基于SDO的异构数据集成研究与应用

郑 垒，曹宝香
（曲阜师范大学 计算机科学学院，山东 日照 276826）

摘 要：为解决现存的异构数据集成解决方案中对异构数据处理没有统一的标准、硬编码多、集成系统扩展性差等问题，提出了基于SDO规范的异构数据集成方案。设计实现了一个统一的数据访问界面；利用数据访问服务封装了各种异构数据源，并将结果以同一种格式暴露给集成系统，实现了对数据源中数据访问、操作方式的统一；在查询分解方面，给出了基于数据源配置文件的查询分解方法，集成系统根据配置文件就可以与相应的数据源取得连接，而且只要修改相应的配置文件，就可以实现数据源的灵活修改；最后将该方案在基于WEB的PLM系统中进行了应用，验证了方案的可行性。实现结果表明该方案开发量小、扩展性好、效率高，能够很好地满足企业异构数据集成的需要。

关键词：服务数据对象；数据集成；异构数据源
中图分类号：TP311 **文献标识码**：A **文章编号**：1673-629X(2009)11-0163-04

Research and Application of Heterogeneous Data Integration Based on SDO

ZHENG Lei, CAO Bao-xiang
(Computer Science College of Qufu Normal University, Rizhao 276826, China)

Abstract: In order to resolve the problems of the existing solutions of heterogeneous data integration, including no uniform standard for the treatment of heterogeneous data, too much hard-coded and poor extendibility, propose a solution based on SDO for heterogeneous data integration. It designs and achieves a unified data access interface; Data access service encapsulates multiple data resources and returns results using the same format, which unifies the methods of data access and data manipulation; A query decomposition method based on configuration files of data sources is given in query decomposition respect. By the configuration files, the integration system can obtain connections with the corresponding data sources. And as long as the configuration file is modified, the system can implement the flexible modification of data sources. Finally, the integrated solution is applied to the PLM system based on WEB project, and its feasibility is verified. The experimental result illustrates that this solution can reduce the ease of development, and has good extendibility and high efficiency. It can very well meet the needs of heterogeneous data integration.

Key words: service data objects; data integration; heterogeneous data sources

0 引 言

随着网络的发展和信息化的逐步深入，大多数企业都实现了信息的计算机化管理，数据是企业最重要的信息资产，在大多数企业中，数据大都以不同的格式分布在不同的系统中，这些数据受数据模型及存储方式的差异，具有明显的异构性、分布性和自治性。如何将这些不同来源、格式和质量的数据进行有效的集成，消除“信息孤岛”[1]，实现企业级数据的全面共享，就是异构数据集成所要解决的问题。

目前，大多数企业普遍采用联邦数据库[2]、数据仓库[3]或基于中间层[4]的方法来集成异构数据源，这些方法虽然都能在一定程度上满足数据集成的需要，但它们都存在一定的不足，如在对数据的处理上没有统一的标准，开发人员需要熟悉各种技术的API(如JDBC, JCA等)才能对各种异构数据源中的数据进行访问，而且只能访问异构数据源中的数据而不能对其进行操作，并且构建的系统具有重用性不高、数据集成效率低、高耦合性等缺点。

文中提出将SOA(Service Oriented Architecture，面向服务的架构)的新型数据编程规范SDO(Service Data Objects，服务数据对象)用于数据集成，就能很好地解决上述问题，而且基于SDO的异构数据集成方案中的

收稿日期：2009-03-09；修回日期：2009-06-30
基金项目：山东省教育科研计划项目(J05G03)
作者简介：郑 垒(1984-)，女，山东泰安人，硕士研究生，研究方向为数据库技术与系统集成；曹宝香，教授，硕士研究生导师，研究方向为数据库技术与系统集成。

数据源可以实现灵活的添加、修改和删除,具有良好的扩展性和实用性,SDO 还有丰富的开源[5]可以使用(如 Eclipse 项目下的 EMF,Apache 的 Tuscany 开源项目等),极大地减少了系统开发的工作量,提高了数据集成的效率。

1 服务数据架构

1.1 SDO 简介

2005 年,IBM 联合 BEA、甲骨文、SAP 等公司共同发布了针对 SOA 的重要编程规范——SCA(Service Component Architecture,服务构件架构)和 SDO。其中 SDO[6]是一种统一各种数据源类型中的数据编程的编程模型规范,为常见的应用模式提供全方位的支持,允许应用程序、工具和框架更加容易地查询、绑定、更新和内省数据。SDO 提供了一种独特的模型来存放结构化的和相互关联的复合对象,使应用程序可以使用这些对象来保存信息,而且对种类繁多的数据源和业务提供了一个统一的数据访问,还可以在业务处理和数据源之间实现解耦合。从某种意义上讲,SDO 框架可以简化和统一 SOA 中的数据应用程序开发,使开发者更专注于业务逻辑而不必关心底层技术,可以很好地满足企业异构数据集成的需要。

目前,SDO 规范已经到了 2.1 版本,其中包括了对各种常用语言的支持:SDO for Java[7] and C++,SDO for PHP,SDO for C,SDO for COBOL 等,有越来越多的公司加入到了 SDO 规范的制定中,并在他们的产品中加入了对 SDO 的支持,如 IBM 系列产品中的 WebSphere Application Server 和 Rational Studio 工具。

1.2 SDO 架构及关键组件

SDO 架构基于断开数据图的概念,如图 1 所示。其关键组件包括:

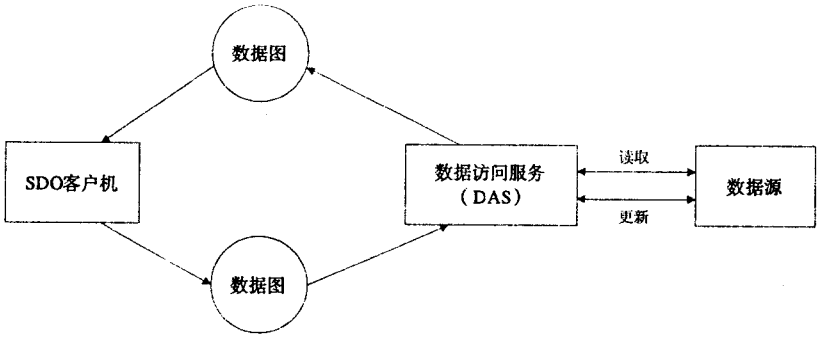

图 1 断开的数据图架构

(1)数据访问服务(Data Access Service,DAS)。

在 SDO 架构中只能由 DAS(Data Access Service,数据访问服务)与数据源进行交互。DAS 是为特定数据源处理技术细节的服务,对各种数据源的操作由 DAS 组件完成,DAS 负责提供某些方法创建包含数据对象的数据图,也负责将对数据的修改应用到数据源。通常在运行时提供 DAS 的实现,而应用程序开发工具提供对数据图的支持,DAS 总是以同一种格式(数据图)的形式返回信息,从而隐藏了实际的数据存储信息,屏蔽了数据源的异构性。通常,不同的数据源对应不同的 DAS,如对关系数据库进行读写的关系 DAS,与 EJB 进行交互的 EJB DAS,对 XML 数据源进行读写的 XML DAS 等等。

(2)数据对象(Data Object)。数据对象用于描述业务数据,由属性的键/值对组成,每个值可以是原始的数据类型,也可以是另一个数据对象。数据对象提供了易于使用的创建和删除方法(createDataObject 和 delete 方法)和获得自身类型(实例类、名称、属性等)的反射方法。数据对象包含在数据图中,是可序列化的。

(3)变更摘要(Change Summery)。

变更摘要表示对 DAS 返回的数据图的修改,变更摘要提供了数据图中被修改的属性(包括原来的值)、新增和删除的数据对象的列表。变更摘要最初是空的,只有当变更摘要日志功能被激活时,才会将数据图中所有数据对象的历史更改信息添加到数据图的变更摘要中。在后台更新时,DAS 使用变更摘要将修改应用于数据源。变更摘要还提供了让 DAS 打开和关闭日志功能的方法。

(4)数据图(Data Graph)。

数据图是一个描述数据的分层结构,包括一个数据对象树和一个变更摘要。数据图通常是系统中组件之间的传输单元,由 DAS 生成,供客户端使用,数据图修改后会被回传给 DAS,然后由 DAS 完成对数据源的更新。一般来说,一个数据图对应一个 XML 文档,当进行信息传输时,数据图被序列化为 XML,SDO 规范提供了序列化的 XML Schema。

2 基于 SDO 的数据集成方案

2.1 体系结构

基于 SDO 的异构数据集成系统体系结构如图 2 所示,异构数据集成平台是数据集成系统的关键部分,它除为用户提供一个数据访问的统一接口外,还由数据对象管理模块、控制模块和 DAS 三个功能模块组成。

数据访问统一接口 用户通过异构数据集成平台提供的统一的数据访问接口提交自己的访问请求,数据集成平台接收到用户的请求后,先对用户进行身份认证,根据用户的权限对用户的行为进行约束,从而保

证了平台的安全性,然后将访问信息提交给控制模块,该模块还将访问结果以统一的形式(表格或者 XML 网页)呈现给用户。

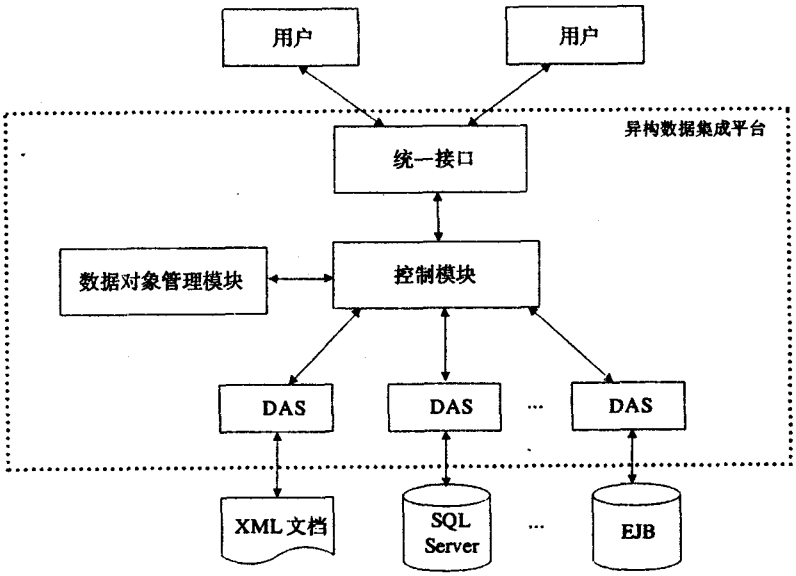

图 2　基于 SDO 的异构数据集成系统体系结构

数据对象管理模块 数据对象管理模块提供了数据源的配置信息(以 XML 格式存储),主要包括数据源的种类和位置、数据访问信息及数据对象映射文件(规定数据对象属性与数据库表中字段对应关系的 XML 格式的文件)的保存路径;通过该模块还可以自动生成/更新数据对象,更新数据对象映射文件,数据对象生成后自动向使用它的业务系统发布更新信息;该模块还向控制模块提供了验证服务。

控制模块 控制模块接收到的数据访问请求信息后,向数据对象管理模块验证操作的合法性,如果验证通过,控制模块将根据请求信息将用户的全局访问请求转换为对相应数据源的本地访问,分配给相应的 DAS 进行数据源的操作。当本地访问结束后,返回一个或者几个 XML 文档(从各数据源抽取的数据),控制模块接收 DAS 返回的结果,并向数据对象管理模块获取相应的数据对象,将结果进行合并,然后将 SDO 标准格式(数据图)的信息进行返回。

DAS DAS 根据数据源的格式、操作类型和数据访问内容,与相应的数据源进行连接,然后对数据源中的数据进行操作,并将结果返回给控制模块。

2.2　系统实现

文中将提出的异构数据集成方案,应用于山东省教育厅的"基于 WEB 的 PLM 系统的研究与实现"项目,如图 3 所示,验证了方案的可行性和正确性。在系统开发中开发工具使用 Eclipse,开源使用 Apache Tuscany 的 SDO 和 DAS 软件包,应用服务器使用 Tomcat,主要使用的技术有 JSP,Java,XML[8],XSD 等。利用该平台实现了对 MySQL 类型的 USER 数据库中 USER 表的操作。下面介绍系统关键部分的实现过程:

(1) 统一访问界面设计。

集成系统提供了一个对异构数据源进行统一访问的界面,在用户看来,各个分布异构的数据源在逻辑上是透明的,好像是在对一个单一的数据源进行操作。该系统支持多条件检索(并含/或含),如图 4 所示。

(2)数据源的配置信息。

异构数据集成平台接收到来自业务系统的请求信息后,需要根据数据源的类型、路径等将请求信息转换成对相应数据源的操作,因此异构数据集成平台需要保存联合的各业务系统的数据源配置信息,包括数据源的类型、全局名称、存储地址、用户名、密码等信息,是一个由数据对象管理模块配置的 XML 文档(userConfig. xml),当进行数据源的添加、删除、修改时,只需更改配置文件中的信息即可。本例 USER 数据库的配置信息如下:

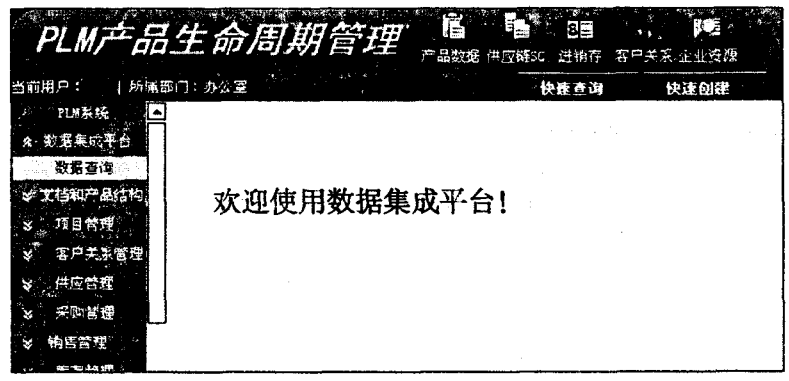

图 3　PLM 系统界面

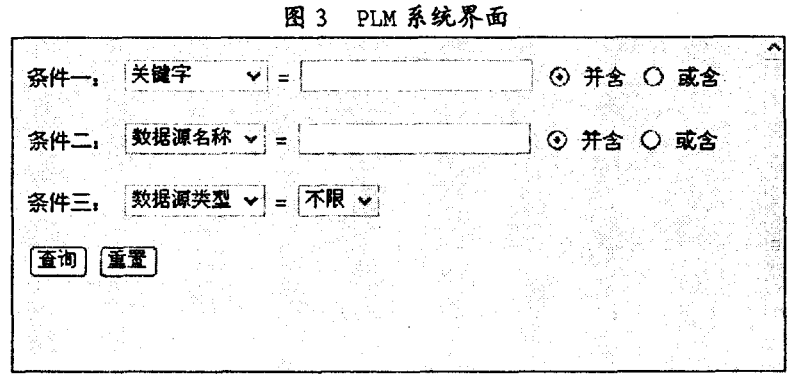

图 4　查询界面

```
<? xml version="1.0" encoding="UTF-8"? >
<Config xmlns="http:///org. apache. tuscany. das. rdb/config. xsd">
<! - ConnectionInfo 描述数据源的连接信息,包括数据源类型、全局名、用户名、密码等信息 - >
<ConnectionInfo>
<ConnectionProperties
driverClass="com. mysql. jdbc. Driver"
databaseURL="jdbc:mysql://localhost/user"
userName="root"
password="root"
loginTimeout="600000"/>
</ConnectionInfo>
<! - Command 描述 SQL 操作的名称、内容、类型等信息 - >
<Command name="AllUsers" SQL="select * from user" kind="Select"/>
<! - Table 描述了表信息,包括表名、主键等 - >
<Table tableName="USER">
<Column columnName="ID" primaryKey="true"/>
</Table>
</Config>
```

(3)根据配置信息创建 DAS。

由 DAS 的工厂类中的 DAS. FACTORY. createDAS()方法创建。

(4)DAS 与数据库的连接。

DAS 根据配置信息获取数据源的连接信息,然后与相应的数据源进行连接,由 UserDatabaseInitializer(String configFile)方法和 MySQLSetup(String databaseInfo)实现。本例中数据库连接的代码:

```
if(this. platformName. equals(MYSQL)){
databaseURL = databaseURL + "? user = " + userName + "&password = " + password + "&createDatabaseIfNotExist = " + "true";//databaseURL, user, password 三个变量由 java. util. StringTokenizer 中的 nextToken()方法获得
connection = DriverManager. getConnection(databaseURL);
}
```

(5)对数据库的操作。

DAS 与数据库连接后就可以对数据源中的数据进行增加(set 方法),删除(delete 方法),修改(set 方法)和查询(getUser 方法)等操作了。

3 结束语

文中将 SDO 规范用于数据集成,提出了新的异构数据集成解决方案,并将其应用到实际项目的开发中。SDO 提供了统一的数据应用开发框架,统一了对多种企业信息系统(EIS)的数据访问,通过使用 SDO 这种独特而简单的模型,应用程序能够摆脱使用多种 API 和框架进行数据访问的复杂工作,在开发过程中只需使用一种 API(SDO API)便可操作各种异构数据源。该文提出的基于 SDO 的异构数据集成系统,还可以方便地进行数据源的添加、修改和删除,具有很好的扩展性,因此将 SDO 用于异构数据集成具有一定的理论和实践意义。

参考文献:

[1] 周 运,牟占生,徐久成. 基于 XML 虚拟数据库的异构数据源集成模型研究[J]. 计算机技术与发展,2008,18(4):84-87.

[2] Fong J, Wong H K, Cheng Z. Converting relational database into XML documents with DOM[J]. Information and Software Technology, 2003,45(4):335-355.

[3] 钟 华,冯文澜,谭红星,等. 面向数据集成的 ETL 系统设计与实现[J]. 计算机科学,2004,31(9):87-89.

[4] 袁晓洁,于士涛,李志梁. 基于 Mediation 的异构数据集成系统 HDIS 设计与实现[J]. 计算机工程与应用,2006,42(1):162-165.

[5] 倪志刚,洪 玫,刘 佳. 基于服务数据对象的异构系统数据集成方案研究[J]. 计算机应用,2007,27(6):21-23.

[6] Service Data Objects, WorkManager, and Timers[EB/OL]. 2003-11. http://www.ibm.com/developerworks/library/specification/j-commonj-sdowmt/

[7] Service Data Objects For Java Specification, Version2.1.0 (pdf)[EB/OL]. 2006.11. http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-sdo/SDO_Specification_Java_V2.01.pdf.

[8] 陈 洋,罗四维. 异构数据库数据集成的研究与实现[J]. 计算机技术与发展,2006,16(7):192-194.

(上接第 114 页)

2006(2):169-170.

[5] 何 雄,方金云,唐志敏. 织女星地理信息系统 VegaGIS 中的空间数据引擎 CoSDE[J]. 计算机应用,2005,25(7):1587-1589.

[6] 杨超伟,李 琦. Web 空间信息发布研究[J]. 北京大学学报:自然科学版,2001,37:413-419.

[7] 汉语分词系统 ICTCLAS[EB/OL]. 版本 3.0.2006-04. http://ictclas.org/index.html.

[8] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. [s.l.]:Addison-Wesley,1999.

[9] 潘明远,董刊生. 织女星地理信息系统 VegaGIS POI 搜索引擎的设计与实现[R]. 北京:中科院计算所技术报告,2007:4-8.