

# 基于 Web Service 的异构集成查询方案

张 峰, 邵秀丽, 李耀芳

(南开大学信息技术科学学院, 天津 300071)

**摘 要:** 提出一种基于 Web Service 的异构集成综合查询系统, 采用 B/S 和 C/S 相结合的多层分布式体系结构, 能够有效提高系统的实用性, 并屏蔽数据的异构性和各种系统模式的差别, 仿真实验结果表明, 该设计方案提升了异构系统的可集成性, 且能实现用户透明查询的功能, 保证数据源的安全。

**关键词:** 异构; 集成; 查询

## Heterogeneous Integrated Scheme Based on Web Service

ZHANG Feng, SHAO Xiu-li, LI Yao-fang

(College of Information Technical Science, Nankai University, Tianjin 300071)

**【Abstract】** An heterogeneous data integration enquiry system based on the Web Service is proposed, which uses B/S and C/S combination of multi-storey distributed architecture to improve the system's practicality. The heterogeneous data and differences of system patterns are shielded. Simulation experimental results show this design scheme can achieve the fact that much more heterogeneous systems can be integrated, the functions of transparent user query are implemented, and the safety of the data source is ensured.

**【Key words】** heterogeneous; integration; query

本文研究的异构数据集成查询问题源于天津港信息化建设面临的实际问题, 也是国家“十一五”项目“数字化口岸关键技术研发及应用”的主要内容。天津港综合物流信息服务平台涉及到跨部门使用跨系统的异构资源, 系统之间不能进行信息交互, 造成许多的信息孤岛, 且数据不能共享, 造成重复性的数据录入、数据冗余、工作效率不高。因此, 如何有效解决异构数据的整合, 建立统一的数据集成查询平台, 是亟待解决的问题。

### 1 异构集成综合查询系统的总体解决方案

#### 1.1 异构数据集成方法

在异构数据集成技术方法中, 常见的 2 种方法是数据仓库法和虚拟法。天津港综合物流信息服务平台需整合的异构数据源具有跨地域、跨系统、跨数据库等特点, 异构集成综合查询系统设计的出发点就是要能够屏蔽这些异构模式, 满足全局用户透明查询所需数据的需求。单独采用数据仓库法或虚拟视图法都不能满足上述平台建设的要求。因此, 本文吸取数据仓库法建立一个全局性的数据仓库, 将参加集成的各数据源的数据副本, 按照一个集成、统一的视图要求, 转换成符合数据仓库的模式, 并存入数据仓库的优点, 另外, 借鉴虚拟视图法仅提供一个虚拟的集成视图以及对这个集成视图的查询处理机制, 用户针对集成视图提出查询的优点, 结合 2 种方案, 提出一个异构数据集成综合查询的解决方案, 采用推数据的方式, 从局部数据源采集需集成的数据, 并将数据在中央服务器进行集成并接受查询。这种方法既避免了传统多数据库获取信息相对困难的缺点, 也提高了系统的可用性和查询工作效率<sup>[1]</sup>。

#### 1.2 系统的解决方案

本文提出的异构集成综合查询系统, 从业务系统的实际使用情况出发, 采用 C/S、B/S 相结合的架构方法, 采取数据

源向集成查询平台推送数据的方式构建而成, 有效解决了各个单位间数据的集成查询, 具有较强的针对性。

异构集成综合查询系统主要分为 2 个部分: 集成查询子系统和通用的数据采集子系统。数据采集子系统利用虚拟视图定时从关系数据源采集数据提交给集成查询子系统, 该方法不需要改变原始数据的存储和管理方式, 只是将数据源的数据封装为 XML 格式传送到服务器端存储, 为下一步的集成做好数据的准备。集成查询子系统将异构的数据进行集成, 处理为统一数据集, 并提供给用户一个全局模式的查询接口。

数据的采集和集成采用 C/S 架构, 集成查询子系统负责数据的集成, 消除多数据源提交的数据间的异构, 以 B/S 架构对外提供的查询服务及数据分析。在异构数据源端, 数据采集子系统定时定期采集数据源更新的数据, 并将数据统一转换为 XML 格式, 提交给集成查询子系统。数据采集子系统与集成查询子系统之间通过 Web Service 进行命令通信和数据文件的传输, 利用 Web Service 技术可以克服平台、系统间的异构。

#### 1.3 系统的逻辑层次

在数据集成应用中, 一个设计良好的平台应该把与具体应用无关的公共部分同与业务相关的功能分离开来, 使其具有可重用性<sup>[2]</sup>。因此, 本文采用面向对象的分析设计方法结合分层思想、面向服务的架构和构件化技术来设计和实现该

**基金项目:** “十一五”国家科技支撑计划基金资助重大项目(2007BAH10B01); 天津市软件专项基金资助项目(07FZRJGX03200, 06FZRJGX01600, 07FZRJGX04600)

**作者简介:** 张 峰(1983—), 男, 硕士研究生, 主研方向: 数据集成, 网络技术; 邵秀丽, 教授、博士; 李耀芳, 硕士

**收稿日期:** 2008-10-30 **E-mail:** superzf@126.com

信息集成平台。从逻辑功能上看,异构集成查询系统可以抽象为 5 个层次,各层及具体功能如下:

- (1)表示层: 为用户提供统一的查询接口;
- (2)全局数据集成层: 它是系统的核心,系统通过此层集成各异构数据源提交的数据,消除各异构数据源提交数据间的异构;
- (3)通信层: 负责数据中央服务器同各个数据采集子系统之间命令消息和数据的传输;
- (4)数据采集层: 主要功能是将数据统一转换为 XML 格式的数据文件,消除数据间的结构异构;
- (5)数据源层: 它属于参与集成查询的自治异构数据源的集合。

## 2 异构集成查询系统的设计

### 2.1 设计思想

利用 XMLSchema 在数据的结构、类型和约束的强大表达能力中提取数据的关系模式和数据,以形成各自的 XML 模式,定时定期采集数据源更新的关系数据,并将关系数据转换为符合源 XML 模式的数据,然后推送至数据集成端。在综合查询端,即数据集成端根据不同的数据集成查询主题构建相应全局数据模式,然后配置全局数据模式和源 XML 模式间的映射规则<sup>[3]</sup>。根据全局模式和局部模式间的映射规则,集成异构数据源提交的局部数据为符合全局数据模式的数据,消除数据间的异构,接受用户的查询。

### 2.2 系统总体设计

异构集成综合查询系统包括集成查询子系统和数据采集子系统 2 个部分。

为便于适应各异构数据源,分别为集成查询子系统和数据采集子系统构建 Web Service 接口,封装集成查询子系统和数据采集子系统具体的业务功能。利用 Web Service 实现集成查询子系统和数据采集子系统间命令消息和数据文件的传输。异构集成查询系统总体结构如图 1 所示。

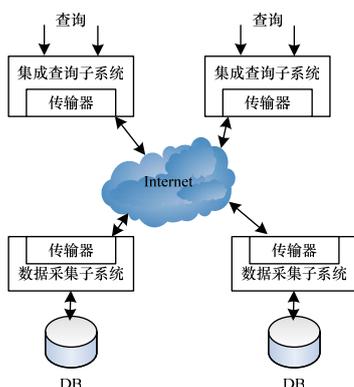


图 1 系统总体结构

每个数据采集子系统可向多个集成查询子系统提供数据,每个集成查询子系统可获取多个数据采集子系统提交的数据。集成查询子系统可根据需要部署在不同的地方,便于快速构建不同目的的异构数据集成查询系统。

数据采集子系统负责连接各异构的数据源,将源关系模式转换为 XML 局部模式,并定期采集数据源更新的数据,将其封装为源 XML 数据,提交给集成查询子系统。集成查询子系统是异构集成查询系统的核心,它负责接受数据采集子系统的注册,数据采集子系统注册完毕后,为每个数据采集子系统建立一个模式库,接收并存储数据采集子系统提交

的模式信息,然后为每个数据采集子系统创建一个数据存储库,合并、存储数据采集子系统定时提交的更新数据。

数据采集子系统和集成查询子系统间通过 Internet 连接,为了便于 2 个子系统间命令消息和数据文件的安全、正确传输,每个数据采集子系统和集成查询子系统均包含一个传输器,用于命令消息和数据文件的传输。

### 2.3 数据采集子系统的设计

数据采集子系统主要实现采集各个异构数据源相关的数据,将数据统一转换为 XML 格式,通过 Internet 将转换后得到的 XML 数据传输给集成查询子系统。

#### 2.3.1 数据采集子系统的结构设计

通过对数据采集子系统所需实现的功能的研究和分析,并结合天津港综合物流信息服务平台的实际要求,设计数据采集子系统的结构如图 2 所示。

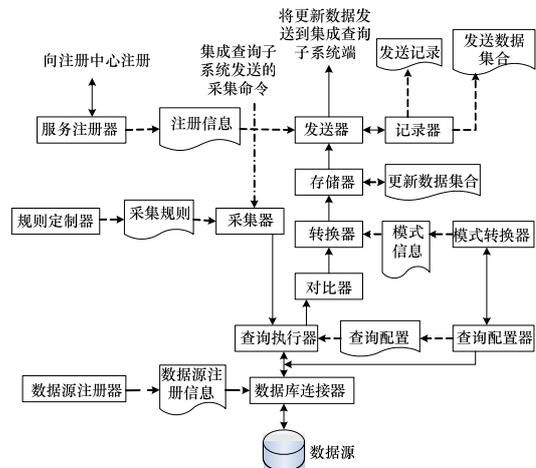


图 2 数据采集子系统结构

对图 2 的说明如下:

#### (1)服务注册器

其功能是向集成查询子系统的服务注册中心进行注册,提供本数据采集子系统的访问方法、数据模式信息、数据采集方式、数据提交时间间隔等信息,同时获取集成查询子系统 Web Service 接口的访问方法。

#### (2)规则定制器

其功能是定义数据采集方式和数据发送时间间隔,采集方式包括定时采集、定期采集、根据集成查询系统采集命令采集 3 种方式。

#### (3)数据源注册器

其功能是注册局部数据源,选择数据源的类型,配置数据源访问方法,包括地址、用户名、密码等信息。

#### (4)采集器

数据采集器主要包括一个定时器,通过规则定制器定义的数据采集方式,当到达数据采集时间或接收到集成查询系统的数据采集命令时,发采集命令给查询执行器。

#### (5)查询执行器

其功能是接收数据采集器的采集命令,根据查询配置器定义的数据库查询语句和数据源注册器中的数据源信息,通过数据库连接器连接数据源,查询并从数据源获取返回的查询结果,将获取的查询结果返回给数据集对比器。

#### (6)数据源连接器

其功能是连接异构的数据源,为查询执行器提供统一的查询接口,便于查询执行器连接异构的数据源,执行数据查

询操作。

(7) 对比器

其功能是记录上次采集数据的结果，通过比较本次查询结果和上次查询结果，获取数据库在采集时间间隔内的数据更新信息，并将数据更新信息提交给数据转换器。

(8) 转换器

其功能是根据模式转换器定义的模式信息，将从数据集对比器获得的数据更新信息转换为 XML 格式的数据，这样做可以将异构数据源中的数据转换为统一的格式，便于集成查询端对数据的处理。

(9) 存储器

其功能是获取数据转换器转换的数据更新信息，并将本次获取的数据更新信息和已获取未提交的更新数据信息合并，合并后的数据存放更新文件中。

(10) 发送器

其功能是根据数据发送方式，到达数据发送时间或数据采集器接收到集成查询端的采集命令获取更新数据完毕后，将所有未提交的更新信息提交给传输层的传输器，发送给集成查询子系统。更新文件发送成功后，则清空原更新文件。

(11) 记录器

其功能是记录每次发送数据的时间、发送的数据记录，若传输器发送数据文件失败，可以从日志记录器获取文件重新发送。

(12) 查询配置器

其功能是通过数据库连接器获取数据库的库表结构，由管理员根据集成主题，确定所需的数据表及字段，设定表间的联接关系及生成结果记录中的主键列，生成数据库查询语句，并将查询语句提交给模式转换器。

(13) 模式转换器

其功能是获取查询配置器生成的查询语句，通过分析查询语句中的信息，将关系数据模式转换为 XML 数据模式，该模式信息将用于关系数据向 XML 数据的转换。

2.3.2 数据采集子系统的工作流程

数据采集系统的工作流程如图 3 所示。

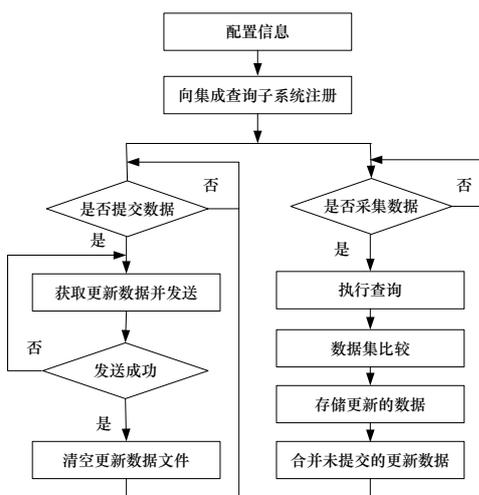


图 3 数据采集子系统工作流程

对图 3 的说明如下：

(1) 配置数据采集信息

1) 配置数据库连接信息，包括数据库类型、数据库地址、用户名、密码等；

2) 配置数据库查询语句，指定能够唯一标识查询结果记录的主键信息；

3) 根据查询语句，自动获取所包含的关系模式，并将其转换为对应的 XML 模式；

4) 定义数据采集方式和数据提交时间间隔。

(2) 向集成查询子系统注册

该模块通过向集成查询子系统的注册中心提交注册信息，包括提供数据采集子系统的访问方法、数据采集方法、XML 模式信息，并在注册后生成数据采集子系统和集成查询子系统之间通信的密钥。

(3) 执行查询

按照数据采集方式，到时、到期或接收到集成查询子系统数据采集命令，根据查询语句执行查询，获取查询结果数据集。

(4) 数据集比较

比较本次采集的数据集与上次采集的数据集，提取更新的数据集。

(5) 存储更新的数据

将更新的数据转为 XML 数据文件，存储在数据存储器。

(6) 合并更新未提交的数据

由于每次获取采集的时间与数据发送时间不一致，因此每获得一次更新数据，需要与以前获取的更新的数据进行合并操作。

(7) 提交数据

向集成查询子系统提交更新的数据，判断是否提交成功，提交成功后，清空更新数据文件，不成功则重新获取更新数据文件并进行提交操作。

2.4 集成查询子系统的设计

集成查询子系统主要功能是对各异构数据源传输来的数据按照用户的查询期望进行集成，形成了统一的数据访问接入点<sup>[4]</sup>。

2.4.1 集成查询子系统的结构设计

集成查询子系统是集成查询的服务器端应用程序，主要的功能是接收各个数据采集子系统传来的数据文件，集成数据采集子系统提交的数据，制定全局查询接口，接受用户的查询，具体结构如图 4 所示。

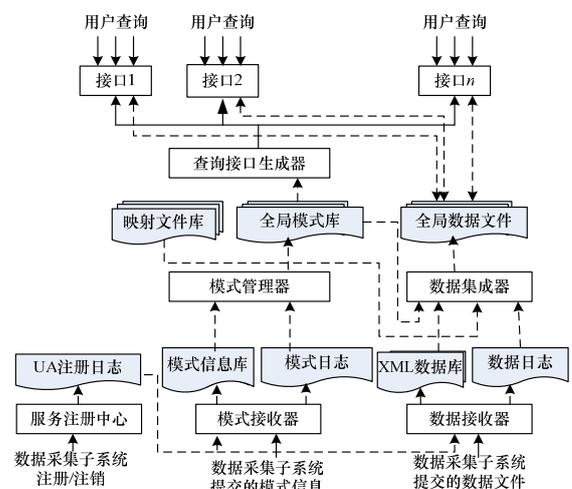


图 4 集成查询子系统结构

集成查询子系统根据用户的查询期望定义 XML 格式的全局模式，然后配置全局模式与局部模式之间的映射关系，

整合局部数据为一个整体的数据文件，等待用户查询。该部分主要包括：服务注册中心，模式接收器，数据接收器，模式管理器，数据集成器，查询接口生成器 6 个功能模块：

(1)服务注册中心：其功能是接受数据采集子系统的注册和注销，记录数据采集子系统的配置信息，包括数据采集子系统的访问方法、数据采集方式、数据采集时间间隔。只有在服务注册中心合法注册的数据采集子系统，模式接收器和数据接收器才接收其提交的模式信息和数据文件。

(2)模式接收器：其功能是为每一个注册的数据采集子系统建立一个模式库，接收数据采集子系统提交的模式信息，并存储到该数据采集子系统对应的模式库中，并标记在模式日志文件中。

(3)数据接收器：其功能是为每一个在服务中心注册的数据采集子系统建立一个数据存储库，接收并存储数据采集子系统提交的数据文件。

(4)模式管理器：其功能是由管理员根据查询主题、数据采集子系统提交的模式信息，定义全局模式，配置全局模式和局部模式间元素的映射关系、转换规则，为集成数据采集子系统提交的数据，消除数据间的异构铺平道路。

(5)查询接口生成器：根据管理员定义的全局模式信息，设定全局模式中哪几项元素作为查询关键列，由查询接口生成器根据全局模式和关键列自动生成相关的查询接口。

(6)数据集成器：其功能是按照管理员定义的全局模式和全局模式与局部模式间的映射关系，把不同数据采集子系统提交的数据文件合并，组成一个完整的数据文件，根据转换规则消除数据间的异构，得到的数据文件等待查询接口查询。当数据采集子系统提交更新数据时，更新相应的全局数据。

#### 2.4.2 集成查询子系统的流程设计

图 5 为集成查询子系统的工作流程。

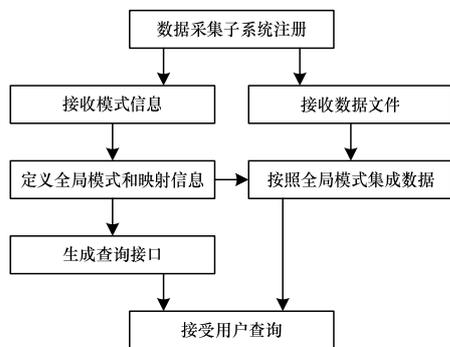


图 5 集成查询子系统工作流程

## 2.5 传输器的设计

传输器的功能是在集成查询子系统与数据采集子系统之间传输命令消息和 XML 数据文件。为保证数据传输的安全，需要对传输的数据和消息进行加密，同时为提高数据传输的效率，减少网络负担，信息的发送方需要对传输的数据进行加密压缩，接收方需要对接收到的数据进行解密解压缩，如图 6 所示。

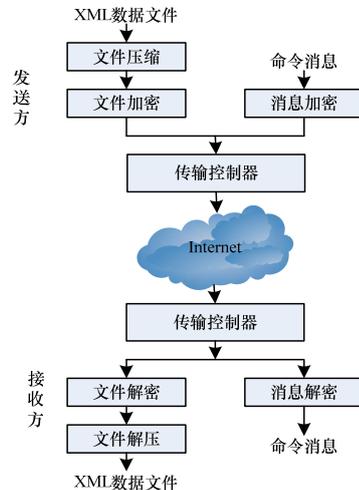


图 6 传输器工作流程

## 3 结束语

天津港业务处理过程中所用系统都是异构的，数据源也各异，开发环境和存储结构也往往不同，造成对数据的综合使用的困难，本文设计开发的异构集成查询系统是为用户和应用系统提供统一、透明、高效地查询位于网络环境中分布的异构数据源的能力和高效的查询处理方式，解决数据集成中可能会出现的平台异构、结构异构、语义异构等问题。

### 参考文献

- [1] 邵秀丽, 张琳, 田振雷. PostgreSQL 在异构数据集成中间件中的应用研究[J]. 计算机工程与设计, 2006, 27(21): 3982-3986.
- [2] 邵秀丽, 韩建彬, 阎仲蹊. 基于 XML 的异构数据源间数据交换的实现研究[J]. 南开大学学报, 2007, 40(3): 9-12.
- [3] 甄玉钢, 刘璐莹, 康建初. 基于 XML 的异构数据库集成系统构架与开发[J]. 计算机工程, 2006, 32(2): 85-87.
- [4] 谢川. 基于 B/S/S 的异构数据源集成系统设计[J]. 计算机应用, 2007, 27(2): 436-441.

编辑 陈文

(上接第 49 页)

### 参考文献

- [1] 温衍博, 吴泉源. MDA 的代码生成技术在电子商务开发平台中的应用[J]. 计算机工程, 2005, 31(20): 104-106.
- [2] Frankel D S. 应用 MDA[M]. 北京: 人民邮电出版社, 2003.
- [3] 崔萌, 史耀馨, 李宜东, 等. 基于 MDA 的 PIM 到 J2EE 平台 PSM 的转换方法[J]. 计算机应用与软件, 2005, 22(1): 1-2.
- [4] 侯金奎, 万建成, 杨潇. 支持 MDA 的 J2EE PSM 模型描述及代

码生成方法[J]. 计算机工程, 2007, 33(15): 79-82.

- [5] Bajec M, Krisper M. A Methodology and Tool Support for Managing Business Rules in Organizations[J]. Information Systems, 2005, 30(6): 243-443.
- [6] 李群明, 张士廉, 王成恩. 可重构的企业管理信息系统研究[J]. 信息与控制, 2001, 30(7): 630-635.

编辑 张帆