DESIGN AND PRACTICE ON METADATA SERVICE SYSTEM OF SURVEYING AND MAPPING RESULTS BASED ON GEONETWORK

ZHA Zhuhua[a], ZHOU Xu[b]

[a]Natioanal Geomatics Center of China, 28 Lianhuachi West Road, 100830, zhazh@nsdi.gov.cn

[b]Zhou Xu, Natioanal Geomatics Center of China, 28 Lianhuachi West Road,100830, zhouxu@nsdi.gov.cn

KEYWORDS：Metadata, Standards, Architecture, Web based, Performance, Distributed

ABSTRACT:

Based on the analysis and research on the current geographic information sharing and metadata service，we design, develop and deploy a distributed metadata service system based on GeoNetwork covering more than 30 nodes in provincial units of China..

By identifying the advantages of GeoNetwork, we design a distributed metadata service system of national surveying and mapping results. It consists of 31 network nodes, a central node and a portal. Network nodes are the direct system metadata source, and are distributed arround the country. Each network node maintains a metadata service system, responsible for metadata uploading and management. The central node harvests metadata from network nodes using OGC CSW 2.0.2 standard interface. The portal shows all metadata in the central node, provides users with a variety of methods and interface for metadata search or querying. It also provides management capabilities on connecting the central node and the network nodes together.

There are defects with GeoNetwork too. Accordingly, we made improvement and optimization on big-amount metadata uploading, synchronization and concurrent access. For metadata uploading and synchronization, by carefully analysis the database and index operation logs, we successfully avoid the performance bottlenecks. And with a batch operation and dynamic memory management solution, data throughput and system performance are significantly improved; For concurrent access, , through a request coding and results cache solution, query performance is greatly improved. To smoothly respond to huge concurrent requests, a web cluster solution is deployed. This paper also gives an experiment analysis and compares the system performance before and after improvement and optimization.

Design and practical results have been applied in national metadata service system of surveying and mapping results. It proved that the improved GeoNetwork service architecture can effectively adaptive for distributed deployment requirements, performance improvement and optimization of the system guarantee its continuous and stable running on the internet.

1 METADATA SERVICE

Metadata service is an important part of building geographic information data sharing service system[2], and a specific one pattern of geographic information network distribution service[5]. Europe, United States and other developed countries establish geographic information distribution service web portal through the integrating geographic information metadata input, query, management and switching nodes, to provide one-stop geographic information query, browse and access services for user[2].

OGC CSW standard is an interface realization standard of geographic information catalog service binding HTTP protocol,

it can used to publish, search metadata of spatially referenced data, service or related resources[3,8].

GeoNetwork is an open source project for geographical spatial metadata service, and it is used widely in the fields. It is an OSGeo incubation project, supportting OGC CSW 2.0.2. It is a standard based and decentralized spatial information management system, designed to enable access to geo-referenced databases and cartographic products from a variety of data providers through metadata query and access, enhancing the spatial information exchange and sharing between organisations and their audience . It can provide access service for customers with a convenient and variety of source spatial data and thematic maps. The main goal of the software is to increase collaboration within and between organisations for reducing duplication and enhancing information consistency and quality and to improve the accessibility of a wide variety of geographic information along with the associated information, organised and documented in a standard and consistent way [1,4]. It is used widely as spatial information management system in the United Nations system such as UNSDI and other international organizations like NSDI, INSPIRE and GEO, etc. Its technical features are: Java architecture, Web Service and Servlet technology, using JDBC to connect database, using XML technology for metadata, using XSLT technology to convert XML, supporting remote access and internationalization.

2 DESIGN OF NATIONAL METADATA SERVICE SYSTEM OF SURVEYING AND MAPPING PRODUCTION

National metadata service system of surveying and mapping results is based on OGC CSW 2.0.2 standard. It is a centralized metadata service system with metadata information model extended from ISO 19115/19119, called SMMD, its namespace is http://data.sbsm.gov.cn/smmd/2007. The information model references to the ISO 19115/19119 and Catalog Services Specification, we define the relationship between the attributes elements of each query and response and metadata information model elements.

The system is composed of network nodes, central node and the web portal. Network nodes are the direct data sources of system metadata, distributed in every province , each node maintain a metadata service system, responsible for metadata uploading, publishing and management. Central node can harvest metadata from network nodes, using OGC CSW interface. The portal showcases all the metadata in central node harvested from network nodes, provides web interface and a variety of ways to query, it also provides management interfaces for central node and network nodes interconnection.

All nodes in the system use GeoNetwork 2.1 as metadata service, include uniform metadata information model to manage own metadata of surveying and mapping results. The metadata service on all nodes support unified standards, metadata information management can use it, any generic CSW client can use the standard to query metadata information. The system realizes the Harvest interface for interoperability. The central node can use Harvest interface to complete near real-time synchronization, and then leads to one-stop query on the portal to achieve the national surveying and mapping results.

The portal offers a variety of ways to query surveying and mapping results, provides full-text search function, advanced search function and map-based search function. Each search function will encode request by GetRecords or GetRecordById interface in OGC CSW specification, the service system will parse request and return the results, the client web page will display these information. In addition to searching, the central node also provides network nodes registration management and metadata access. Network nodes register into central node as information source, central node get the update information of network nodes, and harvest metadata in time. At present central node manages more than 1.1 million metadata which harvested from more than 30 network nodes.

3 PERFORMANCE OPTIMIZATION

The metadata system on results of surveying and mapping is r-unning on internet, it will be influenced by hardware, software,

bandwidth, user accessing and other factors. So, we need optimize GeoNetwork from the following aspects.

3.1 Software Upgrade

Software upgrade is usually the most convenient and effective measures for performance optimization. The most important influence in GeoNetwork is the index library maintenance, including index and search. GeoNetwork 2.1 uses Lucene 1.4.3 as index tool, and Lucene upgraded to 2.9 in 2009. In the wiki of Lucene website on "how to make searching / indexing faster":" Make sure you are using the latest version of Lucene"[11].

3.2 Batch Processing

Operations on GeoNetwork`s data and index are based on single record. It is almost no influence for small amount of metadata. When the amount increases to ten thousand, several hundred thousand or even millions, the impact is very large, the system will be surprisingly slow. When the amount of metadata is large, the index is also becoming large, we operate one single metadata, like insert, update or delete operation, the system will modify the index library, and then optimize it for effective management and high speed search on index. It will take several minutes to complete the operation on single metadata. This kind of operation involves data batch upload, data publish and sites synchronization, like CSW harvest.

Batch processing is a effective and time saving solution to resolve this kind of repeat operations. Each operation first writes the modified metadata to database, and records the metadata id, when all metadata writing complete, the system will rebuild these metadata index once. It will save a lot of time.

3.3 Cache

Cache technology has been considered one of the effective way to reduce server load, network congestion and customer accessing delay[9]. In the field of geoinformation service, web cache technology is also widely used. Each big electronic map website, use tiles based cache technology for map service. A large number of cache using in client side and server side to avoid map redraw on map server. It consumes the processing time for request to the server, and enhance the client`s response. OGC also release WMTS 1.0.0 implementation standard, which can be used to develop scalable and high performance services which WMS can not.

Cache technology can be used in metadata service system. On one hand, like in Table 1, the number of user querying is times more than system metadata and index update. On the other hand, users are usually compare the query results, even repeat query, so the results are repeatable.

Date Range	Query	Udpate
20100101-20101231	121655	29
20110101-20110630	137950	2

Table 1. System query and update frequency

We design the result cache technology for GeoNetwork based on database. When the system gets the first query request, it performs coding algorithm(such as MD5 algorithm), the query string encoded as a unique value, then writes query string, coding value and query result into database. When server gets the same request again, it encode the query string to a value, find the value in the database, and returns result as response. Here we can build index for the encoded value, it is unique, to speedup query and select efficiency.

3.4 Web Cluster

Web cluster technology is the important method in solving the capacity and scalability of web server system[10]. Dispatcher based request dispatching mechanism is our metadata service system's load balancing mechanism.

The metadata service system on surveying and mapping results run on a "4+1" service cluster, shown in Figure1. The system is deployed on a Dell PowerEdge R900 Server(Intel® Xeon® CPU E7420,2.13GHz X 16 CPU, 64G memory), we build 5 virtual machine, which 4 for normal use, 1 as a backup, when any one of the 4 normal crashing, the 1 backup will be instead.

The metadata service system use dispatcher based request dispatching mechanism. The front-end node server uses Nginx as request dispatcher which is a reverse proxy server. As the service system and portal use session for service, Nginx uses ip_hash as load balancing mechanism. Each request will be dispatched to a fixed server by Hash result of access IP, so it can effectively solve the session problem.

The advantages of this technology are: it can ensure the system performance and service capabilities, it is extensible, it can overcome the Java limits on a single machine. System service capability is related to the number of machine in cluster. The disadvantage is that the background data synchronization is more complex, we need synchronize several times.

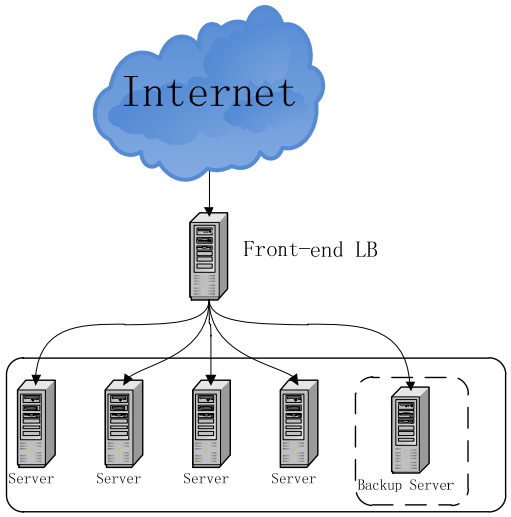

Figure1 “4+1” cluster

3.5 Crash Handling

The crash handling here means that the metadata service system has software crash or can`t serve normal, the natural disasters, infrastructure failure and other human factors are not involved. GeoNetwork uses JDBC to operate database, when the number of metadata is large in one operation, memory overflow may be happen, it can leads to system crash or can`t respond to requests; user concurrent access may also lead to system can`t support, it is needed to design an appropriate program to help the system return to normal state as soon as possible.

The metadata service system uses monitoring keywords for restarting service method to restore. GeoNetwork uses Wrapper to install as a Windows service, we can use filter mechanism on Wrapper. In the filter, we can use monitoring keywords as trigger string, like”RESTEART NOW”, and the trigger action can set to service restart. After the system running, we can throw the keywords when we need, it can be monitored by Wrapper, and then Wrapper can restart the system. Our system based on GeoNetwork can throw the keywords when catch the memory overflow exception, and Wrapper monitors the string, trigger the filter, and act to restart the service system. This method can also be used for service remote management, like restart to apply new settings.

Actually previously mentioned “4+1” model is also a kind of crash handling scene. When any one of the 4 normal server crashed, the front-end dispatcher can monitor and dispatch new request to the 1 backup server, so the cluster can remain stable service capability.

3.6 Experiment

To verify the effectiveness of the performance optimization solutions, we have an experiment on Dell Precision T3400(OS: Windows XP sp3, JVM parameter is set to “-Xms48m –Xmx1024m”). We select GeoNetwork 2.1 and our improved software (called GeoNetwork+) to compare efficiency.

First,we choose the metadata batch import function to compare. Metadata from the metadata service system of surveying and mapping results are extracted to import into the two software, the system efficiency is compared in Table 2. As we can see in the table, first, the efficiency of batch import function through optimized software by batch processing compared to the original GeoNetwork 2.1 can increase 2-10 times, and the greater the amount of metadata, the higher the efficiency; second, in GeoNetwork 2.1 metadata batch import function, its time consumption growth rate is far greater than the amount of data, while the GeoNetwork+ on the contrary, the time consumption rate is less than the amount of data about growth rate. So, GeoNetwork+ will be more responsive to the amount increasing of metadata size.

Software\amount	100(S)	1000(S)	10000(S)	100000(S)
GeoNetwork2.1	5.62	68.074	2313.427	*
GeoNetwork+	3.692	31.182	300.411	2980.333

Table 2 upload efficiency contrast on different amount of metadata

*means can not be imported on the amount of metadata

4 SUMMARY AND FUTURE DIRECTIONS

With the increasingly strong demand on public services of surveying, mapping and geoinformation, to efficiently and comprehensively display and query their metadata is the important way to apply and serve the results of surveying, mapping and geoinformation. Based on the architecture of open source catalog application known as GeoNetwork managing spatially referenced resources, we design and development further a national metadata service system on surveying and mapping in this paper. As an internet application, we also analysis system performance bottlenecks, optimize the function efficiency and load capacity, propose performance optimization solution, efficiently improve the system performance. Next we can improve the system performance on these fields: memory cache and distributed cluster technology and so on.

5 REFERENCES

[1] Jeroen Tichler, Jelle U. Hielkema. GeoNetwork opensource Internationally Standardized Distributed Spatial Information Management [J].OSGeo Journal.2007(2)

[2] Gong Jianya，Du Daosheng，Gao Wenxiu，Xu Feng，Zhou Xu. Technology and Standards of Geographic Information Sharing[M]. Beijing：Science Press，2009

[3] Zhuhua，ZHOU Xu，LIU Ruomei, JIA Yunpeng, LU Ping.On the Realization of Specification OGC CSW[J]. Bulletin of Surveying and Mapping，2009（7）：12-15

[4] GeoNetwork OpenSource, User Manual 2.6.4 [OL]. http://geonetwork-opensource.org/manuals/2.6.4/users/index.html

[5] JIN Zhi-guo, SHOU Chun-fa, LI Cheng-ming, YIN jie. A discussion of the mode of urban geoinformation distribution service based on network [J]. Science of Surveying and Mapping. 2008,33(6):196-198

[6] DU Yunyan, FENG Wenjuan, HE Yawen, XIAO Rulin. Geographic Information Services Integation with Web Services[J]. Geomatics and Information Science of Wuhan University. 2010,35(3):347-349

[7] P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, and F. Farazi. A semantic geo-catalogue implementation for a regional SDI[OL]. http://disi.unitn.it/publications/16959

[8] OGC. OpenGIS Catalogue Service Implementation Specification[OL]. http://www.opengeospatial.org/standards/cat

[9] HE Chen, CHEN Zhao-xiong, HUANG He-yan. Summary of Web Caching Technology [J]. MINI-MICRO SYSTEMS. 2004,25(5):836-842

[10] Li Shuangqing, Gu ping, Cheng Daijie. Analysis and Research on Load Balancing Strategy in Web Cluster System [J]. Computer Engineering and Applications. 2002(19):40-42

[11] Lucene FAQ[OL], http://wiki.apache.org/lucene-java/LuceneFAQ.