基于数据中央开采器为核心的数据仓库建造

孟志清　(湘潭大学计算机科学系　411105)

摘要:本文对于一类中小型数据仓库(Ddat Warehouse)提出了一种基于数据中央开采器(Data Center Mining Unit)为核心的开发数据仓库建造思想,这种设计方法适用于在较低软硬件环境下开发,以提高系统的建造和运行效率。

关键词:数据仓库　数据开采　决策支持　数据库　数据中央开采器

一、引言

众所周知,目前已开发和正在开发的许多管理信息系统存在着一个重要问题,就是缺乏对大量历史信息的处理和加工能力,为了解决这一问题,90 年代在数据库领域里出现了一个新的技术:数据仓库技术—即面向历史数据库处理为经营或决策者提供综合信息和决策支持的技术,成为解决这一问题的有效工具。近几年,在数据仓库领域里掀起了产品开发和理论研究的热潮,许多学者发表了有关数据仓库方面的大量学术文献,为这一领域的理论研究和应用奠定了一定的基础。

根据国内外的技术资料,对数据仓库的定义和研究存在着许多差异和不同的理论,但给我们展现的数据仓库是技术复杂、投入资金大、建造周期长的框架,W. H. Inmon 曾指出数据仓库管理系统应包括三个部分:数据源、后端加工和前端服务。数据源提供原始数据;后端加工实施数据的后台处理(包括接收、析取、汇总、变换、打包和存储等);前端服务面向最终用户。大致的过程可以理解为下面两个主要过程:1. 数据采掘(数据开采)将各种数据源加工成元数据,2. 数据联机分析将元数据进行有效的再现为用户提供数据分析和决策支持,然而实现这一过程是一项非常复杂的技术。

为了加快数据仓库的理论和应用研究,避免购买国外昂贵的商业数据仓库开发产品,开发出适合国内的数据仓库产品,我们进行了在低资源低环境下实现高效率的数据仓库开发的理论和应用的研究。本文将以商场销售分析数据仓库系统建造为例,借助现有的软硬件资源,论述了一种建造数据仓库的理论和方法,为快速开发数据仓库提供了一个新的途径。

二、系统总体结构设计

1. 硬件设计

本文的目的是在较低的硬件资源下,实现数据仓库的建造。例如,目前一些大型商场实现了微机网络的日常事务处理系统,POS 机的应用随处可见并联接与服务器作为商场的日常业务管理。因此,仅投资购进几台微机,可实现商场小型数据仓库的建设。如果商场还没有建立事务处理联机系统,不防将日常业务处理系统作为数据仓库的一部分进行建设,该系统由一台服务器和若干台工作站与 POS 机联结。

数据仓库硬件系统利用的日常业务处理系统服务器(称为源数据服务器)作为源数据采集中心,并联接一台高性能微机工作站(配有大容量的存储器)(称为数据仓库管理工作站)用于数据仓库管理,数据仓库管理工作站联接到另一台高性能微机服务器(称为数据仓库用户服务器)用于数据仓库前端用户系统,并且有若干台微机工作站联与数据仓库用户服务器用于用户对数据仓库的查询和分析。如图 1 所示:

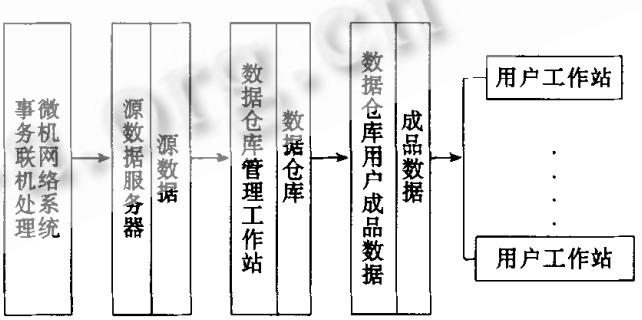

图 1　数据仓库的硬件结构

2. 软件设计

为了提高数据加工和事务联机分析的效率,目前一些建造数据仓库技术均采取了对数据源预加工和数据联机分析分开处理的办法,但是这种预处理加工技术如何组织是提高整个系统建造和运行效率的关键。这里我们提出了以数据中央开采器(DCMU)为核心的建造数据仓库方法和思想,数据仓库的主要作用是将数据仓库中的

源数据，经过加工形成可供用户使用的成品（综合）数据，问题可简化为：

源数据→数据加工→成品数据→用户（数据分析与决策支持）

因此数据仓库本质上是一个数据加工工厂，建造数据仓库问题的核心是数据加工，我们将所有用于数据加工的如分类、统计、预测等处理均汇集在 DCMU 中，这样所有源数据都经过 DCMU 的加工形成成品数据传送到用户服务器网供用户联机分析，用户通过浏览器和决策器实现数据的分析和决策支持。

整个数据仓库管理与应用系统（SDMAS）有以下三个子系统构成：日常事务管理系统、数据仓库管理系统（预处理）和数据仓库应用系统（联机分析）。日常事务管理系统在 SDMAS 中的作用是采集数据源和事务联机处理，并定期向数据仓库管理系统提供数据源，数据仓库管理系统作用是通过 DCMU 对源数据进行综合、变换等工作，产生成品数据，并向数据仓库用户系统提供成品数据和查询知识集，数据仓库用户系统作用是将成品数据供用户进行数据分析和查询等决策支持工作。整个系统按流水线工作，但是每个系统可独立工作（见图 1）。

软件支撑环境可由数据库管理系统软件构成，如 VFP、Oracle、Sysbase、SQL 等，网络环境可以是 Netware 或 Unix 或 Windows NT，开发方法利用现代编程技术，如面向对象技术、可视化技术，并采用客户/服务器结构技术实现，编程工具可用 Power Biulder、Borland Dephi、VC ++ 等。

三、系统详细设计

这一节我们将详细地介绍 SDMAS 各子系统的构成和作用。

1. 日常事务处理系统

该系统主要作为用户日常处理信息的管理，如果用户单位已有该系统，那么只要给它增加一个数据传送功能，将数据源定期传送到数据服务器；如若用户单位没有该系统，那么可作为数据仓库开发的一部分，进行普通的管理信息系统开发，其中包括数据传送功能。数据传送将数据加上时间日志放到另一个标准库中，等待数据仓库读入功能取走，一旦取走后在标准库中清空。该系统作为一般业务管理信息系统开发。

2. 数据仓库管理系统

该系统是 SDMAS 的主要系统之一，功能是对各种数据源进行管理和处理，将成品数据提交数据仓库用户系统。该系统由数据模型管理器（DMMU）和数据中央开采器（DCMU）组成，下面分别介绍。由于数据仓库的数据复杂，我们首先介绍一下数据仓库的数据模型。

（1）数据模型：数据仓库中主要涉及的是元数据，元数据是一个高度综合的概念，为了讨论方便，这里将所涉及的数据分为四大类数据：原始数据（数据源）、成品数据、开采功能数据、功能应用数据，相应有四大类数据库：

①原始数据库—来源于日常事务处理系统。如可从源数据服务器上复制来的，如商场中的一个主要源数据库是每一种商品的成交信息和商品信息。

②成品数据库—来源于数据中央开采器对数据源的加工生成。例如商业销售分析数据仓库系统的成品数据库包括：类码、统计、预测、差异、趋势、关联、规律、规则等成品数据库。成品数据库涉及四个主要数据类：类码（如品名）、位置码（如柜台号）、时间码（如日、月、年）和计算码（如利润和销量）。成品数据库具有两个重要的作用，一是为用户提供成品数据的浏览分析，另一个主要作用是为用户提供决策支持。

③开采功能数据库—是数据中央开采器的信息库，存放数据开采功能的信息返回信息，如开采功能的联结、数据调用、模型和算法等信息，为建立和生成 DCMU 提供必要的数据，该库的产生由人工确定，其数据结构基本上是固定的。

④功能应用数据库—是由数据中央开采器通过的数据开采而自动产生，为前端用户系统对成品数据分析提供必要的检索信息，它有成品数据知识库、查询功能库和分类查询结构库构成，为了保持良好的通用性，库的结构是固定的。它是解决数据仓库大量而复杂的查询重要手段，为提供通用的查询和标准的 SQL 查询程序，减少查询程序的编制工作起到至关重要作用。

（2）数据模型管理器（DMMU）。数据模型管理器是用来管理数据仓库的数据模型的定义、修改、索引和删除等功能，它的主要功能 ：原始数据读入、创建和重建索引、四类数据的浏览与删除。尽管数据仓库数据复杂，但大多成品数据库的结构是自动产生的，并且几乎所有的数据添加记录和索引的更新均是自动的，因此 DMMU 的功能不多，而且基本为定期的例行维护。DMMU 由一个人机交互界面的窗口进行对数据模型的管理。

（3）数据中央开采器（DCMU）。数据仓库的主要目的是为用户提供综合信息，这些信息不仅反映了用户经营的状况和未来的发展趋势，而且能帮助用户制订未来

的经营策略，如商场经营管理中的综合信息是对销售的状况的数据详细了解程度。数据开采是数据仓库应用的一个主要手段之一，开采的目的是从大量的数据中发现有用的信息，近年来数据开采的算法和软件很多，但是如何将一个数据仓库系统中所要的所有开采算法进行有效的使用，是一项复杂困难的问题。为了避免数据分析人员去掌握大量的开采算法，提高分析和决策人员的效率，在这里我们提出以数据中央开采器为核心的提供综合信息的方法。

我们将数据仓库的对数据源加工的所有算法汇集在一个自动处理的软件系统中，(注意这里数据开采的概念是狭义的，与一般的数据开采概念不仅相同，我们把对数据源进行预加工成成品数据的过程称为数据开采)，这个软件系统称为数据中央开采器，数据中央开采器由三个部分：开采功能管理、数据开采器生成、自动数据开采构成，下面分别详细介绍。

①开采功能管理。开采功能管理为数据开采器提供开采模型和算法，并管理开采功能数据库的开采功能的添加、浏览和修改，模型和算法由一系列程序构成，和调用信息一起存放在开采功能数据库库中，供自动调用。开采功能可由人工逐渐向开采功能数据库增加，开采功能越多，DCMU 的数据开采功能越强。如一个开采功能的增加过程为：编制模型和算法程序形成一个函数或对象放入开采功能数据库，并放入调用的信息和需要返回的信息。

②数据开采器生成。开采器的生成由一个可视化人机交互窗口操作，按一定次序计算，其工作过程主要是选择一个开采功能，然后根据提示输入调用信息，逐个从开采数据库中联结开采功能，并生成相应的成品数据结构库直到全部联结完成，然后编译形成一个自动开采可执行程序。由于开采功能可以更新，所以数据开采器可根据实际需要不断地更新升级。有关成品数据库结构按年度建立，每年生成新的库结构。

③自动数据开采。一旦数据开采器生成，我们就可以自动地对原始数据进行加工，并自动产生相应的成品数据和功能应用数据。

例如在商业销售分析数据仓库系统中，其 DCMU 的模型和算法按分类、统计、预测、差异、趋势、关联、规律、规则等分为 8 类。每一类至少提供一个通用模型和算法。开采计算流程如图 2：

整个开采过程是：首先由原始数据库建立分类库，然后建立差异、趋势和关联等库，最后由它们进一步建立规律、规则库。

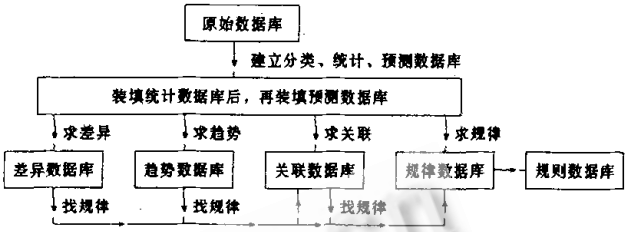

图 2 开采计算流程示意图

3.数据仓库用户系统

数据仓库用户系统是一个客户/服务器模式的多用户系统，它由成品数据应用器(服务器端)、用户联机数据分析浏览器、用户决策支持器和成品数据维护器构成，它利用数据仓库管理系统提供的成品数据和功能应用数据为用户实现数据分析和决策支持。

(1)成品数据应用器(PDAU)：

查询功能管理—为用户联机数据分析浏览器提供对成品数据库的通用查询程序(对象)，将所有的查询程序放在查询功能库中，供自动查询调用该功能管理查询程序的创建、修改、删除等工作。每一个查询程序由三个部分：调用参数、执行程序、返回数据。每一个查询程序的工作原理大致为：当启动一个程序时，根据调用参数打开成品数据库，读入所需的成品数据，并将它返回到用户联机数据分析浏览器供数据显示分析的标准数据库中。

联结查询功能—为成品数据知识库和分类结构查询库中填入查询功能号，将查询功能库中的查询功能号与成品数据库形成一个映射，通过人机交互联结，以实现自动查询。

自动查询—这时一个可执行的通用程序，根据功能应用库我们可以为用户联机数据分析浏览器实现对成品数据库中所需的数据进行快速调用到供数据显示分析的标准数据库中。自动查询提供三种方式：

①浏览查询—用户按次序输入(选择)查询类别和计算域、时间域或位置域，然后在屏幕上显示类码数据库，选择主题码，找到成品数据库文件名，打开分类结构查询库得到查询功能号，由查询功能库调用查询程序，返回成品数据。

②主题查询—用户直接输入查询类别、类别、时间、位置、计算码等，程序到成品数据知识库去按主题词输入

进行模糊匹配，找到相应的查询功能号和调用参数后，到查询功能库中调用查询程序，返回成品数据。

③自然语言查询—用户直接输入含有查询类别、类别、时间、位置、计算码等的自然语言，程序到成品数据知识库去按输入进行成品数据语义匹配，找到相应的查询功能号和调用参数后，到查询功能库中调用查询程序，返回成品数据。

(2)用户联机数据分析浏览器(UOLPAU)：用户联机数据分析浏览器是实现数据仓库对历史信息的综合分析的窗口，它的主要功能是对成品数据进行可视化，以便进行数据分析。通过表格、图形、对比、旋转等可视的方法，把成品数据中的统计、预测、差异、趋势、关联和规律等数据按不同的时间或空间表示出来，以达到对数据分析和帮助决策的目的。

浏览器提供三种对成品数据调用的方式：浏览查询、主题词查询、自然语言查询，并且提供成品数据库的预报，为了做到简单、方便、通用，浏览器设有便于显示标准的数据库格式，所有显示程序和界面窗口程序尽量做到不依赖于成品数据结构。浏览器通过成品数据应用器工作，其工作流程图 3 如下：

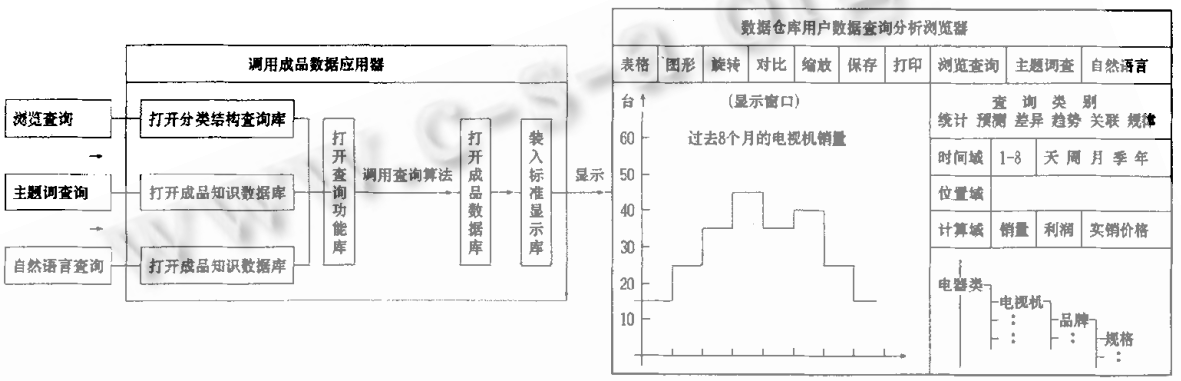

图 3　浏览器工作示意图

成口数据预报功能—这一功能的目的是为用户将成品数据库中存在的最显著性信息找出来，如差异库中利润最大的商品品牌，那类品牌商品销量近一个月一直保持领先地位。当用户从浏览器中启动这一功能时，通过一个通用算法从差异、趋势、关联和规律库中，读入用户浏览器显示。这样做可以为用户提供最显著的综合信息，避免用户花大量时间在浏览器中寻找最显著信息。

(3)用户决策支持器(UDSU)：决策支持器是另一个面向用户的数据分析器，其主要功能是一些专用数学模型或推理算法通过成品数据库进一步加工为用户提供决策支持方案或综合信息。用户决策支持器也是一个专用器件，根据用户需要开发，它是否存在不影响其他的器件开发和正常使用。它主要由两个部分：决策模型程序与人机交互窗口构成，决策模型主要含有(非)线性决策、目标决策、多目标决策、群体决策随机决策和推理模型等，每一个模型对成品数据库专门开发。

为什么这里单独设计一个决策支持器？表面上看，中央数据开采器所得到的成品数据库也为用户提供了一定的决策支持，但是它仅提供了决策支持的较原始的依据和信息，没有提供决策的方案或行动，以及所采取决策的行动或方案可能达到的效果。当然，我们也可以把决策支持模型放入数据中央开采器中，但是决策支持不需产生大量的信息，关键一点是，中央数据开采器主要产生事实和预测数据，数据具有确定性和连续性，而决策支持产生的是不确定的数据(方案、行动、后果等)。因此，从数据的处理方式、规模、结果和性质上看，决策支持器是不同于数据中央开采器。

(4)成品数据维护器(PDPU)：成品数据维护器是用来维护成品数据库的备份、删除、重建索引等工作。

参考文献

[1]　黄璜，数据库技术的发展方向，计算机工程与应用，1995,31(5),1－5

[2]　王晓军，数据仓库在决策支持系统中应用，计算机工程与应用，1997,33(12),43－45

(来稿时间：1999 年 1 月)