Cloud GIS: Theory, Method and Practice

Y. Wang, J. Wang ,C. J. Li, X. B. Yan

School of Computer Science, China University of Geosciences, 430074

Telephone: 13808669081, 15827030409, 13618603396, 15527495676

Fax: -

Email: giswy@126.com, jjseen@163.com, cuglicj@126.com, 814443137@qq.com

1. Big Data Characteristics

We are living in Big Data age and facing massive data in its most raw form or a semi-structured or unstructured format [Ghemawat Gobioff and Leung 2003]. We even don't know how to process or analyze this data using traditional tools or ways for reasons its three main characteristics: volume, variety and velocity [Zikopoulos and Eaton 2011]

The volume of data is exploding in past 10 years. It will expand 40 times from 2000 to 2020 and will reach 35 zettabytes (ZB) by 2020. All this data can't be stored and processed in traditional systems.

With the explosion of sensors and smart devices, as well as social collaboration technologies, data has became much more complex than ever. Traditional analytic platforms can't handle the variety of Big Data which brings along challenges to us for trying to deal with it.

Velocity namely the speed at which the data is flowing requires that we perform analytics against the volume and variety of data while it is still in running, not just after it is at rest. It considers how quickly the data is arriving and stored, and the rates of retrieval.

2. Cloud Computing

Cloud computing therefore come into sight for solving above all difficulties. Cloud computing composed of hundreds of thousands machines with scaling down and up easily offers almost infinite storage spaces and crucial capability for computing. Cloud computing entrusts remote services which compose of user's data ,software and computation. Cloud computing system requires typically three main participants: IaaS, PaaS and SaaS[Foster 2008].

IaaS(Infrastructure as a Service) provides whole hardware layer of services such as computing units, storage, networking, firewall etc. A key technology contributing to IaaS is virtualization which allows servers and storage devices to be shared and utilization be increased. Applications can be easily migrated from one physical server to another.

PaaS(Platform as a service) provides facilities for the deployment of applications without the cost and complexity of managing the underlying hardware and software. These facilities allow customization of existing SaaS applications---software deployment and configuration settings. PaaS should offer application hosting and a deployment environment, along with various integrated services with varying levels of scalability and maintenance.

SaaS(Software as a service) is a software delivery model in which software and associated data are centrally hosted on the cloud. SaaS known as centralized hosting of business applications is typically accessed by users using a web browser. PaaS along with its underlying structure---IaaS come together to be the fundamental of SaaS referred conventionally to applications on Cloud.

3. What Is Cloud GIS

Geographic Information System (GIS) deals with complex Big Data divided mainly into two categories: raster data (satellite/aerial digital images), and vector data (points, lines, polygons). These types of data as we called spatial data generated periodically via all kinds of sensors will be up to so gigantic magnitude soon after the system was founded as to can't be stored and processed by traditional ways such as RDBMS or general distributed file systems.

Cloud GIS which has come into use in recent years changes our traditional perspective on GIS. With Cloud GIS, everyone can easily access GIS applications to store, analyze, visualize, share, and manage GIS assets involving maps, spatial data and variety of information. Cloud GIS also called as cloud computing in GIS transforms the way that information flows throughout a organization or the world.

The development of Cloud GIS claims us an infrastructure to host large volumes of data, a parallel computing framework to apply spatial analysis on data, and a configurable architecture to meet demands for the deployment, extension and supervision, etc [Kang 2011]. Hence a Cloud GIS system can be broken up into three components: IaaS, SaaS and PaaS, as it does in cloud computing system.

4. Cloud GIS with Virtualization

A hybrid of physical servers and virtualization technology is an implementation method of IaaS for Cloud GIS is to adopt. Virtualization is perceived as the simulation of the software and/or hardware upon which other software runs. The usual goal of virtualization is to centralize administrative tasks while improving scalability and overall hardware-resource utilization.

With virtualization, several operating systems can be run in parallel on a single CPU. Instead of relying on the old model of one server with one application that is prone to underutilized resource, this parallelism of virtualization tends to reduce overhead costs and ultimately improves the efficiency and availability of resources.

A full virtualization technology is Virtual Machine (VM). Each VM has private virtualized hardware: network card, disk, graphics adapter, etc. In practice we entrust

KVM (Kernel-based Virtual Machine---an open source software) virtualization solution for Linux. Together with VM we also relies upon a kind of multi-tenancy technology, the LXC (LinuX Containers) which provides lightweight virtualization that lets us isolate processes and resources subtly to avoid the heavy costs using VM alone.

5. Cloud GIS with Hadoop

The Apache Hadoop software library is an open-source software framework that allows for the distributed processing of large data sets across large clusters of commodity hardware [White 2009]. Hadoop library enables applications to work with thousands of computation-independent computers and petabytes of data with capabilities of detecting and failing over at the application layer, so as to deliver a highly-available service on top of a cluster of computers, each of which may be prone to failures. Hadoop is widely used by many large corporations to reduce the difficulty of developing distributed and parallelized applications.

The entire Hadoop ecosystem consists of the Common kernel, HDFS (Hadoop Distributed File System), MapReduce [Cary Sun Hristidis and Rishe 2009] , as well as a number of related projects - among them we specially care about HBase. All of them are designed so that node failures are automatically handled by the framework.

HDFS provides a distributed file system that stores unstructured data on the compute nodes with very high throughput across the cluster.

Hadoop implements a computational paradigm for parallel processing of large data sets named MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster.

HBase is a scalable, column-oriented, distributed database that supports structured data stored in large tables. Use HBase we can random, realtime read/write access to big data. HBase leverages the distributed data storage on top of HDFS.

HDFS is used to store data especially for the semi-structured or unstructured data. Structured data like DBMS tables can be converted and stored into HBase. Also, spatial data such as vector data and image data can be stored in HBase . HBase provides an indexing mechanism by rowid for fast searching and acquiring of rows.

Due to the large size and complexity of spatial data, traditional sequential computing models may take excessive time to do spatial analysis. While a parallel computing models, MapReduce provides the capabilities of scaling down spatial data processing. For example, the parallel construction of image pyramid and other image processing by MapReduce from the boundless image data stored in HBase is widely used.

As we can use KVM and its associates to construct the IaaS layer of cloud computing, the solidity and stability of Hadoop greatly simplifies the development tasks of SaaS for Cloud GIS. As a result, many SaaS layer of Cloud GIS is comprised of Hadoop or its equivalent.

A key component not mentioned in this section is PaaS for Cloud GIS. It is very technical about how to deploy IaaS, SaaS for scaling up or down and closely

connected with the scalability and maintenance of system. PaaS can also integrate various elementary services for internal development of SaaS, for instance, the logging service and cloud file storage service.

6. Conclusion

This work gives an overview on Cloud GIS theory and talks about implementation method of the system in practice. We don't classify public cloud or private cloud computing system for the reason that they are essentially the same thing.

A Cloud GIS system is inherently cloud computing system. A cloud computing system on which GIS services or applications hosting might be recognized as Cloud GIS system. Cloud GIS provides not only the infrastructure services, but also full functionalities of GIS spatial algorithms.

More and more GIS spatial algorithms are migrating from traditional system to cloud computing platform. Many of them are not compatible for parallelism and should be redesigned. However, big challenges still remain to us.

7. Acknowledgements

The project was supported by the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) under grant CUGL110228 and supported by the high-performance computing platform of China University of Geosciences.

8. References

Cary A, Sun Z, Hristidis V and Rishe N, 2009, Experiences on processing spatial data with MapReduce. In: the 21st International Conference on Scientific and Statistical Database Management, Urbana, USA.

Foster I,2008,Cloud Computing and Grid Computing 360-Degree Compared. In: Grid Computing Environments Workshop (GCE '08), Austin, TX, USA ,1-11.

Ghemawat S, Gobioff H, Leung ST, 2003, The google file system. In: 19th Symposium on Operating Systems Principles, NY, USA, 29-43.

Kang C, 2011, Cloud computing and its applications, Clark university, USA.

White T, 2009, Hadoop:The Definitive Guide. O'Reilly Media, Sebastopol.CA, USA.

Zikopoulos PC and Eaton C, 2011, Understanding big data: analytics for enterprise class hadoop and streaming data . McGraw-Hill, NY, USA.