基于异构数据资源整合的方法和系统实现

徐立新
(广东电网电力科学研究院,广东 广州 510080)

摘 要:随着电力行业信息领域的不断发展,信息化的不断深入,使得电力企业在对信息化过程中积累的大量异构数据源的处理变得尤为重要。基于 Web 技术的成熟,产生了将这些异构数据整合起来的可能性,这已成为当前十分热门的课题。文中在基于数据库系统开发的技术上,提供了一种数据资源整合的方法,同时在这个方法的基础上实现了一整套数据整合的解决方案。在采用基于 J2EE 标准规范和 B/S 模式的基础上把数据源的定义,数据集的提取,映射关系定义、目的数据源定义和数据加载的流程通过工作流的方式耦合,灵活地解决了资源整合的问题。

关键词:异构数据;数据整合;方法;实现
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)12-0172-04
doi:10.3969/j.issn.1673-629X.2014.12.040

Method and System Implementation Based on Heterogeneous Data Resources Integration

XU Li-xin
(Guangdong Grid Electric Power Research Institute, Guangzhou 510080, China)

Abstract: With the continuous development of information field of the electric power industry, the deepening of the information, make the electric power enterprise in the process of informatization in the processing of the accumulation of a large number of heterogeneous data sources is becoming particularly important. Based on Web technology mature, have the possibility to integrate these heterogeneous data, which has become the current very popular topic. Based on the development technology of database systems, provide a method of data resources integration, and on the basis of this method implement a set of data integration solutions. On the basis of B/S model and J2EE based standard specification, couple the definition of the data source, the extraction of data sets, mapping relationship definition, objective data definition and data loading process through workflow, flexibly solving the problem of resource integration.

Key words: heterogeneous data; data integration; method; implementation

0 引 言

随着信息化的推进,电力企业和政府部门等单位在分析报告撰写和数据决策分析的需求中,需要把远程的一些异构数据作为分析的基础。对于远程的这些数据库系统各不相同、存储方式上存在大量数据错误以及存储信息的数据结构上存在很大的差异的异构数据,由于缺乏统一的数据描述标准,使得这样的数据阻碍了单位主体信息化和数字化的进程。在目前的方法和实现方案中主要是基于 C/S 结构的数据整合方式,这种方式下需要在用户机器上安装客户端,对用户机器的要求很高,同时客户端软件维护难度大,缺乏对工作流程的支持,而且不支持电力行业需要大系统整合的特殊情况,从而导致整合数据和数据使用分离在不同的系统中,不利于资源的共享与检索。针对这些问题,文中阐述了一种数据资源整合的方法,同时在这个方法的基础上实现了一整套解决方案。在采用基于 J2EE 标准规范和 B/S 模式的基础上把数据源的定义,数据集的提取,映射关系定义、目的数据源定义和数据加载的流程通过工作流的方式耦合,灵活地解决了资源整合的问题。

1 异构数据整合系统架构

系统框架是系统的核心部分,系统架构层负责与数据库的交互,为平台业务支撑层的协调和指挥处理

收稿日期:2014-01-15 **修回日期**:2014-04-21 **网络出版时间**:2014-10-23
基金项目:中国南方电网科技计划项目(K-GD2012-400)
作者简介:徐立新(1967-),男,湖北罗田人,硕士,高级工程师,从事科技情报研究工作。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1052.014.html

多个应用的运行提供支持。为了充分利用电力企业已有的网络硬件环境,本系统采用基于 J2EE 标准规范和 B/S 模式,以 Oracle 网络数据库管理系统为系统的数据库核心,采用通用的微软平台与 IE 浏览器作为与客户端交流的平台,建立一个基于异构数据整合技术,实现异构资源的统一管理,结构化数据和非结构化数据统一搜索的综合系统。

1.1　J2EE 平台规范

J2EE 是一个标准的体系结构,它主要面向基于 Web 的企业应用开发方面,与企业密切相连,程序设计语言为 Java。许多应用都是按照 J2EE 规范来开发的,在这种情况下,基于 J2EE 体系的系统之间的兼容性很高,配合密切,能够方便的融合,因此,这些应用在各 J2EE 服务器之间是可以无障碍移植的。J2EE 作为一种支持性技术,与 Internet 联系紧密,并能够为 Web 应用程序提供更强的稳定性、有效性。当其应用于电力企业时,能够实现电力企业数据管理的稳定性、实时性、兼容性以及有效性。

1.2　Web 数据库数据存储

Web 数据库作为 J2EE 架构的核心内容,在电力企业异构数据资源整合系统中也十分重要。其数据库主要有 SQL Server、MYSQL 和 Oracle,它们都是使用广泛、很受用户欢迎的数据库。这三种数据库使用非常简捷方便,有很强的可扩展性、适应性和可移植性,并且稳定可靠,不容易被干扰,在国内知名度很高,得到了很多用户的高度认可,应用范围也十分广泛。在本系统中采用了 Oracle 作为其整合数据存储的数据库,Oracle 是一种支持对象关系的模型,主要面向网络计算机,作为目前广受欢迎的服务器/客户机结构的数据库之一,它有着很强的稳定性和可靠性。适用于电力企业的数据存储和访问。

本系统数据存储实现海量数据资源的存储管理,在自建资源库、互联网采集资源库、外购资源库,以及其他交换数据的基础上,构建出索引库和平台基础库,通过全文检索适配器网关,实现索引库和基础库的同步更新,从而实现对各类资源进行有效整合和管理。

1.3　Web 的 B/S(Brower/Server)体系结构

B/S 模式工作原理是:电力企业用户通过浏览器将请求发送给 Web 服务器,Web 服务器起一个中转的作用,再将请求发送给数据库服务器来处理,数据库服务器将请求数据进行合理的处理后将结果再返回给 Web 服务器和浏览器。较为流行的 B/S 模式的网络结构如图 1 所示。

B/S 结构模式的主要优点有:

(1)其界面相对比较统一,容易进行操作;

(2)能使系统具有很强的开放性,在这种情况下,没有对用户的数量进行限制;

(3)使 Internet 与所有局域网进行顺利的连接,异构系统的连接将很容易实现。

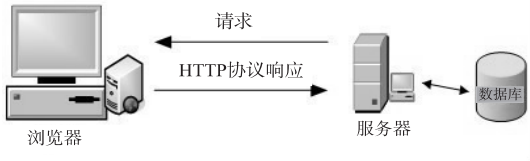

图 1　浏览器/服务器系统结构

2　异构数据源的整合

2.1　异构数据源

所谓数据源异构指的是数据源具有不同的类型、数据源在存储上具有不同模式以及数据语义的差异这三个方面。在不同的存储模式中,使用对象模式、关系模式为其存储模式的系统具有较好的兼容性,但在某些系统中,即使使用同类存储模式,其模式结构的差异也会造成系统资源的异构性;所谓语义的差异指的是结构相同的数据形式解释不同语义或者同一语义由不同形式的数据表示。

从数据的来源来分,异构数据可以分为互联网数据资源、自产资源和外购资源三种。异构数据是通过各种数据源途径获得的,数据源为数据的加工、处理提供源数据;也为系统中所涉及的跨库检索和异构数据整合提供了必要的基础。

电力企业自产资源即为企业内部的各类简报、简讯、报告、文档、汇编专辑、交换资源和内参、视频文件、技术论坛数据等。外购资源为通过有偿的方式从各资源商手中购置的数据资源,主要有:成果库、中文论文库、标准库、外文论文库、机构库、专利库、图书等数据。互联网数据资源是通过网络爬虫从互联网上定点采集的信息资源[1-4]。

对于源数据的存储分为数据库数据存储和原文文件存储。数据库数据存储是将从多种途径获得而来的源数据,通过排重、分类等多种形式加工处理,然后存储到 Web 数据库中。文件存储是将系统中的所有资源文件通过一定的分类规则统一存放到 Web 数据库中存储。

2.2　异构数据整合方法

通过对数据仓库、相关的多个数据库以及数据集构成的异构数据资源的整合,实现不同数据库和系统之间的资源共享和透明访问,对于组成异构数据库的不同数据库和数据集,由于它们在整合之前各自拥有自己的 DBMS,所以在整合之后各组成部分仍然具有自身独立的自治性和应用特性,这样的机制使得在实现数据共享和互访的同时,增强各部分数据的安全性。

异构数据整合完成以后可以实现资源的透明访问和资源共享[5-7]。

电力企业内部的资源建设涉及到多种外购资源和自产资源,这些资源数据的来源和格式均不相同。在保证各原始数据系统的完整性和可访问性的前提下,如何将这些数据进行整合,形成一个整合的资源库,以向专业系统提供数据共享和一站式跨库检索的数据支撑是解决异构数据整合到透明访问的核心。其具体的整合方法如图 2 所示。

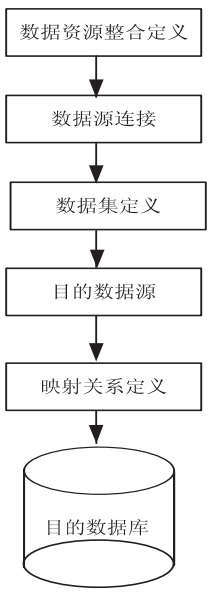

图 2 异构数据资源整合的方法

(1)数据资源整合定义过程。

定义一个具体数据整合的任务,在定义中指定整合的名称、映射关系和执行的调度规则等信息。数据资源管理整合的模块通过属性的绑定形式,将执行过程中的信息关联起来,可以灵活搭配具体的信息。

(2)数据源连接过程。

主要是定义和维护需要提取的数据的来源信息。包括数据库所在的 IP 地址、数据库类型、数据库名称、数据库用户名和密码等信息。通过这种定义方式,可以灵活地切换连接的方式。同时在数据库连接的实现中加入连接验证的功能,可以校验连接的有效性。

(3)数据集定义。

主要是定义需要提取的原数据的数据集合,其定义是建立在数据源连接的基础上。包括数据集名称定义,具体提取的数据库表和数据库表中列。

(4)目的数据源。

主要是定义和维护整合后资源导入到的具体目的数据库和表格。包括对目的数据库的定义和数据表的指定。

(5)映射关系。

在此过程中定义了源数据和目的数据之间的一个映射关系,主要包括字段连接、字段截取、字段求和、字段取整、字段求平均值,字段类型转换和字段格式转换等。样例如下:

字段连接:字段 Name = “中国”,字段 Province = “广东”,连接后目的字段 address = “中国广东”。

字段截取:字段 Company = “中国微软亚洲研究院”,截取后目的字段 Company = “微软亚洲研究院”。

字段求和:字段 Salary = {5 000,10 000,20 000},求和后目的字段 TotalSalary = 35 000。

其他操作如上。

(6)执行过程。

通过前面过程的实现,通过执行过程可以异步的工作流形式提取数据源的数据,根据映射关系存放到目的数据库中。

通过以上方法和实现方案可以看出,数据整合方式的定义方式串联上整个数据资源的整合流程,一个整合定义,关联到具体的映射关系,映射关系中需要指定数据的来源和目的数据源,这样在具体执行中可以以一种工作流的方式异步整合数据。而且通过把数据资源整合的实现集成到数据的使用系统中,这样无缝地解决了数字鸿沟的问题,也避免了操作的跨系统问题。在 B/S 构架的基础上把数据源的定义,数据集的提取,映射关系定义、目的数据源定义和数据加载的流程通过工作流的方式松耦合,灵活地解决了资源整合问题,使用表明在电力行业的报告撰写中提供了很好的支撑,这种方法在其他领域也具有很好的推广性。

2.3 数据库整合方案

要完成数据库全文检索的功能,对不同数据库之间的整合尤为重要,文中通过自定义的全文检索数据库作为资源数据的统一存储数据库,使用全文检索网关来实现多种关系型数据库的数据到全文检索数据库的转换和同步共享。其工作方式如图 3 所示。

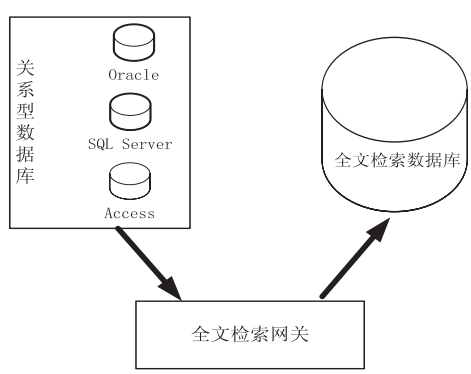

图 3 关系数据库网关示意图

全文检索网关的作用是数据更新代理,保证关系数据库中的数据发生变化时,数据所对应的全文索引可以及时更新,考虑到数据量和性能的影响,数据更新代理必须能够实现索引的增量更新。在这种系统结构

下,实现非结构化数据的检索应用包括如下步骤:

(1)用户提交非结构化数据的检索请求;

(2)应用层将用户检索条件提交给专业数据库进行检索,专业数据库返回命中记录的主键信息;

(3)应用层根据专业数据库返回的主键信息在关系数据库中找到相应命中的记录;

(4)将最终检索结果以用户期望的表现形式返回给最终用户。

在典型的应用中,全文检索网关系统的定位在于将关系数据库处理结构化数据的优势和专业数据库处理非结构化数据的优势结合起来,同时在应用层进行无缝集成。其实现原理如图 4 所示。

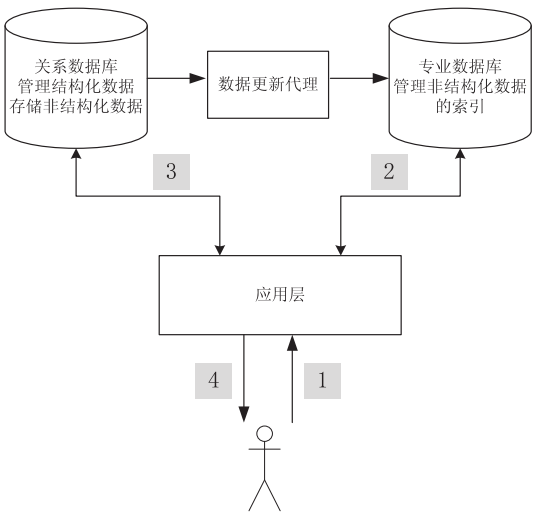

图 4　典型应用的系统结构示意图

以上所述实例仅表达了该技术的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对该技术范围的限制。应当指出的是,对于该领域的普通技术人员来说,在不脱离该技术构思的前提下,还可以做出若干变形和改进,这些都属于该技术的保护范围。因此,该技术保护范围应以所附权利要求为准。

对于电力企业,完成图 5 所示的数据整合即可。

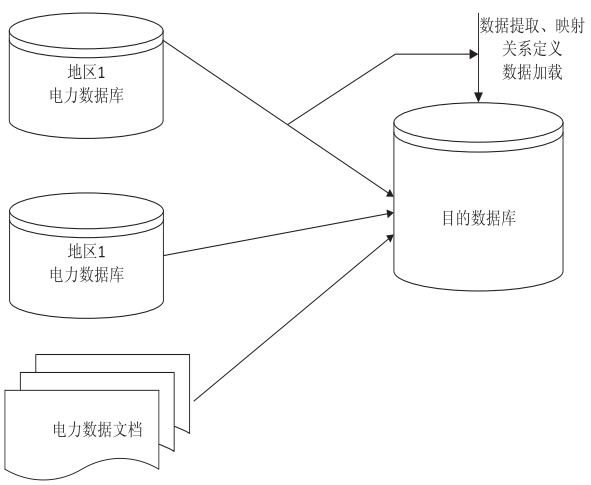

图 5　电力企业数据资源整合的整体结构图

3　电力企业异构数据资源透明访问设计概述

3.1　异构数据资源透明访问总体思路

对于电力企业异构数据资源共享和透明访问设计的总体思路是基于在服务器层的改进设计来完成的。具体来讲,在各个不同数据源工作的区域设计一个服务器层,用该层来屏蔽数据库底层的各不同资源的异构性,服务器层在屏蔽的同时还能为高层的访问提供标准的接口以便统一查询,在该层接口的设计上将其封装为标准的 Web 服务。用户在访问数据库时,调用 Web 服务,Web 服务可以为客户方提供一个多样、全面、统一的查询界面,在 Web 服务中,界面的设计灵活多变,可以满足用户的不同需求。基于该思路设计的资源透明访问机制有很多优势:

(1)不需建立庞大的数据库,且界面集成度高;

(2)便于以后的数据维护和索引;

(3)检索过程趋于简单快速且精确[8-11]。

3.2　异构数据源透明访问架构设计

用户方便、高效地使用资源数据库来对异构数据进行透明的访问取决于异构数据源透明访问的架构设计,对于同一领域或者不同领域采用不同的数据格式对数据资源进行解释说明,在资源进行利用、检索和描述时,使用不同的数据格式来描述这些功能的完成,这样在技术上产生了源数据的相互操作的问题。电力企业的信息系统在异构资源整合以后,为用户提供了统一、友好、高效的查询检索界面,在资源入口上得到了统一的接口访问[12-13]。

4　结束语

异构数据源的整合可以屏蔽各种数据结构的异构性,提供访问异构数据源的服务,不需要改变底层数据的存储和管理方式,即可实现分布异构数据的互操作。数据库名称和访问权限信息;配置数据集的维护,提取具体的数据表格或具体数据文件的指定;配置目的数据源,指定需要导入到哪个具体的目的数据库表中,映射关系,指定数据源中的数据导入到目的数据源的转化关系;根据所述映射关系,将所述数据集整合至所述目的数据源。在 B/S 架构的基础上把数据源的定义,数据集的提取,映射关系定义、目的数据源定义和数据加载的流程通过工作流的方式松耦合,灵活地解决了资源整合问题。在电力行业的数据撰写中提供了很好的支撑,在其他领域也具有很好的推广性。

参考文献:

[1]　张保军,刘高军. 基于 TMN 的电信网管数据集成研究与应

(下转第 179 页)

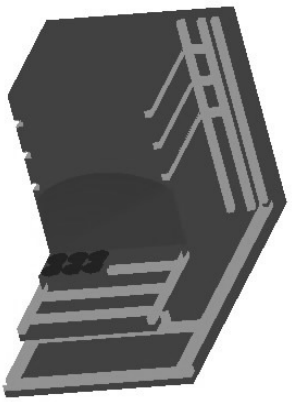

图 2　有底柱阶段自然崩落法三维模型

3　结束语

有底柱阶段自然崩落采矿法系统是由可视化编程语言 VB 来实现的,设计人员只需要输入相应的崩落法参数信息并读入预先生成的矿体信息文件就可以查看基于该矿体的有底柱阶段自然崩落采矿法的二维图形和三维模型的模拟图。但是,由于有底柱阶段自然崩落采矿法对矿体因素要求比较高,且矿体在开采过程中的参数比较复杂,因此,在参数的提取时还有待于进一步改进。

参考文献:

[1] 王家臣,王炳文. 金属矿床露天与地下开采[M]. 徐州:中国矿业大学出版社,2008:306-311.

[2] 陈国山,翁春林. 金属矿地下开采[M]. 北京:冶金工业出版社,2008:115-116.

[3] 王贺伟. 单层长壁式崩落采矿法系统管理与设计[J]. 信息技术,2011(4):172-173.

[4] 沈南山,顾晓春,尹升华. 国内外自然崩落采矿法技术现状[J]. 采矿技术,2009,9(4):1-4.

[5] 郭金峰. 我国地下矿山采矿方法的进展及发展趋势[J]. 金属矿山,2000(2):4-7.

[6] 王永光. 自然崩落法研究现状综述[J]. 有色矿山,1985(12):1-3.

[7] 谭光伟. 铜矿峪矿 5 号矿体自然崩落规律研究[J]. 有色金属:矿山部分,1997(5):8-12.

[8] 李翁然,周叔良. 自然崩落采矿法的发展现状[J]. 世界采矿快报,1996,12(3):12-15.

[9] 王福坤. 自然崩落法在中厚矿体中的应用研究[J]. 矿业研究与开发,1994,14(3):21-24.

[10] 袁海平,曹　平. 我国自然崩落法发展现状与应用展望[J]. 金属矿山,2004(8):25-28.

[11] 张树茂. 铜矿峪矿自然崩落法回采实践[J]. 金属矿山,2003(2):12-14.

[12] 陈国山. 地下采矿技术[M]. 北京:冶金工业出版社,2008:181-185.

[13] 张树兵,戴　红,陈　哲. Visual Basic 6.0 中文版入门与提高[M]. 北京:清华大学出版社,1999:100-107.

[14] 张　亮,杨　青,王　振. 基于采矿 CAD 的硐室参数绘图系统设计及实现[J]. 计算机技术与发展,2012,22(2):195-197.

[15] 王海燕,高江凤,杨金超. 井下爆炸材料库 CAD 系统的研究与实现[J]. 计算机技术与发展,2013,23(2):222-224.

[16] 高江凤,王海燕,杨金超. 基于采矿 CAD 沉井井筒支护系统的设计及实现[J]. 计算机技术与发展,2013,23(4):181-183.

[17] 张　岩,廖士中. 二维凸包 Graham 算法的设计与实现[J]. 牡丹江师范学院学报(自然科学版),1999(2):1-2.

[18] 程鹏飞,闫浩文,韩振辉. 一个求解多边形最小面积外接矩形的算法[J]. 工程图学学报,2008,29(1):122-126.

(上接第 175 页)

用[J]. 计算机技术与发展,2006,16(6):40-42.

[2] 李登元,胡素芳,周　毅. 基于异构数据源整合与集成的信息系统集成技术研究与应用[J]. 军械工程学院学报,2004,16(6):62-64.

[3] 余腊生,李　徐. 基于 Web 服务的跨网络异构数据交换技术[J]. 计算机应用,2005,25(12Z):9-11.

[4] 魏东平,潘向阳. 基于 XML 的异构数据的整合与集成模式探讨[J]. 内蒙古科技与经济,2005(5):87-88.

[5] McBrien P J, Poulovassilis A. Data integration by bi-directional schema transformation rules[C]//Proceedings of 19th international conference on data engineering. [s. l.]:[s. n.], 2013.

[6] Friedman M, Levy A, Millstein T. Navigational plans for data integration[C]//Proc of the 16th national conference on artificial intelligence. [s. l.]:[s. n.],1999.

[7] Ullman J D. Information integration using logical views[C]// Proc of the 6th international conference on database theory. [s. l.]:[s. n.],1997.

[8] Abiteboul S. Querying semi-structured data[C]//Lecture no-tes in computer science. New York:Springer-Verlag,1997:1-18.

[9] 张　岩,周明全,焦翠花. 网络科技资源中异构数据库访问技术的研究[J]. 计算机系统应用,2008(11):87-89.

[10] 杨金莹,刘明生. 基于 Web Services 的异构数据资源透明访问技术[J]. 石家庄铁道学院学报(自然科学版),2009,22(1):72-76.

[11] 王霓虹,张光磊. 基于 XML 的异构数据库集成的研究[J]. 信息技术,2006,30(5):173-176.

[12] 李宇翔,李端明. 电子商务中 XML 数据交换技术的应用研究[J]. 商场现代化,2008(28):55-56.

[13] 魏东平,潘向阳,孙东海,等. 基于 XMLSchema 的异构数据源集成技术研究[J]. 微计算机应用,2008,29(4):92-94.

基于异构数据资源整合的方法和系统实现

作者： 徐立新， XU Li-xin
作者单位： 广东电网电力科学研究院,广东 广州,510080
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(12)