一种面向领域 WEB 服务的数据中心模型①

沈 燕, 雷 蕾

(南阳理工学院 软件学院, 南阳 473000)

摘 要: 提出一种通用的面向领域 WEB 服务的数据中心模型. 利用语义集成、数据映射、数据仓库及其他数据集成技术, 不仅完成分布式异构数据源的无缝数据集成, 而且实现数据源与数据中心之间的数据共享和透明数据交换, 为领域 WEB 服务提供统一数据服务. 在"油气生产系统软件集成平台"中采用该模型, 组建了中国石油油气井生产领域数据中心, 构建一个面向油气井生产领域, 集生产管理、设备管理、工作流程控制、优化设计、故障诊断、辅助决策等功能为一体的 WEB 服务平台, 解决了海量、分布式异构数据源的有机集成和无缝共享问题. 从而验证了该模型的正确性及可行性.

关键词: 数据中心; 语义集成; 数据映射; 数据仓库; 分布式异构数据源集成

Field-Oriented Data Center Model for WEB Service

SHEN Yan, LEI Lei

(Software institute, Nanyang Institute of Technology, Nanyang 473000, China)

Abstract: Presents a general Field-Oriented Data Center (FODC) for WEB services. Using the semantic integration, data mapping, data warehouse and other data integration technologies, not only achieves seamless integration of distributed heterogeneous data sources, but also realizes data sharing and data exchanging between distributed heterogeneous data sources and the FODC. Therefore, the model provides a unified data services for WEB services. In an application of the model—Oil and Gas Production System Software Integration Platform, established the Science Data Cloud in the field of the oil and gas well production of China, solved the massive, distributed heterogeneous data organic integration and seamless sharing problems, and presented a WEB service platform for oil and gas production, with production management, equipment management, workflow control, optimal design, fault diagnosis, decision support and other functions. The case verified the correctness and feasibility of the model.

Key words: data center; semantic integration; data mapping; data warehouse; distributed heterogeneous data integration

随着计算机信息技术的飞速发展, 许多领域内部遗留下来的分布异构业务应用系统变成了一个个"信息孤岛"[1], 随着 Internet 的日益普及, 迫切需要建立面向领域的 WEB 服务平台[2,3], 集成诸多分散的业务系统, 集成分布式异构数据源, 在这些"信息孤岛"之间共享数据和交换数据. 因此, 我们提出一种通用的面向领域 WEB 服务的数据中心模型[4,5]: Field-Oriented Data Center (FODC) for Web service 是针对某一具体领域分布的、异构的数据源, 利用语义集成、数据映射技术, 创建全局统一的数据视图, 利用数据仓库等技术, 实现分布异构数据无缝集成、共享和透明的数据交换, 为领域 WEB 服务提供统一数据服务.

创建领域数据中心是解决分布式异构数据源集成问题的关键[6,9]. 数据中心主要解决数据密集领域的资源共享问题. 按照实现技术原理数据中心可以分为两类: 虚拟数据中心和物理数据中心. 虚拟数据中心数据源扩展比较容易, 但转换时非常复杂, 系统管理更困难. 物理数据中心是利用数据仓库以及 ETL 技术, 集成分布式异

① 收稿时间:2012-12-11;收到修改稿时间:2013-01-11

构数据源，创建一个物理数据中心，并与数据源同步. 物理数据中心实现起来相对简单，数据处理速度较快[8]. 文章主要研究物理数据中心模型的建立和应用.

1　一种面向领域WEB服务数据中心模型框架

面向领域 WEB 服务数据中心模型整体框架包括客户端、WEB 服务、数据中心和数据桥(实现分布、异构数据源集成及共享)四个部分，如图 1 所示.

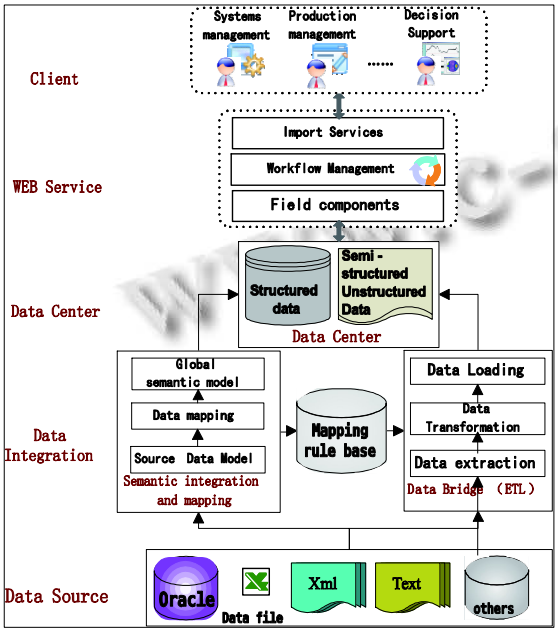

图 1　面向领域 WEB 服务数据中心模型框架图

1) 客户端

基于网络浏览器或其它应用程序，对外提供统一应用服务，用户提交查询或其它操作，用户不必关心底层数据源的分布情况和数据模式等差异. 最终只返回满足用户最初查询要求以及数据接口描述的数据.

2) WEB 服务

WEB 服务器为用户提供工作流定制和环境检测，用户可以根据需要添加、修改或删除领域服务. WEB 服务器和数据中心有直接数据访问接口，只要遵循接口规范，即可以有效地、透明地操作底层各个数据源，而不考虑系统底层异构数据源的具体实现[6].

3) 数据中心

数据中心是系统数据服务管理平台，接收 WEB 服务提交的查询等数据操作语句，解析并根据数据映射规则将其转化为针对数据中心的直接查询，返回用户查询结果，并转化为最初用户的数据接口描述形式.

4) 数据桥

数据桥连接数据中心与多种异构数据源(关系数据库、面向对象数据库和非结构数据等)，提供通用数据访问接口，能屏蔽各种异构数据源之间的差异，跨平台、跨网络访问数据源，为上层提供统一数据视图.

2　面向领域WEB服务数据中心模型实现流程

数据中心为领域 WEB 服务提供数据服务，必须对底层各个异构数据源进行抽象、归纳与综合，形成局部语义模型，利用语义集成技术，合并局部模型，消除语义异构和模式异构，构建全局应用视图，形成数据中心全局模型，这是整个数据库中心设计的关键. 然后利用全局模型物理设计并优化数据中心. 通过数据桥工具装载底层数据源数据到数据中心，为用户提供统一、快速的数据服务.

2.1　利用语义集成技术创建系统全局数据模型

数据集成问题是为用户提供统一的视图以访问驻留在分布、异构、自治数据源上的数据[6]. 语义数据集成系统 SI 是一个五元组：SI=<GS, LS, DS, Mgs, Msd>[7]，其层次结构如图 2 所示.

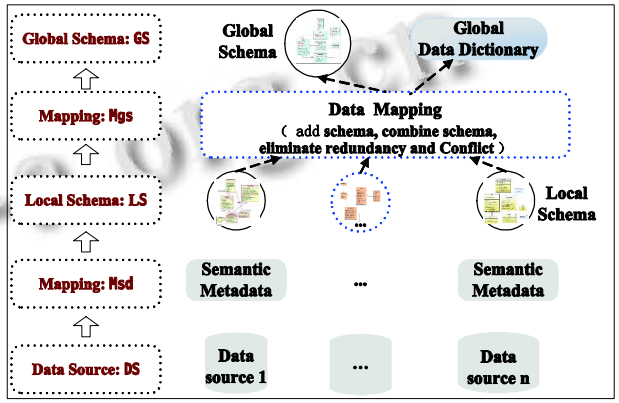

图 2　数据集成系统的语义层次图

GS 代表领域的全局概念模式，表示全局的语义. LS 是局部概念模式，是局部数据源的语义描述. DS 是各种异构数据源. Mgs 是 GS 与 LS 之间的映射，描述两个模式之间的联系. Msd 是 LS 和 DS 之间的映射，描述了来自数据源的数据模式与局部概念模式之间的联系.

异构数据集成是以语义模型驱动方式进行的，其特点就是语义模型先于模型与数据之间的映射而存在，

数据到模型的映射依据语义模型中的知识完成[5].

语义模型是由对问题域进行抽象所形成的类结构树和作用于类结构树上的语义关系、映射关系组成.

语义模型在应用框架中可以看作是一个在语义层面上的数据协调器. 在语义模型部署及获得外部应用请求之后, 首先通过模型访问接口对类结构树进行遍历并得到所关心的树节点信息; 然后提取数据映射关系, 再建立属性映射关系, 同时对其所涉及的语义关系进行解析, 进而获得分布于异构数据库中的数据信息; 最后返还给应用的是语义一致的数据结果集.

语义模型的实现和数据映射是密不可分的, 利用语义模型中的知识进行推理, 实现异构数据源到全局数据模型的映射是面向领域异构数据源集成的重要过程. 数据映射应该贯穿于始终, 起着元数据的作用.

2.2 依据系统全局数据模型, 创建物理数据中心

首先依据系统全局数据视图, 确定系统全局逻辑模型: 选择 DBMS, 分析主题域, 确定当前要装载的主题, 确定粒度层次划分, 确定数据分割策略, 定义关系模式, 把全局概念视图转换为全局逻辑模型定义要装载的主题的逻辑实现, 记录数据仓库的元数据. 然后依据全局逻辑模型设计系统全局物理模型, 确定数据的存储结构、索引策略, 确定数据存放位置, 确定存储分配. 依据存取时间、存储空间利用率和维护代价, 确定数据的存储结构. 按数据的重要程度、使用频率以及对响应时间的要求进行分类, 并将不同类的数据分别存储在不同的存储设备中.

确定数据中心实现的物理模型, 必须全面了解所选用的数据库管理系统, 特别是存储结构和存取方法. 必须了解数据环境、数据的使用频度、使用方式、数据规模以及响应时间要求等, 这些是对时间和空间效率进行平衡和优化的重要依据. 必须了解外部存储设备的特性, 如分块原则, 块大小的规定, 设备的 I/O 特性等.

在实施过程中, 为了提高数据中心的自适应性及独立性, 不影响数据中心对外提供的服务, 将数据中心从逻辑上分为三层: 基础层、分析层、结果层.

基础层数据保留了来自源系统的原始数据, 存储在二维数据表中. 可以提高数据仓库的提取效率, 降低对源系统的资源占用. 分析层按照领域知识库(又称中间规则库)的抽取、清洗及整合规则, 把基础层相关数据整理统一放在二维表中. 优点是能够将基础层的原始数据自由组合, 满足多变的业务需求. 结果层辅助决策、多维分析及其他 WEB 服务应用将在结果层上建立模型, 该层一般使用星型数据模型, 创建不同的分析纬度, 提高上层应用的效率.

数据中心利用虚拟化技术对众多存储设备进行整合, 降低系统的复杂程度, 将大量存储设备整合到一个较大的虚拟存储池内, 使用户通过统一的视图来管理整个系统的存储资源, 降低维护成本和提高存储设备性能和利用率. 数据中心是一个集成的信息存储仓库, 既具备批量和周期性的数据加载能力, 也具备数据变化探测、新数据的连续加载和更新能力, 并能结合历史数据和新颖数据实现查询分析和自动规则触发, 从而为上层 WEB 服务提供快速、高效的数据服务.

2.3 构建数据桥

数据桥作为数据中心与异构数据源连接的桥梁, 支持数据的多向传递和集中处理, 可满足数据跨域集中、历史数据迁移和远程数据共享及同步等应用需求. 数据桥利用 ETL 技术完成数据源发现、数据源分析、数据标准化、数据映射、数据抽取、数据清洗、数据转换、数据加工、数据加载这一系列数据处理过程, 加载数据到领域数据中心, 并提供数据共享和同步服务.

ETL 即数据抽取(E)、转换(T)、加载(L)过程, 屏蔽了复杂的业务逻辑, 为基于数据中心的分析和应用提供统一的数据接口, 常应用于信息系统中数据的迁移、交换和同步. ETL 有两种处理流程. 一种是异步(Asynchronous) ETL 方式, 也称为文本文件(Flat file)方式. 另外一种是同步(Synchronous) ETL 方式, 也称为直接传输 (Direct transfer) 方式. 数据桥设计和实施过程当中, 考虑到底层数据源的实际情况, 分布在不同地域的、要通过网络传输的异构数据源多采用异步 ETL 方式, 本地局域网内的数据源一般采用同步 ETL 方式.

总的来说, 数据桥利用 ETL 技术提供了一种异构数据源集成的通用解决方案, 可以避免手工操作的繁琐, 提高异构数据抽取、转换和转载的效率, 且不影响原来领域应用系统的使用.

2.4 构建 WEB 应用服务层

WEB 应用服务层采用 SOA 架构(Service-Oriented Architecture)[9], 为各种 WEB 应用定义通用接口和契约, WEB 应用程序的不同功能单元(称为服务单元)通过这些定义和契约联系起来. 这些服务单元层是基础, 可以直接被应用调用、组合. WEB 应用服务层基于 SOA 体系架构集成业务功能模块, 建立可重用且性能可靠的系统架构, 解决了应用功能变化, 带来解决方案变化, 从而导致系统架构变化的问题.

WEB 应用服务自上而下可以分为服务接入层、工作流层和领域组件层三个层次.

服务接入层提供了多种服务方式供用户访问. 使用户可以直接访问平台所提供的应用, 并且可以利用 Mashup[1]技术快速构建其自己的应用. 同时该层提供 WEB 服务接口, 可以让用户访问和添加领域的服务. 工作流层提供工作流的表示、存储、运行以及实现机制, 向用户提供工作流定制和检测环境. 领域组件层主要定义组件的接口规范和元数据表示方法, 构建组件库管理各种类型的软件组件, 并建立动态的组件查询和发现机制. 利用 web service 技术对领域组件进行封装, 形成领域组件库, 支持业务组件的远程安装与维护.

模型 WEB 应用层以 WEB 形式对外提供统一服务接入、工作流程管理、业务组件定制、多维数据分析、辅助决策等功能, 模型 WEB 应用层利用 SOA 架构保证系统可扩展性, 并使系统能够跨网络、跨平台、支持庞大用户群体.

3　面向领域WEB服务数据中心应用案例

“油气生产系统软件集成平台”是针对油气勘探开发和油田管理决策的信息化系统, 系统设计开发实现过程中需要集成各个油田公司原有业务系统数据源, 例如 A2、PDPMIS、OPRS 等数据源, 由于油田公司数据的分散和异构特点, 我们无法简单的把各个数据源合并. 系统实现过程中采用上述面向领域 WEB 服务数据中心模型, 建立了针对中国石油领域油气井勘探开发、生产管理的数据中心、石油领域语义集成知识库、各个油田数据源到数据中心的映射规则库, 解决了中国石油领域海量、分布式异构数据源的有机集成问题. 同时设计实现了领域数据桥, 解决了中国石油领域各油田公司数据平台的无缝连接、数据共享及同步问题.

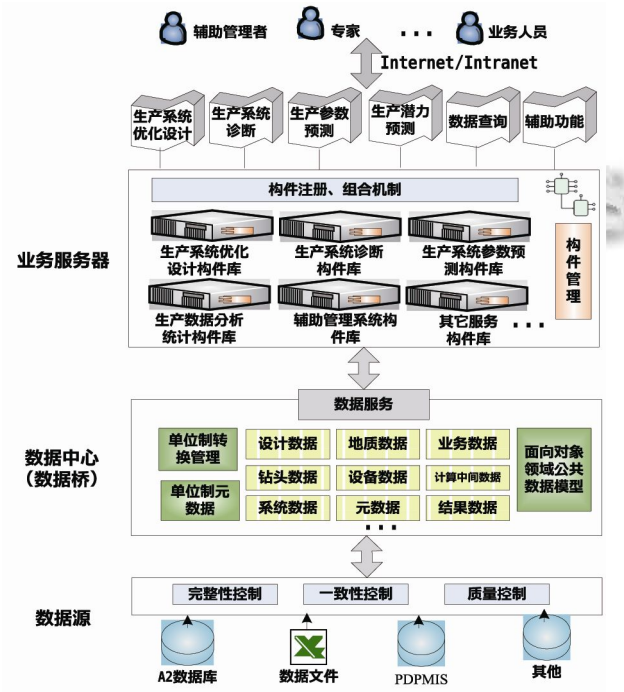

图 3　油气生产系统软件集成平台解决方案

平台系统构建了一个面向油气井生产领域, 集生产管理、设备管理、工作流程控制、优化设计、故障诊断、辅助决策等功能为一体的 WEB 服务平台, 该平台使用 ASP.NET、J2EE 等开发工具在 oracle 数据库下开发, 系统解决方案如图 3 所示.

4　结论

提出了基于语义集成模型和数据桥技术的面向领域 WEB 服务数据中心模型, 给出了该模型的实现流程, 并在“油气生产系统软件集成平台”中进行了应用, 解决了油气井生产领域的海量、分布式异构数据源的有机集成和无缝共享问题. 实践证明该模型是行之有效的. 面向领域 WEB 服务的数据中心可以看做领域“云存储”的具体实现, 为领域“私有云”” 的实施提供了数据基础.

参考文献

1 Poggi A. Structured and Semi-Structured Data Integration [Ph.D. Thesis]. Joint work between the University of Rome “La Sapienza”:(Italy) and the University of Paris-Sud (France), 2007.
2 OKBC. Open Knowledge Base Connectivity. http://www.ai.sri.com/～okbc/[2004-10-10].
3 岳昆,王晓玲,周傲英.Web 服务核心支撑技术:研究综述.软件学报,2004,15(3):428–442.
4 王克飞,张树生,周竞涛.语义模型驱动的数据属性匹配技术研究.计算机应用研究,2006,23(10):39–40.
5 林子雨,杨冬青,宋国杰,王腾蛟,唐世渭.实时主动数据仓库中多维数据实视图的选择.软件学报,2008,19(2):301–313.
6 张晓明.领域科学数据语义集成模型及映射[学位论文].北京科科技大学,2009.
7 王静,孟小峰.半结构化数据的模式研究综述.计算机科学,2001,28(2):6–10.
8 李文贵,张超英,等.基于共享数据中心的学分制收费模型的研究与应用.计算机工程与设计,2010,31(7):1623–1626.
9 邵永春,邓晓华,苏志文,李世友.分布式空间数据中心建设研究.微计算机信息,2008,01-1:233–234.