Research on some key technologies of Hadoop's cloud GIS

Xinyong Wang a, Qin He b
China Mobile Information Technology Company Limited, Guangzhou 510000, China.

a wangxinyong@chinamobile.com, b heqing@chinamobile.com

Abstract: Cloud computing is the product of the development of Internet computing to a certain stage, is the latest result of the evolution of parallel computing, grid computing and other new computing methods. The infinite expansion of cloud computing storage technology can meet the rapid growth of spatial data storage space requirements, powerful computing capabilities for spatial information retrieval, processing, analysis and other high-speed services to ensure. Aiming at the problems of massive data storage, processing, analysis and continuous service faced by GIS, this paper constructs the application of Hadoop-based GIS, and studies some key technologies.

Key words: Cloud computing, cloud GIS, Hadoop, key technologies.

1. Preface
The advantages of cloud computing are also applicable to GIS information processing. Because of the large amount and variety of data involved in GIS, it is difficult for traditional data processing and sharing of GIS data. By virtue of the distribution and heterogeneity of multi-data in cloud computing, the problem can be simplified, so that GIS has supercomputing service ability and higher service efficiency for the future. The development of cloud GIS and its standardization provide the basic technical conditions. This paper focuses on the combination of cloud computing with GIS, using the cloud computing technology of massive data computing capabilities, combined with hadoop's own characteristics to build a cloud GIS framework. Because the system is based on the open source cloud computing platform Hadoop, it runs on a cluster of multi-distributed database applications composed of a large number of hardware devices. Hadoop simulates and implements the main technology of Google cloud computing. The core technology of Hadoop is HDFS, and its MR (MapReduce) and HBase correspond to MR and Bigtable in the core technology of Google cloud computing, GFS, respectively. At present, the platform is an open source distributed cloud computing software platform, making the calculation method has a bright future.

2. Cloud GIS framework based on HADOOP platform
The cloud GIS system architecture can be divided into three layers: the first layer is the cloud GIS service part, the second layer is the cloud GIS architecture part, and the third layer is the physical part. The frame structure is shown in Figure 1 below.

Cloud GIS services are based on three services provided by cloud computing, namely, software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS), and are determined according to the characteristics of geographical information system (GIS). The 3 service is to inherit the
corresponding services provided in cloud computing. That is, geographic information infrastructure as a service (GLLaas) corresponds to infrastructure as a service, and geographic information platform as a service (GLPaaS) corresponds to platform as a service and geographic information software as a service (GISaaS) corresponds to software as a service. Among them, geographic information as a service (GIaaS) provides map query, route planning and other online services. These online services adopt a service-oriented architecture, and encapsulate their services under this framework to become a standard web service. D z7Web Service, these services are managed and applied or shared under this framework. Geographic information software as a service provides spatial analysis function, format conversion and spatial data publishing function through network means. These services are the implementation goals of building the cloud GIS framework, which is based on the cloud GIS architecture.

The cloud GIS architecture is mainly composed of 3 parts: middleware, HADOOP platform and virtual resource pool. The middleware layer is the scheduling function of resource management and concurrent processing of massive data in cloud GIS during task execution, which can make resource efficient application. In the process of task scheduling, the interaction of cloud GIS data can be realized by geographical markup language. The middleware includes system user management, which manages the user's privileges when using the system services, manages the application configuration and data security protection of the system, and billing the user using cloud GIS services. The task management is mainly to balance the load of the running tasks of the system and to realize and manage the map slicing. The overall security of cloud GIS system can be realized by means of firewall and network encryption. The HADOOP platform is mainly built on the environment, that is, the operation platform. The platform is built by installing Xen Virtualization in Linux environment, installing and configuring Eclipse in creating virtual machines, and then adding Hadoop plug-ins to Eclipse and configuring it. Virtual resources include virtual software resources and virtual hardware resources. The virtual resources in the framework are composed of technical resources, network resources, data resources and software resources. Hadoop platform uses its own characteristics to manage cloud computing resources, so that the platform can schedule a large number of GIS applications efficiently, and is responsible for load balancing in the process of concurrent operation of cloud GIS tasks, and the massive data of GIS is operated by MR model. Virtual resource pool is a large amount of resource data storage such as technical resources, data resources, network resources and software resources.

The physical part of the cloud GIS framework is the foundation of the whole framework. When building the cloud GIS framework, these physical devices can be purchased or leased.
3. Analysis of several key technologies of Hadoop's cloud GIS

3.1. Cloud GIS data storage technology based on Hadoop
Vector data storage based on Hadoop can be generally divided into vector data storage based on distributed file system HDFS and vector data storage based on distributed database HBase. The storage mode based on HDFS is generally combined with the simple element’s specification of OGC to design the vector storage format, which stores the vector data into HDFS file according to the set KeyValue format, and establishes a certain spatial index to realize the retrieval of vector data. The vector data storage method based on HBase database also designs the corresponding KeyValue format, and stores the vector elements in the corresponding tables. It also needs to establish a certain spatial index.

Spatial information multi-level grid can solve many problems such as efficient storage and index of massive geospatial data, integrated expression of sphere and plane, high performance parallel computing.
and so on. It can also solve many practical problems such as integrated organization, open sharing, interoperability and rapid distribution of multi-source and multi-scale spatial information.

3.2. Parallel computing of spatial data based on MapReduce
The main overhead of the MapReduce parallel computing model should include the algorithm communication overhead and disk I/O overhead besides the algorithm execution efficiency.

There are two reasons to focus on communication overhead and disk I/O overhead: first, tasks performed in spatial processing algorithms generally have a linear relationship with the amount of spatial data input; second, if the intermediate results of Map task calculation exceed a certain threshold, they need to be written to disk, and the Reduce phase needs to read the middle of Map phase calculation. As a result, it is unlikely that the Reduce task will be executed on the same computing node as the Map task, requiring data to be read from the data node that stores the intermediate results. Communication may be achieved through cluster interconnection rather than memory-to-disk transmission, which increases network communication overhead and disk I/O overhead.

Through the above analysis, the time cost of the MapReduce parallel computing model includes "spatial computing algorithm"+ "disk I/O"+ "communication overhead". And spatial analysis may produce a large number of intermediate results, such as buffer analysis, intermediate results and final results may be larger than the actual amount of input data, which requires more attention to disk I/O and communication overhead.

3.3. Spatial information service based on Hadoop

3.3.1. HTTP Agreement. Hypertext transport protocol is the most widely used protocol in the Internet, and it is also the underlying transport protocol that network services mainly use. It specifies how the client sends requests to the server and how the server responds to requests. The hypertext transfer protocol provides eight kinds of request methods, which are OPTIONS, GET, HEAD, POST, PUT, DELETE, TRACE and CONNECT respectively, the most commonly used method is GET, POST and HEAD.

3.3.2. Servlet. Servlet is a protocol independent, cross-platform server component, integrated into the server, can realize the remote dynamic loading on the network. In the architecture of Web applications, Web Server is located at the bottom of the abstract layer, which directly handles HTTP requests and responses, and Servlet has the functions of session management, lifecycle management, and basic error handling on Web Server. The workflow is shown in figure 2.

![Figure 2. Client HTTP request to Web server process](image)

3.3.3. Spring lightweight Web application framework. The Spring framework is a lightweight open source framework for layered Java SE/EE applications, created by Rod Johnson, with IoC (Inverse of Control) and AOP (Aspect Oriented Programming) as the kernel, providing a number of enterprise-level applications such as Spring MVC and Spring JDBC in the presentation layer, and Spring JDBC in the
persistence layer, and transaction management in the business layer. The framework of the technology is shown in Figure 3.

### Figure 3. A complete Web application architecture based on Spring framework

| Component | Description |
|-----------|-------------|
| **Spring AOP** | Source-level Metadata AOP infrastructure |
| **Spring ORM** | Hibernate Support IBatis Support JDO Support |
| **Spring Web** | WebApplicationContext Multipart resolver Web utilities |
| **Spring Web MVC** | Web MVC Framework Web Views JSP/VelocityEngine PDF/Export |
| **Spring Context** | Application context UI support Validation JNDI EJB support and Remodeling Mail |
| **Spring Core** | Supporting utilities Bean container |

Spring Core is the most basic part of the Spring Framework. It provides dependency injection features to manage beans. BeanFactory is the core of any Spring application; Spring Context is the Spring context that provides access to beans and adds functions for resource binding, event migration, and resource loading; Spring Web is built on The application context module provides context for Web-based applications and provides Web-oriented application integration capabilities; Spring Web MVC provides an MVC-based Web application framework that can easily be integrated with other MVC frameworks.

### 3.3.4. JAXB

JAXB (Java Architecture for XML Binding) is a component used to implement the binding of structured objects in Java to XML elements. JAXB solves the difficulty of converting Java classes and objects to XML Schemas by providing standard mapping specifications between them.

The architecture of JAXB includes three main components: the Schema generator, the Schema compiler, and the runtime binding framework. The Schema generator is responsible for generating the corresponding XML-Schema class files, and the generation rules are determined by the JAXB annotations annotated in the Java code; the Schema compiler is used to generate the corresponding Java class files from the XML-Schema, and an object factory class is responsible for instantiating the generated classes; the runtime binding framework provides serialization (marshalling) and anti-serialization (unmarshalling) operation, and is responsible for checking the contents of XML files. The steps of JAXB data binding are shown in Figure 4. The steps are as follows:
4. Concluding remarks
In the field of GIS research, cloud computing architecture encapsulates various functions of cloud geographic system (CGIS) into commodity function blocks, and then uses these encapsulated function blocks as commodities on the network for users to use on demand. Under the combination of traditional GIS and HA-DOOP, the data of GIS can be stored in the cloud computing platform with distributed database, which makes the resource sharing, improves the computing efficiency and reduces the waste of resources. Cloud computing is the result of the comprehensive evolution of a variety of technologies, and is still in the development stage. With the development of cloud computing technology, the design of cloud GIS architecture needs to be constantly improved to make it more optimized.

References
[1] Zhu Lin Yun GIS platform architecture and key implementation technology research [J]. Electronic Science and technology, 2016,03 (05): 634-637.
[2] Research on the Key Technologies of Constructing the Cloud GIS Platform [J]. Surveying and Spatial Geographic Information, 2016,39(03): 32-33+36.
[3] Chang Yan. Hadoop cloud computing and its key technology [J]. software guide, 2015,14 (09): 7-9.
[4] Zhao Xiaojuan, Chen Tao. Exploration and Research on Building Cloud GIS Platform Based on Hadoop[J]. Guangdong Science and Technology, 2014,23(20): 140-141.
[5] Fan Jianyong, long Ming, Xiong Wei. Research on cloud GIS architecture based on Hadoop [J]. surveying and mapping bulletin, 2013 (11): 93-97.