地学空间数据仓库的构建技术

王永志[1],高光大[2],杨毅恒[3],陈 苗[4]
WANG Yong-zhi[1], GAO Guang-da[2], YANG Yi-heng[3], CHEN Miao[4]

1. 吉林大学仪器科学与电气工程学院,教育部地球探测重点实验室,吉林 长春 130026;
2. 中国地质大学 计算机学院,北京 100083; 3. 北京信息科技大学理学院,北京 100083;
4. 中国地质科学院矿产资源研究所,北京 100037
1. *College of Instrumentation Science and Electrical Engineering, Jilin University, Key Laboratory of Earth Exploration of Ministry of Education, Changchun* 130026, *Jilin, China*;
2. *College of Computer, China University of Geosciences, Beijing* 100083, *China*;
3. *College of Science, Beijing Information Science and Technology University, Beijing* 100083, *China*;
4. *Institute of Mineral Resources, Chinese Academy of Geological Sciences, Beijing* 100037, *China*

摘要:为了将中国多源、异构、分散的地学数据集中到一起,为资源评价提供有效的数据供应,将地学空间数据仓库作为实现数据集成的解决方案。首次提出了符合中国国情的具有数据源、空间ETL、空间数据存储、基于SOA的应用服务和客户应用的5层地学空间数据仓库的体系结构。根据中国地质行业行政划分和数据的分布情况,设计了能够实现地学数据集成的国家、大区所、省三级管理的地学空间数据仓库系统的物理部署方案。这是一套符合中国地学实际情况且完整可行的地学数据集成方案。

关键词:地学空间数据仓库;数据集市;空间数据抽取、转换和集成;面向服务的体系结构
中图分类号:P5 **文献标志码**:A **文章编号**:1671-2552(2008)05-0713-06

Wang Y Z, Gao G D, Yang Y H, Chen M. Technology for the construction of the geoscience spatial data warehouse. *Geological Bulletin of China*, 2008, 27(5):713-718

Abstract: The authors took the geoscience spatial data warehouse as a scheme of data integration in order to integrate multi-source, heterogeneous and disperse geological data of China and provide effective data for resource assessment. They for the first time present a geoscience spatial data warehouse architecture that conforms to China's national conditions and have five levels, i.e. the data source, spatial ETL, spatial data storage, application service based on SOA and client application. The authors designed a three-level (state, administrative regions and provinces) physical deployment scheme for the geoscience spatial data warehouse system according to the administration regions of China's geological work and distribution of data. It can realize the objectives of geoscience data integration. Research results show that this is a complete and feasible geoscience data integration scheme that conforms to the actual situation of geoscience of China.

Key words: geoscience spatial data warehouse; data mart; spatial data extract, transfer and integration; Service-Oriented Architecture

地质调查和战略性矿产资源勘查是国家的基础性、公益性工作。多年来积累了大量的全国性空间和非空间数据,主要包括区域、遥感、水文、环境、海洋、农业、城市、灾害等地质调查数据和矿产资源、地球物理、地球化学等地质勘查数据,以及地质成果资料数据、境外地质矿产资源数据等,共有9大类60多种

收稿日期:2007-10-24;**修订日期**:2008-03-07
地调项目:国土资源部地质调查重大专项(编号:1212010633901)、金土工程子项目(编号:JTXM-DW-KZ4)资助。
作者简介:王永志(1974-),男,在读博士,讲师,从事分布式计算、地学数据仓库和空间数据挖掘技术研究。
E-mail:iamwangyongzhi@126.com

数据库,数据总容量近5TB。为了有效地利用已有的地学数据进行矿产资源潜力评价、矿产资源战略勘查、矿产综合利用与开发等各种空间和非空间的数据分析或联机分析处理,甚至通过地学空间数据挖掘,从海量地学数据中寻找有用的、有价值的模式,必须保证数据来源的质量。由于地学数据具有空间性、海量性、多源性、异构性、分散性等特点,中国地学数据目前尚未实现真正意义上的无缝集中或分布式存储,迫切需要构建一个科学、规范、专门管理地学空间数据、为地学分析与决策提供数据源的地学空间数据仓库系统(GSDWS—Geoscience Spatial Data Warehouse System),并开发相应的管理系统,彻底解决地学领域的"信息孤岛"问题。

数据仓库是一个面向主题的、集成的、非易失的、随时间变化的、用来支持管理人员决策的数据集合[1]。它是近年来为OLAP分析、数据挖掘等处理提供海量数据存储、数据组织的容器和解决数据集成问题的关键技术[2]。

本文结合地学数据的特性和国家矿产预测的应用需求,首次提出将数据仓库技术与GIS技术、地学数据集成与共享相结合,创造性地提出构建适合中国国情的地学空间数据仓库体系结构(GSDWA—Geoscience Spatial Data Warehouse Architecture)和物理逻辑结构,它可以将多个异种源的地学数据融合,产生高质量的规范化的数据,能很好地满足国家、大区所、省三级管理体系中不同层次的地学数据管理、地学分析和战略决策对数据环境的需求,为地学数据一体化集成与共享、地学空间数据挖掘、地学空间数据一站式服务(Geospatial One-Stop Service System)等进行了数据和技术上的支撑准备。

1 地学空间数据仓库

1.1 地学空间数据仓库的体系结构

地学空间数据仓库是面向主题的、集成的、时变的、相对稳定的、海量的地学空间数据和非空间数据的集合,是地学空间数据决策支持的基础平台[3]。地学空间数据仓库可以将多个异种的、自治的、分布的信息源有机地组织起来,采用分布式空间数据存储对象概念进行存储,并提供对空间和非空间数据简便、有效的访问[4-5]。

地学空间数据仓库从逻辑上可以划分为数据源、空间数据转换、地学空间数据存储、空间应用(计算)服务、前端空间数据分析工具5层结构(图1)。它对于用户来讲就是一个大的数据仓库,实际上是由多个分布式的地学数据集市(Data Mart)和数据仓库组成的一个虚拟的海量数据源。

地学空间数据仓库体系结构的特点主要有:(1)经过ETL处理并存储在数据仓库中的数据与数据源无关;(2)数据模型规范、统一;(3)数据存储管理集中;(4)数据存储和地学空间数据处理之间松耦合;(5)数据含义与表达分离,故同一内容可以用文字、表格、图形、图像等多种形式呈现;(6)系统应用基于SOA构建,故应用扩展性和组件重用性好;(7)使用存储网格和应用网格技术,所以数据存储能力和承担地学计算能力扩充性强;(8)存储、地学计算和应用均是位置透明的;(9)编程语言无关、数据库无关、平台无关;(10)数据库、应用程序服务器、实现技术等均是目前世界上最流行和最成熟的;(11)可以与一站式服务和数据挖掘有机地结合。

1.2 数据源

数据源是地学空间数据仓库系统的数据源泉。这些数据均来自地质单位应用系统产生的原始空间数据、属性数据、地质成果报告等。数据的存储格式有GIS(MapGIS、ArcGIS等)、XML、文本文件、电子表格、关系或关系-对象数据库(Access、Oracle、SQL Server、DB2等)、栅格影像等。

1.3 空间数据转换层

空间数据转换层(Spatial ETL)是保证地学空间数据仓库数据质量、数据规范和标准化的关键环节。通过对操作型的地学数据进行数据整理和清洗、数据变换、数据集成、数据装载和定期的数据刷新来构造地学空间数据仓库。将数据源中要保存到数据仓库中的数据抽取出来并临时存放在数据准备区里。在数据准备区中进行地学数据的清理、转换、一体化,将净化后的标准地学数据装载到地学数据仓库的数据库中,若有新的地学数据要追加到数据仓库中,可以执行刷新操作[6]。为达到转换前后数据的保真和关联关系不变,采用空间数据转换工具,通过JDBC、ArcSDE、ADO.NET或ODBC等数据库连接引擎、按照转换规则完成ETL过程。采用的工具有ArcGIS SDE、ArcCatalog、Oracle MapBuilder、shp2sdo、Oracle Data Integrator等,以及自编的空间ETL转换工具。采用C/S、B/S两种模式分期分批进行。

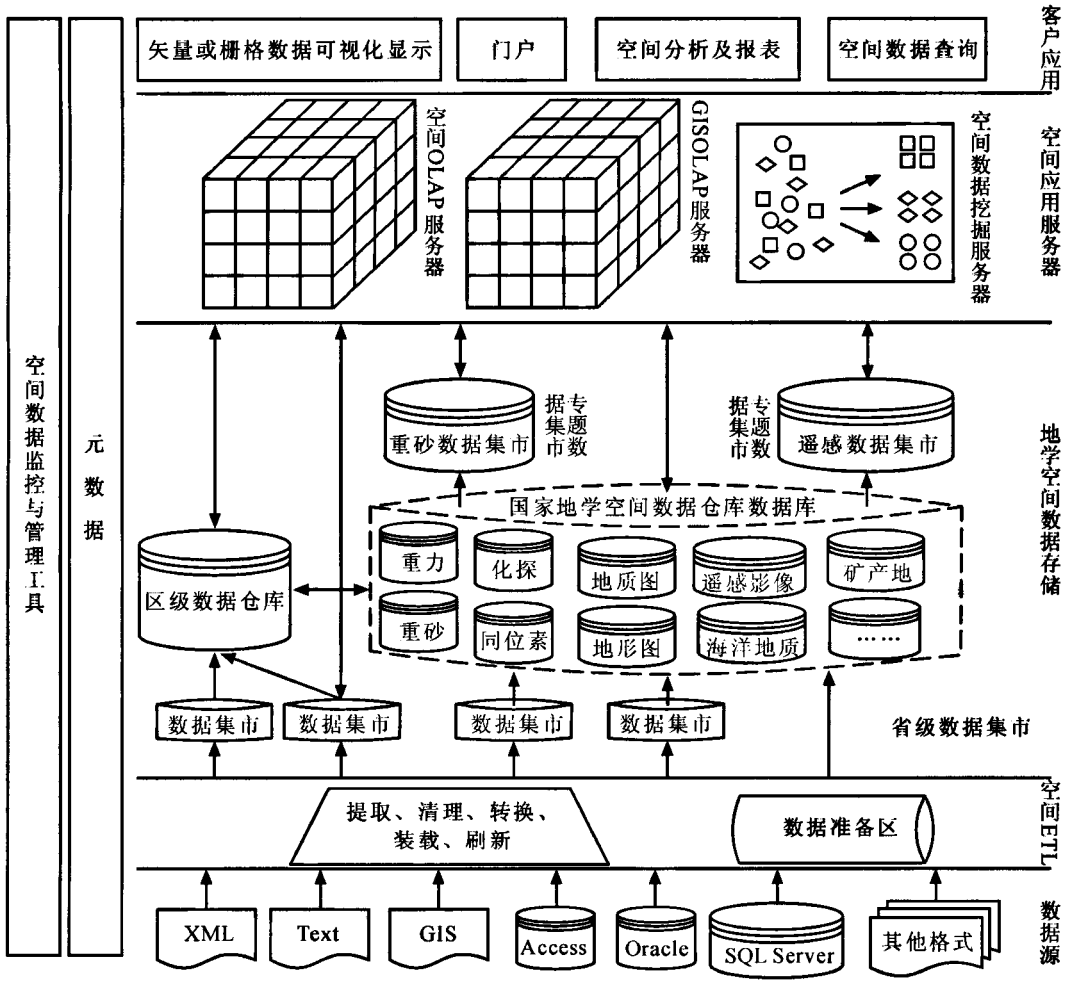

图1　地学空间数据仓库的体系结构

Fig.1　Architecture of the geoscience spatial data warehouse

1.4　数据仓库存储层

数据仓库存储层是整个地学空间数据仓库系统的核心部分，在实际存储时采用分布式和集中式共存模式,对用户是位置无关和透明的,就像一个可以源源不断供应各种地学空间数据的超级虚拟数据仓库存储网格，可以从根本上满足地学海量数据存储的需求。数据存储有数据仓库和数据集市2种模式，在省级采用数据集市的形式存储地学数据，在区级和国家级则采取数据仓库来存储数据。操作型地学数据经过空间ETL操作之后可以直接进入各省级的数据集市进行存储，也可以在基于角色的权限管理范围内和业务规定的情况下直接上载到国家级的数据仓库中[7]。省级数据集市的数据可以在符合权限和业务规则的前提下上载到国家级数据仓库中,也可以直接上载到区级数据仓库中，再从区级数据仓库上载到国家级数据仓库中。数据仓库存储层为应用层提供的数据可以从国家级数据仓库、国家级数据仓库形成的专题性数据集市、区级数据仓库、省级数据集市通过数据访问引擎提取数据。

地学数据仓库中的数据围绕重力、化探、区域地质等大的主题进行存储，每个主题数据又按矿种、储量、矿床、岩石等细粒度的主题进行组织。数据仓库中存储相关主题不同粒度的汇总数据(如基础数据、中间信息数据、最终成果数据等)。另外,由于地质成果报告多以Word、Excel等非结构形式存在，且数据量也较大，故采用支持全文搜索的Oracle内容数据库(OCD-Oracle Content Database)对地质成果进行统一管理[8],并且将成果数据与空

间、属性数据相关联。

存储数据库采用业界公认的、支持空间数据存储和计算的、基于标准的对象-关系数据库Oracle10g数据库，尤其突出使用Oracle10g数据库的关系-对象特性和支持GIS应用程序的高级空间处理能力、基于位置服务和企业级空间信息系统的Spatial选件。属性数据采用关系-对象特性存储；存储点、线、面矢量格式的地学空间数据采用Oracle Spatial的SDO_GEOMETRY类型，对于缓冲区、距离等计算使用Oracle的空间计算函数；采用GeoRaster进行栅格数据管理[9-11]。在设计中充分使用Oracle数据库的继承、封装等面向对象的特性。国家地质数字中心的数据仓库在存储时(内容数据仓库、空间数据仓库)采用了网格存储技术，实现时使用表分区(Table Partition)、空间索引(Spatial Index)、网络数据模型(Network Data Mode)、拓扑数据模型(Topology Data Model)等[12]。

1.5 应用服务层

应用服务层是进行地学数据的具体处理，调用各种空间OLAP、GISOLAP和空间数据挖掘等计算算法和函数，这些算法和函数通过JDBC、ADO.NET等数据访问引擎从地学数据仓库中分期分批地提取要处理的多维数据并进行具体计算，将计算结果以XML形式返给客户。本层采用世界上目前最流行的面向服务体系结构SOA(Service-Oriented Architecture)框架进行搭建[13-14]，部署的组件是封装了地图服务、地学空间查询、OLAP分析、聚类分析、因子分析等数据挖掘算法的Web服务组件。这些组件可以集中存储在国家级数据仓库应用程序服务器上，也可以注册到集中应用程序服务器上，而实际计算由分布在世界各地的Web组件协同完成。由于地学空间数据处理要比常规商务数据处理复杂得多，而且海量数据在Internet网上传输也不实际，故尽可能少地产生中间数据，将处理交给本地自治服务器和数据存储交互。为了有效地保证计算速度，在这一层采用了Web缓存技术、网格计算和均衡负载技术以提升计算质量，但这些对客户均是透明的，客户只是觉得有一个具有超级空间计算能力的Web服务器。由于Oracle数据库将OLAP和数据挖掘、数据存储无缝地集成在一起，故从性能、安全、管理等方面达到了理想的效果。组件部署和应用程序服务器使用Oracle SOA套件，地图显示服务使用ArcGIS Server和专门显示Oracle空间数据的Oracle MapViewer[15]，发布影像使用ArcGIS IMS Server。

1.6 前端应用层

前端应用层向集成了空间OLAP或空间挖掘等组件的应用服务层发送基于HTTP协议、XML格式的空间查询及分析的请求，并接收和显示计算返回的结果。主要有各种报表、查询、数据分析、门户网站、数据挖掘工具、矢量、栅格格式的地学数据可视化。

1.7 元数据管理

地学空间数据仓库中的元数据管理采用基于Oracle10g的XMLDB格式存储的地学空间公共仓库元模型GSCWM (Geology Spatial Common Warehouse Metamodel)[16]，它是地学数据仓库体系中的关键组件，贯穿于整个地学数据仓库系统的设计、开发、运行、应用和维护的全过程，包括数据仓库结构的描述、操作元数据、汇总算法、ETL规则、与性能有关的数据等。保存在国家级地学元数据服务器上，主要包括联机事务处理OLTP(On Line Transaction Process)数据的元数据、地学空间数据仓库的元数据、空间ETL元数据、地学空间数据集市元数据、地学空间OLAP元数据和空间数据挖掘元数据。采用模型形式化方法，并采取适配器原理设计访问模式[17]。

1.8 监控及空间管理工具

监控工具主要对各个地学数据仓库、地学数据集市、应用服务器的运行状态进行远程监视、分析和管理。为保证地质数据的机密性，数据访问是基于严格的角色授权管理的。空间数据管理工具是管理空间数据仓库和数据集市的数据ETL、定期刷新、查询、服务管理、备份的管理系统，为面向一般决策过程的数据仓库服务[18]。

2 地学空间数据仓库的物理结构

从国家地质行业行政划分和数据处理的实际需要出发，地学空间数据仓库的物理结构采用集中式和分布式2种模型，总体部署策略采用省、区、国家三级同时存在的结构，每一级可以有基础数据、中间信息数据和最终成果数据(图2)。在每一处均按企业级架构进行部署，即数据(空间和属性数据、成果文件等)保存在数据服务存储层、组件和服务部署在Web服务组件服务器上、Web应用部署在Web应用程序服务器上。客户均可以通过Internet/Intranet访问共

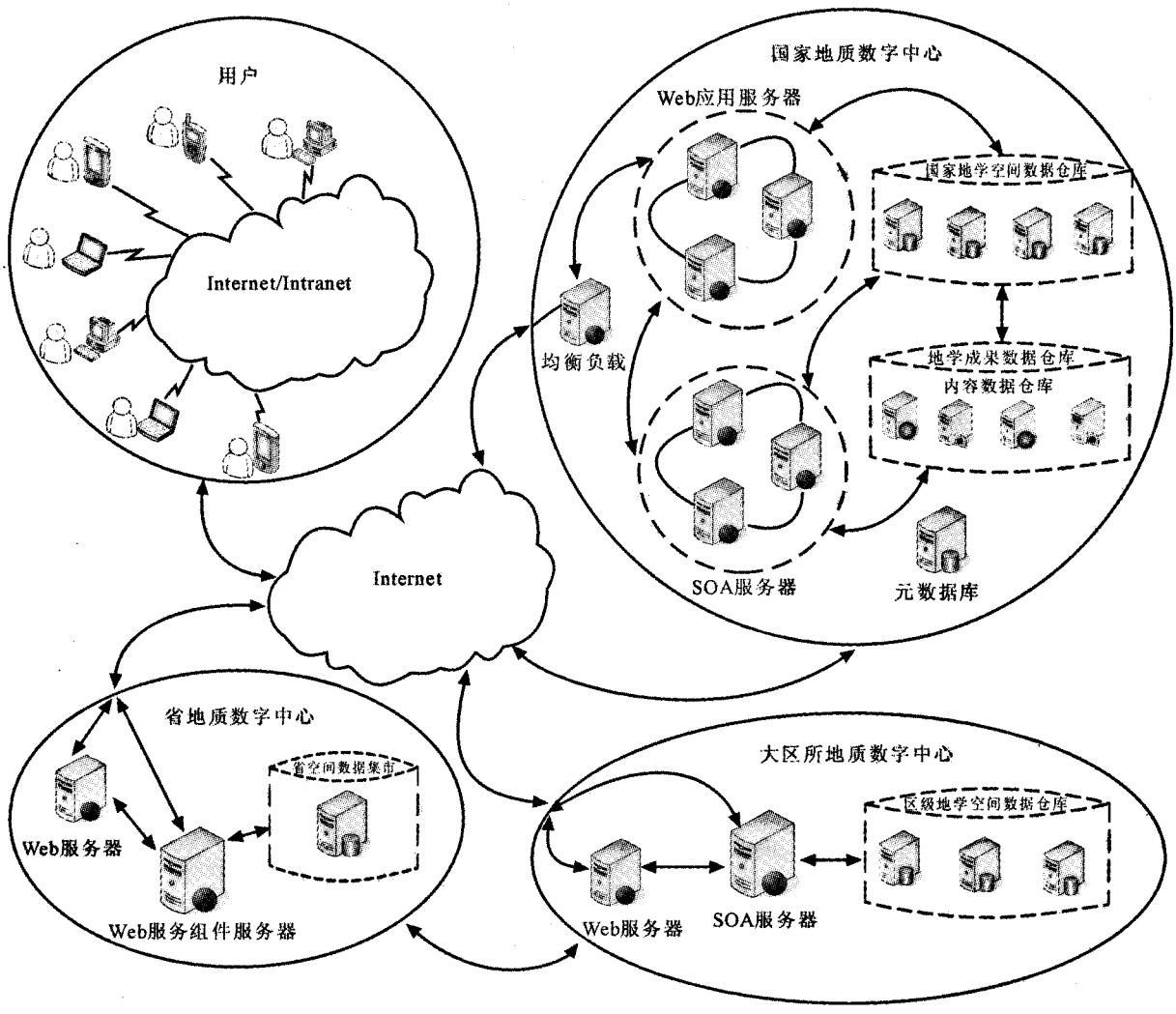

图2　国家地学空间数据仓库系统的物理逻辑结构

Fig.2　Physical logic structure of the national geoscience spatial data warehouse system of China

享的服务去操作数据仓库中的多维数据。

地学空间数据集市、数据仓库三级存储的访问对于客户是位置透明的。客户的请求将由国家、区、省地学数字中心的Web服务组件与相应的地学空间数据仓库/集市中的数据交互处理得出。

省级地质数字中心有Web应用服务器、Web服务组件服务器、数据集市服务器(保存本地元数据);区级地质数字中心具有Web应用服务器、SOA服务器、区级数据仓库服务器(保存本地元数据);国家地质数字中心具有均衡负载器、Web应用服务器集群、SOA服务器集群、属性和空间数据仓库服务器集群、保存地质成果的内容数据仓库集群、元数据库等。省、地区、国家地质数字中心通过Internet互相连通,而数据也通过Web服务器在权限允许的范围内进行网络ETL或通过Web服务组件访问,数据和服务器的管理采用Oracle的企业级网格管理系统进行统一管理。这样各自的数据既保持地方自治,又能通过SOA组成一个大的、具有超级计算能力的、存储海量数据的虚拟网格计算服务和空间数据仓库。

3　结论与讨论

(1)构建了具有数据源、空间数据转换、地学空间数据存储、空间应用(计算)服务、前端空间数据分析工具5层结构的地学空间数据仓库体系结构,提出

了实现地质领域真正意义上的数据集成与共享、一体化存储的解决方案，是结束地学数据长期分而治之的局面和抢救国家地质信息财富的一种有效的技术手段。

(2)设计了符合中国地学管理实际的省、区、国家三级交互的物理逻辑部署方案。

(3)实质性的数据整合可为地学计算服务共享、地学空间分析及地学空间数据挖掘打下良好的数据基础。

为了保证地学数据处理和服务的高质和可持续性，应在开发模式、服务安全策略、空间访问并发控制、响应性能优化、流程规范化等方面进行深入的研究与实践。

致谢：感谢中国地质调查局发展研究中心李景朝教授提供项目支持，并对文章内容给予详细指导。

参考文献：

[1]William H I. Building the Data Warehouse [M]. Indianapolis:Wiley Publishing,2005:29-33.

[2]王晓明，高勇，刘玉玲.面向水环境管理的空间数据仓库构建[J].计算机应用研究，2005，11：195-197.

[3]李德仁，王树良，李德毅.空间数据挖掘理论与应用[M].北京：科学出版社，2006：128-132.

[4]朱庆，周艳.分布式空间数据存储对象[J].武汉大学学报(信息科学版)，2006，31(5)：391-394.

[5]崔晓军，薛永生.数据仓库集成环境研究与实现[J].计算机应用研究，2006，12：178-184.

[6]Paul L. Oracle Data Warehousing Guide 10g Release 2[OL]. Oracle Corporation, http://download.oracle.com/docs/cd/B19306_01/server.102/b14223/toc.htm,2005.

[7]李海波，王丽珍，杨莉.统一空间数据仓库权限管理分析与设计[J].计算机工程，2006，32(19)：54-57.

[8]Marla A. Oracle Content Database Administrator's Guide[OL]. Oracle Corporation,http://download.oracle.com/docs/cd/B32119_01/doc/contentdb.1012/b31268.pdf,2006.

[9]Chuck M. Oracle Spatial User's Guide and Reference 10g Release 1 [OL]. Oracle Corporation,http://download.oracle.com/docs/html/B10826_01/toc.htm,2003.

[10]姚力波，王仁礼.基于Oracle Spatial空间数据库的GIS数据管理[J].测绘与空间地理信息，2006，29(2)：82-86.

[11]Chuck M. Oracle Spatial GeoRaster[OL]. Oracle Corporation,http://download.oracle.com/docs/pdf/B14254_01.pdf,2005.

[12]韦波，李景文.基于Oracle 10g拓扑数据模型的空间管网信息系统WebGIS实现[J].计算机应用，2006，26：115-116,121.

[13]Ken C, Orlando C. Oracle SOA Suite Developer's Guide[OL],Oracle Corporation,http://download.oracle.com/docs/cd/B31017_01/core.1013/b28764.pdf,2006.

[14]魏东，陈晓江，房鼎益.基于SOA体系结构的软件开发方法研究[J].微电子学与计算机，2005，22(6):73-75.

[15]Chuck M. Oracle Application Server MapViewer User's Guide [OL]. Oracle Corporation,http://download-west.oracle.com/docs/cd/B14099_19/web.1012/b14036/toc.htm,2005.

[16]John P,Dan C,Douglas T. Common Warehouse Metamodel Developer's Guide[M].Indianapolis:Wiley Publishing,2003：431-554.

[17]赵晓非，黄志球.基于CWM的元数据集成中形式化推理技术的研究[J].计算机科学，2006，33(12):177-182.

[18]王宁，王延章，于森.面向一般决策过程的数据仓库系统研究[J].计算机集成制造系统，2006，12(1):139-143.

[19]邹逸江.多维空间分析的关键技术——空间数据立方体[J].地理与地理信息科学，2006，22(1)：13-16.

[20]林杰斌，刘明德，陈湘.数据挖掘与OLAP理论与实务[M].北京：清华大学出版社，2003：24-31.