文章编号:1000 - 582X(2004)05 - 0034 - 04

基于 Web Services 的数据采集*

胡泽军,李 华,吴中福
(重庆大学 计算机学院,重庆 400030)

摘 要:以异构的、自治的、分布的数据库系统构建数据仓库是个挑战,必须解决两个问题:一是采取有效的措施从各分布的异构数据源采集数据,二是对收集来的数据进行清理和格式转换。笔者分析了 Web services 的开放性和互操作性,提出了以 SOAP 协议和 HTTP 协议连接 Internet 的异构数据源,实现基于 Web Services 的数据采集系统。基于组件技术,提出了一种通用的数据采集器结构,用于数据清理和数据转换。并探讨了数据采集器以 XML 方式实现数据转换和数据装载的核心技术。

关键词:Web Services;数据采集;SOAP;数据仓库;异构数据源;XML
中图分类号:TP393 **文献标识码**:A

许多商业集团经过多年的发展,积累了丰富的数据,可供决策用的数据资源越来越多。这些数据是存放在多个异构的、自治的、分布的信息系统中。各数据库结构的差异,操作平台的异构性以及混乱的概念和术语,成为共享的数据资源的障碍。将来自多个异构数据源的信息进行复制、预处理、集成、注释、汇总后,存储于一个语义一致的数据仓库,解决了操作异种数据库的数据问题,也为数据挖掘创造了条件[1]。以异构平台上的数据构建数据仓库,必须解决两个关键问题:一是采取有效的措施从各数据源收集数据,二是将收集来的数据进行处理,转换数据格式、剔出无效数据,最后将数据加载到数据仓库中。目前采用的办法是为每个需要集成的企业资源或外部资源编写连接代码,提供访问界面。由于每个应用的基础架构都不尽相同,在这些应用上修改和维护系统极为困难。针对这些些问题,笔者论述了如何利用现有工具和 Web Services 核心技术,低成本地连接异构数据源、为数据仓库采集数据。

1 基于 Web Services 的数据采集系统结构

为解决异构数据源的数据采集,提出以 Web Services 作为基础架构的数据采集系统(如图 1 所示)。

图 1 描述了一个营销集团构建的一个数据采集系统,系统包括多个分布的异构子系统(在图中是各分公司)、一个采集器和注册机构。每个异构子系统都有一个 Web Services 服务器,Web Services 屏蔽异构子系统的内部细节,向外公布它的服务接口,能响应采集器的请求、并提供服务。UDDI(Universal Description Discover and Integration)注册中心是在 Internet 上为 Web Services 提供注册服务的机构。客户能通过 UDDI 注册中心查找可用的 Web Services。采集器的功能是从各个 Web Services 收集数据,处理后向公司总部的数据仓库中加载数据。采集器通过 SOAP/http(Simple Object Access Protocol)协议同 Web Services 间交换 XML 文档数据。

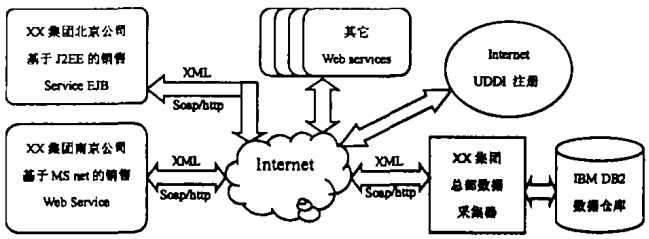

图 1 基于 Web Services 数据采集系统

1.1 数据采集系统的工作流程

Web Service 系统的基本工作模式是用户首先从 UDDI 注册中心查得 Web Services 的服务接口,然后按接口要求格式向 Web Services 请求服务。在图 1 中的采集系统,Web Services 是服务端,数据采集器是作为各个分公司的 Web Service 的客户端来使用。总部的

* 收稿日期:2003 - 12 - 28
作者简介:胡泽军(1972 -),男,四川资中人,重庆大学硕士生,主要研究方向:计算机网络与通信、计算机安全、数据挖掘。

数据采集器与各分公司进行交互时遵从相应分公司的 WSDL(Web Services Description Language)的描述格式。当采集器需要采集数据时，将采集请求翻译成 XML 文档格式的查询消息，向分公司的 Web Services 提出服务请求。分公司服务器响应采集器的请求，并传回采集器所要的数据。采集器收集来自多个异构系统的 Web Services 数据，它们的语境、语义不可能完全一致。因此采集器把源数据装入数据仓库前，首先对数据进行变换，主要是用共同的语义和句法将输入的 XML 转换成标准的格式。例如，如果货币单位不同、计量单位不同、商品标识不同，则要在采集器中转换成统一的标准形式。然后，采集器对数据进行清理，主要清除不感兴趣的噪声或不一致的数据。最后采集器将加工后的源数据装入数据仓库中。

1.2 采集系统的优越性

用 Web Services 集成原有异构系统时，构建速度快，费用省。Web Services 不是取代已有的系统，而是继承已有的企业软件资源，通过用少量代码封装已有的异构系统，就可以为外界提供标准网络服务接口。为了构建数据采集系统，也可以考虑在所有的分公司配置相同的操作系统、数据库系统和应用软件。但这种方案几乎不可行，企业不但要对原系统的用户进行再培训，需要时间和精力，而且在安装、配置、再训练中的费用也不菲。

基于 Web Services 数据采集系统运行费用省。该系统运行在 Internet 上，不用专门租用线路，因而运行费用低。

2 用 Web Services 封装异构数据源

使用 Web Services 的目的是封装各子系统异构平台，为数据采集器提供具有标准接口的数据源。如果某分公司要实现 Web Services，一般要经过如下步骤。首先，该分公司定义一个 WSDL 文件。WSDL 文件按 WSDL 规范描述了 Web Services 所提供的服务以及用户调用该服务时应遵从的消息格式。第二，实现 WSDL 描述的服务，Web Services 接口一般是简单地调用原有系统已经存在的功能。第三，为了便于 Internet 上的客户发现服务，应当对 WSDL 进行注册。如图 1，也可直接把 WSDL 发给集团总部的数据采集器。

2.1 Web Services 的核心技术

Web Services 消息能够穿越不同的程序语言、操作系统、防火墙、硬件平台同另外的应用系统通信，完全归功于 Web Services 采用的技术标准[2]：

1）系统间通信采用标准 XML 形式；

2）一般用 HTTP 作为应用层通信协议；

3）在 HTTP 基础上用 SOAP 消息的交换协议；

4）WSDL 被用来提供接口输入输出参数的元数据描述；

5）使用 UDDI 来发布所有的注册 Web Services 接口。

XML 即可扩展标记语言，是一套规范，允许编程人员自行定义如 HTML 般的标注，以方便数据存取和处理、交换、转换等[3]。XML 的主要特点是使用有意义的标记，这个特性使计算机可以理解数据含义[4]。XML 可以在任何系统、应用程序、任何平台上运行。XML 数据的传输以 Internet 为基础，传输费用便宜。

SOAP 用作承载 XML 的协议。SOAP 主要作用在于穿越防火墙、实现远过程调用。SOAP 由 4 个部分组成：SOAP 信封，它构造了一个整体的表示框架，可用于表示消息中是什么，谁处理它。SOAP 编码规则，它定义了一个数据的编译机制，通过它来定义应用程序中需要使用的数据类型。SOAP RPC 表示，它定义了一个表示远端过程调用和响应的约定。SOAP 绑定，它定义了一个使用底层传输协议来完成在结点间交换 SOAP 信封的约定。

2.2 用 WSDL 定义 Web Services 接口

WSDL 文件是 XML 实例文档。本文写作时，WSDL 1.2 是 W3C 的一个工作草案，它定义了一套 XML schemas，用于指导如何创建 WSDL 实例文件[5]。所有 WSDL 结构都被定义为 < definitions > 元素的子元素。WSDL 语言提供了一种模式和格式来描述 Web Services。WSDL 使得对 Web Services 的抽象功能描述同具体的实现细节相分离。WSDL 首先描述服务提供者和请求者之间消息格式，消息本身被抽象描述，然后捆绑到具体的网络协议和消息格式。消息由若干类型的数据项组成。服务者和请求者之间的多个交换消息被描述为操作。操作的集合称为端口类型。端口的集合称为服务类型，其中的每个端口执行一个端口类型，而服务类型包含了服务交互的所有的细节。

2.3 对 Web Services 的接口功能进行实现及发布

由于 Web Services 要向外提供的功能在系统中已经存在，当前的任务就是编写少量的代码，为原有系统加装一个壳，将原系统封装起来。在程序中主要实现两个功能。其一是对 Web Services 客户端传来的请求消息进行解析，用解析所得参数调用相应的组件功能。其二是将系统中组件函数的调用结果按照消息的格式，用 XML 语法进行封装(WSDL 定义的格式)，最后传送给 Web Services 的客户端。

Web Services 的功能实现后，需要对其服务进行发布。将要发布的功能在 UDDI 注册中心进行注册，然后客户就可在 UDDI 注册中心发现自己所要的服务。[6]。对于本应用而言，XX 的分公司完全可以将 WSDL 直接发送给总部的数据采集器，也就是所谓静态绑定。Web Services 的功能实现和接口发布后，客户便可通过 Internet 对其访问了。

3 用数据采集器采集数据

3.1 数据采集器的结构

数据采集器的结构如图 2 所示。采集器由 4 个部分组成：采集组件、转换组件、清理组件、和装载组件。采集组件使用适当的方法在各组件和系统之间传送数据。这个采集器能够合并各种来源的数据，转换并提供数据仓库系统所需的数据源。这个采集器模型是一个通用的模型，也可用于类似的需要转换异构数据源的地方。

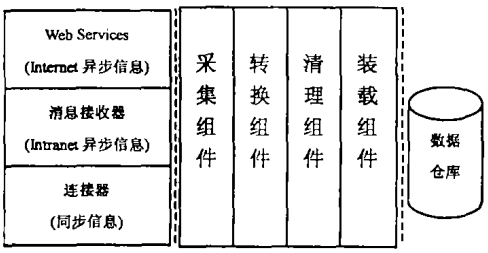

图 2 数据采集器的结构图

在整个数据采集器的工作流程中数据的格式都是 XML 的形式，这样有几个好处：一是免除 XML 数据格式同其它格式的来回转换；二是可以利用 DOM 来对 XML 进行有效的处理；三是可以利用 Schema 文档对 XML 数据作有效性的效验，而不用编写定制代码。由于几大数据仓库供应商现在都支持用 XML 直接访问数据仓库，所以软件装载组件也是用 XML 格式访问数据仓库。

3.2 数据采集

采集组件可以用 3 种方式获取数据：Web Services、消息接收器、连接器。Web Services 用于从 Internet 上接收 XML 格式的异步信息；消息接收器用于从 Intranet 上接收异步信息；连接器可以从 Intranet 上接收同步信息。在 XX 公司的这个应用中，采集组件用 Web Services 模式来完成，使用 XML 文档进行异步交换。

当采集器需要采集数据时，将要采集的内容目录翻译成 XML 文档的查询消息，向分公司的 Web Services 提出请求，其消息的请求格式要符合 WSDL 的描述。Web Services 接受请求后会返回响应消息，响应消息以一个 SOAP 信封进行封装。

3.3 数据转换及清理

采用 Web services 模式仅仅是从不同结构的数据源收集了可用数据，然而不同的分公司系统可能采取了不同的计量单位、不同的标识（如：字段名），Web Services 在封装原有系统时并没有被标准化，因此采集而来的数据可能存在语境、语义上的差别。数据转换组件完成的主要功能是标准化收集到的 XML 文档信息的语境和语义，比如统一单位和标识，这也是采集数据中要解决的难点。数据的转换处理可以利用现有的工具进行。例如，定义好指导转换的 XSLT（Extensible Style Language Transformations）文件，然后利用 MSXML4.0 就可以实现源 XML 文档到目标 XML 档的转换[7]。

清理组件用于对 XML 文档信息中的标准格式数据进行分析处理，是消除噪声数据、去除冗余数据、解决冲突功能部件。由于数据仓库是按照一定的主题来采集数据，清理组件要作对商业决定有用的分析。

3.4 数据装载

数据装载组件把整理好的数据写入数据仓库，即是一个简单的数据仓库访问组件。该组件若是以 XML 方式直接访问数据仓库则可大为简化，近来各数据仓库厂商纷纷支持 XML 方式访问数据仓库，这种方式现已可行。利用 XML 文件，以及用于映射 XML 中的元素到数据仓库相应字段的 schema 文件，就可实现对数据仓库的装载。例如，对 SQL Server2000（需要安装 Microsoft SQLXML 3.0 Service Pack1）的访问，下列语句可以把上面接收到的消息插入数据仓库 vendingif 表中：

```
<ROOT xmlns: updg = ..... >
<updg: sync mapping - schema = " vendingif.xml" >
<updg: before >
</updg: before >
<updg: after >
<vendingif tradename = " TV DC2908" ... />
</updg: after >
</updg: sync >
</ROOT >
```

对数据仓库的操作要依赖 mapping schema，它建立了 XML 中数据元素和数据仓库表的字段的对应关系。上列中 vendingif.xml 文件就是用于将数据映射到 vendingif 表的 mapping schema。

4 结 论

在开发中采用 Web Services 标准更容易解决异构平台的数据交换问题。因为程序员可以采用通用的工具(各种 Web Services 标准、软件商提供的开发工具)，这些工具的使用免除了访问异构平台的苦恼，让程序员将精力集中于要解决的语义转换上。SOAP 是基于文本的协议，建立在广为人知的 HTTP 协议之上，开发者容易上手。SOAP 比其它方案中使用二进制的协议要求更高的带宽，这个问题对联机事务处理有影响，但对数据采集的解决方案来说不重要。

参考文献：

[1] HAN JIAWEI, MICHELINE KAMBER. Data Mining: Concepts and Techniques [M]. SAN FRANCISCO: Morgan Kaufmann Publishers, Inc. 2001.

[2] 柴晓路，梁宇奇. Web Services 技术、架构和应用[M]. 北京：电子工业出版社，2003.

[3] 硕网资讯. 洞悉 XML[M]. 北京：北京大学出版社，2001.

[4] W3c. Extensible Markup Language (XML) 1.0. Second Edition[EB/OL]. http://www.w3.org/TR/2000/REC-xml-20001006.2000.

[5] ROBERTO CHINNICI. Web Services Description Language (WSDL) Version 1.2[EB/OL]. http://www.w3.org/TR/2003/WD-wsdl12-20030303, 2003.3.

[6] TOM BELLWOOD. UDDI Version 3.0 Published Specification[EB/OL]. http://uddi.org/pubs/uddi-v3.00-published-20020719.htm.

[7] MICROSOFT XML Core Services 4.0[EB/OL]. http://msdn.microsoft.com/downloads/default.asp,2002.

Data Collection Based on Web Services

HU Ze-jun, LI Hua, WU Zhong-fu

(College of Computer, Chongqing University, Chongqing 400030, China)

Abstract: To construct data warehouse with heterogeneous, self-governing and distributed database system is a challenge. Two problems have to be solved: one is how to collect data from heterogeneous data sources by an effective method, the other is how to clear up and change the format of the data collected from data sources. The opening and the interact-operation property of Web Services is Analyzed. A Data Collection System based on Web Services, which communicates with heterogeneous data sources with SOAP and HTTP, has been designed. A Data Collector is introduced to clear up and to transform data with composition technology. The pivotal XML technology of transforming and loading up the data are discussed.

Key words: web services; data collection; SOAP; data warehouse; heterogeneous data source; XML

(编辑　吕赛英)