基于 Hadoop 的高性能 GIS 云计算平台研究

李 茜

（广西经济管理干部学院 计算机系，广西 南宁 530007）

[摘 要] 文章在研究当前主流 GIS 平台云计算体系架构的基础上，提出利用 Hadoop 这一开源的分布式计算环境设计一种高性能的 GIS 云计算平台，并予以实现。该平台利用 HBase 分布式数据库和 HDFS 分布式文件系统对空间数据进行存储，并使用 MapReduce 技术对 GIS 的空间分析任务进行集群化和分布式处理，大大提高了高密度计算环境下海量矢量空间数据的计算效率，为构建大规模访问下的公共 GIS 服务平台提供了一种可行的实现办法。

[关键词] Hadoop；GIS；云计算；MapReduce；服务平台

当前，云计算已经成为时下国内外主流的地理信息系统平台提供商，如 ESRI、MapInfo、Intergraph、北京超图（SuperMap）、武汉中地（MapGIS）、武大吉奥（GeoStar）等争相布局成为 GIS 产业研发与应用的重要领域。

从目前各大厂家推出的产品来看，其体系结构仍旧是在传统的 GIS 平台上，围绕着集群服务改良和成熟 IaaS 平台的虚拟资源调用这一思路包装相应的云计算产品，其数据存储形式和事务处理方式并未与以往的产品有较大的改进，不能充分满足和适应大规模公共 GIS 服务环境下和空间大数据分析环境下的空间数据存储和分布式空间分析计算、空间数据挖掘等新应用环境下对资源管理、数据共享服务和计算能力的要求。

Hadoop 是一个成熟、可靠、高效、跨平台的具备弹性横向扩展能力的分布式系统基础架构。笔者提出在此基础上构建一种高性能的GIS云计算平台，以提高海量空间数据的计算、存储和访问效率，为未来构建大规模公共 GIS 服务平台提出新的思路。

1 国内外主流应用现状

云计算目前提供 4 种主流类型的服务：基础设施即服务（IaaS）、平台即服务（PaaS）、软件即服务（SaaS）和数据即服务（DaaS），这 4 种服务统称为“ XaaS ”。

随着云计算从概念到应用的逐渐落地，GIS 行业也逐渐提出了“ GIS 云 ”或“ 云 GIS ”的概念。全球 GIS 平台市场占有率第一的 ESRI 公司自 2010 年推出的 ArcGIS 10 系列产品，再到 2015 年最新的 10.3.1 版本，都已经逐渐将云计算技术纳入其核心功能内容，提出了公有云环境下依托 Amazon EC2 等公有云服务商提供 GIS 服务器集群镜像，私有云环境中具体实现为 GIS 服务资源弹性调度的概念和相应的部署技术。该技术将所有的 GIS Server 以 P2P 技术进行集群互联，对外以 Server Site 群集的形式统一对外提供 GIS 服务 [1]。在集群资源不足的情况下通过 CloudBuilder 组件动态调动 IaaS 层的虚拟机资源满足计算节点不足的问题。但是这一技术并未提供对大量空间数据的分布式存储和对于单一巨事务类

[作者简介] 李茜（1980-）女，广西桂林人， 2009 年毕业于武汉理工大学，计算机科学与技术专业，主要研究方向为计算机应用技术和地理信息系统技术。副教授。

型或者大型空间分析计算任务能够拆分为多个任务块分配到不同计算节点进行分布式处理以提高计算效率和运算吞吐能力的功能。从实际来看，仍未是一个真正意义上的“云 GIS”系统，系统的瓶颈永远存在于集群中的某个节点服务器上。

国内北京超图（SuperMap）的 SuperMap GIS 7C（最近发布了 8C 版本）产品线，提出了基于 SuperMap iPortal、SuperMap iServer 和 SuperMap iExpress 的 Cloud Platform“云 GIS”平台。其核心 SuperMap iServer 软件的云计算功能实现也类似于 ArcGIS 10 系列产品，除了新增 Geo-CDN（GIS 数据分发服务）和集成各类云服务的 Portal 产品外，也未具备采用分布式计算和分布式存储方式处理空间分析事务的能力，也不能认为是真正意义上的基于云的系统。近期超图与浪潮集团携手提出的云平台解决方案，笔者尚未见到更多深入的技术探讨和对应报道。

国内专家学者在此方面也提出了关于 GIS 云计算架构的不同理解，并开展了相应的研究。

方雷在其博士论文 [2] 中提到将“云 GIS”分为 6 个层次，分别为物理层、虚拟层、数据资源层、云计算支持平台及服务组件层、服务层和应用层，并在此基础上应用微软的 Dryad 平台讨论了空间数据存储的负载均衡，但未涉及到采用分布式系统环境来管理和存储数据。

范建永等人提出了一个在开源平台 Hadoop 上搭建的开源的“云 GIS”体系架构 [3]，这一观点对笔者的研究具有很大的参考价值。此平台将数据按照比例尺和分辨率不同放置于 HBase 相应的表中，极大地提高了空间数据访问的效率，但验证的数据规模相对有限，也未对空间、非空间混合数据交叉关联查询进行分析研究，商用化程度不足。

顾荣等人在 Hadoop 上，提出了一个对 MapReduce 编程框架模型在处理短事务中的性能进一步优化的方法 [4]，这对笔者设计的技术框架有较大的启示。

周鹏等人提出结合云计算的基本概念和 GIS 的软件工程特点 [5]，从 GIS 的软件结构、开发组织和部署管理等方面探讨云计算下的 GIS 软件开发，但是没有具体落地的实例。

笔者基于 Hadoop 的技术特点，结合 GIS 处理海量空间数据的方式方法，设计一个高性能的 GIS 云计算平台，该平台可满足海量矢量数据实时显示和实时空间分析。

2　基于 Hadoop 的高性能 GIS 云计算平台结构

2.1　设计目的

根据应用环境的不同，“云上”的应用平台的物理架构和相应的功能具有较大的差异。笔者研究的基于 Hadoop 的高性能 GIS 云计算平台，主要是从面向海量 GIS 显示效率的提升、GIS 数据更新和数据分析处理 3 个方面开展相关工作。

（1）提升矢量数据的显示效率。传统的矢量数据必须进行切片才能显示，否则就会显示缓慢甚至无法显示。

（2）便于数据更新。传统的数据展示，矢量采用切片的方式进行展示，因此导致每次数据更新的时候都需要重新切片来进行数据显示的更新；如果切片更新不及时就会出现查询等操作的结果与实际数据展示效果不一致的情况。

（3）数据的分析处理。传统的 GIS 分析处理方式，都是采用单节点多核多线程或者多集群单节点的方式来解决大任务和长事物的处理问题，而没有使用网络内资源（多机器协同）来解决大任务的分解处理的问题。当面对单个任务的处理内容或复杂度超过机器的负载能力时，就会出现处理缓慢甚至无法处理的情况。

2.2　体系结构

基于 Hadoop 的高性能 GIS 云计算平台的整体结构如图 1 所示，一共包括 4 层，自下而上分为数据存储层、数据管理层、服务层和应用层。

（1）数据存储层。此层是平台体系结构的基础层，它提供分布式文件型空间数据的存储和索引，分布式数据库的空间数据存储和索引，空间数据操作日志记录，空间数据备份安全管理，数据读写优先级设置，操作错误保护等功能。在这一层中，主要是使用 Hadoop 架构中的 HDFS 和 HBase 进行数据存储与管理，前者用于最终数据的存储，后者用于空间数据索引表、空间数据表的组织与构建。

（2）数据管理层。此层是整个体系架构的核

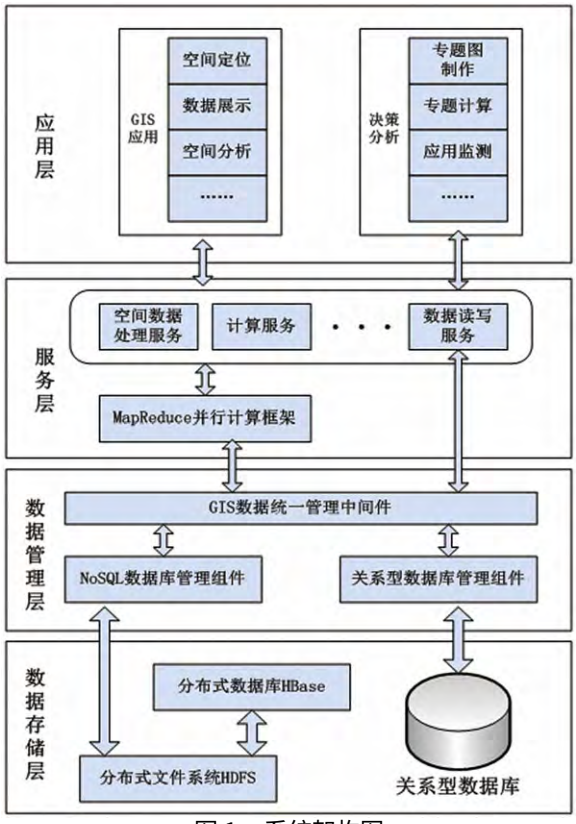

图 1 系统架构图

心，主要提供空间数据相关的 NoSQL 的优化，统一 NoSQL 和传统数据库的访问接口，不同类型库、跨库与文件系统的操作优化等功能。这层提供主要的数据访问接口，如多分辨率影像地图服务的数据访问接口、矢量数据分布式存储接口、元数据存储接口和其他业务数据存储接口等。

（3）服务层。这一层是实现基于 Hadoop 的高性能计算的基础，利用 MapReduce 技术将处理和分析的任务并行分配到不同的云中节点上处理，提升处理性能。该层主要实现的功能是提供数据操作的服务接口，在线制图、配图接口，空间分析、查询、处理接口，算法模型接入接口等。

（4）应用层。这一层主要是面向用户提供相应的数据分析服务和应用功能搭建能力，辅助用户构建专题应用系统，以满足不同用户的实际需求，如提供在线制图、配图和空间分析、查询、处理功能等。

2.3 技术特点

（1）基于 Hadoop 的多分辨率影像数据和多比例尺的矢量数据的存储管理。

（2）支持多业务、多部门数据的分类管理。不同部门的业务数据可以按照多租户的模式存储在 HBase 表中，以满足不同用户相互隔离的应用。

（3）基于 MapReduce 编程模型的高性能计算实现，充分利用基础设施的计算资源。

（4）自定义业务功能服务发布。用户可以利用应用层发布的服务调用接口，自行开发相应的 GIS 应用服务。

3 系统实现及测试

此研究的体系结构环境实现及测试环境主要由连接在千兆以太网交换机上的 1 台物理主机上部署的 5 个虚拟主机构成。其中，1 台物理主机作为主节点运行 NameNode 和 JobTracker，4 个子节点运行 Data Node 及 Task Tracker。同时，需要对 Hadoop 和 HBase 的配置文件进行相关的配置。物理主机的硬件配置为 Xeon E5405 2.0GHZ×4；操作系统为 OpenSUSE Linux 11.2；内存为 4GB。

实验数据为 10GB 左右的谷歌全球矢量数据。

开展的实验类型为基于 MapReduce 的矢量数据构建显示效率。

经过测试，当使用谷歌全球的 10GB 矢量数据时，2 节点的实验环境数据显示时间约为 12s，兴趣点检索时间约为 200ms；4 节点的显示时间约 5 s，检索时间约为 100ms。相比较同等配置的 ArcGIS 10.1 的 2 节点集群处理，显示全图的时间大约为 2.37m，性能提升较为明显。

此平台还利用 REST 接口开发了符合 OGC 标准的矢量数据发布显示功能，浏览效果如图 2 所示。

4 结 语

GIS 云计算和云平台是整个 GIS 产业围绕信息技术发展的重要方向。通过采用云计算的相关应用技术，可以大大优化 GIS 空间数据的存储与管理能力，并可以在此基础上，通过分布式技术，充分发挥每个节点的能力，提供极高的性能。笔者基于 Hadoop 的高性能 GIS 云计算平台，采用 MapReduce 框架解决海量空间数据的分布式计算效率问题，并采用 HBase 等分布式存储技术提升了数据

（下转第 32 页）

（3）加大对成熟模式的推广力度，但要因地制宜。各地石漠化治理模式的成功，其根本原因是因地制宜，找到适合自身发展的道路。这些模式是“样板工程”、“示范工程”，有很强的可推广性，但并非简单的复制。石漠化治理需具体情况具体分析，可以以小流域治理为重点，可以以保水保土为重点，还可以通过生态移民，减少当地生态压力为重点，等等。

（4）加强石漠化治理基层人才队伍建设。当前，我国石漠化面积较广，分布较散，而石漠化治理人才较为紧缺，特别是在基层，石漠化治理大多靠经验积累，缺乏有效的科学支撑，导致一些石漠化治理项目功效不大，造成浪费。而技术专家又不可能长期驻守基层，因此，要加强石漠化治理基层人才队伍建设，可通过举办培训班、定期组织专家队伍巡回授课、免费发放宣传资料等方式，将专业的治理知识传送到基层。

（5）形成学科合力，共同攻克石漠化治理难题。长期以来，人们都有一种误区，认为石漠化治理属于自然科学的领域，与人文社会科学关系不大，这种思路不利于石漠化治理的有序开展。石漠化治理固然离不开自然科学的技术支撑，如地质勘察、项目设计、工程建设等，但人文科学的治理理念也必不可少，如投入产出分析、土地权属变动、融资等。治理的最终目的是实现人与环境的和谐发展，人是一切经济社会活动的核心。只有将自然科学和人文科学共同应用于石漠化治理，形成合力，才能达到更好的治理效果。

（6）加快石漠化监测体系建设，实现对石漠化地区全域动态监测。要重视对监测网络的资金投入和建设，优先对重点地区实行全时监测，继而完善全域监测网络。要加强对监测结果的运用，加快科技成果的转化，而不应仅用于发论文、写报告，要切实用于指导当地石漠化治理工作的开展。要破除地方保护主义，依托大数据平台，整合监测数据，实现区域内监测数据的共享，真正推动石漠化集中连片地区共同整治、联动发展。N

致谢：感谢广西马山县人民政府、马山县发展和改革局给予本文的帮助。

（上接第 28 页）

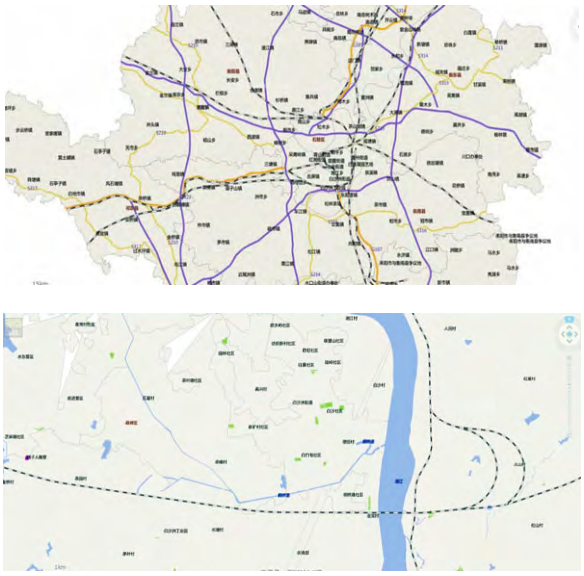

图 2　矢量数据发布浏览效果图

检索和显示的效率。整个平台架构已经过验证，具有可操作性。N

[参 考 文 献]

[1] 朱剑 . 基于虚拟云计算架构的 GIS 服务资源弹性调度应用研究 [J]. 测绘通报，2013，（5）:92-95，107.

[2] 方雷 . 基于云计算的土地资源服务高效处理理论框架及其平台关键技术探索与研究 [D]. 杭州：浙江大学，2011：94-97.

[3] 范建永，龙明，熊伟 . 基于 Hadoop 的云 GIS 体系结构研究 [J]. 测绘通报，2013，（11）:93-97.

[4] 顾荣，严金双，等 .Hadoop MapReduce 短作业执行性能优化 [J]. 计算机研究与发展，2014，51(6):1 270-1 280.

[5] 周鹏，尹菲 . 基于云计算技术的 GIS 软件工程模式 [J]. 测绘通报，2010，（11）:22-24.