数据整合的利器——数据仓库

刘 学 风

(天津科技大学图书馆 天津 300222)

摘 要 阐明了数字资源整合的必要性和层次模型,分析了利用数据仓库进行数据整合的可行性。介绍了数据仓库的关键技术数据抽取、转换和加载,最后利用Oracle数据仓库构建器和实验数据进行数据仓库的构建试验。

关键词 数字资源整合 数据仓库 数据整合

中图分类号 G250 **文献标识码** A **文章编号**:1002-1965(2009)0107-03

1 背 景

随着以计算机技术为代表的信息技术的迅速发展,图书馆数字信息资源的建设成绩斐然,作为一个专门搜集整理保存并传播信息的机构,图书馆拥有非常丰富的数字化信息资源,这些数字资源既包括图书、期刊、报纸、标准、专利产品等全文数据库;又包括如文摘、题录和书目数据的数据库和相关系统。

从计算机应用系统的角度来看,这些数字资源系统的硬件平台、数据库管理软件、数据库元数据、用户端应用程序都不尽相同,因而具有分布异构性。从用户的角度来看,各个数字资源系统中数字资源内容交叉重复,影响用户对信息的选择与获取;数字资源间的知识关联程度低,知识结构体系遭破坏等。

然而用户希望在统一的检索环境和检索界面下,以最小的时间和精力实现"一站式"的文献检索、浏览和使用。图书馆也希望其用户的信息需求得到充分满足,并更好地体现其"以人为本"的服务理念,提高数字资源的利用率。因此,无论是用户方面还是图书馆方面都迫切地需要解决由于"信息爆炸"和"信息污染"给信息利用带来的不便。数字资源整合在这种情况下应运而生。

2 数字资源整合的层次模型

从计算机信息系统集成的角度来看,根据数字资源的分布异构性,数字资源整合从低级到高级可以分为四个层次:网络整合、系统软件整合、数据整合和应用整合[1~2],如图1所示。

网络和系统软件的整合是基础,目前各数字资源系统基本上都运行在集成的网络和系统软件基础上,而数据整合和应用整合是数字资源系统能否有效整合的关键。数据整合就是采用合适的技术手段将数字资源系统中的异构数据按一定的规则组织在一起,方便用户的有效访问。除数据库技术和较常见的中间件技术外,数据仓库技术是解决数据整合问题的重要技术。利用中间件整合异构数据库并不需要改变原始数据的存储和管理方式,中间件位于异构数据库系统(数据层)和应用程序(应用层)之间,向下协调各数据库系统,向上位访问整合数据的应用提供统一的数据模式和数据访问的通用接口。各数据库仍完成各自的任务,中间件主要为异构数据源的高层次的检索来服务。这种数据整合的思路是,通过模式翻译器将局部数据库模式以某种公共数据模型为基础映射成局部集成模式,然后通过模式集成器将各个局部集成模式按需要采用全局数据模型来定义,最终成为全局概念模式。

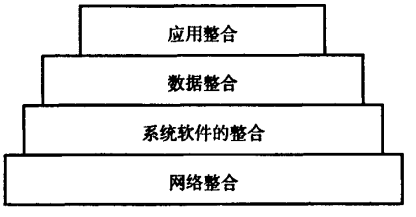

图1 数字资源整合层次图

3 利用数据仓库进行数字资源整合的可行性

数据仓库创始人W.H.Inmon于1991年在《Building the Data Warehouse》将"数据仓库"(Data Warehouse,DW)定义为"一个用以更好地支持企业或组织的决策分析处理的、面向主题的、集成的、非易失的且随时间变化的数据集合"[3]。元数据是数据仓库的核心。多个数据源的异构数据向数据仓库集成时,多个数据源的局部数据模式映射到数据仓库的全局数据模式,并对各个数据源数据进行抽取、转换和清洗、装载等处理来为数据仓库应用程序使用。

虽然最早数据仓库的提出是用于企业的联机分析系统和决策支持系统,但分析数据仓库的基本特征,如集成性、稳定性和随时间不断变化的特点,对于数字资源系统数据层的整合来说,数据仓库可以将各个异构数字资源系统数据库中

收稿日期:2009-03-07
作者简介:刘学风,男,1977年生,硕士。

的数据集成到数据仓库中,将各数字资源系统中的变化按一定的周期更新到数据仓库中,并保持数据仓库中集成数据的相对稳定。因此数据仓库是实现图书馆数字资源数据层整合的有利工具。

4 数据仓库的关键技术:ETL

创建数据仓库最重要的步骤是将数据从各种操作型数据系统抽取出来,排除数据中缺陷,完成一系列转换,最后将数据加载到数据仓库。这个过程称为数据抽取、转换和加载即(Extraction, Transformation and Loading,简写为 ETL),这项工作一般要占到整个数据仓库创建工作量的 75% 到 80%。

数据抽取即从数据源系统抽取数据仓库系统需要的数据,并不是源数据库的所有细节数据对于数据仓库的主题都有用,必须根据已确定主题的需要,从原有操作型数据库中抽取相关数据到数据仓库。

数据清洗就是将错误的、不完整的、不一致的数据在进入数据仓库之前进行更正或者删除,以免影响数据的质量。由于待整合的各系统的数据存在很多问题,比如:滥用缩写词、惯用语、数据输入错误、数据中内嵌的控制信息、重复记录、丢失值、拼写变化、不同的计量单位和过时的编码等,我们必须针对系统的各个环节通过试抽取,将有问题的记录先剔除出来,根据实际情况调整相应的清洗工作。数据转换是将源数据变为目标数据的关键环节,它包括存储格式转换、类型修正、字段运算、信息合并等。严格来说,数据清洗可看作是数据转换的一种。

数据加载是将从数据源系统中抽取、转换后的数据加载到数据仓库系统中。数据的加载可使用数据仓库厂商提供的数据加载工具进行数据加载,也可通过数据仓库厂商提供的 API 编程进行数据加载。

5 利用数据仓库进行数字资源数据整合的试验

利用数据仓库对数字资源进行整合,关键在于怎样将各个数字资源管理系统中的数字资源整合到数据仓库中,即数据仓库系统的设计和实现问题;并保证数据仓库中数据的完整性和正确性、无重复等。而数字资源整合数据仓库设计的主要任务是将各个数字资源系统中的数据进行抽取、转换、清洗、加载,形成完整一致的数据,以向用户提供一站式的检索服务。本实验将利用 Oracle 数据仓库解决方案和试验数据进行数字资源数据仓库的构建试验。

5.1 **试验平台** 本次试验的硬件环境:CPU2.0GHz,内存:512M,硬盘 40G。系统的软件环境:操作系统为 Windows 2000 Server with SP4,数据库系统 Oracle 9.2.0.1.0[4],数据仓库工具采用 Oracle Warehouse Builder 10g release 1(10.1)[5]。

5.2 **源数据准备** 书目数据来源于北京大学出版社 marc 数据下载论坛,通过书目数据软件 Marc Converter 1.5(试用版)将书目数据的书名、作者、出版社、ISBN、中图分类号、版本等信息转换成以逗号分隔的文本文件格式。此外,还有出版社、中图分类的数据以文本文件的格式存储,它们构成数字资源整合数据仓库的源数据[6~7]。

5.3 **数据仓库设计与实现** 利用 OWB Client 连接到设计元数据库(Design Repository)即可进行元数据的设计,即数据仓库逻辑模型的设计。首先创建源模块 LIB_SOURCE 和目标模块 LIB_TARGET,模块是用来标记数据存储的物理位置;然后将源数据文件的元数据导入到源模块中;在目标模块中分别创建三个目标表 TB_PUBLISHER、TB_CATALOG、TB_CLASSIFICTION 用来存放从源文件抽取过来的源数据;最后创建 MAP_INITI_CLASSIFICTIONTAB_LOAD、MAP_INITI_PUBLISERTAB_LOAD、MAP_INITI_CATALOGTAB_LOAD 三个映射并部署和执行,将源数据从源文件经过 ETL 过程到目标表中,数据加载后即可供数据仓库应用程序编程使用。关键步骤如图 2 至图 6 所示(见下页)。

经清洗、转换加载后的数据仓库数据,可以供应用编程使用,即数据仓库应用的开发。这里,仅用 Oracle 的查询工具 SQL PLUS 来查看。如查询题名等于"财务管理分析"记录的 ISBN 号、出版日期、出版者、题名等的执行结果见图 7(见下页)。

6 结束语

由于源数据的不完全性,在数据仓库建模试验过程中仅限于源数据能够提供的信息,而在实际数字资源整合的过程中,应该根据图书馆数字资源管理系统的提供源数据的实际情况,综合考虑数字资源整合的标准化和用户查询的需要,遵循数字资源的相关标准如中国数字图书馆标准与规范基本元数据及扩展集标准等来建立整合数据仓库的数据模型。

此外,由于时间和试验条件如网络、计算机性能、数据的可获得性等的限制,只能利用少量试验数据来模拟数据仓库的建立过程,但实际数字资源整合数据仓库的建立将处理更复杂的关系、更大的数据量和数据更新等问题,所以实际过程比模拟过程复杂,这些都有待于进一步的研究。

参 考 文 献

1 邓 苏,张维明,黄宏斌等.信息系统集成技术[M].北京:电子工业出版社,2004:15-16.

2 朱建刚,杨春梅.利用中间件进行异构数据库互访[J].福建电脑,2005(11):6-7.

3 W H Inmon 等著;王志海等译.数据仓库[M].北京:机械工业出版社,2002:20.

4 Oracle Corporation. Oracle9.2.0.1.0 数据库管理系统[CP].2009-1-20. http://www.oracle.com/technology/software/products/oracle9i/htdocs/winsoft.html.

5 Oracle Corporation. Oracle Warehouse Builder 10g Release 1 (10.1.0.4)[CP].2009-1-20. http://www.oracle.com/technology/software/products/warehouse/index.html.

6 PChome 下载中心. Marc Converter 1.5(试用版)[CP].2009-1-21. http://download.pchome.net/industry/financial/detail-29029.html.

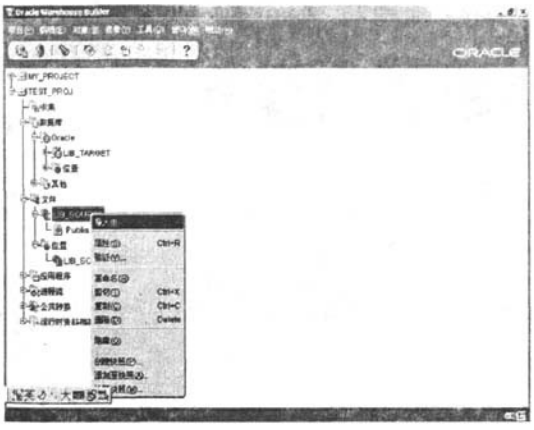

图 2　导入平面文件的元数据

图 3　平面文件元数据导入结果

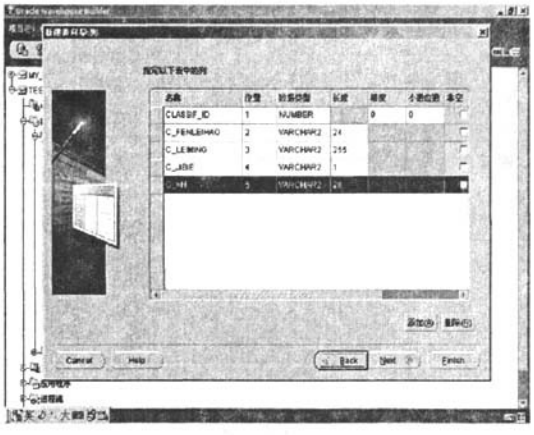

图 4　目标表 TB_CLASSIFICTION

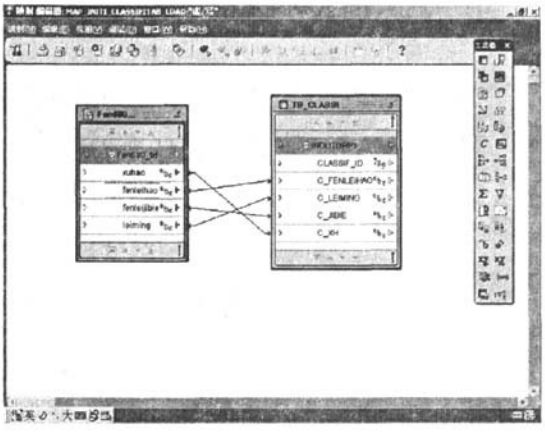

图 5　映射 MAP_INITI_CLASSIFITAB_LOAD

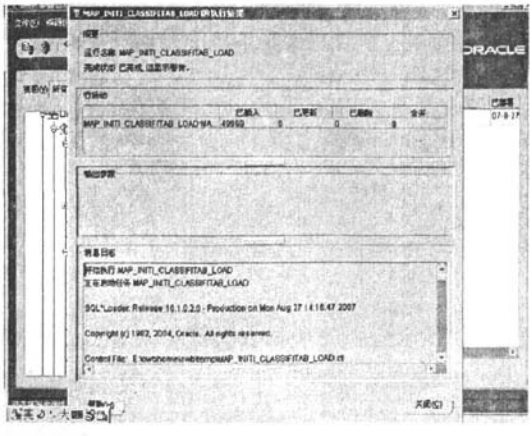

图 6　映射 MAP_INITI_CLASSIFITAB_LOAD 执行结果

图 7　SQL PLUS 查询结果

7　http://www.pupbook.com/forumdisplay.php?fid=74.2009-1-10.

(责编:贺晓利)