海洋环境数据在线服务系统设计与实现

宋晓[1,2]，姜晓轶[1,2]，韩璐遥[1,2]，王漪[1,2]

（1.国家海洋信息中心　天津　300171；2.国家海洋局数字海洋科学技术重点实验室　天津　300171）

摘要：文章针对海洋数据信息孤岛、服务对象单一等问题，设计了海洋环境数据在线服务系统。采用并行数据库技术实现数据的快速组织整理、查询检索，解决数据高耦合、高冗余等问题，提高对数据的快速检索能力。利用虚拟化技术完成存储、服务器、网络设备资源整合与集群建设，解决物理设备分布散乱、利用率低的问题，提高资源利用率，节约成本。利用域控管理机制，实现系统信息安全与数据在线服务。

关键词：虚拟技术；并行数据库；海洋环境数据

中图分类号：P717；TP391　　**文献标志码**：A　　**文章编号**：1005－9857(2015)04－0019－05

1　引言

海洋环境数据在线服务系统建设是集海洋科学、地理信息系统与计算机科学的综合性技术。由于服务对象与服务目的不尽相同，各单位和科研院所建设了许多海洋环境数据和应用系统，这些系统之间不可避免地造成了数据冗余和资源浪费，同时也导致信息孤岛和重复建设等问题，不利于海洋数据的共享与服务[1－4]。

传统的海洋数据服务一般是专项专建、专人专用，针对人群比较单一、数据类型比较简单，而且在项目结束后通常建设的数据库和应用系统由于没有后期的经费支撑而停用。本研究提出的海洋环境数据在线服务系统（以下简称系统），涵盖了多专项、多学科的数据，在原有数据库、应用系统、专网基础上进行系统集成、数据库扩建，为海洋局属各单位提供共享服务。

系统是运用面向服务架构的设计思想搭建应用系统。采用并行数据库技术实现大数据量的存储、加载、更新、查询等操作，利用 ETL 调度工具实现源数据库到并行数据库的数据抽取、转换和转载，减少重新建库的工作量。采用虚拟化技术整合存储、服务器、网络资源，建设数据中心集群，提高资源利用率，采用域控管理机制实现数据安全管理，权限管理。采用 VPN 认证管理机制，保障系统安全正常运转。

2　系统设计

系统通过面向服务的总体架构，以数据的汇集、处理、应用为基础主线，采用高速并行技术，结合虚拟化技术等先进 IT 技术，设计系统的逻辑架构、功能架构、物理架构与技术架构。

2.1　逻辑架构

系统总体架框架由数据层、管理层和应用层 3 部分构成，数据层是指通过对历史收集、专项调查、在线传输等方式收集，采用数据集、数据库方式进行数据存储与管理；管理层是指对使用系统的用户进行统一认证、用户管理、数据授权等实现用户有效可控的管理；应用层是指为用户提供数据的在线查询检索、数据时空分布检索、产品加工处理等应用服务，满足用户多样化的需求。应用层与管理层通过内网和专网访问数据层，实现数据的管理、查询、处理等服务。系统总体逻辑框架如图 1 所示。

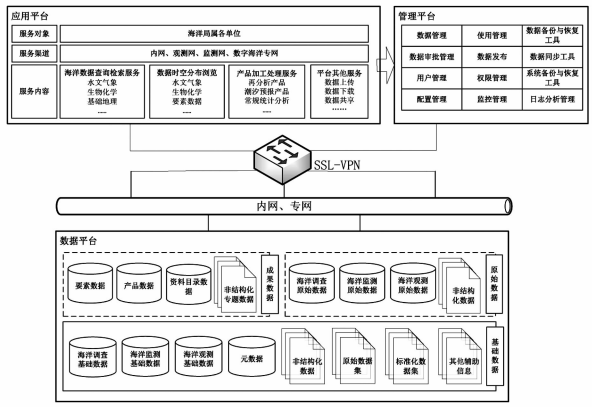

图 1　系统总体逻辑框结构

（1）数据层。数据层主要由原始数据、基础

数据和成果数据 3 部分组成。

原始数据是指海洋仪器现场采集的原始资料、现场汇交的纸质或者电子资料等。原始数据采用文件方式存储，基于原始资料清单和数据库文件目录等方式进行管理。类型包括海洋观测原始资料、海洋监测原始资料、专项调查汇交资料等。

基础数据是指对原始数据进行整理、排重、质量控制等处理之后形成的标准化数据。内容主要包括专项调查数据、观测实时资料数据与国际业务化数据等，专项调查数据包括水文、气象等 9 个学科，观测实时资料数据包括海洋站、雷达、浮标等。基础数据采用数据库存储方式，根据基础数据的资料类型、资料格式、数据观测频率、数据传输频率、数据量等设计数据库结构。

成果数据是指经过信息提取、多源数据融合、数值模型分析、统计分析等手段处理后形成的数据。成果数据由要素数据、成果专题数据、资料目录数据组成，采用数据库存储方式。要素数据是以基础数据为基础，根据数据的专题应用保障和服务需求，按照时间、空间、专题要素等进行组织的数据。成果专题数据主要包括数值型产品和图形产品，涵盖海洋再分析产品、实况分析产品、潮汐预报产品和海洋专题产品等。资料目录数据主要包括原始数据集目录索引、标准数据集目录索引、产品数据目录索引等。

(2)管理层。管理层主要负责系统的用户管理、资源管理、业务流程管理和运行监控管理等内容。用户管理包括用户的创建、更改和删除、角色管理、功能授权与数据授权；资源管理包括目录索引管理、数据导航管理、信息发布管理与信息资源管理；业务流程管理包括数据申请、虚拟机管理、数据审批管理等；运行监控管理包括运行环境监控、数据资源监控与用户行为监控。

(3)应用层。应用层依托于中心内网和海洋专网，基于并行数据库技术和虚拟化技术，实现海洋局属单位间的数据在线服务。应用层主要包括：数据时空分布展示、数据查询检索服务、数据共享虚拟环境、产品制作与产品导出功能。

数据时空分布展示是利用数据的经纬度、时间范围、站次数等关键信息，通过统计计算数据量，依据色彩图例，进行时空分布展示。

数据查询检索服务包括数据库查询检索和数据集查询检索。该服务可提供基于矢量地图及影像地图的地图显示控件的数据查询服务，以及使用关键字对数据进行查询。

产品制作是指对资料进行整理、标准化处理，开展数据识别、解码等预处理操作，利用数据统计分析工具进行产品的加工制作。

产品导出是指对用户加工制作产生的产品成果提供数据的导出功能，实现数据从虚拟机到本机的导出服务。

2.2 物理架构

按照系统设计，对系统运行硬件环境进行搭建，硬件环境涵盖原始数据文件存储区、数据库存储区、数据处理区、数据服务区。按照网络布局可化为中心内网和海洋专网，内网为中心内部用户提供在线服务的入口，专网主要包括海洋观测网、海洋监测网、数字海洋网；数字海洋网为海洋局属单位提供在线服务的入口，用户经由内网/数字海洋网通过 VPN 身份认证后方可进入用户主页，通过登录进入个人虚拟工作环境(即用户虚拟机)，用户可在虚拟机中对数据进行查询、处理和产品制作(图 2)。

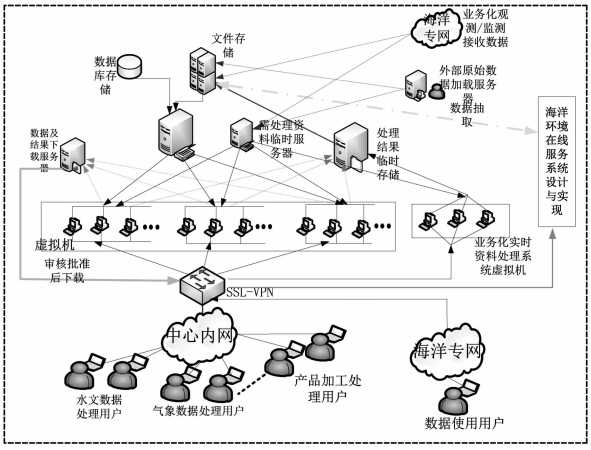

图 2 系统物理结构

系统经由海洋观测网和海洋监测网接收实时、延时观测和监测的海洋数据，并发送到系统的文件存储区和处理资料临时存储区，由存储管理系统进行数据的接收、存储和管理。利用用户授权管理将数据分发到数据处理用户的虚拟机中。数据处理用户通过中心内网登录到虚拟机后，开展数据整理、标准化处理工作后，将处理结果按照指定的

路径存放。由数据传输系统同步传输到产品制作用户的虚拟机中,用户可开展产品加工制作并将成果按照指定的路径存放。最终由数据交换系统存储到统一的资料存储管理区。ETL 处理系统经过数据抽取、清洗、转换等处理,将数据处理结果和产品加载入库,最终经由中心内网和海洋专网为海洋局属单位提供数据共享服务。

3　系统功能实现

系统通过用户唯一入口登录,保证数据安全;开发数据处理系统,完成数据格式化转换;利用 ETL 处理系统,完成并行数据库的数据处理与调度,包括数据抽取、数据转换与清洗及数据加载;开发数据库检索、数据集检索、文件输出审批和文件导出等应用程序;开发系统运行监控管理系统,对系统的运行环境、数据状况和用户行为进行监控和管理。系统主要功能模块如图 3 所示。

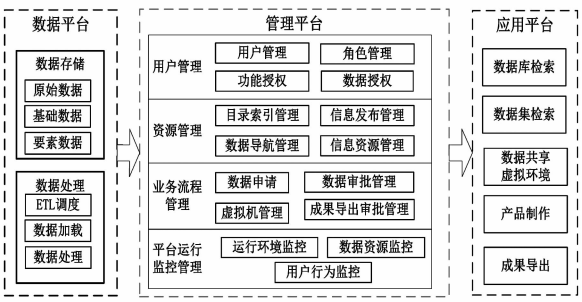

图 3　系统功能框图

3.1　数据处理分系统

3.1.1　实时数据处理子系统

根据海洋环境数据观测的采集规范和编码规定,对接收、收集和整合的大量海洋调查、业务化观测/监测等资料,按照资料类型、观测仪器、观测手段、要素内容等特点,开展数据识别、解码、数字化、数据项检查、代码检查等预处理,按照时间、空间和观测资料类型进行排重、排序和初步质量控制,剔除异常数据,依据数据来源、时间、地点等信息对数据文件进行挑选、过滤、分类存放,同时完善和新建相应的海洋环境数据存储标准,对资料进行标准化格式转换。

3.1.2　历史数据处理子系统

系统根据海洋环境数据观测设备性能、仪器订正参数、资料种类、观测要素类型、观测方式、资料时空分布、要素数据经验范围等特点,配置质量控制参数,采用相应的质量控制方法,对各类海洋环境数据进行精细化的计算机自动质量控制和人工审核。质量控制方法包括范围检验、非法码检验、相关检验、季节性检验、一致性检验、着陆点检验、梯度检验、尖峰检验、气候学检验和极值检验等。

3.2　数据库加载分系统

数据库加载系统包括通用数据库加载系统与并行数据库加载系统。通用数据库加载系统是通过加载文件清单的方式进行数据管理,清单文件是对每类数据的特征描述,包括文件类型、文件名、调查机构、绝对路径、备注等信息,通过一条记录就可以确认数据类型并找到数据存储位置。清单文件的组织结构与数据库表结构一致,且加载系统可实现清单列名与数据库列名对应关系的动态调整,清单配置文件设置完成后,单击上传,将清单的记录入库,加载过程中可通过状态条查看加载进度。

并行数据库加载系统先按照数据库结构利用 ETL 处理系统通过抽取数据文件的相关信息形成库文件,将库文件存放在规定的目录下,并查看库文件的文件表结构,创建相应的数据库表,创建 shell 脚本并制定源文件和目标文件,最后写入数据库。

3.3　数据查询检索分系统

系统主要分为两大模块:关键字查询和图形化检索。系统界面左侧显示海洋资料体系结构,右侧用于经纬度区域选择地图和查询结果浏览。用户首先在左侧选择相应的航次,然后在右侧地图圈定需求的区域,再输入关键字,查询该区域的特定信息,或查询特定区域的所有信息,或查询所有区域的特定信息,并能够对查询结果进行统计、排序、固定格式表格的导出。

3.4　运行监控管理分系统

通过建立运行环境监控信息数据库,确定数据库中各类监控信息表、监控要素字段、监控状态字段、表关系和数据字典等,实现运行环境监控、数据监控与用户行为监控的实体建设。

3.4.1　运行环境监控与管理子系统

运行环境监控与管理子系统包括硬件环境监控和软件环境监控两部分。硬件环境监控是通过对系统局域网硬件设备运行的日志信息进行提取、分析,实现对服务器、存储阵列、交换机、路由器、防火墙等设备故障诊断、告警等功能。软件环境监控是通过研制各商业软件(操作系

统、数据库软件等)与各业务系统(数据处理软件等)运行日志读取接口,实时读取日志信息并加载运行环境监控信息数据库。

3.4.2 数据资源监控与管理子系统

数据资源监控与管理子系统通过对数据汇集状态实时监控,实现信息反馈、到期告警、汇集情况季报与年报输出等功能,实现对海洋数据处理和质量情况的实时监控和预警、数据处理任务。调度管理;通过提取用户登录日志、数据库与数据集访问日志、数据申请信息进行分析,实现数据的服务内容、服务对象、应用领域情况的实时监控。

3.4.3 用户行为监控与管理子系统

用户行为监控与管理子分系统实时对用户的登录、数据资源访问、外部设备使用、软件安装预警和设备接入等行为进行监控,具有终止用户操作、告警提示、季度分析报告输出等功能,在提供用户方便使用的前提下保障系统的稳定运行。

4 关键技术

根据系统总体功能定位,在已有的工作基础之上,以数据的汇集、处理、存储、管理、服务过程为主线,采用操作系统、数据库、数据管理与共享3层软件体系,集成各类自主研发功能,构建灵活、稳定的架构模式。架构主要基于虚拟化技术、并行处理技术、数据检索并行处理技术与J2EE技术等关键技术。

4.1 虚拟化技术

由于用户对处理器、内存等硬件和操作系统需求不同,用户工作使用的数据处理软件、资料质量控制软件和产品制作软件不尽相同,为满足用户需求,同时提高服务器、存储阵列等资源的利用率,采用服务器虚拟化技术实现满足不同用户需求的虚拟机,同时消除服务器与存储阵列对应用系统的物理局限性。

服务器虚拟化技术是将一个物理服务器虚拟成若干个服务器使用,使得单个物理服务器上可以运行多个虚拟服务器,并对虚拟服务器的硬件资源如处理器、内存、I/O设备等进行配置管理,提供统一的指令集和设备接口。系统利用服务器虚拟化技术可实现多客户操作系统,不同硬件配置与软件环境的虚拟机,根据用户需求分配相应的虚拟机资源,并可对服务器、存储阵列、虚拟机进行统一的配置和管理。

服务器虚拟化是通过虚拟化软件向上提供对硬件设备的抽象和对虚拟服务器的管理,利用CPU虚拟化、内存虚拟化、设备与I/O虚拟化技术对硬件资源进行虚拟化,采用虚拟机实时迁移技术实现动态资源整合。系统选用VMware ESX Server虚拟化软件,实现对硬件的抽象,资源的分配、调度和管理[5-6]。

4.2 并行处理技术

利用高速并行处理引擎,完成多层次海洋数据体系动态更新的ETL(抽取、转换、加载)并行处理,实现整个系统的数据处理与调度,包括数据抽取、数据传输、数据转换与清洗、数据加载以及调度监控。

4.2.1 数据抽取

数据抽取的方式包括:全表刷新、时间戳增量、日志增量和时间戳比较。系统采用时间戳增量方式完成数据的抽取,时间戳增量方式是通过记录时间将增量数据从源数据抽取出来,以附加的方式加载到高速数据存储中,完成源数据中的记录定期更新。时间戳增量方式是在源系统需要抽取的数据表中增加时间戳字段,用以表示数据的修改或新增时间,在数据抽取时通过它来识别和抽取增量数据。

4.2.2 数据转换

由于海洋数据通过调查、汇交、网载等多种手段获取,每种手段来源的数据存在定义不规范、格式不统一等情况,导致系统的源数据存在重复、错误、格式不一等情况。数据转换是将多来源、多调查手段、多要素和多格式的数据进行转换,形成格式统一、实用性强的数据存储层。

4.2.3 数据加载

将业务系统和源数据库层抽取、转换后的数据加载、更新到目标数据库中。根据业务数据的实际情况,对不同业务系统的数据采用不同的加载周期;根据数据的抽取策略以及业务规则确定,采用直接追加、全部覆盖、更新追加等多种方式进行处理。

4.2.4 高速并行调度

利用高速并行ETL调度,按照既定步骤完成数据抽取、转换、加载的全部时间和流程的调度任务。调度的内容包括:从各业务系统到数

据层的调度，实现多来源数据的提取、转换和加载；从数据层到数据存储的调度，实现了原始数据、基础数据、产品数据的高速并行存储；从数据存储到应用层的调度，实现数据的并行查询检索[7－9]。

5　结束语

海洋环境数据在线服务系统实现了内部资源整合和数据业务流程的规划设计，完成了海洋数据从接收、整理、标准化处理到产品加工的一体化管理与服务。但是系统仍存在很多不足，如数据加载程序中间过程仍需要人工干预，数据三维可视化方面存在不足。因此其进一步改进目标是实现数据的自动化加载，开发信息可视化展示系统。

参考文献

[1]　韩春花，张俊明，梁建峰，等．侧扫声呐探测数据管理系统设计与实现[J]．海洋通报，2011，30(2)：187－191.

[2]　傅世锋，蒋金龙，李胜睿，等．福建省海洋环境保护规划信息系统设计与实现[J]．热带地理，2011，31(6)：593－597.

[3]　杨勇，高金耀，杨春国，等．基于 GIS 的海洋地震数据管理系统的设计与实现[J]．海洋通报，2011，30(4)：414－418.

[4]　陈宏文，王刚龙，邵长高，等．基于 C/S 模式的海洋地质调查数据保密框架设计与实现[J]．海洋技术，2010，29(3)：131－133.

[5]　孙晨阳．服务器虚拟化技术与应用[J]．科学大众：科学教育，2014(3)：169－170.

[6]　王博．应用虚拟化技术在海上平台的应用[J]．网友世界，2014(2)：24－25.

[7]　刘豹．一种分布式 ETL 工具的设计与实现[J]．软件，2013，34(10)：73－77.

[8]　夏魏，邵清．ETL 在超市大数据量中的应用研究[J]．信息技术，2013(11)：117－120.

[9]　李晨翔，何刚，孙莉．基于 Hadoop 平台的分布式 ETL 系统设计与实现[J]．福建电脑，2013(11)：111－114.