一种数据仓库的通用框架——参照结构

田继华，　郑全弟
(空军工程大学 导弹学院，陕西 三原　713800)

摘　要:论述了一种数据仓库设计与开发的通用框架——参照结构。重点描述了参照结构的原理及各部分构成,最后举例说明了它在数据仓库分析中的两种应用方式——垂直剖析和水平剖析。

关键词:数据仓库;参照结构;垂直剖析;水平剖析
中图分类号:TP392　　**文献标识码**:A　　**文章编号**:1009-3516(2002)02-0068-03

数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合,用于支持经营管理中的决策制定过程[1-2]。

随着数据仓库技术的兴起,基于数据仓库的各类应用系统、软件、产品应运而生,但同时也有许多这样的系统半途夭折了。原因是多方面的,但系统在设计与分析时就不够完善却成了众多数据仓库应用系统失败的最主要原因。从根本上说是缺少合理、实用的构架[3]。本文提出一种数据仓库通用框架——参照结构以帮助解决这一问题。

1　数据仓库参照结构

1.1　什么是参照结构?

在拥有多种技术、多个制售商、多种定义和术语的世界里,需要有一种通用的结构来进行比较、作出选择、评估风险、建立标准,这种结构就称为参照结构[4]。数据仓库参照结构就是把数据仓库方案的各组成部分进行分离的通用框架。

1.2　使用参照结构的好处

1)提出了一种通用的框架。使得数据仓库应用系统的投资者、开发者和使用者在讨论、评估以及实现中有通用的结构可供使用。

2)实现不同的组成方案。例如,有的企业可以只建立数据站场,而以后再逐步建立数据仓库。

3)帮助用户按阶段完成目标。参照结构为用户提供一个持久的框架,以便用户按阶段完成预定目标。

4)对制售商提供的各种构件和工具进行选择。通过参照结构可以使用户对不同的解决方案进行精心比较从而选择合适的构件和工具。

1.3　参照结构总体介绍

数据仓库参照结构把数据仓库应用系统的组成部分划分成块和层。块主要有数据源、数据站场、数据仓库、存取与使用等。层主要有数据管理、元数据管理、传输和基础结构等。块与特定的数据仓库相关,而层则表示用于实现块的环境[5],如图1所示。具体来说:

数据源模块:主要为数据仓库提供原始数据。可以分为内部数据、外部数据和元数据等。

数据仓库模块:是结构的主要支撑部分。分为求精、重构和数据仓库三部分。求精部分负责标准化、净化、过滤与匹配,以及为所抽取数据的原始信息附加时间标记,还包括一些元数据的抽取和创建等。重构部分负责检验数据是否满足用户分析的需求,完成数据的集成与分割、概括与聚集、预算与推导、翻译与格式

收稿日期:2001-01-11
作者简介:田继华(1976-),女,江苏南京人,硕士生,主要从事计算机仿真研究.

化、转换与映射以及元数据的创建等。数据仓库部分是这一模块的主体部分,包括建模、概括、聚集、调整与确认、建立结构化查询、创建词汇表等。

数据站场模块:该块与数据仓库模块的主要区别在于最终用户的侧重点上。数据站场主要偏重于用户的商业目标。

数据仓库存取和使用模块:该块帮助数据仓库发挥作用,提供各种对数据仓库进行操作的工具及应用程序。它包括:存取与检索、分析与报表两部分。存取与检索部分负责对数据进行检索及多维转换等操作。分析与报表部分是一组作用于数据仓库或数据站场的工具和应用程序集合,比如,报表工具、分析工具、决策支持工具、数据挖掘工具等。

数据管理层:主要完成数据的抽取、添加、恢复等任务。

元数据管理层:主要负责管理数据仓库所使用的元数据。

传输层:主要完成不同模块间数据的传输任务。比如客户端和服务器端数据按 TCP/IP 协议进行网络传输等。

基础结构层:包括一些为数据仓库服务的管理和应用系统。比如系统管理、工作流程管理、存储管理等。

1.4　数据仓库参照结构的应用

如同前面提到的,数据仓库参照结构为数据仓库应用程序的开发者提供了一种搭建各部件的参考框架,并可以为系统的设计、分析及模拟运行提供便利。下面通过将其运用于系统分析来具体展示它很好的作用和易用性。

有两种方法将数据仓库参照结构用作系统分析——垂直剖析与水平剖析。

1)垂直剖析。垂直剖析就是把块分割开,把各层分为段以形成多个"分割"。在同一"分割"中的块共享同一层,也就是共享同一环境。因此垂直剖析创建了硬件和软件平台的边界。垂直剖析非常有用,因为根据它可以制定许多重要决策。比如,数据源与数据仓库是否应运行于相同平台上?数据站场有没有存在的必要?

图 2 是某单位对参照结构的垂直剖析方案。该单位将数据源和数据仓库划分在同一分割中,数据站场、存取与使用则处在另外两个分割中。如同前面说过的,在同一分割中的块共享同一层,也就是共享同一环境。因此,数据源与数据仓库必须处于同一平台上,并且共享相同的传输和基础结构机制。如果数据源程序运行于主机的 DB2 之上,那么数据仓库也必须运行于基于相同 DB2 的平台上。使用这种结构时,主机功能必须足够强大以适应数据仓库运行的需要。

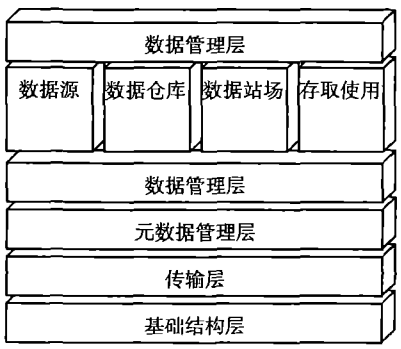

图 1　数据仓库参照结构

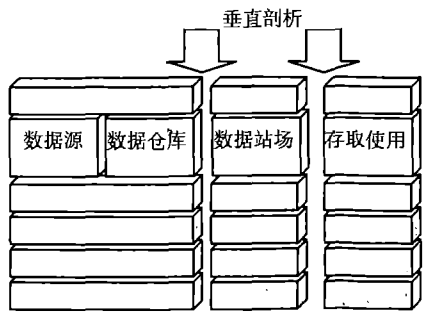

图 2　数据源与数据仓库位于同一平台的垂直剖析

图 3 是另一种垂直剖析。它把数据源、数据仓库、数据站场和存取与使用全部分割开。这便是常用的基于主机和客户/服务器的数据仓库应用系统结构。数据源及相应程序环境位于宿主机上,而数据仓库则存在于能提供充分处理支持的大型服务器上。数据站场则位于同一台或另一台服务器上,负责管理数据站场构件。存取与使用则位于客户的工作站上。这种结构适合于主机功能不很强大,但拥有数据服务器以服务多个客户程序的情况。

由此可见,通过对数据仓库参照结构进行适当的垂直剖析可以明确系统建立的分配方式、使用硬件和平台、功能划分等问题。如何选择合适的垂直剖析方式,要根据单位以及使用者的具体要求和现有条件来决

定。

2)水平剖析。一旦建立了垂直剖析,就必须考虑许多与实现有关的问题,如创建数据仓库需要什么样的人员、为数据仓库的各个构件选择什么产品和软件等。这时就需要使用水平剖析。水平剖析就是将参照结构分成不同的"片",每一"片"完成各块和层的一部分任务。

图 4 是支持项目成员和技术的参照结构水平剖析方法。它可以帮助单位了解数据仓库应用系统实现过程中个人的作用,以及个人需完成的任务。比如,数据仓库的创建任务就需要处在不同片中的技术人员完成,包括数据管理人员、建模人员、报表分析人员等。

图 3 基于主机和客户/服务器的垂直剖析

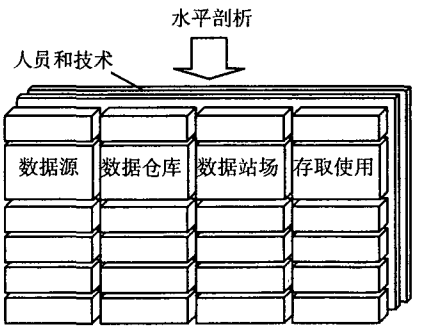

图 4 基于工程人员及技术的水平剖析

2. 结束语

以上论述了参照结构这一数据仓库通用框架,对它进行一定的操作,比如垂直剖析或水平剖析等,便可以极大地方便数据仓库开发以及系统分析人员的工作,是设计开发数据仓库应用系统的有效工具。

参考文献:

[1] Tom Hammergren. 数据仓库技术[M]. 北京:中国水利水电出版社,1998.
[2] Inmon W H. Building the Data Warehouse[M]. New York: John Wiley & Sons, 1996.
[3] Harjinder S Gill. 数据仓库—客户/服务器计算指南[M]. 北京:清华大学出版社,1997.
[4] 张水平,郑雪雁. 数据仓库技术研究[J]. 空军工程大学学报,2000,1(3):68-71.
[5] 王 珊. 数据仓库技术与联机分析处理[M]. 北京:科学出版社,1998.

(编辑:田新华)

A Kind of Universal Framework of Data Warehouse - Reference Structure

TIAN Ji-hua, ZHENG Quan-di
(The Missile Institute, Air Force Engineering University, Sanyuan 713800, China)

Abstract: This paper presents a kind of universal framework of data warehouse - reference structure, and mainly describes the principle and each component of the reference structure, finally illustrates two ways of applying the reference structure to data warehouse analysis, i.e. vertical analysis and horizontal analysis.

Keywords: data warehouse; reference structure; vertical analysis; horizontal analysis