分 3 部分 17 章。第一部分主要研究数据仓库的设计,对数据仓库的设计难点进行深入地剖析。第二部分全面讲述数据仓库的使用,对各种工具进行逐个地比较、分析和评价。第三部分重点放在如何优化和维护数据仓库的性能,对涉及数据仓库性能的问题和手段进行了完整的论述。

本书内容详实、示例丰富、结构合理、语言简洁、图文并茂。作为一本数据仓库设计和使用指南,本书主要面向数据仓库和数据库的系统管理人员和从事数据仓库系统应用开发的专业人员,可以指导他们完成数据仓库的管理工作,解决管理过程中碰到的各种技术难题。

版权所有,翻印必究。
本书封面贴有清华大学出版社激光防伪标签,无标签者不得销售。

书　名: Microsoft SQL Server 2000 数据仓库设计和使用指南
作　者: 康博创作室　主编
出版者: 清华大学出版社(北京清华大学学研楼校内,邮编 100084)
http://www.tup.tsinghua.edu.cn
印刷者: 北京市丰华印刷厂
发行者: 新华书店总店北京发行所
开　本: 787×1092 1/16　印张:　　字数:　　千字
版　次: 2000 年　月第 1 版　2000 年　月第 1 次印刷
书　号: ISBN 7-302-01175-3/TP.2225
印　数: 0001～0000
定　价: 00.00 元

前　言

当前信息技术正在从数据处理向数据使用方向转变,数据库技术正在从联机事务处理(OnLine Transaction Processing,OLTP)向联机分析处理(OnLine Analysis Processing,OLAP)转变。数据仓库是一种把收集的数据转变成有意义的信息技术,它通过使用多步进程来执行,这些进程包括收集数据、净化数据和存储数据,以便把这些数据用在分析和报告应用程序中。这种数据可以来源于许多不同的数据源,包括不同的数据库系统,甚至来源于不同的操作系统。数据分析的主要目标是确认数据结构和趋势,帮助制定有关企业经营的更好的决策。

创建数据仓库的原因是因为许多企业越来越依赖于从信息系统中收集信息。因此,为了企业的运转就需要一个信息数据仓库。企业的各种客户也希望能够在许可的情况下访问企业的有关数据。数据仓库可以为各种研究商业运作规律的分析人士提供足够的信息,这些信息在通常的业务数据库中是很难做到的。这是因为,在大多数情况下,企业会将历史数据从业务数据库中转移到备份系统中,这使得许多分析人士无法深入分析数据,难于做出有竞争性的决策。为了更好地管理数据,保持数据的一致性,以及从企业的角度分析数据,这才引出了数据仓库的概念。数据仓库允许企业的各个部门之间共享数据,为企业更快、更好地做出商业决策提供更加准确、完整的信息。

本书是第一本对 Microsoft SQL Server 2000 数据仓库技术进行全面阐述的著作。它具有内容全面、技术权威、结构合理等特点。本指南既涉及了数据仓库的理论,又结合具体的产品实例进行研究,可以指导各种用户迅速掌握数据仓库设计、使用和维护技术。从本书的内容上看,由于作者一直从事数据仓库技术的研究,并且多次与微软(中国)有限公司技术人员进行技术合作,因此所介绍的内容具有权威性。在结构编排上,按照设计、使用、维护的方式进行组织内容,适合读者的思维和学习习惯。

在本书的编写过程中,得到了微软(中国)有限公司有关技术人员的大力支持,在此表示感谢。

本书由康博金典科技开发有限公司策划,闪四清博士主编。

作者

2000 年 11 月

第 1 部分　设计篇

第 1 章　数据仓库技术概述

第 2 章　数据仓库生命周期管理

第 3 章　确定用户需求

第 4 章　设计和建立数据仓库

第 5 章　提取和加载数据——数据复制技术

第 6 章　提取和加载数据——DTS 技术

第 1 章　数据仓库技术概述

信息技术发展的浪潮把人们带到了信息爆炸的时代。每天、每时、每刻潮水般的信息拥到了人们的面前,需要人们去处理去应用。信息的价值在于用户通过使用这些信息从中得到收益。计算机硬件技术和软件技术的飞速发展使得数据处理更加容易,数据处理已经不是今后信息技术的重点,代之而起的信息技术重点是如何使用这些数据信息。如何快速有效地使用这些比大海之水还要多的信息是今后信息技术需要解决的一个难题。数据仓库作为一种高效地解决数据使用的技术,正在越来越多地应用到传统的数据库技术的领域,数据仓库一词也频频出现在有关信息技术的报刊、杂志、书籍、网络上。从现在开始,本书将全面研究使用 Microsoft SQL Server 2000 建立数据仓库的技术。

本章的要点:

数据仓库的概念和特征
DSS 和 OLTP 的区别
数据仓库的构成
数据仓库的工具
建立数据仓库的过程

1.1　数据仓库概述

过去,商业计算主要是为自动化生产、精简工作任务和高速采集数据而服务。这些用途准确地来说应该归之于数据处理。然而在从 20 世纪 90 年代开始,信息技术将取代数据处理而成为计算的重点。在今后的 10 年中,作为全球经济及技术管理者的信息工作者必须重视信息的策略性价值,所面对的问题不是简单地处理数据而是如何使用数据。传统上,数据库系统仅仅是用来辅助商业数据处理的,许多数据库管理系统的研究工作也是集中于这个方面。但是,最近数据库系统的研究出现了许多应用,特别是在数据仓库(Data Warehouse)方面。下面谈谈数据仓库的概念和特征。

1.1.1　数据仓库的概念和特征

数据仓库涉及在关系数据库中存储数据和处理这些数据,使数据成为查询和决策支持分布的更加有效的工具。当前,数据传输和分析进程最流行的说法是商业智能。换句话说,数据仓库是一种把收集的数据转变成有意义的信息技术。数据仓库通过使用多步进程来执行,这些进程包括收集数据、过滤数据和存储数据,以便把这些数据用在分析和报告应用程序中。这种数据可以来源于许多不同的数据源,包括不同的数据库系统,甚至来源于不同的操作系统。数据分析的主要目标是确认数据结构和趋势,以助于制作有关企业经营的更好的决策。

需要全本请访问 www.ertongbook.com

创建数据仓库的原因是因为公司更加依赖于从信息系统中收集信息。因此,为了公司的运转就需要一个信息数据仓库。客户也希望能够经许可后访问公司的数据。数据仓库为商业分析家所提供的信息,在过去的业务数据库中是很难得到的。大多数情况下,公司会将历史数据从业务数据库中转移到备份系统中,这使得用户无法深入分析数据,难于做出有竞争性的决策。对更好地管理数据,保持数据的一致性,以及从企业的角度分析数据的需求引出了数据仓库的概念。数据仓库允许企业的各个部门之间共享数据,为企业更快、更好地做出商业决策提供更加准确、完整的信息。

在企业中收集和汇总存储在许多位置和有许多格式的数据是一个不小的任务。这就是数据市场所在。数据仓库指企业级标准化和汇总化信息的存储,而数据市场是指这种目标信息的子集。数据市场只是包含了描述企业某一个特定方面的信息。例如,数据市场的一个常见用途是存储仅与一个特定商业单位有关的分析信息。

首先,还是让我们看看数据仓库的概念。什么是数据仓库呢?不同的研究人员有不同的答案。现在,比较一致的定义是,数据仓库就是面向主题的、一致的、不同时间的、稳定的数据集合,用于支持经营管理中的决策支持过程。也就是说,数据仓库是一个处理过程,该过程从历史的角度组织和存储数据,并能集成地进行数据分析。简而言之,数据仓库就是一个大的数据库,存储了公司的所有业务数据。数据仓库允许企业的各个部门之间共享数据,为企业更快、更好地做出经营决策提供更加准确、完整的信息。数据仓库除了具有传统数据库管理系统(DBMS)的共享性、完整性、数据独立性外,还具有自己的一些特征。为了理解数据仓库,下面从数据仓库的四个特征方面对数据仓库的概念进行解释。

数据仓库是面向主题的。传统的数据库是面向应用设计的,它的数据只是为处理具体应用而组织在一起的。应用是客观世界既定的,它对于数据内容的划分未必适用于分析所需。而主题是一个在较高层次将数据归类的标准,每一个主题基本对应一个宏观的领域,基于主题组织的数据被划分为各自独立的领域,每一个领域都有自己的逻辑内容而互不交叉。面向应用是经常会随着各种经营环境的变化而发生变化,而面向主题则因为比应用具有更高的层次而比较稳定。数据的产生都是基于应用而产生,因此,在数据进入数据仓库之前,必然要经过加工和集成,将原始数据结构做一个从面向应用到面向主题的大转变。在一个公司中,各个职能部门的经营数据源保存着公司经营的方方面面的大量数据,例如客户记录、产品信息、销售情况等。但是,在传统的数据库中,这些数据没有被仔细地与公司的全面经营状况进行相关分析,例如如何从客户记录中找到哪一类客户与自己公司的经营效益紧密相关等。数据仓库通过组织这些关键的公司信息,可以大大地简化对公司经营状况的分析,以便为公司领导进行正确的经营决策。我们举一个例子说明数据仓库的这种面向主题的特征。在一个公司的销售部门中,如果只是记录原始的销售数据,那么使用传统的数据库技术就可以解决这个问题。但是,如果希望对这些数据进行分析,找到哪些区域的用户最喜欢公司的产品、哪些年龄段的用户最喜欢公司的产品、什么时间公司的销售情况最好等诸如此类的问题,那么就得借助于数据仓库技术。

数据仓库是一致的数据的集合。在数据仓库中,使用一致的命名规定、度量、物理属性和语义来统一来自多种数据源的经营数据。例如,在许多公司中,应用程序常常以不同的格式使用类似的数据:日期按照“年月日”或者“月日年”格式存储,真/假数据可能表示为 1/0、ON/OFF、TRUE/FALSE、正/负等。不同的应用程序可能使用不同的术语来描述同样的数据,

例如一个应用程序可能使用“平衡表”来描述银行帐户,另外一个应用程序可能使用“总量表”表示银行的帐户。有时候一个公司的名称可能使用了不同的表示方式,例如,微软(中国)有限公司可能表示为“微软”、“微软公司”、“Microsoft”、“Microsoft Company”、“Microsoft Co .”、“中国微软”等。在数据仓库中存储的数据需要一种单一的、可以由企业分析人员接受的格式存储,而不是依据这这些数据来自的各种外部数据源存在格式。这种统一性允许数据来自不同的企业和不同的部门,例如公司以前保存的数据、各种报表中的数据、甚至来自 Internet 上的数据。这些数据可以统一在数据仓库中,以便使公司的经营人员更好地分析和理解企业的经营状况。

存储在 OLTP 系统中的数据可以正确地表示任何时间的任何值。OLTP 系统经常只是包含当前的数据。例如,处理定单的应用程序总是显示库存的当前值,而不显示在过去某一时间的库存值。不同时间的查询一般地会得到不同的查询结果。然而,存储在数据仓库中的数据,可以准确地表示过去某一时刻的数据,因为这些存储的数据表示了历史信息。存储在数据仓库中的数据一般表示过了一段比较长的时间的数据,一般是五年到十年。数据仓库存储了企业一段时间的快照形式。这些数据一般是不改变的。与此不同的是,业务数据库通常只保存有用事务数据 30-90 天。数据经集成进入数据仓库后是极少或者根本不更新的。因此,常用的操作是追加操作和历史查询。

因为存储在数据仓库中的数据表示是某一时刻点的数据,所以在数据仓库中,不允许删除、插入和修改。当数据移动到数据仓库之后,一般是不改变的,除非存储的数据是不正确的。一般地,在数据仓库中发生的操作是建立数据仓库时的加载数据和查询数据。因为数据在加载之后不再修改,所以对数据仓库的设计可以通过使用索引、预先计算的数据和物理的数据库正规化来优化查询的性能。万一数据仓库中的数据需要修改,那么可以使用 OLAP 工具来管理静态的数据仓库数据和动态的数据。

1.1.2　OLTP 和 DSS 的差别

一般地,数据库系统也称为联机事务处理(OnLine Transaction Processing, OLTP)系统,其主要特征是由多个并发用户修改系统中的数据,典型的 OLTP 运行示意图如图 1-1 所示。OLTP 事务是一个工作单位,通常可指在很短的时间段内完成,因为它采用实时或者在线方式处理数据库,换句话说,这些事务不停地更新基于当前最新信息之上的数据库,所以下一个用户便可以利用这些反映当前最新状况。这类事务处理系统的一个例子是订购货物系统。在该例中,与系统相关的所有信息都分散保存在磁盘系统的各个表中。而且,系统机构中任何一个用户都可以访问这些信息。像 Item _ table 和 stock _ level _ table 这样的表存放所有销售货物类型与数量的最新信息。当出现某种货物的订单时,该数据库必须知道:为了防止脱销,库存中需要保存这种货物的数量。

OLTP 系统的主要作用为:

支持大量并发用户增加和修改数据库中的数据;

只存储组织的当前数据,不存储历史数据;

包含了大量的验证事务数据是否正确的数据;

拥有复杂的数据结构;

负责回应组织的事务活动;

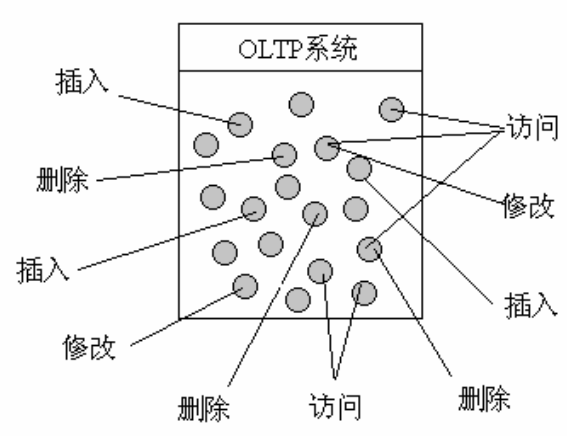


图 1-1 典型的 OLTP 运行示意图

为组织的日常经营提供技术基础。

现在人们需要的另外一种应用系统类型是决策支持系统(Decision Support System, DSS)。这种类型的系统一般用于为管理层提供信息,以便做出有关促进企业增长的决策。例如,可以使用这样的决策支持系统向有关领导和专家通报复杂系统的数据情况。在决策支持系统中,很少对数据库进行写操作,除非在对数据库进行更新或者装入时,典型的 DSS 运行示意图如图 1-2 所示。如果把 OLTP 系统应用于决策支持活动,那么就会面临许多困难,例如:

对于复杂的结构,需要创建用于分析的特殊查询语句,而这些工作只有数据库技术专家才能做好;

对系统中大量数据的分析汇总影响在线事务的处理性能;

当执行复杂的查询时,由于速度过慢而影响决策的执行;

由于数据经常改变而影响数据分析的一致性;

安全性过于复杂。

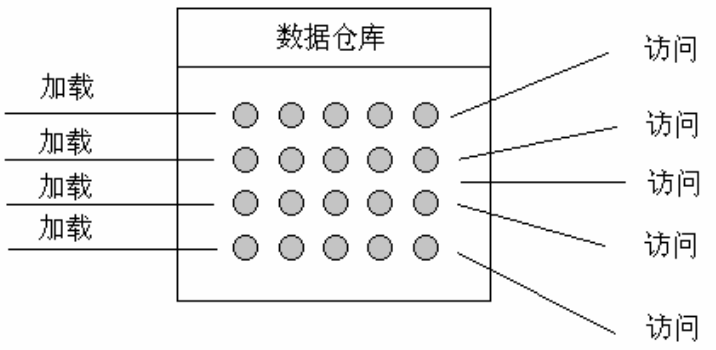


图 1-2 典型的 DSS 运行示意图

数据仓库就是为了满足用户进行决策支持的需要,它提供了解决上述问题的方案。一般而言,在 DSS 系统中,使用数据仓库可以执行下面工作:

可以把来自异构数据源中的数据整合成单一的数据结构;

把数据按照简单的方式进行组织就可以完成高效率的分析查询;

包含了用于分析的有效的、一致的、集成的、格式化的经过转换的数据;

提供了业务历史的稳定的数据;

根据确定的周期加载数据,而不是由频繁的事务修改数据;

安全性要求比较简单。

在联机事务处理(OLTP)系统中,事务的吞吐量比率通常使用每秒钟完成的事务处理数

(TPS)或者每分钟完成的事务处理数(TPM)来测定。对于决策支持系统(DSS)而言,吞吐量通常用每小时处理的查询数(QPH)来测定。QPH 表明,这些查询数量庞大,他们在完成之前占有绝大部分的机器资源。几乎在所有情况中,联机事务处理系统和决策支持系统的事务处理之间的比率是数百个数千个联机事务处理系统处理事务比一个决策支持系统处理事务。一个联机事务处理系统,即使很大,通常也不过是 300GB 左右,而一个大型的决策支持系统的规模可以达到 1TB(1TB = 1000GB)。

表 1-1 汇总了 OLTP 和 DSS 两个系统之间的特征。

表 1-1　OLTP 和 DSS 的特征比较

特征	OLTP	DSS
数据量	小	大
能够修改数据	能	否
数据结构的复杂程度	复杂	简单
数据是否稳定	不稳定	稳定
数据一致性	低	高
安全性	复杂	简单
事务吞吐率	高	低
作用	事务处理	分析

1.1.3　数据仓库的组成部分

对于一个典型的数据仓库,包括数据市场、关系型数据库、数据源、数据准备区、各种表示服务等组成部分。

数据市场是数据仓库的子集。如果说数据仓库是整个企业组织的所有主题数据存储区,那么数据市场就是某一个主题数据的存储领域。虽说如此,但是数据市场也有着非常重要的作用。在某种程度上,由于数据仓库的庞大数据可能使某些决策支持工具难以很好地发挥性能,数据浏览过于困难,这样可能会使一些用户放弃数据仓库的使用。而数据市场由于包含的数据比较小,因此可以很方便地建立起来,使用户可以非常方便地使用。数据市场既是数据仓库的一种实现,也是实现数据仓库的重要过程。

虽然关系型数据库是 OLTP 的基础,但是也是数据仓库的一个非常重要的组成。关系型数据库管理系统提供的强大功能可以直接应用到数据仓库的应用过程中,例如为数据仓库提供存储引擎、快速传输和修改、灵活的索引机制、高效的数据检索,可以包含和管理数据仓库中的海量数据,可以支持数据市场的建立。关系型数据库管理系统 Microsoft SQL Server 2000 为数据仓库提供的强大工具包括数据转换服务(DTS)、多维在线分析服务(Analysis Services)、使用自然语言的英语查询服务等。

数据仓库的目的是向决策者提供所需的各种决策信息。这些决策信息来自多个数据

源,既可以是关系型数据库中的数据,也可以是电子表格的数据,还可以是文本数据。数据仓库必须具有能力把这些异构数据源中的数据收集和整理好,然后使这些数据准确地反映出企业组织的经营状况和历史状态。一个企业可能有多个 OLTP 系统,每一个 OLTP 系统可能由不同的部门设计和开发,这些系统所涉及的数据结构和数据元素随着数据库的不同而不同。这些数据虽然不能直接用于决策支持,但是这些长期积累下的数据是建立数据仓库重要的基础和数据来源。

数据仓库的数据从各种数据源提取之后,然后进行精炼和格式化,最后转换到数据仓库结构中。数据准备区是一个关系型数据库,从其他数据源中提取的数据首先放置在这个关系数据库中,然后转换成统一的数据仓库要求的格式,在这里,对转变后的数据进行一致性和参考完整性检查,为加载到数据仓库数据库中做好准备。在某些情况下,如果不影响数据仓库的性能,那么数据准备区和数据仓库数据库可以合并起来。但是,大多数情况下,数据准备区和数据仓库数据库是分离的,这样可以提高数据仓库的数据质量和查询性能。

数据仓库的目的是为决策者提供必要的决策信息。根据的决策者的要求,为了提供决策者所需的信息,数据仓库必须提供一些实现这些要求的服务功能。这些服务功能包括预定义的报表、在线分析处理、数据挖掘等。通过访问关系数据库和多维数据立方体,可以定制用户需要的各种报表,通过使用这些预定义的报表,决策者可以进行日常的决策。但是,预定义报表只能满足特殊的需要,而不能满足决策者由于复杂的决策环境变化而引起的动态分析的需要。这时,需要使用在线分析处理工具。为了从浩如烟海的杂乱数据中提取发现有价值的规律,可以使用数据挖掘工具执行数据挖掘工作。

1.2 Microsoft SQL Server 数据仓库的工具

在 Microsoft SQL Server 2000 系统中,提供了大量的数据仓库设计、建立、数据加载、数据使用、数据挖掘等工具。这些工具不仅可以用于数据库应用的支持,而且还可以提供数据仓库的支持,并且有些工具专用于数据仓库。表 1-2 列出了一些数据仓库中经常用到的工具,还有许多非常专业的数据仓库工具在本书后面的介绍中提及。

表 1-2 常用的数据仓库工具列表

工具	描述
关系型数据库	作为数据仓库设计、构造和维护的基础
数据转换服务(DTS)	用于向数据仓库中加载数据
数据复制	用于具有分布式数据仓库分布数据的需要
OLE DB	应用程序与数据源的接口
Analysis Services	用于采集和分析数据仓库中的数据
English Query	提供使用英语语言查询数据仓库的需要
Meta Data Services	浏览数据仓库中的元数据

工具	描述
PivotTable Service	用于定制操作多维数据的客户端接口

关系型数据库是设计、构造、维护数据仓库的基础技术。Microsoft SQL Server 2000 是一个关系型数据库管理系统,它提供了强大的数据库引擎和许多工具,这些工具可以用于数据库的设计和操纵。这些数据库引擎和工具,不仅可以用于数据库的管理,而且许多工具可以用于数据仓库的管理,是数据仓库的技术基础。

数据转换服务(DTS)提供了数据转换功能,例如数据引入、引出以及在 SQL Server 和任何 OLE DB、ODBC 或者文本格式文件之间转换数据。利用 DTS,通过交互式地或者按照规划自动地(无须人工干预)从多处异构数据源输入数据,这样便有可能在 SQL Server 上建立数据仓库和数据市场。最大的问题是保持数据不断更新以与操作系统相适应。数据的来源特性允许用户输入查询数据的输入时间、地点及其计算方法。数据的引入引出是通过以相同的格式读写数据,在应用程序之间交换数据的过程。例如,DTS 能够把数据从 ASCII 文本文件或者 ORACLE 数据库中引入到 SQL Server 中。反之,数据也能够从 SQL Server 中引出到一个 ODBC 数据源,或一个 Microsoft Excel 表单中。一次转换是指在源数据被存储到目的地之前,对其进行的一系列操作。例如,DTS 允许从一个或多个源域中计算新值,或将一个单独的域分成多个域,然后存储进不同的目标列中。转换特性使得在引入引出中完成复杂数据的验证,清除及功能增强变得非常简单。定制转换对象可以被创建并集成进第三方产品中。DTS 支持多步封装,其中大量文件能被单独处理,然后最终被集成为一个单一的文件。在目的地,该文件的记录能被系统分为多个记录,或相反,多个记录也被集合为一个单一的记录。DTS 也与微软的中心库集成在一起,微软数据中心库存储数据元、数据传送包和数据源特性,这包括所有数据的来源。许多独立的软件经销商支持和共享这个数据元模型。数据转换服务可以在 SQL 服务器间移动数据结构、数据、触发器、规则、默认、约束和用户定义的数据类型。只有数据结构和数据可以在不同的异构数据源中间传递。

数据复制是一种实现数据分布的方法,就是指把一个系统中的数据通过网络分布到另外一个或者多个地理位置不同的系统中,以适应可伸缩组织的需要、减轻主服务器的工作负荷和提高数据的使用效率。数据复制的过程类似于报纸杂志的出版过程,即把信息从信息源迅速传送到信息接收处。出版服务器包含了将要复制的数据所在的数据库,该服务器负责制作将要出版的数据,然后将这些出版的数据的所有变化发送到分布服务器中。订阅服务器是数据复制的目标地,负责接收复制过来的数据,并且负责接收从出版服务器上传送过来的全部变化。分布服务器在出版服务器和订阅服务器之间起一个中介作用。分布服务器负责从出版服务器中接收复制过来的数据,然后在该服务器上的分布数据库中存储这些变化,最后按照指定的时间间隔推向相应的订阅服务器中。在复制过程中,把将要复制的数据也称为出版物和文章。文章是复制的基本单元,是组成出版物的最小单位。文章是与数据库中的表对应的。文章既可以对应一个表中全部数据,也可以只包括表中的一部分数据。单个文章是不能复制的,文章必须组成出版物才能出版。出版物是一个完整的将要复制的内容,它包括一篇或者多篇文章,是订阅服务器订阅的对象。

OLE DB 是数据存取机或服务支持器,也是这些应用程序将会用到的高度交互的基于部

需要全本请访问 www.ertongbook.com

件环境的事务逻辑部件。OLE DB 是微软最近开发的数据访问编程界面。微软已经定位 OLE DB 是 ODBC 的继承者。ODBC 主要是用于处理关系型数据。ODBC 是基于 SQL 的,它可以很好地用于关型数据库访问,它不能用于非关系型数据源。像 ODBC 一样,OLE DB 也提供了对关系型数据的访问,但是 OLE DB 扩展了由 ODBC 提供的功能。OLE DB 的设计目的用于所有数据类型的标准界面。除了关系型数据库的访问,OLE DB 提供了对各种各样数据源的访问,包括像 Excel 电子表格的表格数据、像 dBase 的 ISAM 文件、电子邮件、新的 NT 与 Active Directory 和 IBM 的 DB2 数据。使用 OLE DB,用一个界面就可以访问许多不同的数据源。OLE DB 是 Microsoft 的数据访问策略,称为万能数据访问的基础。万能数据访问指的是一组通用界面,它用来表示来自任何数据源的数据。OLE DB 是使万能数据访问成为现实的技术。万能数据访问和 OLE DB 的另一端,所有的对象都应该在数据库中维护。不是企图把商业要求的不同的数据移动到一个面向对象的数据库中,OLE DB 创建和理解企业数据并且在各种数据源中维护。OLE DB 提供了针对各种数据的类似界面。OLE DB 可以用来访问任何可以用基本的行和列格式表示的数据。

Analysis Services 服务。在数据仓库或者数据市场中的信息由在线分析进程(OLAP)处理。OLAP 可以有效地以由维度和度量组成的立体方式查看数据。维度是一种描述性分类,例如,维度可能是地理位置,也可能是产品类型。度量是一种定量数值,例如销售额、库存量或者总费用。来自于原始数据源的合计存储在每一个立体单元中。这种组织数据的方法可以轻易地过滤数据、有效地和加恢数据的查询。然而,存在着这种平衡:虽然 OLAP 合计是在数据仓库中可以达到的查询性能的关键因素,但是存储这些合计数据的成本是磁盘存储量。事实上,合计数据量可以轻易地超过原有的数据量。另外,当维度和合计量提高时,所要求的 OLAP 数据存储量也大大地提高。这种对存储量的极大地要求一般称为数据爆炸。OLAP 描述的是一种多维数据服务(这里指的维是指人们观察客观世界的角度,例如时间、地域、业务等),这种服务的设计目的是保证分析员、经理和决策者针对特定的问题,通过快速、一致、交互式的实时数据访问和分析,获得有创意的发现。在 SQL Server 2000 中,有三种用于存储数据仓库中维度数据的方法,每一种方法都随其数据存储要求和其数据检索速度而变化。(1)MOLAP。多维型 OLAP 在一个用于压缩索引的永久数据存储中存储维度数据和事实数据。合计存储用来加快数据访问。MOLAP 查询引擎通常是专有的,并且优化成由 MDLAP 数据存储使用的存储格式。MOLAP 提供了比 ROLAP 更快的查询处理速度,并且要求更少的存储空间。然而,它不能较好地伸缩,并且要求使用单独的数据库来存储。(2) ROLAP。关系型 OLAP 在关系型数据表中存储合计。ROLAP 的关系型数据库的应用允许其利用已有的数据库资源,并且允许 ROLAP 应用程序很好地伸缩。然而,ROLAP 使用表存储合计一般要求比 MOLAP 更多的磁盘空间,速度相对比较慢。(3) HOLAP。正如其名称所示,混合型 OLAP 介于 MOLAP 和 ROLAP 之间。像 ROLAP 一样,HOLAP 使主数据存储在源数据库中。像 MOLAP 一样,HOLAP 把合计存储在一个永久性数据存储的地方,它与主关系数据库分开。这种混合形成使 HOLAP 可以提供 MOLAP 和 ROLAP 两者的优点。然而,不像 MOLAP 和 ROLAP,要求 HOLAD 按照已定义的标准,没有统一的实现形式。

English Query 服务。微软的英语查询环境允许设计者将他们的关系数据库转移到英语查询上来。这给最终用户提供了使用英语提出问题而不是用 SQL 语言进行查询的能力。英 English Query 服务是利用英语查询域编辑器创建的,提供了数据库的有关信息,从而英语

查询可用于处理关于特殊表、域和数据的英语问题。

在 Microsoft SQL Server 2000 中,许多工具把元数据存储在了位于 msdb 系统数据库中的集中式中心仓库中。这里提到的元数据是数据的数据。Meta Data Services 提供了对这些元数据的浏览功能。数据中心库提供了一个通用的位置,可以用来存放对象和对象之间的关系。通过使用一些软件工具,来描述面向对象的信息,例如:软件工具使用各类对象组成的信息;对象接口界面;界面由属性、方法和集合组成;集合包含了对其他界面的关系,以及关联对象。

PivotTable 服务和 Analysis Services 服务一起为用户提供客户端对 OLAP 数据的存取。PivotTable 服务运行于客户端工作站商,它使得其部门可以利用 Visual Basic 或其他语言来开发用户程序,这些程序可以利用 OLE DB 技术,并使用 Analysis Services 服务中的 OLAP 数据或直接取自关系数据库的数据。当它和 OLAP 服务一起使用时,PivotTable 服务可以自动的将进程和缓冲内存分配到最合适的位置,并且允许多个客户动态存取同一个立方体。PivotTable 服务也能在本地客户机上存储数据,从而使用户可以在不连接 Analysis Services 服务的情况下对数据进行分析。这个移动式的解决方案允许分析者将数据带回家或在路途中进行分析。为最终用户提供的 OLAP 数据分析和描述工具可以利用 PivotTable 服务进行开发。微软的 ActiveX 控件技术及其 Office 组件正在被创建以产生强大的图形用户界面。PivotTable 服务也提供了开放的界面,独立的软件销售商可以利用它来开发第三方应用产品。

1.3　数据仓库实现步骤

一般地,实现数据仓库的步骤包括明确用户的需求、设计和建立数据仓库、使用数据仓库、维护和优化数据仓库的性能等,这些步骤的示意图如图 1-3 所示。下面简单介绍这些步骤。

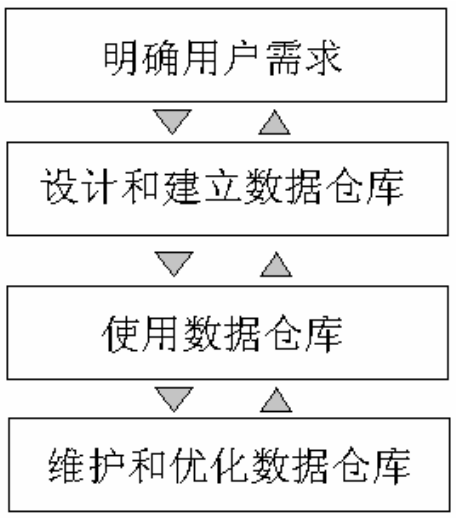


图 1-3　实现数据仓库的步骤示意图

1.3.1　明确用户需求

确定终端用户的需要,为数据仓库中存储的数据建立模型。通过数据模型,可以得到企业完整而清晰的描述信息。数据模型是面向主题建立的,同时又为多个面向应用的数据源

的集成提供了统一的标准。数据仓库的数据模型一般包括:企业的各个主题域、主题域之间的联系、描述主题的码和属性组。

深入地分析企业的数据源,记录数据源系统的功能与处理过程。一般地,设计数据仓库最重要的一步便是要理解商业动作的规律,只有了解数据是如何被处理的,才能分解商业处理过程,从中获取数据元素。

利用现有系统的信息,确定从源数据到数据仓库的数据模型所必须的转化/ 综合逻辑。这涉及到应该合并转化多少数据;是综合所有的数据文件还是综合发生变化的操作系统文件;转化/ 综合过程应该多长时间执行一次等问题。决定数据转化与更新频率是重要的商业事件。无论数据仓库的更新是采用事件驱动还是时间驱动,都必须让数据仓库知道当某种事件发生时就需要更新数据。

在数据仓库建立之前,应该写一个详细的方案和实现规划。这种方案和实现规划包括:建立商业案例、收集用户需求、确定技术需求。建立商业案例包括由该方案解决的商业需求、方案的成本和投资的收益。收集用户需求主要是调查用户建立数据仓库的意图。用户需求可以确定这些内容:数据需求(粒度级)、企业操作系统包含的数据、这些数据遵循的商业规则、需要提供给用户的查询、用户需要的标准报告、将要使用的客户应用程序工具。确定技术要求包括下列内容:硬件体系结构和框架(例如,链接到数据市场所在的地理位置)、备份和恢复机制、安全性限制、从操作系统到数据仓库加载数据和转换数据的方法。

1.3.2 设计

设计和建立数据库是成功地创建数据仓库的一个关键步骤。这一步通常由有经验的数据库设计人员使用,因为这一步涉及的数据来自多种数据源并且要把它们合并成一个单独的逻辑模型。不像 OLTP 系统那样以高度的正规化形式存储数据,数据仓库中存储的数据以一种非常非正规化的形式存储数据以便提高查询的性能。数据仓库常常使用星型模式和雪花型模式来存储数据,作为 OLAP 工具管理的合计基础,以便尽可能快地响应复杂查询。

在转换 OLTP 数据库模式到星型模式时,涉及的步骤如下:确定事实表和维度表;设计事实表;设计维度表;实现数据库设计。

把操作系统中数据提取出来然后加载到数据仓库中,随着复杂性的变化而变化。如果在数据源中的数据和将要出现在数据仓库中的数据是直接关联,那么这个进程非常简单。这个进程也可能非常地复杂,例如数据源的数据驻留在多个异构系统中,并且在加载数据之前需要大量的转变格式和修改。提取和加载数据的进程如下:校验操作系统中的数据;从操作系统中提取数据;净化数据;把数据转换和迁移到数据仓库中。

1.3.3 使用数据仓库

数据仓库的传统作用是采集合组织历史数据,对这些历史数据进行各种分析,为决策者提供依据。在这些数据仓库中,只有数据库专家才能使用这些数据仓库中的数据,因为只有这些专家才能写出决策者需要检索、汇总、格式化数据的复杂的查询语句。随着数据仓库技术的发展和利用数据仓库执行决策的层次降低,对数据仓库的直接访问需求变得越来越重要。

在 Microsoft SQL Server 2000 中,使用数据仓库中的数据可以采用多种方法,例如 Trans-

act-SQL 语句、Analysis Services、脚本语言等。图 1-4 是一个使用 Analysis Manager 查看数据仓库的示例。本书后面将对这些方法和技术进行详细的研究。

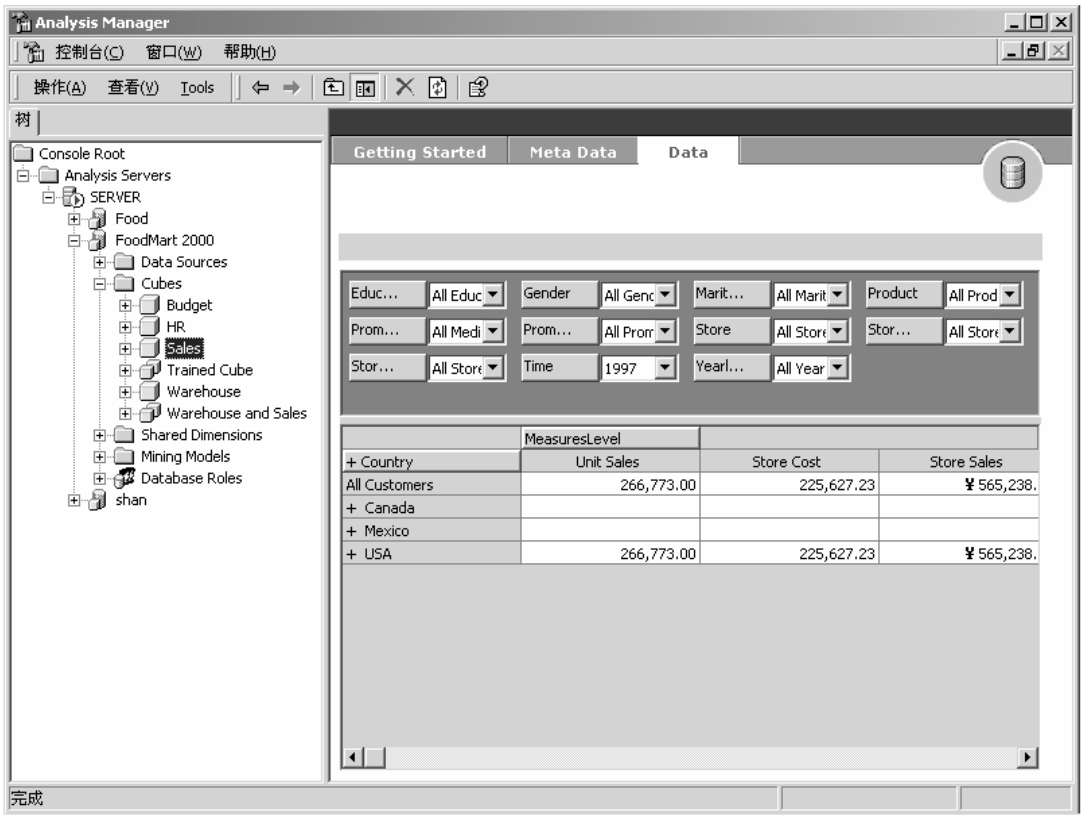


图 1-4　使用 Analysis Manager 查看数据仓库

1.3.4　维护和优化数据仓库

数据仓库的目的是采集合组织历史数据以便分析，支持决策者进行决策。为了实现这个目的，当创建数据仓库时，加载已有的历史数据到数据仓库中。以后，随着时间的推移，应该使用操作系统中的数据周期地修改数据仓库中的数据。因此，周期地加载或修改数据仓库中的数据是非常重要的维护工作。

为了确保数据仓库数据在系统被破坏之后，可以顺利地恢复到破坏前的状态，需要对数据仓库中的数据进行备份。数据仓库的备份不同于 OLTP 系统数据库的备份。数据库是一个动态的数据容器，数据不停地在发生着变化，因此数据库的备份是一个复杂的过程。而数据仓库是一个相对静态的数据的容器，其中的数据很少发生变化，所以对数据仓库的备份可以采取比较简单的方式进行。

为了提高数据仓库查询数据的效率，需要为数据仓库建立完善的索引机制。这样可以加快数据的检索速度。

为了发现影响数据仓库系统性能的平静因素，可以使用性能监视工具监视系统的运行状态，及时调整运行参数，提高数据仓库的运行效率。

1.4　小结

通过与微软的其他产品的紧密结合，Microsoft SQL Server 2000 提供了一整套的功能强大的创建完整的决策支持系统的工具。这种决策支持系统能够以方便的访问格式向使用者提

供所需要的准确的决策信息。如果数据仓库规划和设计合理,以及决策者对所创建的决策支持系统有清醒的认识,那么 Microsoft SQL Server 2000 数据仓库能够对如何提高组织的运营效率、优化业务活动的决策提供巨大的帮助。

第 2 章　数据仓库生命周期管理

数据仓库的创建不同于 OLTP 系统中数据库的创建,因为数据仓库的创建和管理是一项复杂、庞大的工程。对于数据仓库的管理应该采取工程项目的管理方式,确保数据仓库成功运行。本章将从工程项目管理的角度研究如何管理数据仓库的创建。

本章的要点:

数据仓库系统框架
影响数据仓库项目成功的因素
数据仓库的技术
数据仓库项目所需的角色

2.1　数据仓库系统框架

数据仓库是企业系统项目的重要组成部分。如果数据仓库的规划、设计和创建缺乏经过精确定义的框架,那么项目可能归于失败。所谓框架,是指可以据此创建一个完整系统的结构规范。建立数据仓库的合适框架应该具有三个基本特点:

拥有为项目做出技术决策的观察点;
拥有确定项目是否能够准时完成的计量参照点;
拥有以灵活方式反复运用现有知识的能力。

Microsoft Solutions Framework(MSF)就是具有这三大特点的一种框架。一般地,通过 MSF 框架建立的系统应该具有以下一些特点:

系统可以被分解成若干个小系统,这些小系统既可以逐一单个运行,也可以同时运行。因此,系统可以延续很长时间,但是短期内也可以获得结果。

系统是按照不同的组成部分创建的。因此,既可以一部分一部分地创建,也可以对各部分并行地创建。

系统是分多个阶段创建的,每一个阶段都建成系统的一个部分。

整个系统项目的完成,依靠多个工作小组协调完成。

MSF 框架把系统项目分解成了设想、规划、开发、稳定四个阶段。这四个阶段不是简单地循环往复,而是呈螺旋趋势上升的。这是因为,在一般情况下,任何企业系统都不可能在一个阶段中完成,相反,经过多个阶段,每一个阶段都会给系统增加新的功能。从这个角度来看,各阶段的工作永远也不会有终结,数据仓库等系统一直会处于一个设想、规划、开发、稳定、再设想……的循环往复之中。这四个阶段的示意图如图 2-1 所示。

2.1.2　里程碑

在 MSF 框架的设想、规划、开发、稳定阶段中,每一个阶段工作的完成都是整个系统项

需要全本请访问 www.ertongbook.com