HIERARCHICAL OPTIMIZATION MODEL ON GEONETWORK

ZHA Zhuhua[a, *], JIANG Jie[a], ZHOU Xu[a]

[a] National Geomatics Center of China, 28 Lianhuachi West Road, 100830, Beijing, China - (zhazh, jjie, zhouxu)@nsdi.gov.cn

Commission IV, IV/5

KEY WORDS: GeoNetwork, HOM, Performance, Metadata; Web based; Performance;

ABSTRACT:

In existing construction experience of Spatial Data Infrastructure (SDI), GeoNetwork, as the geographical information integrated solution, is an effective way of building SDI. During GeoNetwork serving as an internet application, several shortcomings are exposed. The first one is that the time consuming of data loading has been considerately increasing with the growth of metadata count. Consequently, the efficiency of query and search service becomes lower. Another problem is that stability and robustness are both ruined since huge amount of metadata. The final flaw is that the requirements of multi-user concurrent accessing based on massive data are not effectively satisfied on the internet. A novel approach, Hierarchical Optimization Model (HOM), is presented to solve the incapability of GeoNetwork working with massive data in this paper. HOM optimizes the GeoNetwork from these aspects: internal procedure, external deployment strategies, etc. This model builds an efficient index for accessing huge metadata and supporting concurrent processes. In this way, the services based on GeoNetwork can maintain stable while running massive metadata. As an experiment, we deployed more than 30 GeoNetwork nodes, and harvest nearly 1.1 million metadata. From the contrast between the HOM-improved software and the original one, the model makes indexing and retrieval processes more quickly and keeps the speed stable on metadata amount increasing. It also shows stable on multi-user concurrent accessing to system services, the experiment achieved good results and proved that our optimization model is efficient and reliable.

1. INTRODUCTION

In Spatial Data Infrastruture(NSDI) implementation fields, metadata service is an important part , it can be used to build geographic information data sharing service system, and a specific pattern of geographic information network distribution service (Jin et al., 2008). 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(Gong Jianya,2009).

1.1 GeoNetwork

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, supporting 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 organizations 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 organizations 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, organized and documented in a standard and consistent way (Jeroen Tichler,2007). 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.

There are some shortcomings are exposed when using it as a web application which has huge users. During GeoNetwork serving as an internet application, several shortcomings are exposed. The first one is that the time consuming of data loading has been considerately increasing with the growth of metadata count. Consequently, the efficiency of query and search service becomes lower. Another problem is that stability and robustness are both ruined since huge amount of metadata. The final flaw is that the requirements of multi-user concurrent accessing based on massive data are not effectively satisfied on the internet.

1.2 Aims

A metadata service system need be constructed as SDI part, It is running on internet, has more than 30 nodes, the size of metadata for loading is more than 1,000,000, publishing and searching, We use system to harvest metadata of surveying and mapping results in China, and serves for people. Users can search results of surveying and mapping which they interested on this system, they can know how to get the result, where it is and call the number showed in the metadata.We use GeoNetwork to build our metadata service system as our system prototype.

GeoNetwork has some shortcomings when using as an internet application. The first one is that the time consuming of data loading has been considerately increasing with the growth of metadata count, because it loads metadata and builds index one by one, as building and optimization index may take

more time. It even can not load metadata more than 100,000 one time, because it may result in"out of memory" error .
Consequently, the efficiency of query and search service becomes lower. GeoNetwork use lucene index engine for querying and searching, but optimization and acceleration strategy are not enough. When the metadata amount is big enough, the query efficiency become slow. Meanwhile, stability and robustness are both ruined since huge amount of metadata.
The final flaw is that the requirements of multi-user concurrent accessing based on massive data are not effectively satisfied on the internet.
The original GeoNetwork 2.1 can not satisfy these requirements, so we need an optimization solution to improve.

2. HIERACHICAL OPTIMIZATION MODEL

Hierachical optimizatioin model(HOM) consists of software level and deploy level. The software level means the methods can be taken in some software, maybe GeoNetwork itself. The deploy level means that these methods can be taken when deploy the metadata service system . The software level needs modify the GeoNetwork project source code, it is inside. The deploy level needs construct a smart web deploy solution and uses some specific software to get some excellent function, like disaster recovery. It is outside.

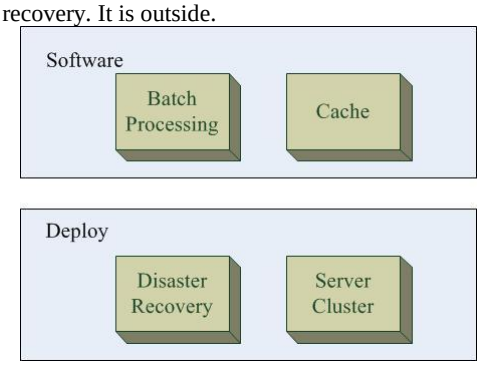

Figure 1.HOM

2.1 Software Level

1.Batch Processing
GeoNetwork`s data and index Operations 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, the operation on one single metadata, like inserting, updating or deleting, 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.
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.

2 .Cache
Cache technology has been considered one of the effective way to reduce server load, network congestion and customer accessing delay(HE Chen,2004).In the field of geo information 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, the number of user querying is times more than system metadata and index updating. On the other hand, users are usually compare the query results, even repeat query, so the results are repeatable.
We design the result cache technology 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.

2.2 Deploy Level

1.Web Cluster
Web cluster technology is the important method in solving the capacity and scalability of web server system(Li Shuangqing,2002).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 Figure2.The system is deployed on a hardware server, 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 system uses 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.

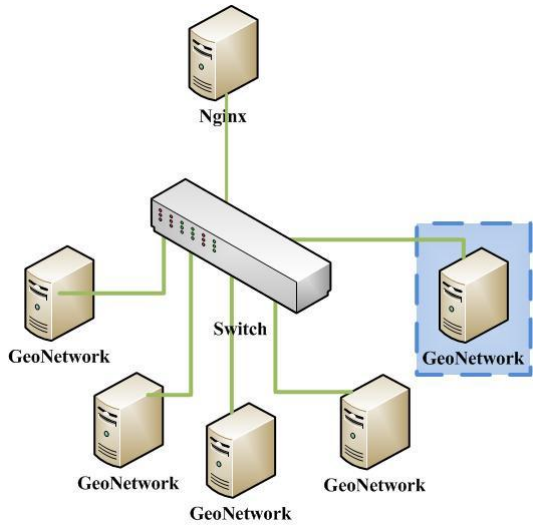

Figure2.Web cluster

2.Disaster Recovery

Disaster recovery here means that the metadata service system crashed or can`t serve normal, the natural disasters, infrastructure failure 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 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 disaster recovery 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. RESULTS

An experiment on Dell Precision T3400(OS: Windows XP sp3, JVM parameter "-Xms48m –Xmx1024m") has been done to verify the effectiveness of HOM-improved solution. The system efficiency is compared in Table 1 and Figure 3.

The header of the table is the volume of metadata, and the unit is seconds, which means the computer need the time to finish the metadata volume.

Software\volume	100(S)	1000(S)	10000(S)	100000(S)
GeoNetwork2.1	5.62	68.074	2313.427	x
HOM-improved	3.692	31.182	300.411	2980.333

x: can not be imported one time on the amount level

Table 1. Import efficiency contrast on different amount of metadata

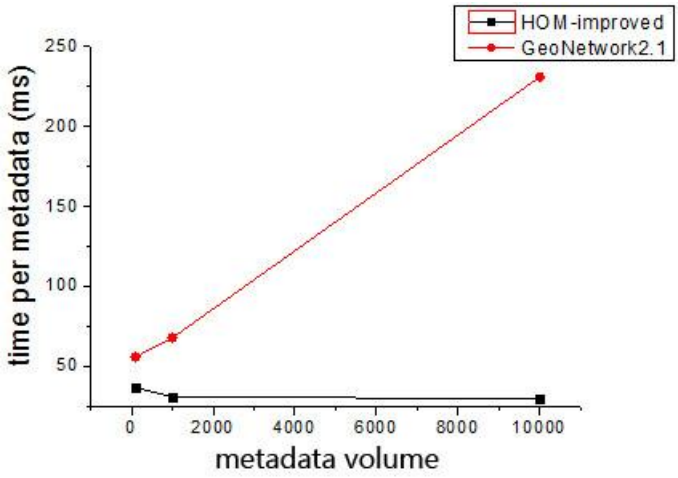

Figure 3. Time per metadata contrast on data importing

From Table 1, we can see the efficiency of batch import function through HOM-improved software compared to the original GeoNetwork 2.1 can increase 2-10 times, and the greater the amount of metadata, the higher the efficiency.

From Figure 3, in GeoNetwork 2.1 metadata batch import function has time consumption growth rate is far greater than the amount of data, while the HOM-improved on the contrary, the time consumption rate is less than the amount of data about growth rate.

In summary, HOM-improved solution will be more responsive to the amount increasing of metadata amount, it is adaptive to our metadata service system.

4. CONCLUSIONS

GeoNetwork as a geographical spatial metadata service, can used to publish, search metadata, is the base software for our metadata service system. The Hierarchical Optimization Model has been presented for preventing the original GeoNetwork 2.1 shortcomings when serving as an internet application. Based on the HOM-impoved solution, we break through the bottlenecks, efficiently improve the function efficiency , load capacity and the system performance. Next we will submit our model and source code to GeoNetwork project.

REFERENCES

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

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

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

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

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