

文章编号:1009-2722(2018)03-0021-05

分布式多源异构海洋地质 数据管理与发布

林 峰¹, 郭家树², 宋怀荣¹

(1 中国地质调查局青岛海洋地质研究所, 青岛 266701; 2 中国石油大学(华东)计算机与通信工程学院, 青岛 266555)

摘 要:多部门合作产生的海洋地质数据有着分布式、多源、异构的特点,在构建现实的海洋地质数据库系统过程中,通过解析海洋地质数据特征,在不改变原始数据的存储和管理方式下,提出一种基于统一数据模型的异构数据集成模式和基于 WCF 的数据与应用发布方式,从而实现多源异构数据统一、透明的集成访问,构造数据服务的基础。

关键词:海洋地质;分布式异构;数据整合

中图分类号:P628+.4

文献标识码:A

DOI:10.16028/j.1009-2722.2018.03004

多源异构空间数据的集成是空间信息领域重要的研究方向,是针对同一对象的多源异构空间数据进行集成,使之最终达成统一管理,集中服务于用户。从事海洋地质调查和研究的各单位和部门,在调查和研究活动中不断产生着大量的数据,各海洋地质部门在数据库建设和信息化应用上各自购置软硬件、独立研制、开发不同的数据库和地理信息应用系统,以满足各单位和部门特定需求。这样操作的结果,实际上是形成了部门内部相对独立的数据体系,数据库相互之间的可操作性差,从而产生信息的“孤岛”,不利于数据整合集成和提供统一服务。随着信息化的进程和大数据、云服务的需求,如何整合集成这些分散在各单位和部门之间的数据,成为亟待解决的关键问题。

在构建专业海洋地质数据共享与服务平台的过程中,笔者通过解析各部门的海洋地质数据特点和数据模型,寻求一种海洋地质空间地理信息的集成和发布方式,达到大数据的整合和真正共

享服务的目的。为此通过基于建立统一的数据逻辑模型,利用 Nhibernate 对逻辑模型到物理模型进行映射,最后达成异构数据的整合,并通过 WCF (Windows Communication Foundation) 实现了数据和应用的集成和发布。

1 研究现状

随着数据库技术和网络技术的快速发展,越来越多的数据库系统应用需要访问一些分布式的、异构的和自治的数据库。这些数据库在数据模型、查询语言、系统结构等方面存在差异,用户不能以统一的模式访问获取分布式异构数据,需要解决不同种类的数据管理系统之间的互操作问题^[1]。

从以往的数据库建设实践来看,数据库采取的是集中统一式的管理模式,即在数据中心构建一个统一的数据库系统,各种来源的数据在此提交汇总,从物理存储层面和逻辑层面,这个数据库是集中统一形式的。实际情况是牵涉到海洋地质数据的生产和研究部门,甚至各个项目,都各自建有独立的数据库系统,这些系统所用系统平台、软硬件设备乃至数据格式、存储模式等等,都存在

收稿日期:2018-01-08

基金项目:国家专项(GZH201100313)

作者简介:林 峰(1963—),男,硕士,研究员,主要从事数据库建设及地理信息系统方面的研究工作。E-mail:hsy2003@163.com

或多或少的差异,而且从网络结构上,各自是相互独立的,这就形成了各自为政的一个个“数据孤岛”的局面。

为解决“数据孤岛”的问题,国内对数据整合与集成技术进行了许多有益的探索。李军等^[2]对地球空间数据集成研究现状及存在的问题进行了详细的分析;赵辉等^[3]提出了一种基于XML为中间件的异构数据源集成的方法,并予以实现;李广建等^[4]在对国外143个整合系统进行统计研究的基础上,探讨数字整合系统的实现机制和每一种机制涉及的主要技术;唐健等^[5]以ETL技术框架为基础,提出国土资源数据的一个整合方案;王娟等^[6]针对数字化油田建设中异构数据源的信息提取问题,提出使用数据服务总线(DSB)技术,并给出了以数据总线为基础的异构数据源整合和集成技术方法。这些在各个领域中对于数据整合技术的探索和实践都相当程度地促进了多源异构数据集成技术的发展。

在海洋地质领域,从2010年开始的“数字海洋地质计划”其中一个重要目的,就是要为解决海洋地质“数据孤岛”提供一个切实可行的系统框架和整合服务平台。为此,项目组通过对海洋地质数据特点的剖析以及各单位各部门数据库建设的现状,并借鉴国内已经得到运用的数据整合技术和可利用的平台技术,提出了分布式多源异构海洋地质数据整合、管理与数据发布的一套方案。

2 分布式数据管理平台构架

分布式数据管理平台自下而上包括了4个逻辑层次:物理数据层、数据服务层、数据服务平台层和应用层(图1)。

(1)物理数据层 是实际的物理数据存储层。主要是已有的各类业务数据,通常这些数据都是存储在关系型数据库中,如Oracle、SQL Server、Access等关系数据库,以关系表、字段、记录等方式提供数据。

(2)数据服务层 数据服务层是整个系统的基础,主要解决异构数据的统一访问。以服务的形式提供业务数据查询,完成异构数据的统一访问,通过ORM将物理数据转换为统一的基于对象的模型,根据用户的请求返回数据结果。



图1 分布式数据管理平台的总体构架

Fig. 1 Framework of the distributed data management platform

(3)数据服务平台层 数据服务平台层是整个系统分布式的核心,主要完成数据的分布式访问。

(4)应用层 各种对数据进行访问的应用。该体系框架的设计主要考虑了系统设计的高内聚和低耦合的原则,应用层获取数据只需要通过数据服务平台层(也可直接访问数据服务层),无需关心数据的位置与物理存储,提交服务请求,数据服务平台层通过数据服务层从物理数据库层中获取数据并以一致的数据模型提供给应用对象。

从系统总体框架来看,分布式数据管理平台的总体构架直接针对目前面临的实际问题:可以禁止用户直接访问数据库的信息,可以有效地将业务应用层与物理数据库层解耦,同时基于Nhibernate,在实现了对象持久化同时,自动生成SQL,从而提高了开发效率。而利用WCF服务进行异构数据集成能够满足各种分布式信息资源之间的数据集成,已有的信息系统和数据能够在保持自治的基础上被充分利用起来,这样不仅节省了资源,而且也为信息化的不断发展提供了空间。

3 统一逻辑模型设计

3.1 逻辑模型构建

构造一个统一的数据模型是对多源异构数据进行集成发布的基础。在这方面借鉴了油田引进的国外 POSC 数据模型,但 POSC 模型本身是一套相当庞大复杂的体系,在应用当中保留了 POSC 核心的面向对象/业务流程,参照了 POSC

模型^[7]中对专业数据的相关定义和描述,在对海洋地质数据业务剖析的基础上,设计了统一数据模型需要的模型体系架构,以支持分布式异构数据的整合。

具体地说,通过分析海洋地质数据业务流和数据流,建立起模型标准、描述方法与规范;通过海洋地质数据模型标准的建设,建立起一个面向对象的、相对稳定的、可扩展的数据一体化逻辑模型(图 2)。根据不同的应用目的对逻辑模型进行投影生产不同的应用数据模型。

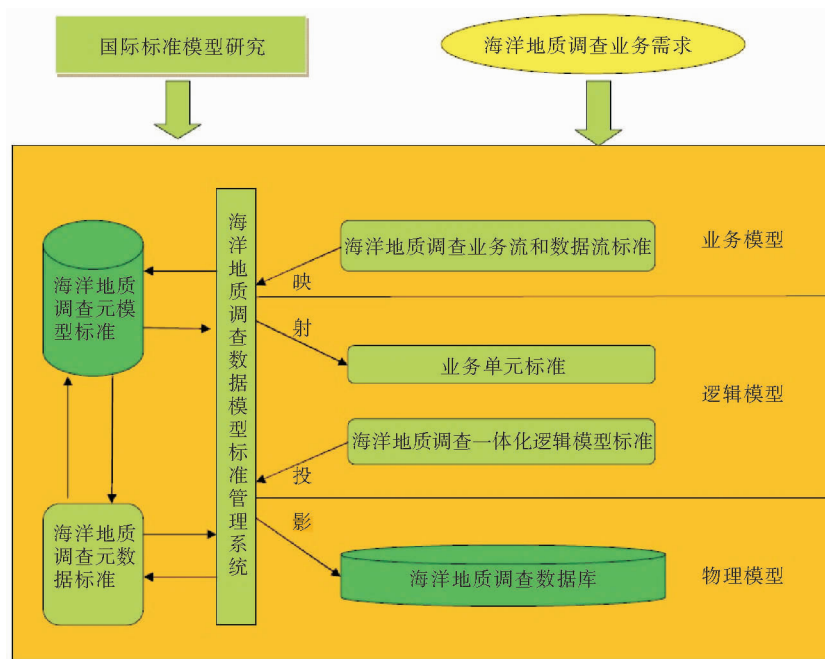


图 2 海洋地质数据模型体系架构

Fig. 2 Marine geological data model architecture

设计中根据海洋地质调查与研究的业务流程,对需要管理的数据对象按照业务域进行划分,业务域是从整体上针对海洋地质调查主要业务进行划分。它不是对现有架构的照搬,也不仅仅是基础业务整理,而是综合了海洋地质专业、海洋地质调查生命周期以及数据生产流程来划分数据标准的“业务域”,再细分出一级到三级业务。做到不同业务域间的业务不重复,并保证能覆盖所有的海洋地质调查和研究业务。

3.2 逻辑模型的映射

基于 .Net 框架下的 Nhibernate 关系/对象

映射模型来实现物理层的关系数据库到逻辑模型对象的映射,这种映射包含了逻辑模型中对象之间的一对一、一对多、多对多映射模式。

NHibernate 是一个基于 .Net 的针对关系型数据库的对象持久化类库,根据 Nhibernate 映射规则,可对逻辑模型中的统一对象与数据库中的关系数据进行映射^[8],在此以钻孔数据逻辑模型中的“钻孔基本信息”和“钻孔回次信息”为例,进行一对多映射,最终形成 XML 格式的映射文件。逻辑模型如图 3 所示。

为了成功实现对象与关系之间映射,必须建立映射类(*.CS)、映射文件(*.hbm.xml)以及

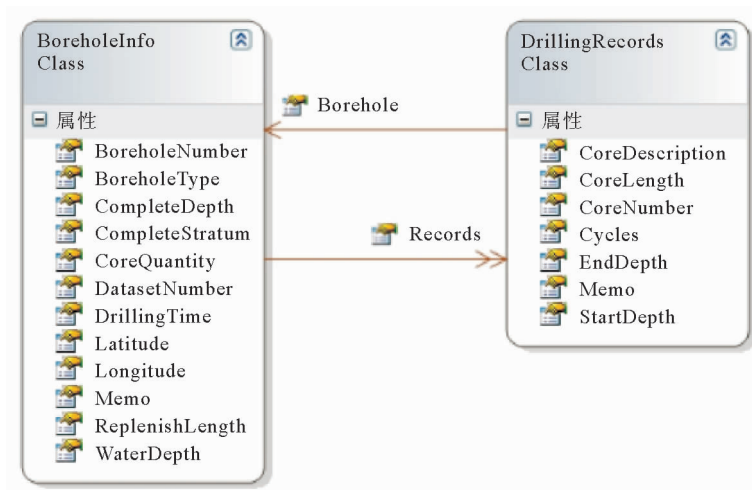


图3 逻辑模型的映射

Fig. 3 Mapping of logical models

数据库配置文件(*.dg.xml);映射类:包含业务属性的业务类,Nhibernate把表中的记录映射成为类的对象;映射文件:指定业务类和数据库表之间的关系,如业务类和数据库表的对应关系、表字段和类属性名称的对应关系等;配置文件:设置与数据库连接时需要的连接信息,比如数据库、用户名、密码以及连接字符串等。

在这3种主要的文件中,映射类为普通.NET源文件、映射文件为XML格式、数据库配置文件为XML格式。

4 基于WCF的数据服务的发布

WCF(Windows Communication Foundation)是Microsoft平台上的SOA架构,用于构建分布式和可交互操作的应用。笔者运用前面的统一数据模型,通过DataContract与DataMember属性将模型数据类(Class)定义为数据契约,逻辑模型中类的属性在实现时定义为C#的Virtual属性。通过get和set关键字实现属性的定义。

为了实现对逻辑模型简单、一致的数据访问接口,基于HQL(Hibernate Query Language)设计了统一的数据服务契约。基于泛型的Nhibernate通用查询方法SelectAll Hql查询语句,查询获取结果数据集合,通过TCP协议,对结果进行平台间的传递。

分装对象的跨平台传递是先将数据序列化,

将类对象转换为可保持或传输的格式;与序列化相对的是反序列化,它将流转换为对象。这两个过程结合起来,可以通过网络存储和传输数据,从而实现对象的跨机器传递。序列化有两种类型:①二进制序列化:完整的序列化整个对象,具有类型保真度,这对于在应用程序的不同调用之间保留对象的状态很有用。例如,通过将对象序列化为网络数据,可在不同机器的应用程序之间共享对象。可以将对象序列化到流、磁盘、内存和网络等等。远程处理使用序列化在计算机或应用程序域之间传递对象的“值”。②XML序列化:XML是一个开放式标准,可以按照这个标准将需要共享的数据序列化成XML通过Web传输。XML序列化仅针对公共属性和字段,且不保持类型保真度数据服务平台实现分布式数据访问和发布还需要提供对数据服务的管理,主要是实现对数据服务的注册、查找、更新的功能。我们对于数据和应用服务设计了统一的访问接口,通过访问接口接收缓存数据服务列表,实现对数据服务的管理,主要是服务位置的管理及服务URL地址的管理。

数据服务列表可以通过序列化保存到数据服务源配置文件dataProvider.config中,也可以从dataProvider.config中反序列化为服务列表。

多数据服务的分布式访问就简化成按照服务列表,数据服务平台对已经注册的数据服务中间件进行访问的过程。关键问题为数据服务中间件的动态的调用及查询结果的综合。系统结构如图

4 所示。

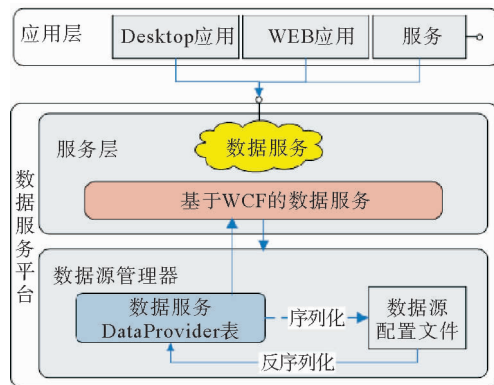


图 4 数据服务管理的方法

Fig. 4 Methods of data service management

5 结论

本文在 .net 平台上,通过建构统一的数据逻辑模型和利用 WCF 服务包装(数据和应用皆包装成服务),构建了一个多源异构空间数据集成模型,实现了多源异构空间数据的整合、序列化传递以及数据服务中间件的建构和管理,克服了传统数据库管理模式数据集成和发布的瓶颈。对

海洋地质多源异构空间数据的集成服务提供了解决方案。利用该技术可以使得在最大限度保持局部数据库目前状态不变的前提下,将海洋地质数据提交到一个逻辑上集中的数据库进行管理、并通过数据中心平台形成数据服务和应用。

参考文献:

- [1] Latwin W, Mark L, Roussopoulos N. Interoperability of Multiple Autonomous Databases [J]. ACM Computing Surveys, 1990, 22(3): 267-293.
- [2] 李 军, 费川云. 地球空间数据集成研究概况[J]. 地理科学进展, 2000, 19(3): 13-21.
- [3] 赵 辉, 陶世群, 李丽萍. 一种基于 XML 的异构数据源集成的方法与实现[J]. 山西大学学报: 自然科学版, 2005, 28(1): 33-36.
- [4] 李广建, 语 宇, 张 丽. 数字资源整合的实现机制及关键技术——对国外数字资源整合系统的实证研究[J]. 中国图书馆学报, 2007, 33(2): 75-80.
- [5] 唐 健, 沈陈华, 周国峰, 等. 国土资源数据整合方案设计及其实现研究[J]. 中国土地科学, 2009, 23(9): 72-78.
- [6] 王 娟, 梁鸿军, 李 良, 等. 油田数字化的异构数据源整合与集成技术[J]. 油气田地面工程, 2014, 33(11): 10-11.
- [7] 雷景生. 基于 POSC 平台的油气勘探数据仓库系统及其应用[J]. 计算机应用, 2003, 23(8): 90-92.
- [8] 杨雪梅, 董逸生, 王永利. 异构数据源集成中的模式映射技术[J]. 计算机科学, 2006, 33(7): 87-91.

MANAGEMENT AND SERVICE OF DISTRIBUTED MULTI SOURCE HETEROGENEOUS MARINE GEOLOGICAL DATA

LIN Feng¹, GUO Jiashu², SONG Huairong¹

(1 Qingdao Institute of Marine Geology, CGS, Qingdao 266701, China;

2 College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao 266555, China)

Abstract: Marine geology data produced by multiple channels are usually distributed, multi-sourced and heterogeneous. In the process of constructing a realistic marine geological database system, we noticed the characteristics of marine geological data, and then proposed a heterogeneous data integration mode based on a unified data model and the way of data and application service based on WCF. The unified logical data model is mainly made with reference to the POSC model which was generally used in the field of oil and gas. In the WCF service architecture, data and applications are packaged into class objects and transferred by object serialization and deserialization. This form of system design has the advantages of on unchanged storage and management mode of original data.

Key words: marine geology; distributed isomerism; data integration