数据仓库集成环境研究与实现*

崔晓军[1,2]，薛永生[2]

（1.襄樊职业技术学院 信息技术系，湖北 襄樊 441050；2.厦门大学 计算机科学系，福建 厦门 361005）

摘 要：引入开放式设计思想，使得数据仓库集成环境具有很强的适应性，该集成环境架构在.NET平台上，采用组件开发技术，使系统具有良好的可靠性、可扩展性和安全性。

关键词：数据仓库；集成环境；XML；体系结构

中图法分类号：TP311 文献标识码：A 文章编号：1001-3695(2006)12-0178-03

Research and Realization of Integrated Enviromment for DataWarehouse

CUI Xiao-jun[1,2], XUE Yong-sheng[2]

(1.*Dept. of Information Technology, Xiangfan Vocational & Technical College, Xiangfan Hubei* 441050, *China*; 2.*Dept. of Computer Science, Xiamen University, Xiamen Fujian* 361005, *China*)

Abstract: The design of EDWIE (Integrated Enviromment for Enterprise DataWarehouse) adopts the open thought and the software is based on the .NET software architecture, so the flexibility, reliability, expansibility and security of this system are favorable.

Key words: DataWarehouse; Integrated Enviromment; XML; Architecture

在Internet飞速发展的网络信息时代，信息资源的经济价值和社会价值越来越明显，许多企业在信息化的过程中积累了大量的数据，这些数据已成为企业一个巨大的“宝藏”，如何从中挖掘出有价值的信息，使其能更好地为企业进行决策支持，已成为企业信息化进程中亟待解决的问题[1]。这种需求导致了数据仓库技术的飞速发展，许多企业纷纷上马建立自己的数据仓库，然而在建设过程中由于资金、技术等各方面的限制，未能很好地进行数据仓库应用系统的开发，建成的数据仓库并没有发挥出“信息宝藏”的作用，更有甚者把数据仓库建成了“数据监狱”。

本文旨在研究一种适合我国国情的企业级数据仓库的构建、部署、管理、使用和维护的集成机制和体系结构，即一个开放式数据仓库集成环境的软件平台。在此数据仓库环境平台上，企业可以很方便地建立自己的数据仓库系统及数据仓库应用系统，降低数据仓库的建立、维护成本，提高数据仓库的应用效益。

1 数据仓库集成环境

1.1 数据仓库

William H. Inmon在《建立数据仓库》中指出，数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合[2]。

1.2 数据仓库管理系统

数据仓库管理系统是负责对数据仓库进行管理的软件，它包括数据抽取、转换、加载、存储、查询和维护以及元数据管理等功能。在实际应用中，通常由一组完成专门功能的数据仓库工具加上数据库管理系统组成。

1.3 数据仓库应用系统

数据仓库应用系统是在建立的数据仓库和数据集市上进行查询、分析、挖掘等操作，是对企业决策支持服务的应用系统，即数据仓库的前端工具[3]。数据仓库应用系统主要由两部分组成：①数据查询。它必须能提供比较快速、灵活的数据查询功能，能保证从数据仓库中有效地获取信息和支持决策。②数据分析。它通常包括OLAP分析、数据挖掘、统计分析和OLAM等技术。数据分析技术不一定需要建立在数据仓库的基础上，但实践证明，建立在数据仓库基础上的数据分析的效率和能力将大大提高。

1.4 数据仓库集成环境

数据仓库的集成环境(Integrated Enviromment for Enterprise DataWarehouse, EDWIE)是集数据仓库、数据仓库管理系统和数据仓库应用系统于一体的集成系统。在该集成环境下，企业可以根据自己的实际情况很方便地进行数据仓库的建模、数据抽取与转换、元数据管理及智能化分析与输出，无须再借助另外的工具软件。

2 体系结构设计

2.1 逻辑结构设计

EDWIE系统在逻辑上分为六层：数据源、ETL层、数据存储层、应用服务层、Web服务层和表示层。其结构如图1所示。

收稿日期：2005-09-26；修返日期：2006-07-16

基金项目：国家自然科学基金资助项目(50474033)；福建省自然科学基金资助项目(A0310008)；福建省高新技术研究开放计划重点项目(2003H043)

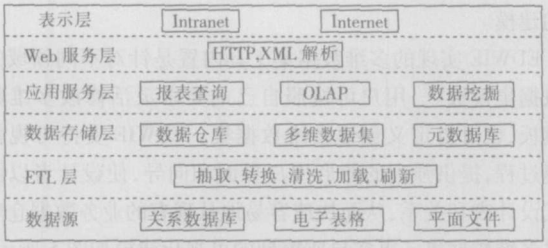


图 1　数据仓库集成环境体系结构图

(1)数据源。数据仓库的数据应包括分析、决策所需的各种数据,既可来源于企业的联机事务处理系统(OLTP),又可来源于企业外部。从数据组织形式上,既有关系型数据库,又有电子表格和平面文件等半结构化的文件。

(2) ETL层。它是实现数据从数据源到数据仓库的工具层,要完成数据的抽取、转换、清洗、装载及刷新[4]。

(3)数据存储层。它是整个系统的核心,以统一的数据格式集中存储企业级数据仓库、多维数据集及元数据库。

(4)应用服务层。它由 XML解析器和查询分析处理两个组件组成,主要功能是接收从客户端经过 Web服务器发送来的 HTTP请求,负责到数据存储层访问相应的数据,实现报表查询、联机分析处理(OLAP)和数据挖掘(DM)等处理,并将结果以 XML的格式返回 Web服务层。

(5)Web服务层。它负责管理客户端与应用服务器之间的信息流,接收客户端的 HTTP请求和 XML消息,将其解析后传递给应用服务器;将应用服务器运行的结果以 XML的形式返回至客户端的浏览器。

(6)表示层,它也称为客户端层。在客户端既可得到 HTML文档,也能得到 XML文档;既能通过 HTML发送查询请求,也能接收 HTML和 XML格式的结果;客户既可来自于 Intranet也可通过 Internet访问本系统。

2.2　模式设计

本系统采用 B/S和 C/S两种模式相结合的方式。数据从数据源经过 ETL(抽取、转换、清洗、加载)至数据仓库的处理过程采用 C/S模式。该过程(图 2)涉及大量的数据转移,对访问速度的要求很高,因此适合采用 C/S模式。为了在数据抽取、转换、清洗时不影响数据仓库的使用,系统采用独立的 ETL服务器进行处理。数据仓库应用服务采用三层 B/S模式,如图 3所示。

(1)客户端通过浏览器将数据查询请求发送给 Web服务器。

(2)Web服务器将客户端的 HTTP请求翻译成 XML格式。

(3)当应用服务器接收到 Web服务器的 XML格式的查询请求时,就将查询请求送到 XML解析器。

(4)XML解析器对查询进行分析,区分是需要访问数据仓库中的业务数据还是元数据。

(5)应用服务器通过 ADO.NET访问数据存储区中相应的数据,在应用服务器中进行计算、查询等操作,操作完成后将需要保存的数据存入数据仓库服务器。

(6)应用服务器的 XML处理器将查询结果转换成 XML格式,发送给 Web服务器。

(7)Web服务器将收到的查询结果转换成浏览器中显示的格式,显示在客户端的浏览器上。

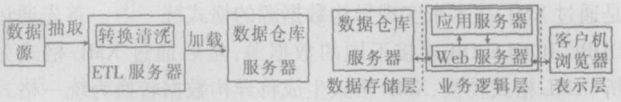


图 2　ETL过程

图 3　数据仓库应用系统结构

3　功能结构设计与实现

3.1　系统管理

系统管理主要负责整个系统的访问控制、日志管理和数据仓库的备份与恢复。①对于系统的访问控制,采用用户—全局组—角色的三级访问策略,即用户通过加入全局组获得权限,而全局组通过加入角色获得权限,角色直接拥有系统访问的权限。②日志是数据仓库保证数据一致性的重要依据。EDW IE将用户对数据仓库所做的表级别的操作都记录下来作为数据仓库的事务日志,结合 DBMS自身的日志功能进行管理。③数据备份与恢复主要实现数据仓库的数据备份和数据恢复。备份策略包括完全备份、差异备份和日志备份;数据恢复根据恢复策略选择不同的备份数据进行操作。

3.2　ETL

ETL是指从数据源获取数据,并经过清洗、转换、集成后,将其加载到数据仓库的过程。其中,清洗是指发现数据、模式的不一致、不兼容并加以消除,提高数据质量[5];转换是将操作数据转换成另一种格式以更加适用于数据仓库设计;集成是将业务数据从一个或几个来源中取出,并将数据映射到数据仓库的新数据结构上。清洗和转换用来保证数据的准确性和一致性,集成用来保证数据的整体性,因此 ETL过程的顺利实施是保证整个数据仓库系统正常使用的关键。

ETL从功能上分为完全 ETL和增量 ETL两种不同的处理。完全 ETL用于建立或重建数据仓库,它由数据导入、数据抽取、数据清洗、数据转换、用户评估和数据加载几个过程组成。其流程如下:通过 ODBC和 OLE DB连接数据源,将异构数据源的数据统一转换为第三方的关系数据库格式(如 SQL Server),存储在 ETL服务器中的数据导入区;然后建立相应的抽取规则,根据抽取规则将导入区的数据抽取到 ETL服务器中的数据准备区,在数据准备区中进行数据清洗、数据转换,对处理后的数据由用户进行评估后加载到数据仓库服务器中。

在此过程中,数据导入是关键。由于数据仓库的数据既可以是企业内部 OLTP的数据,也可以是外来的数据,它们通常由不同的应用系统生成,数据格式各异,如何实现各个异构数据源的数据集成是数据仓库系统必须考虑的问题。EDW IE采用的方法是将异构的数据源统一转换成统一格式的关系型数据库存储在 ETL服务器,然后再进行 ETL处理。其工作原理如图 4所示。

图 4　数据源格式转换

源数据直接进入一个格式转换器,格式转换器的设计思想

是通过 XML技术来实现异构数据源的格式统一[6]。首先通过 ODBC和 OLE DB连接不同的数据源，然后通过 XML格式分析、XML格式转换、XML格式生成将异构数据转换为统一格式的关系数据库。其中 XML格式分析模块是对进入的数据进行分析，判断其数据格式，并将每个数据源的数据格式进行记录；XML格式转换模块是用于将各种异构的数据格式转换为标准的 XML格式，在该模块中加入智能搜索引擎，使其能够自动进行格式匹配和格式转换；XML格式生成模块是将 XML格式的数据转换为特定的 RDBMS的格式，该模块事先存储了各种数据格式间的对照关系，即转换标准，这是整个格式转换器的核心。在整个格式转换过程中，XML起到中间数据表示和消息传输的作用。经过格式转换器处理过的数据，就具有了统一的格式，这样大大简化了随后的数据抽取、数据清洗、数据转换等工作。

增量 ETL用于定期对数据仓库的数据进行刷新，其主要流程与完全 ETL相似，只是每次在数据抽取之前先进行增量数据的判别，将增量数据过程采用两种不同的策略存入到 ETL服务器的数据准备区：① 替换数据准备区中相应表的记录；②直接将增量信息追回到数据准备区中相应的表中。然后对数据准备区中的数据进行自动的数据清洗、数据转换，并将其加载到数据仓库完成增量 ETL的处理。

3.3 元数据管理

元数据是关于数据的数据，被称为数据仓库系统的灵魂。通过分析数据仓库集成环境体系结构中各部分与元数据系统交互情况，EDW IE中将元数据划分为两大类[7]：

(1)基本元数据对象模型。它包括数据源元数据对象、数据仓库元数据对象、多维数据集元数据对象、ETL元数据对象、OLAP元数据对象、数据挖掘元数据对象和系统管理元数据对象。

(2)过程元数据对象模型。它包括 ETL过程元数据对象和数据挖掘过程元数据对象。

EDW IE在数据结构设计上采用一种面向对象的元数据结构，即将各种元数据封装在相应的元数据类中，将这些元数据类的对象实例通过层次结构有机地组织起来，构成一种层次型对象模型。系统通过这些对象对各元数据表进行操作管理。系统通过这种对象模型访问元数据，而不需要直接接触元数据库。在经过良好封装的元数据类中包含各种属性和方法，属性表达了相应的元数据值，而方法定义了对相关元数据的各种操作，并负责维护元数据之间的一致性。元数据的存取、更新和管理通过访问这些属性和方法来实现。

从功能模块的设计上，元数据管理包括元数据查询、元数据编辑、元数据更新和函数管理。元数据查询用于元数据的浏览和条件查询；元数据编辑用于对元数据表记录的增加、删除、修改操作；元数据更新用于数据源表结构发生变化时对元数据进行自动更新；函数管理用于 ETL函数和数据挖掘函数在系统中的注册管理。

3.4 多维数据集管理

数据仓库一般使用星型模型、雪花模型或事实星座模型来组织多维数据，在 EDW IE中采用事实星座模型来实现数据仓库的建模。

EDW IE实现的多维数据集工具内置是针对不同领域的多维数据集模板[8]，用户可按照自己的要求灵活修改多维数据集模板，也可自定义创建多维数据集。EDW IE支持可视化的建模过程，提供强大的图形用户界面和向导，使设计者以图形方式设计表和关系，从而快速容易地从现有的业务数据仓库中创建多维数据集。事实星座模型的可视化建模如图 5所示。

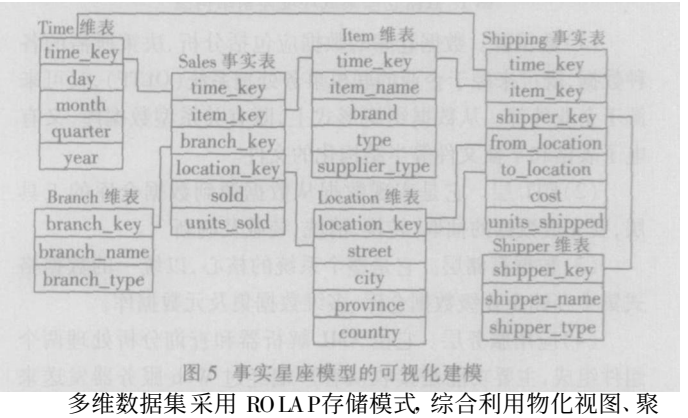


图 5 事实星座模型的可视化建模

多维数据集采用 ROLAP存储模式，综合利用物化视图、聚集表以及链接索引来提高多维数据集响应客户端查询的速度，使其最大限度地满足客户的要求。

多维数据集的更新在数据仓库刷新后自动进行，按照维表、事实表、聚集表和物化视图的顺序依次实现更新。

3.5 联机分析处理

联机分析处理是数据仓库最直接的前端应用。多维数据的展示、报表查询、OLAP操作、自定义计算操作是 EDW IE所提供的基本功能。利用微软的 MS Chart控件实现多维数据的图表展示和树型表格的展示，在多维数据展示基础上还可进行聚合、钻取、切片、切块以及旋转等 OLAP分析动作。OLAP查询包括定制报表的查询、统计查询、自定义查询，并可将查询结果导出为 Excel文件。

3.6 数据挖掘

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的但又是潜在的有用信息和知识的过程。它和 OLAP一起构成数据仓库应用系统。作为 EDW IE重要的前台工具，基于数据仓库的数据挖掘的最大的优势就是数据仓库中数据的高质量，可以利用集成在系统中的 ETL过程实现数据的清洗、变换和集成。

DM从功能上分为新建挖掘模型、历史挖掘模型浏览和挖掘模型更新。新建挖掘模型既可以直接创建，也可以利用已有模型创建，它包含数据准备与数据挖掘引擎两个过程。其中数据准备用于获得符合挖掘算法执行要求的数据，包括数据提取、采样、转换等过程。数据的提取来源主要有数据仓库、多维数据集和外部数据，对于外部的数据，系统将直接利用 ETL过程加载入仓库。数据的采样、转换策略全部通过函数实现，通过元数据管理系统中的函数注册模块进行注册。数据挖掘引擎是整个挖掘系统的核心部分，它包括挖掘算法执行、结果展示和模型存储的功能。挖掘算法也是作为函数进行封装并通过函数注册模块注册到系统中。模型存储了本次(下转第 184页)

4.5 访问结果 Bean

JSP通过访问 QueryProInfo的数据负责为用户创建应答,相关 JSP语句如下:

```
<jsp: useBean id= "resultsBean "scope= "session" class= "QueryProInfo" />
```

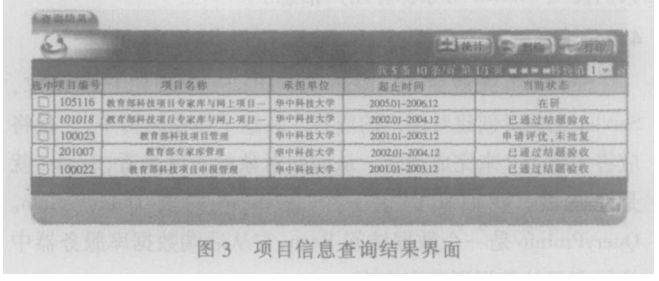
<jsp: useBean>动作在 Scope定义的 Session中寻找 QueryProInfo对象。我们在 Session中放了一个 QueryProInfo Bean的对象,叫做 resultsBean 这个动作可以从用户会话中得到 resultsBean 然后 JSP就可以使用任何一个 resultsBean的 Get方法来访问它的数据:

```
<%= resultsBean getResult( )%>
```

这个 JSP代码可以嵌入在 HTML页面中,应用程序开发者的任务是设计 Servlet数据访问 Bean JSP的变化也不会影响 Servlet

4.6 对用户请求的应答

JSP请求时动态地编译成 Java Servlet并且被应用服务器缓存。用户所接收到的请求是一个 JSP页面,它们包括动态产生的内容,依据 HTML中的诸如 JavaScript脚本将查询结果组成分类显示在用户端浏览器。图 3是进行一个具体查询结果的显示页面。



选中	项目编号	项目名称	承担单位	起止时间	当前状态
□	105116	教育部科技项目专家库与网上项目…	华中科技大学	2005.01-2006.12	在研
□	101018	教育部科技项目专家库与网上项目…	华中科技大学	2002.01-2004.12	已通过结题验收
□	100023	教育部科技项目管理	华中科技大学	2001.01-2003.12	申请评优,未批复
□	201007	教育部专家库管理	华中科技大学	2002.01-2004.12	已通过结题验收
□	100022	教育部科技项目申报管理	华中科技大学	2001.01-2003.12	已通过结题验收

图 3 项目信息查询结果界面

5 结束语

网上项目申报与管理系统,运用工作流的设计方法构建了系统整体功能模块结构,实现了流程的控制管理,提高了网上申报及审批的办事效率;采用表示层、业务层和数据服务层的分层设计体系,使系统层次分明、架构清晰。系统采用 Tomcat服务器,结合 JSP/Servlet技术的解决方案实现了表示层和逻辑层的分离,具有良好的扩展性;JSP页面的动态显示满足了 Web环境下客户对系统的样式需求;基于 Java的平台独立性,使得系统可在不同的操作系统、不同的硬件环境下运行,使得系统具有很好的跨平台性。

该系统作为教育部科技项目专家库与网上项目管理平台的一个子系统,正在被教育部各部门及各大高校广泛使用,并取得了良好的效果。

参考文献:

[1] The Workflow Reference Model[R]. TC00-1003 Hampshire: Workflow Management Coalition 1995.

[2] 张璞,庄成三. 基于 Servlet技术的 Web应用及其实例分析[J]. 计算机工程与科学,2001 23(2):37-39.

[3] 王斌,杨宗凯,吴砥. 基于 J2EE平台的教育资源注册及检索系统[J]. 计算机应用研究,2004 21(6):243-245.

[4] 何川,方兴. JSP编程实践[M]. 北京:清华大学出版社,2002.31-34.

[5] 陈华军. J2EE构建企业级应用解决方案[M]. 北京:人民邮电出版社,2002.102-137.

[6] 陈海山. 深入 Java Servlet网编程[M]. 北京:清华大学出版社,2002.103-181.

作者简介:

王琴(1981-),女,江苏扬州人,硕士,主要研究方向为远程教育;杨宗凯(1963-),男,国家信息技术标准化技术委员会教育技术分技术委员会(CELTS)委员,ISO/IEC JC1/SC36/SP联合主席,国家“十五”攻关网络教育关键技术及其示范工程专家组成员,教授,博导,博士,主要研究方向为网络教育、电子商务、智能信号处理与应用、宽带网络通信技术等;吴砥(1978-),男,主要研究方向为远程教育、电子商务。

(上接第 180页)挖掘任务的数据来源、函数调用及挖掘结果等信息,是本系统数据结构的核心。历史挖掘任务浏览供用户浏览模型中保存的数据挖掘任务及挖掘结果,并可对结果进行评价。挖掘模型更新用于对已有的挖掘模型进行更新。

4 结束语

本文旨在研究一种适合我国国情的企业级数据仓库的构建、部署、管理、使用与维护的集成机制、体系结构及其关键技术。引入开放式设计思想使得数据仓库集成环境具有很强的适应性,为用户提供一个集设计、分析、输出、优化于一体的综合平台,以加快企业级数据仓库的建设过程,提高数据仓库的应用效益,降低数据仓库的维护成本,在中小企业数据仓库的应用和推广中具有实际的价值和重要的意义。

参考文献:

[1] Shim J P, Warkentin Merrill Past, Present, and Future of Decision Support Technology[J]. Decision Support Systems, 2002, 33(4): 111-126.

[2] W H Inmon. Building the Data Warehouse[M]. Beijing: China Machine Press, 2000.

[3] R Kimball, L Reeves, M Ross, The Data Warehouse Lifecycle Toolkit[M]. New York: John Wiley & Sons, 1998. 281-295.

[4] P Vassiliadis, Z Vagena, S Skiadopoulos, ARKTOS: Towards the Modeling, Design, Control and Execution of ETL Processes[J]. Information Systems, 2001, 26(8): 537-561.

[5] 郭志懋,周傲英. 数据质量和数据清洗研究综述[J]. 软件学报,2002, 13(11): 2076-2081.

[6] 关文革,武强,安海忠,等. 基于 Web的分布式数据仓库体系结构的研究[J]. 计算机应用研究,2004 21(6): 64-66.

[7] 贾自艳,黄友平,罗平,等. 面向数据质量的 ETL过程建模与实现[J]. 系统仿真学报,2004, 16(5): 907-911.

[8] 段江娇,黄震华,陈昕. 基于 Web的数据仓库集成研究[J]. 计算机科学,2004 31(10): 114-117.

作者简介:

崔晓军,副教授,硕士研究生,主要研究方向为数据仓库、数据挖掘、XML等;薛永生,教授,主要研究方向为数据库理论与应用、分布式数据库、数据仓库、数据挖掘、网络技术等。