文章编号:1001 - 4179(2013)S1 - 0182 - 03

基于大数据的水利水电云 GIS 平台概述

王喜春[1,2],邹志禹[2]

(1.长江勘测规划设计研究院 长江空间信息技术工程有限公司,湖北 武汉 430010; 2.中国长江三峡集团公司 科技与环境保护部,北京 100038)

摘要:随着信息技术的飞速发展和云计算概念的提出,常规的应用服务平台已远远不能满足水利水电地理信息系统(GIS)的全方位要求,需要有一个强大的基础架构为各种应用服务提供支撑。介绍了大数据的定义及其特点。基于面向服务的体系架构,结合 GIS 和云计算,提出了水利水电工程中云 GIS 平台的体系架构。介绍了水利水电工程云 GIS 平台实现的关键技术,如与传统数据仓库工具的整合技术、实时数据的处理技术等。

关 键 词:大数据; 云 GIS 平台; 信息技术; 水利水电工程

中图法分类号: TP391 **文献标志码**: A

1 大数据简介

1.1 大数据的定义

按照现在较为普遍的定义,大数据是诞生于各类终端中的庞大的分结构化数据,而拥有存储、分析该类数据能力的技术就是大数据技术。大数据技术包括基础架构、数据管理、分析挖掘和决策支持四个层面。大数据的关键是在种类繁多,数量庞大的数据中,快速获取信息。

大数据和云计算密不可分,相互促进发展。云计算首先要进行数据的收集,而大数据是数据分析能力的升级。大数据和云计算都需要有庞大的数据存储和计算能力作为支撑。将计算资源集中以实现规模效应是信息产业发展的必然结果,对于没有大规模数据中心,但又需要对大规模数据进行存储和计算的企业,可以通过使用云计算资源来实现大规模数据的处理。

1.2 大数据的特点

(1)数据体量巨大。实体世界中,数以百万计的数据采集传感器被嵌入到各种设备中,在数字化世界中,消费者每天的生活(通信、上网浏览、购物、分享、搜索)都在产生着数量庞大的数据。IDC 研究表明,数字领域存在着 1.8 万亿 GB 的数据,其数据量正在以 55% 的速度逐年增长。

(2)数据类型繁多。数据可分为结构化数据、半结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据,音频、视频、图片、地理位置信息等类型的非结构化数据量占比达到 80% 左右,并在逐步提升,有用信息的提取难度不断增大。

(3)价值密度低。价值密度的高低与数据总量的大小成反比。以视频为例,一部 1 h 的视频,在连续不间断监控过程中,可能有用的数据仅仅只有一两秒。

(4)时效性高。这是大数据区分于传统数据挖掘最显著的特征。数据的价值除了与数据规模相关,还与数据处理周期成正比关系,而数据处理的速度越快(比如 1 s 内)、越及时,其价值越大,发挥的效能越大。

2 大数据与水利水电云 GIS 平台

2.1 水利水电云 GIS 平台体系架构

目前云计算及物联网技术已经在大数据处理、大规模计算、用户透明、减少系统设备投入和维护等方面体现出无以伦比的优势。当我们要面向各种各样的水利水电应用服务需求以及大数据时,一个常规的应用服务平台是不能满足水利水电地理信息系统(GIS)的

收稿日期:2013 - 01 - 04

作者简介:王喜春,男,高级工程师,博士,主要从事 GIS、云计算、软件工程、摄影测量与遥感等研究工作。E - mail: wangxc73@sina.com

全方位要求的,这时需要有一个强大的基础架构作为各种应用服务的支撑架构。因此基于面向服务的体系架构,结合 GIS 和云计算,建设现代水利水电共性服务的云 GIS 平台,对解决水利水电工程对 GIS 应用的各种需求,具有重要意义。

水利水电云 GIS 平台体系架构如图 1 所示。整个体系结构自底向上分为物理层、虚拟层、数据层、大数据支持平台及服务组件层、服务层、应用层以及横跨多个层次的服务发现、服务监控、资源调度、计量统计等服务。下面对各层的功能进行简单论述。

(1) 物理层。该层是架构的最底层,由计算机硬件(普通 PC 机)和交换机等相关物理设备组成。

(2) 虚拟层。该层由操作系统内核、虚拟机及虚拟化工具组成。

(3) 数据层。该层主要功能是存储基础地理信息数据、环境专题数据、工程管理数据、移民管理数据等各种专题数据及其索引信息并为实际应用提供数据支持。它包括数据库服务器和数据文件服务器,数据库服务器的访问既可基于中间件,如 ArcSDE 的方式,也可以选择直接接口访问的方法。对于大数据量的空间数据采用分布式存储。

(4) 大数据支持平台及服务组件层。该层包含了服务组件,提供了高性能计算、可视化实现、空间处理等服务或功能。该层的支持平台实现分布式计算功能以达到高性能计算的目的。该层还包括了支持 GIS 的运行环境,体现了云 GIS 平台的空间性。此外,服务组件在服务功能和服务质量(QOS)两个方面反映了服务的内在涵义。它还遵循服务组件架构规范(SCA)和服务数据对象规范(SDO)。

(5) 服务层。该层中包含服务组件层中定义的全部服务。水利水电云 GIS 平台服务将各种水利水电相关空间地理信息资源在 Web 上进行注册,形成服务网络,并通过 Web Service 技术对用户提供服务。平台所提供服务的内容包括资源注册服务、空间数据服务、空间信息处理服务等,所有这些服务能够被发现和调用,并能被编排以创建组合服务。服务发现是指为各种服务发布自己的功能元数据、查找其他服务、获得其他服务实例的地址信息,进而与其交互提供支持。服务监控是指在服务发现的基础上,更进一步提供对所请求服务所在节点状态信息的查询,在确保服务可发现的同时,保证服务可用。资源调度则是建立在服务监控基础上,实现了对同类服务按节点状态进行合理选择的调度功能。

通过这些服务,水利水电业主单位、设计单位、地方政府、施工单位等各类用户根据各自的不同需求,能够以云 GIS 平台的网络资源、计算资源以及存储资源为载体,获得各类水利水电空间数据资源;经过专业处理软件提供的各种服务,将空间数据资源转换为空间信息资源;各种服务通过经该服务层明确定义的,可供用户编程的服务接口实现平台的访问功能。

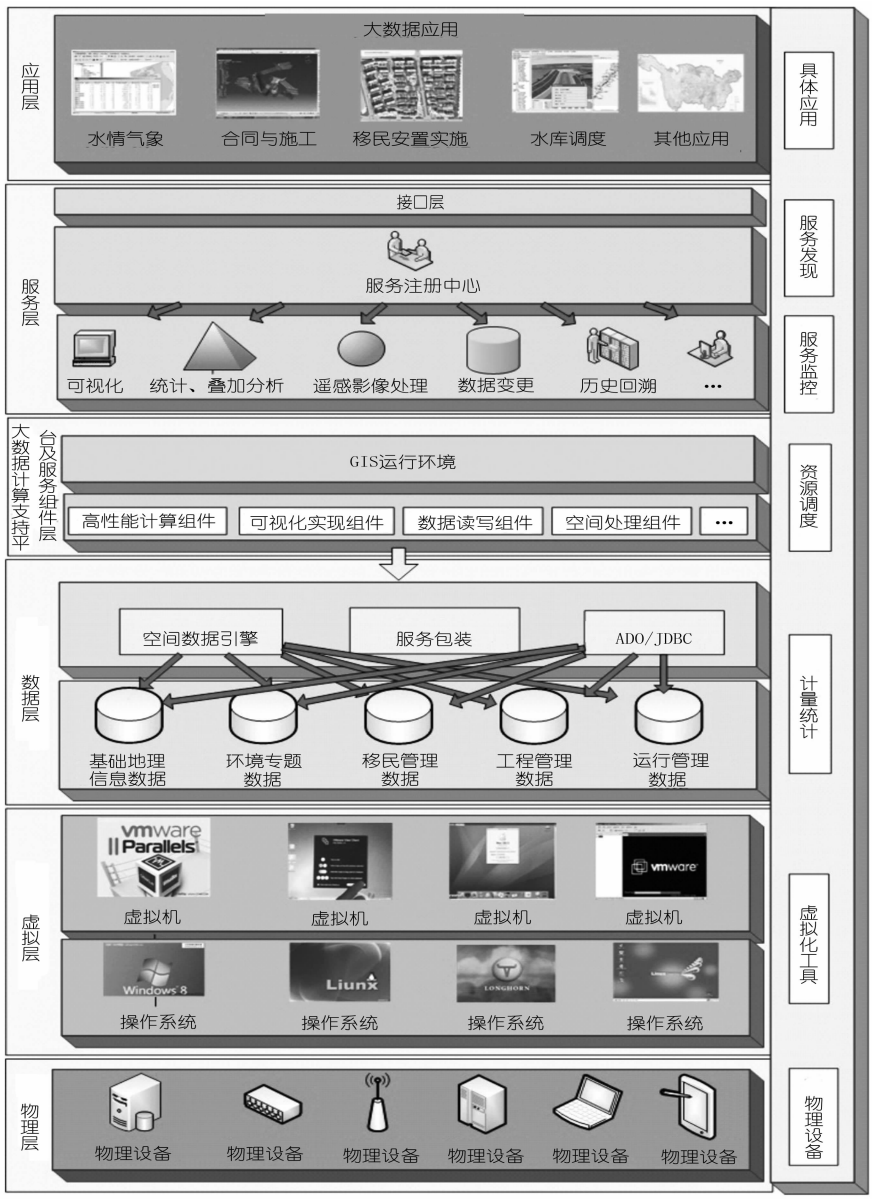

图 1　水利水电云 GIS 平台体系架构

特定业务单元组件和特定项目组件,并以服务描述的形式公开这些组件的部分接口。

(6)应用层。该层为所有架构体系的最上层,其内容为具体的 GIS 应用,其生命周期则涵盖了水利水电勘察设计、工程建设以及运行管理等各个阶段。所有应用包括具体的业务流程和用户客户端,都是依靠服务层提供的服务接口实现的。业务流程为用户提供了服务组件层中具体水利水电业务应用系统的服务化封装,再通过服务的组合和编排、服务链、工作流等技术实现业务流程的动态组合,最终通过服务层提供的服务接口实现。用户客户端则为用户与各种服务的交互提供了图形化和非图形化的图形用户界面。

2.2 水利水电大数据及其应用

水利水电云 GIS 平台贯穿水利水电枢纽工程全生命周期,其平台建设目标是通过建立勘测、设计、施工、环境、移民、设备物资供应等大数据基础架构,对种类繁多,数量庞大的数据进行有效管理。

基础地理信息方面,建立基础地理信息共享平台。将水利水电工程全流域的各种数字正射影像图、数字高程模型、数字栅格图、数字线划图等基础地理信息数据集成管理,为水利水电工程管理、移民管理、运行管理等应用提供基础地理信息。

环境专题方面,从流域角度出发,在基础地理信息共享平台的基础上,将各电站、库区及流域的水文泥沙、水情气象、水土保持、地震监测、地质灾害等生态环境信息集成管理,为灾害防治以及环境管理提供决策支持服务。应用自然语言处理和支持分析等一系列先进的分析方法,通过深层的分析查询,为用户实现环境数据的快速分析、决策和进行反馈,同时降低管理成本,从而充分挖掘大数据的价值。

工程管理方面,建立流域多工程信息管理平台,将前期立项、投资控制、物资设备、合同与施工、项目后评估等信息集成管理,为流域多工程的综合精细化管控提供决策支持服务。建设主要内容包括前期与立项管理子系统、合同与施工管理子系统、物资设备管理子系统、质量安全管理子系统、计划进度管理子系统、财务管理子系统等,对种类繁多,数量庞大的数据进行查询、分析、反馈和直观展示,结合施工进度开展数字仿真与分析计算,实现从设计、计划、施工生产、质量控制的全过程管理。

移民管理方面,建设主要内容包括移民安置规划成果管理子系统、移民安置实施管理子系统、移民投资管理子系统、移民后期扶持管理子系统、移民实物指标管理子系统等,实现移民实物指标可核查、移民资金拨付可追溯、移民安置效果可评估。

运行管理方面,建立运行管理平台,将水库联合调度、电力生产、电力营销、电力安全运营等信息集成管理,为智能化运行提供决策支持服务。建设主要内容包括水库联合调度子系统、电力生产子系统、电力营销子系统、电力安全运营子系统等,通过对来自物联网、云计算、新能源并网、移动互联等技术提供的生产数据和运营管理数据,运用 GIS 云平台强大的数据挖掘、数据分析和决策能力,实现电力企业精细化运营管理,动态优化流域调度信息,最大限度发挥流域梯级水库群的综合效益。

2.3 关键技术

(1)与传统数据仓库工具的整合技术。由于存储能力的增长远远赶不上数据的增长,现有数据中心技术难以满足大数据的应用需求,设计最合理的分层存储架构已成为基于大数据的云 GIS 平台的关键。同时,由于平台的用户将面对不同的数据库和分析环境,因此向外和向上的扩展能力也是非常重要的。在存储方面应考虑与其他数据库和数据管理环境共存,包括标准的关系数据和分析数据仓库。应用创新的压缩和重复数据删除技术是解决大数据问题的关键,通过低成本的服务器集群,进行大规模并行处理来提高数据管理能力与管理效率。

(2)实时数据的处理技术。数据成本下降促使数据量急剧增长,而新的数据源和数据采集技术的出现使数据类型增多,各种非结构化的数据又增加了大数据的复杂性。大数据本身也存在一些其他风险,大数据的集合和高密度的测量将令“错误发现”的风险增加。应重视数据工程而非数据科学,即主要考虑大数据分析算法和系统的效率。未来,大数据研究的任务不是获取越来越多的数据,而是数据的去冗分类、去粗取精。高扩展高可用的数据分析技术、新的数据表示方法、高通量计算机等都是亟待解决的技术问题。

(3)信息资源注册服务技术。重点解决各类水利水电空间数据及服务资源的分布式注册与管理。为使服务使用者能够了解服务提供者提供的服务以及确定哪些服务能够满足自己的要求,服务提供方将相关信息服务的描述和访问入口注册到注册中心,服务使用者通过注册中心查找所需服务。信息资源服务注册中心由多个全局注册中心数据库以及集群注册中心数据库组成,全局注册中心数据库为上一级数据库,集群注册中心数据库为下一级数据库。全局注册中心数据库与集群注册中心数据库在实现分布式注册与管理的同时实现数据同步。

(下转第 196 页)

需要注意的是必须在服务器 A 和 B 上都安装 rsync,其中 A 服务器上是以服务器模式运行 rsync,而 B 上则以客户端方式运行 rsync。这样在 Web 服务器 A 上运行 rsync 守护进程,在 B 上定时运行客户程序来备份 Web 服务器 A 上需要备份的内容。

#rsync - avzP nemo@192.168.10.1:/nemo/backup

程序 rsync 不同于 ftp 传递协议,除首次备份实行完全备份之外,之后的备份可以实现自动增量备份,程序会自动识别修改或增加的文档,实现备份数据与源数据的版本同步。

容易忽略的是,由于虚拟机的 IP 易于变更,会造成不能联机的情况,可以修改 root 家目录下的隐含目录.ssh 中的文件 known_hosts,删除其中对应故障主机的那条记录,故障即可恢复。

在备份机上执行以下命令,可以动态查看备份增量变化情况。

watch -n 60 "du -sh"(每分钟刷新一次动态信息)

4.2 虚拟化实践中应注意的问题

KVM 开源虚拟化技术可以最大程度保障单位对自身解决方案和服务的弹性定制,并可有效实现后期按需扩展性。尤其重要的是,开源的 RHEL 没有强硬的许可证限,这使单位在今后横向扩展新的虚拟化服务器时不会再产生额外成本。正是这种高效率和低成本的虚拟化应用也带来如下一些负面影响。

(1) 过于集中地部署密集型应用程序到虚拟机,造成这些应用程序争夺同一硬件服务器的带宽、内存、处理器和存储等资源,可能会遇到网络瓶颈和性能问题,引起服务器负载过重。最为明显的压力来自于内存,一般情况内存使用不要超过物理内存的 90% 为宜,应用上文提及的 KSM 内存调优也可以改善性能,前提是控制虚拟机总数,往往可以提高虚拟机 10% 的性能。

(2) 多个系统整合在 1 台服务器中,节约资源的同时,也面临一个严重的问题,即一旦服务器出现硬件故障,其上运行的多个系统都将停止运行。虚拟化的服务器合并的程度越高,此风险越大。可以使用异地备份的方式定期备份重要数据;严格控制服务器机房的温度,避免物理机升温过快,造成硬件的不稳定甚至宕机。最终的解决方案是通过双机双存储集群方案提高动态高可用性。

(3) 虽然 RHEL 系统安全性极高,但是超级管理员 root 账号具有至高无上的权限,严格管理账号密码至关重要,严格控制硬件防火墙的的端口,对外仅开放有限的端口号,服务器的 IP 地址使用专有网段,限定超级管理员访问地址,严禁非管理员接近物理机甚至上机操作等,都是很必要的安全措施。

5 结 语

服务器虚拟化已经成为主流技术,并非只有在大型的数据中心才可以应用,上例说明了小型的服务器环境也可以成功应用。服务器虚拟化技术可以帮助单位整合服务器资源,并在一定程度上解决了数据中心空间、电力、制冷不足的问题,具有良好的应用效果。随着虚拟化技术的日益完善,会有越来越多的应用逐渐迁移到虚拟化环境中,各种各样的安全威胁也会不断出现,对政务网站的数据安全及基础架构的安全提出了新的要求。因此,在部署服务器虚拟化的同时,必须完善安全管理策略,严格执行安全措施,从根本上预防虚拟化安全问题,真正实现政务网站安全高效节约成本的最终目标。

参考文献:

[1] 刘亚军,刘延军,李涛.服务器虚拟化技术在报业的应用初探[J].中国传媒科技,2011,(11).

[2] 金天昕.服务器虚拟化管理问题与对策[J].中小企业管理与科技(下旬刊),2010,(12).

[3] 秦学东.开源虚拟化——KVM 的构建[J].现代图书情报技术,2011,(11).

[4] 宋欣.图书馆服务器虚拟化存在的安全风险与防范措施[J].现代信息技术,2011,(4).

[5] 宋晓光,杨晒晒,吕渊鸣.虚拟化技术在数字化校园建设中的应用[J].中国教育网络,2011,(3).

(编辑:邓 玲)

(上接第 184 页)

(4) 管理与数据的安全。水利水电云 GIS 平台中异构网络环境中透明的协同或融合,需要具有良好的可扩展性并能进行网络组件化的即插即用自主管理,减少甚至排除人为的配置与干预,尤其以大规模自组织工作模式的泛在节点与终端自主管理为突出需求。同时还应能针对网络及业务环境的变化做出迅速的反应,保证水利水电云 GIS 平台中传感采集、网络传输、业务认证端到端的信息安全,构建以用户为中心面向水利水电应用的可管可控可信的网络支撑体系。

(编辑:郑 毅)