

异构数据库数据集成的研究与实现

陈 洋, 罗四维

(北京交通大学 计算机与信息技术学院, 北京 100044)

摘 要: 由于企业间数据的交互和共享越来越广泛, 迫切需要对异构数据库数据进行集成。根据异构数据集成需要, 提出了利用 XML 技术集成异构数据库数据源的一个框架。框架核心部分由 3 层结构组成, 综合管理层负责数据抽取和数据交互, XML 数据库层负责数据预处理和查询, 装配管理层根据需要过滤和组装数据, 最终得到需要的数据。用 Java 编程实现了一个原型系统, 对 Oracle 和 SQL server 实际数据源做了数据集成测试。实验结果证明该架构是一个较优的解决方案。

关键词: 异构数据; 数据集成; XML; Xquery

中图分类号: TP311.138

文献标识码: A

文章编号: 1673-629X(2006)07-0192-03

Research and Implementation on Data Integration of Heterogeneous Database

CHEN Yang, LUO Si-wei

(College of Computer Science, Beijing Jiaotong University, Beijing 100044, China)

Abstract: For the more and more extensive alternation and share among the enterprise data, there is exigent need of integrating data from heterogeneous database. Since the need, in the paper, a frame integrating data from heterogeneous database based on XML is designed. The core part is made up of three layers. The Integration Manager layer is mainly responsible for data extracting, the XML Database layer with preceding disposal and querying, and the Assembly Manager layer mainly filtrates and assembles data with the need, then get the data at last. We implemented a simple system with Java and tested it with Oracle and SQL server data. The experiment result proves that it is a good method.

Key words: heterogeneous data; data integration; XML; Xquery

0 引 言

随着企业信息化的加剧, 企业内部各种应用系统呈阶段性逐渐增加, 而对企业的整体管理需要集成各应用系统下数据库的数据信息, 使得原有的各种异构数据库数据能够整合在一起, 提供给用户透明统一的信息平台接口, 用户就像处理一个数据库一样而不必关心其中各种异构数据集成的细节。此外, 企业间数据的交互和共享越来越广泛, 也迫切需要这种数据的集成。

目前数据集成典型的方法主要有模式集成方法和数据复制方法。其中模式集成是指在构建集成系统时将各数据源的数据视图集成成为全局模式, 使用户能够按照全局模式透明地访问各种数据源的数据; 数据复制是指将各个数据源的数据复制到与其相关的其他数据源上, 并维护数据源整体上的一致性, 提高信息共享利用的效率^[1]。

文中提出了一种基于 XML (Extensible Markup Language) 技术的异构数据库数据集成的构架, 属于模式集成

方法。

1 XML 及其相关技术

异构数据集成必须把各种异构数据最终都转化为一种统一的全局数据模式, 以供用户访问。随着 XML 及其相关技术的不断发展和成熟, XML 成为了应用间数据交换的一种标准。它是 W3G 设计的一种可扩展标记语言, 根据其提供的规则, 程序开发人员可以根据自己的需要定义具体的数据结构。这种数据描述是一种能够用通用编辑器读取的文本, 赋予了 XML 跨平台的能力。这样对于不同数据源人们就可以按照一定规则转换成统一的数据模式, 对其进行统一访问^[2]。

对数据库转和 XML 文档之间的转换需要进行解析, 主要解析器有 DOM, SAX, JDOM, JAXP。

文档对象模型 (通常称为 DOM) 为 XML 文档的已解析版本定义了一组接口。解析器读入整个文档, 然后构建一个驻留内存的树结构, 然后您的代码就可以使用 DOM 接口来操作这个树结构。

SAX 解析器在解析开始的时候就开始发送事件。当解析器发现文档开始、元素开始和文本等时, 代码会收到

收稿日期: 2005-10-14

作者简介: 陈 洋 (1982-), 男, 四川泸州人, 硕士研究生, 研究方向为数据库、网格应用; 罗四维, 教授, 博士生导师, 研究方向为计算机并行处理、人工神经网络。

一个事件。您的应用程序可以立即开始生成结果,不必一直等到整个文档被解析完毕。

用 DOM 和 SAX 模型完成某些任务时有困难,于是创建了 JDOM 包。JDOM 是基于 Java 技术的开放源码项目,它试图遵循 80/20 规则:用 DOM 和 SAX 20% 的功能来满足 80% 的用户需求。JDOM 使用 SAX 和 DOM 解析器,因此它是作为一组相对较小的 Java 类被实现的。

尽管 DOM, SAX 和 JDOM 为大多数常见任务提供了标准接口,但仍有些事情是它们不能解决的,于是 Sun 发布了 JAXP(用于 XML 解析的 Java API, Java API for XML Parsing)。

该 API 为使用 DOM, SAX 和 XSLT 处理 XML 文档提供了公共接口^[3]。

对 XML 进行查询,主要使用了两种查询语言: Xpath 和 Xquery。Xpath 是一种语言,它描述了一种通过使用建立在文档的逻辑结构或者层次结构路径上的寻址语法而定位和处理 XML 文档的办法,定义了 XML 文档路径定位查询的语言规范; Xquery 是一种灵活的可以从 XML 文档中抽取数据的查询语言规范,与 SQL 非常相似,构建在 Xpath 规范之上,并能自由地组织返回结果。

2 主要原理和框架

系统框架结构如图 1 所示,主要由 3 部分组成: 综合管理层 (Integration Manager), XML 数据库层 (XML Database) 和装配管理层 (Assembly Manager)。当然除此之外还有原有的企业各异构应用系统层和在整个异构数据集成系统之上的统一管理应用层,这不是笔者在此研究的重点。

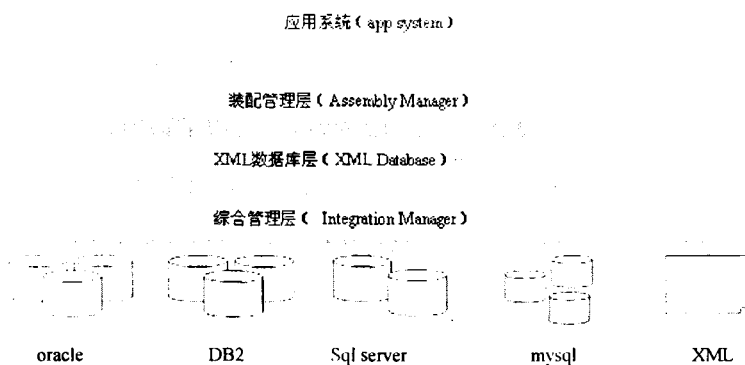


图 1 系统框架

2.1 综合管理层

人们面对的是一系列异构的数据库数据信息,如不同的数据库 Oracle, SQLserver, DB2 等,或者是 XML 文档等,此层主要有两个任务:数据抽取和数据交互。

2.1.1 数据抽取

抽取数据即组成不同的视图(view),视图分为 4 类:

- ①JDBC View: 连接各种支持 JDBC 访问的数据源;
- ②HTTP View: 支持 HTTP 协议方式访问数据;
- ③Web Service View: 支持以 Web Service 协议访问数

据;

④Xquery View: 支持将各种 View 和各种 XML 文档进行进一步加工和拼装。

对数据进行抽取后,就可将各异构数据源看成统一的逻辑数据库,利用 Xquery 查询并集合异构数据源中的数据,逻辑数据库用 XML 格式保存了各数据源位置、数据格式等元数据。只需使用一个查询语句,就可查询所有数据源的相关信息。

2.1.2 数据交互

由于异构数据源随时都有改变的可能,并且不同数据源之间也需要数据的交互,故把数据的交互分为两个步骤: Listener, Converter。其中 Listener 即监听,当外部数据发生改变的时候,Listener 就会做记录并产生相应的动作,Listener 包括 Queue Listener, Topic Listener, HTTP Listener, Directory Listener, Customer Listener。Converter 接受 Listener 的信息并按 XML 格式和要求的步骤改变数据信息,存入目的地。

2.2 XML 数据库层

XML Database 结构如图 2 所示,主要负责数据预处理和查询。

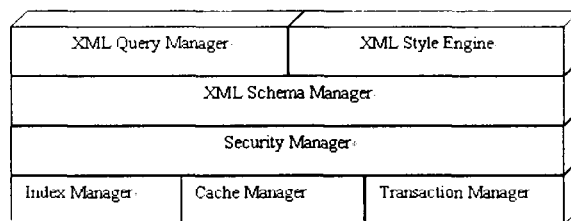


图 2 XML Database 结构

其中预处理包括索引管理(Index manager)、缓存管理

(Cache manager)、事务管理(Transaction manager)、安全管理(Security manager)、格式处理(XML Schema or DTD manager)。索引管理是为了加快查询的速度和效率,对建立的索引的管理。缓存管理是利用缓存机制把需要经常查询或近期查询的异构数据缓存起来,提高下次访问的效率。所谓事务是指一组逻辑操作单元,使数据从一种状态变换到另一种状态。为确保数据库中数据的一致性,数据的操纵应当是离散的成组的逻辑单元:当它全部完成时,数据的一致性可以保持,若这个单元中的一部分

操作失败,所有从起始点以后的操作应全部回退到开始状态,事务管理保证了数据的一致性。安全性管理包括机密性、完整性和可用性,如全局范围的身份验证、访问控制、完整性控制等^[4]。格式处理包括 XML Schema 和 DTD, XML Schema 是用一套预先规定的 XML 元素和属性创建的。这些元素和属性定义了文档的结构和内容模式。相应的一套精巧的规则指定了每个 Schema 元素或者属性的合法用途。DTD 定义了可以在 XML 文档中出现的元素、这些元素出现的次序、它们可以如何相互嵌套以及 XML

文档结构的其它详细信息^[3]。

数据查询主要用 Xpath 和 Xquery 技术,使用查询引擎(Xquery engine)进行查询,返回用户期望的查询结果。

2.3 装配管理层

装配管理层主要使用管道线(pipeline)原理,根据需要过滤和组装数据,最终得到需要的数据,其过程如图 3 所示。

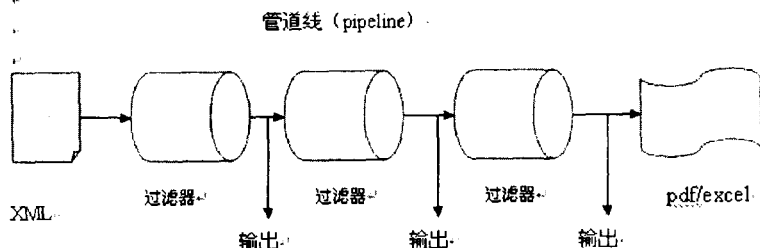


图 3 装配管理过程

输入端为 XML 数据源(从 XML Database 层得到的 XML 文档或 Xquery 结果集),中间通过一系列过滤器(Filter),便可得到相应期望的输出,向上层传递需要的数据。

3 系统分析

整个框架核心部分采用了 3 层结构,使得各层之间分工相对独立,当其中一层负载过重或出现问题的时候,可以针对这一层问题具体解决,减少对其他部分的影响,提高系统效率^[5]。此外,由于 XML 是能被计算机理解的通用语言,不受具体平台的影响,这增强了系统的可移植性。

由于整个框架的设计牵涉到的细节比较多,只实现了

一个原型系统,对一些细节的地方没有面面俱到,如安全管理和事务管理。通过对 Oracle 和 SQL server 实际数据源集成实验,能比较满意地得到所需要提取的数据,由于采用 Java 编程实现,所以效率上受到一定的影响,如果提高硬件配置能弥补这一缺陷。

4 结束语

异构数据库数据集成是企业发展过程中迫切需要解决的问题。文中通过把异构数据库数据统一转化为 XML 全局数据模式,在此基础上提供用户统一透明的数据查询,给出了异构数据源集成的一个统一框架,对框架的几个层次和设计给了详细分析说明,是一种可行的较优的解决方案。

参考文献:

- [1] 陈跃国,王京春. 数据集成综述[J]. 计算机科学,2004,31(5):48-50.
- [2] 陶以政,唐定勇. 基于 java 和 XML 技术的异构信息系统数据集成框架应用研究[J]. 计算机应用研究,2004(5):38-40.
- [3] Jasnowski M. Java, XML 和 Web 服务宝典[M]. 盖江南等译. 北京:电子工业出版社,2002.
- [4] Date C J. 数据库系统导论[M]. 孟小峰,王 珊,等译. 北京:机械工业出版社,2000.
- [5] 杨晓强,陈 冰. 用基于 XML 的中间件访问异构数据库[J]. 计算机应用研究,2004(6):205-206.

(上接第 155 页)

的数据库系统中。

(2) 导入工具。导入工具将文本文件中的数据导入到容灾中心的备份数据库中。同样,导入工具也可以根据配置文件中的配置选项或者用户界面选项将用户数据正确地导入到容灾中心的数据库中。导入工具还可以支持数据割接中的导入和增量导入的功能。

(3) 容灾中心的解决方案。在本地数据库中用导出工具将用户数据按照容灾中心的数据库配置导出用户数据,每个文本文件存放一个号段的用户数据,然后通过高速以太网传输到容灾中心,按号段配置用导入工具导入到数据库中。再通过在业务比较闲的时刻,通过导出工具的自动导出功能和传输过来的文本文件进行比较,记录比较日志,如果有不一致的用户由操作管理员手动更新,以确保容灾中心数据和本地数据的数据的一致性。

4 结束语

随着移动通信业务的发展和用户的不断增加,新业务的不断扩展,大容量、稳定、实时和容灾能力强的 HLR 数

据库系统已经变得越来越重要,以上所讨论的各种容灾方案都各有其优缺点,综合以上各种因素,在本地采用双 PC 服务器外挂磁盘阵列的方式,容灾系统采用 N+M 的容灾方式,本地和容灾中心通过高速的以太网以异步方式同步数据。辅助于导入导出工具的文本格式的数据一致性处理的方案具有较高的使用价值。

参考文献:

- [1] Digital Cellular Telecommunication System(phase2): Operations and Perfomance Management [S]. GSM12.06 (ETS 300612-3),1996.
- [2] 杨留清. 数字移动通信系统[M]. 北京:人民邮电出版社,1995.
- [3] 卢炎生,潘 怡,赵 栋,等. 一个内存库管理系统的数据库组织[J]. 华中理工大学学报,1999,27(10):64-66.
- [4] 昂卫武,杨有庆. 移动通信系统 HLR 容灾技术[J]. 中兴通讯技术,2004(10):51-54.
- [5] 黄东军,陈 刚. WCDMA 归属位置寄存器大容量实时数据容灾技术[J]. 微机发展,2005,15(8):10-12.