

数据仓库研究及其在现代企业中的应用

The Research of Data Warehouse and Its Applications in Modern Companies

张凤荔 葛晓峰 周明天

(电子科技大学计算机学院 成都 610054)

Abstract In this paper, we introduce the origination of the DW ideology and its architecture & principles, as well as relative techniques. With the example that the MIS of ENWEI company, we discuss how to apply the base theories of DW and implement its main techniques in modern companies. At the end of this paper, we propose the ways that how to build an effective Decision Support System based on the MIS of ENWEI company.

Keywords Data Warehouse (DW), On-Line Analytical Processing (OLAP), Decision Support System (DSS), Data Mining (DM)

1 数据仓库与 OLAP

1.1 从数据库到数据仓库

过去几十年中,数据库技术,特别是 OLTP(在线事务处理)发展得较为成熟,它的根本任务就是及时地、安全地将当前事务所产生的记录保存下来。随着社会的发展,人们提出了利用现有的数据,进行分析和推理,导致了决策支持系统(DSS)的产生。随着数据量的迅速增大以及查询要求的复杂化,建立在 OLTP 的数据库上的 DSS,暴露出许多难以克服的困难:

1)缺乏组织性。各个部门抽取的数据没有统一的时间基准,抽取算法、抽取级别也各不相同,并且可能参考了不同的外部数据。

2)效率极为低下,数据难以转化为有用的信息。原始状态中的大量数据,同一字段在不同应用中又可能存在着同名异义、异名同义、类型不同、字长不同等许多难以识别的麻烦。

3)其他困扰着基于传统 DB DSS 的问题。DSS 对 OLTP 性能的影响是难以忍受的。另外,OLTP 的 DB 各个应用领域的数据库保存期限都不一样,难以满足 DSS 需要。

数据仓库最主要的特征是:

1)数据仓库是面向主题的。它与传统数据库面向应用相对应,主题是一个在较高层次将数据归类标准,每一个主题对应一个宏观的分析领域。

2)数据仓库是集成的。原始数据与适合 DSS 分析的数据之间差别很大。因此,在数据进入数据仓库之前,必然要经过加工与集成。

3)数据仓库是稳定的。它反映的是历史数据的内容,而不是处理联机数据。因而,数据经集成进入数据仓库后是极少甚至根本不更新的。

4)数据仓库是随时间变化的。数据仓库内的数据时限要远远长于操作型环境中的数据时限;数据仓库中的数据都是历史数据;数据仓库数据的码键一般都包含时间项,从而标明了该数据的历史时期。

一种典型的数据仓库结构如图 1 所示。它分为

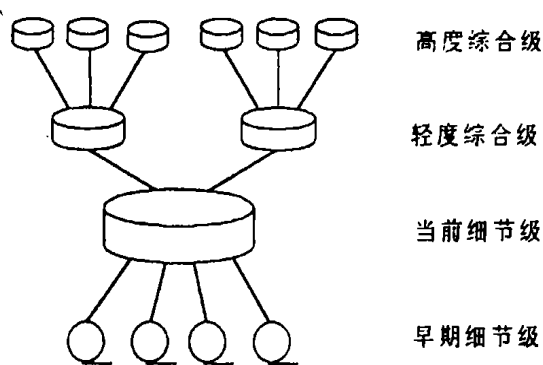


图 1 数据仓库结构

四级:早期细节级、当前细节级、轻度综合级和高度综合级。原始数据经集成以后,首先进入当前细节级,并根据需要进行进一步综合,进入轻度综合级乃至高度综合级,老化的数据则进入早期细节级。数据

仓库中存在着不同的细节级别,称之为“粒度”。粒度越高表示细节程度越低、综合程度越高。

数据仓库的另一个重要技术是如何定期向数据仓库追加数据。常用的技术和方法有:

◇时标方法:如果数据含有时标,只需根据时标判断即可。

◇DELTA 文件:由应用生成,记录应用所改变的所有内容。

◇前后映象文件的方法:在抽取数据前后对数据库各作一次快照,然后比较两幅快照的不同从而确定新数据。

◇日志文件:它不会影响 OLTP 的性能。同时,它还具有 DELTA 文件的优越性质,提取数据只局限于日志文件。

1.2 从 OLTP 到 OLAP

OLAP(在线分析处理)是以数据库或数据仓库为基础的,其最终数据来源与 OLTP 一样,均来自底层的数据库系统。OLTP 面对的是操作人员和低层管理人员,而 OLAP 面对的是决策人员和高层管理人员,所使用的经综合提炼的历史数据均来自 OLTP 所依赖的底层数据库,OLAP 数据较之 OLTP 数据要多一步数据多维化或预综合处理。

数据分析模型的分类如表 1 所示。

表 1 分析模型分类

数据分析模型	绝对模型→解释模型→思考模型→公式模型
处理数据范围	历史数据和当前数据→预测数据、行为
用户—分析人员交互	少———→多
多维分析	少———→多
现有工具支持	多———→少

• 绝对模型(Categorical Model):属于静态数据分析。它通过比较历史数据值或行为来描述过去发生的事实。

• 解释模型(Exegetical Model):用户—分析人员利用系统已有的层次综合路径层层细化,找出事实发生的原因。

• 思考模型(Contemplative Model):属于动态数据分析,它旨在说明在一维或多维上引入一组具体变量或参数后将会发生什么。

• 公式模型(Formulaic Model):是动态性最高的一类模型,表示在多个维上,需要引入哪些变量或参数,以及引入后所产生的结果。

OLAP 是应企业动态分析而产生的,其功能是通过创建、操作、激活及综合来自解释模型、思考模型及公式模型中的信息。它可以识别变量间的新的或不可预知的关联,通过创建大量的维(综合路径)及指定维间的计算条件和表达式来处理大量数据。

数据仓库系统由三部分组成:数据仓库(DW)、数据仓库管理系统(DWMS)和数据仓库工具。从数据仓库体系结构(图 2)可以清楚地看到这三部分的关系。

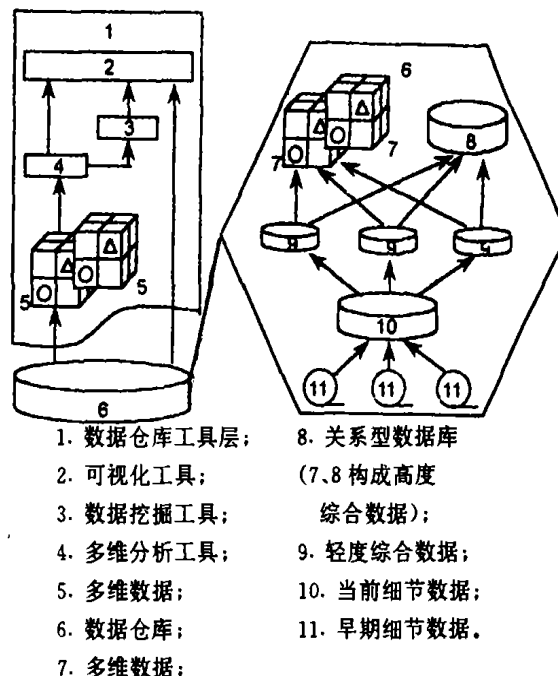


图 2 数据仓库体系结构

在整个系统中数据仓库居于核心地位,是信息挖掘的基础;数据仓库管理系统负责管理整个系统的运转,起着承上启下的作用;而数据分析工具则是整个系统发挥作用的关键,只有通过高效的工具,数据仓库才能真正发挥数据宝库的作用。

信息处理的多层次要求导致了一种新的数据环境—操作数据库 ODS 的建立。ODS 一方面包含细节的、当前或接近当前的数据,可进行联机操作型处理;另一方面,又是一种面向主题、集成的数据环境,可以辅助企业完成日常决策。

2 数据仓库在现代企业中的应用

2.1 一个销售系统的结构与工作原理

恩威公司销售分系统的数据库结构如图 3 所示。其中:各分公司的 OLTP 数据库位于各分公司的高性能 PC 机上,由 SQL Anywhere 管理。主要用于各分公司的日常事务处理,涉及到日常销售、回

款、仓储管理等诸多方面。通过 X.25 或电话拨号上网与总部 OLTP 数据库进行数据交换。

销售总部 OLTP 数据库位于公司总部的数据服务器上,由 Sybase SQL Server 管理。主要用于公司销售总部的日常事务处理,包括销售人员档案管理、业绩考核、日常计划的制定以及销售总部自身的销售业务等。该数据库的地位和作用在很大程度上相当于一个 ODS。

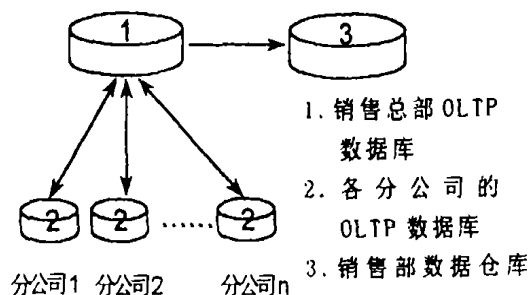


图3 数据库结构

销售数据仓库完整的工作过程如图4所示。其中的 OLAP 工具部分由虚线框界定,表示在该系统中尚且缺乏,目前在这个位置上工作的只是一些模型化的工具。而在 IT 软件市场上已经出现了许多实用的 OLAP 产品,通过考察和引进,相信这个问题在不远的将来即可得到完善的解决。

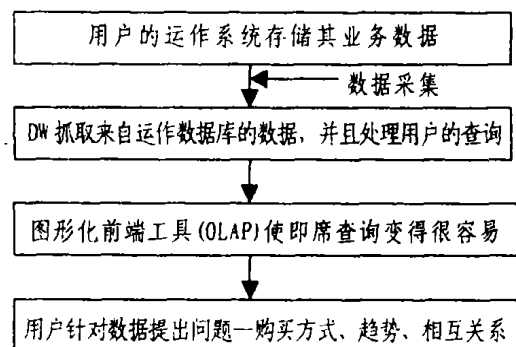


图4 DW 工作流程示意

2.2 销售数据仓库的逻辑设计

在销售数据仓库的逻辑设计中,我们采用了流行的星型模式(Star Schema)。星型模式是基于关系型数据库,面向 OLAP 的一种多维化的数据组织方式。

2.2.1 星型模式的有关概念。关系型结构能较好地适应多维数据的表示和存储。关系数据库将多维数据库中的多维结构划分为两类表:一类是事实(Fact)表,用来存储事实的度量值及各个维的码值;

另一类是维表,对每一个维来说,至少有一个表用来保存该维的元数据,即维的描述信息,包括维的层次及成员类别等。在相关事实表中,这些值会衍生出该维的列。事实表是通过每一个维的码值同维表联系在一起的,该结构即被称为“星型模式”(如图5)。

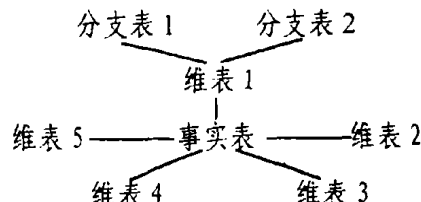


图5 星型模式示意图

有时,对于内部层次复杂的维,可以用多张表来描述一个维。比如,产品维可以进一步划分为类型表、颜色表、商标表等,这样,在“星”的角上又出现了分支。这种变种的星型模型被称为“雪片模式”(Snow Flake Schema)。对层次复杂、成员类较多的维采用多张表来描述,而对于较为简单的维可以用一张表来描述。

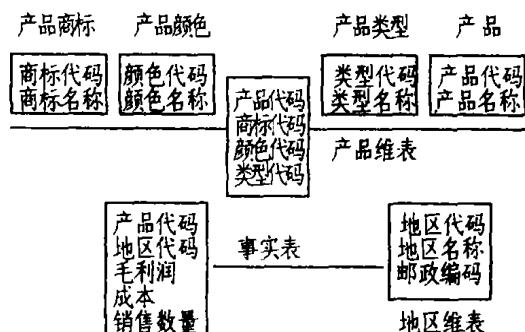


图6 星型模式实例

2.2.2 具体设计方案。基于销售部的日常业务,我们选择了销售金额与回款金额这两个基本事实。并选定了与之相关的四个维——时间、市场(地区)、产品和客户。

于是,事实表的结构可以表示如下:

表2

时间	市场 代码	产品 代码	客户 代码	销售 金额	回款 金额
----	----------	----------	----------	----------	----------

为了提供详尽的销售业务信息,该事实表的时间维采用了比较小的粒度,精确到每天。由于客户和市场是紧密联系的,因此,该事实表的大小仅与产品、客户以及时间的粒度有关。

市场维是最为复杂的一个维,该维是层次结构

Web 客户只需输入前端所在地址(如 <http://pact518/DatabasePublish.html>)便可访问数据库信息。其运行过程是这样的:

1)浏览器获得用户的 URL 请求,通过 HTTP 协议将前端 HTML 页面下载到本地执行。

2)applet 从数据库名输入域获得用户键入的数据库名后,通过 ORB 及 CORBA/IIOP 协议调用后端 DataBase Server 的服务 GetDBInfo。

3)DataBase Server 通过与 SQL Server 交互获得数据库结构信息并返回给 applet,applet 将库结构信息显示在表选择框和域选择框中。

(4)用户根据 applet 提供的信息进行查询命令编辑,然后提交数据库查询请求。

5)applet 调用后端 DataBase Server 提供的服务 GetDBResult, DataBase Server 通过与 SQL Server 交互获得查询结果,并返回 applet,applet 将结果加以显示。

结论 本文主要探讨了基于 CORBA 的数据库发布方法,这种发布方法对数据库发布者来说只需给出数据库地址,免除了发布者编程麻烦。由于应用了 CORBA 和 Java,我们所提供的对象服务 DataBase Server 是生存在 CORBA 总线上的,因此整个数据库发布系统可在多种平台上移植,并且可

被其他的 CORBA 对象调用,具有开放性和可重用性。其次,DataBase Server 是独立于服务器的,因而对 DataBase Server 的修改与升级并不影响 HTTP 服务器的运行。由于采用了面向对象的设计思想,增加一个服务功能,如增加一个数据库口令服务,只需增加一个接口,因而有良好的可扩充性。

由于 CORBA 的产生时间并不长,各项技术都在发展成熟中,基于 CORBA 的数据库发布还有待于进一步深入研究和开发,并有着广泛的应用前景。

参考文献

- 1 Omnipotent Corporations. CORBA Tutorial. Available at: <http://www.omnipotent.com/addvalue/corba.html>
- 2 Object Management Group(OMG). The Common Object Request Broker Architecture and Specification (CORBA),1995
- 3 刘东波. Web 数据库技术进展. 中国计算机用户, 1997(10,上):45~49
- 4 The NSAPI versus the CGI Interface. Available at: <http://home.de.netscape.com/newref/std/nsapi-vs-cgi.html>
- 5 王克宏,丁铨,孙元. Java 语言 SQL 接口——JDBC 编程技术. 北京:清华大学出版社,1997

(上接第 74 页)

06/30 之间”一类范围的过滤即被称为切块。必须仔细地分析各数据维的可操作性以及固有的相关程度,有时统计图类型(折线图、条形图、堆积图、饼图以及散点图等)也不得不作为一个维来对待。

3 数据挖掘与决策支持系统

数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的新技术,它能挖掘出数据间潜在的模式,找出最有价值的信息。数据挖掘的思想依赖于以下三个技术的支持程度:大规模数据采集、功能强大的并行多处理器计算机、数据挖掘算法。

数据挖掘可实现的功能有:自动预测趋势和行为;自动发现以前未知的模式。

虽然数据仓库、OLAP 和数据挖掘最初是作为

三种独立的信息处理技术出现的,但是,由于它们内在的联系和互补性,今天,DW+OLAP+DM 的结构已被业内人士公认为是 DSS 的有效解决方案。这种全新的 DSS 构架的真正意义在于重新展示了信息的本质,表明了信息系统的设计观念从处理驱动到数据驱动的转变。在未来的时代,信息的重点将转移到数据模式分析,处理技术则是应数据分析的需要而产生的。

参考文献

- 1 Hofland P, Utsler J. Data Mining at Your Desk. 1997. 7
- 2 王珊. 数据仓库、联机分析处理、数据挖掘-基于数据库技术的 DSS 解决方案. 计算机世界,1997(1)
- 3 Watterson K. Attention, Data-Mart Shoppers, 1997. 7