Research on Association and Search Services of Massive Geospatial Information Based on Cloud Computing

Huijuan ZHANG[a], Zihui SONG[b], Hongyuan ZHU[c], Fuqing ZHANG[d]

Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, Beijing, 100101, China

[a]email: zhanghj@digitalearth.cn, [b]email: songzh@digitalearth.cn, [c]email: zhuhy@digitalearth.cn, [d]email: zhangfq@digitalearth.cn

Keywords: Cloud Computing; Massive Geospatial Information; Association; Search Services

Abstract. Adapt to the global strategy of the state, the paper builds a geospatial information service platform with a global coverage, high resolution, multi-subject, and the national global strategy formulation and implementation, to provide the fusion and strategic decision-making ability of massive dynamic global geospatial information, and provide an independent global geospatial information base for the country's global strategy development, deployment and the national life. To solve the limitations of system performance, stability, expansibility, etc. of the traditional single node architecture of geospatial information service with mass data storage, multi-user concurrent access, through cloud computing of the fusion of the collection computer cluster technology, data storage technology and high-performance network services, the paper achieves association and search services of massive geospatial information, and provides the basis for establishing a geospatial information service platform with the global coverage, high resolution, multi-subject and adapt to the global strategy formulation and implementation.

Introduction

Adapt to the global strategy of the state, we need a geospatial information service platform with a global coverage, high resolution, multi-subject, and the national global strategy formulation and implementation, to provide the fusion and strategic decision-making ability of massive dynamic global geospatial information, and provide an independent global geospatial information base for the country's global strategy development, deployment and the national life [1,2]. So we begin to research the association and search service of massive geospatial information to establish a geospatial information service platform with the global coverage, high resolution, and multi-subject and adapt to the global strategy formulation and implementation through cloud computing of the fusion of the collection computer cluster technology, data storage technology and high-performance network services [3].

System Architecture and Technology

The data scale of global mass spatial data services to meet small and medium scale processing of spatial data ranges from GB or TB-level scale to dozens of TB, hundreds of TB and PB grade scale. Building a global mass spatial data service cluster, eliminating mass spatial data storage and high performance bottleneck of the spatial information service, achieving high-performance spatial information services [4]. The overall technical routes as shown in Figure 1.

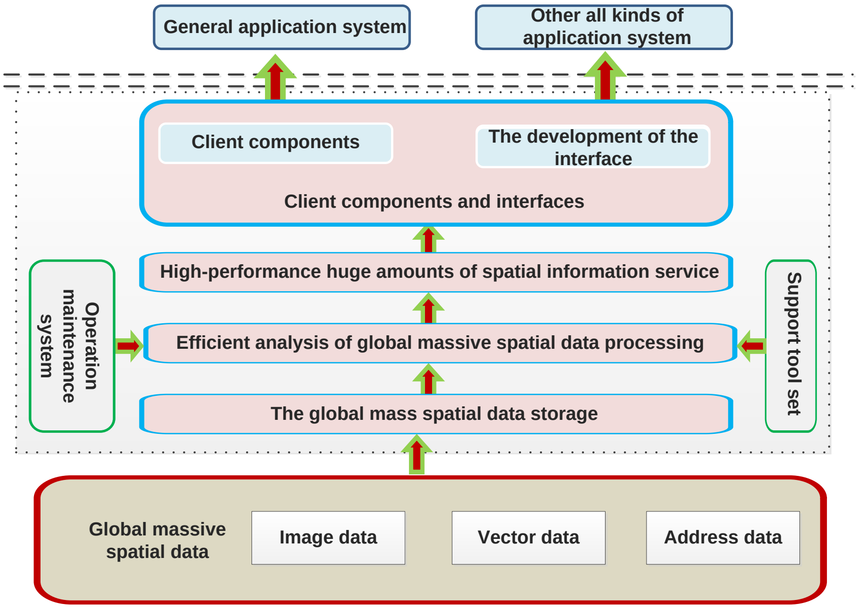

General application system
Other all kinds of application system
Client components
The development of the interface
Client components and interfaces
High-performance huge amounts of spatial information service
Efficient analysis of global massive spatial data processing
The global mass spatial data storage
Operation maintenance system
Support tool set
Global massive spatial data
Image data
Vector data
Address data

Fig.1. The overall technical route of geospatial information cloud platform

Massive place names and addresses are associated with map, to resolve the addresses association with an address spatial position, realize space search problem of address data. Uses the standard address space database and address global unified address space resources spatial database, database, address acquisition, address updates, database management services, sharing five features such as Exchange and operational monitoring [5]. The operation and maintenance process as shown in Figure 2.

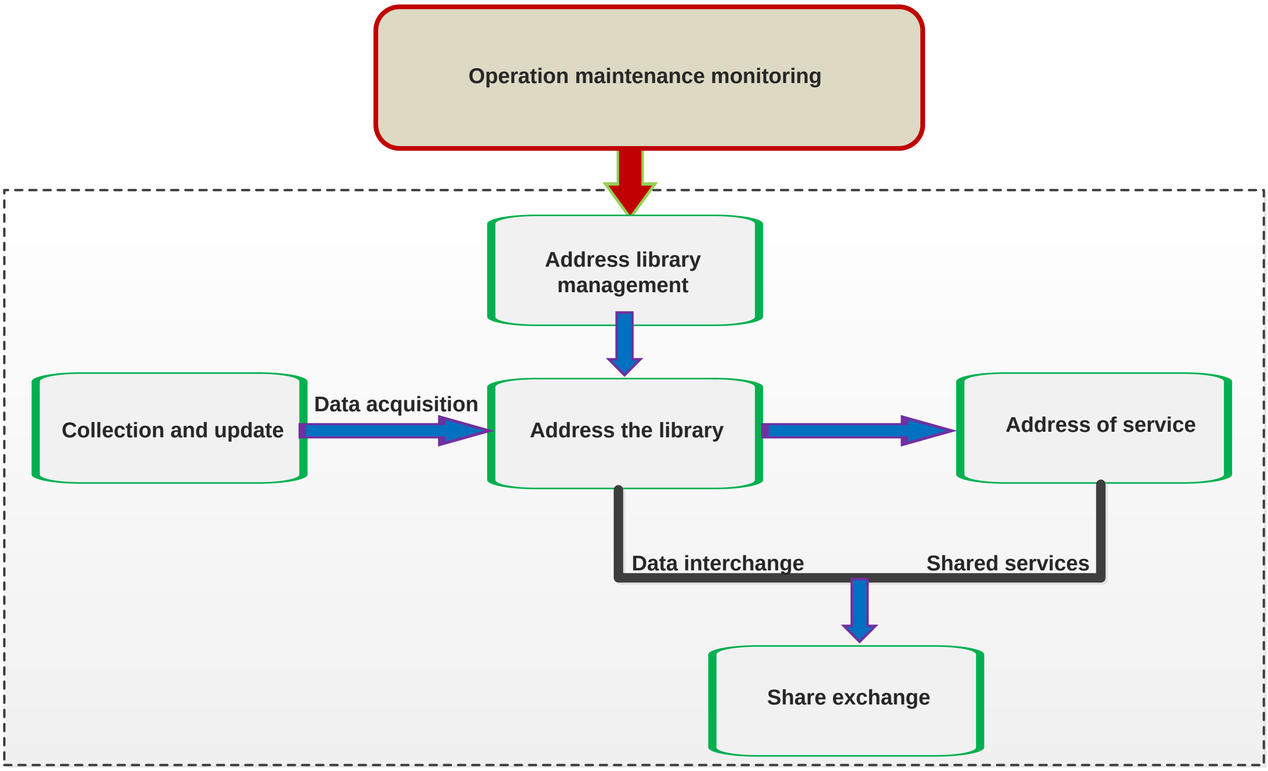


Fig.2. The operation process of address library system

Combining multiple computers by Computer Supported Cooperative Work, acquires far more high performance than a single server on storage, calculation, and the limits of capacity. Because cluster by working together to obtain high performance also led cluster system has a complex and complicated procedures, algorithms, complex problems [6].

HA clusters from the perspective of system stability, addressing cluster node failure failed takeover and failover in case of problems. HA clusters are used for critical databases, file sharing, software applications, network services, and other areas.

LB through the load balancer integrates back-end nodes, forming a giant virtual computer. Load balancer to collect and monitor the load on the back-end nodes (CPU load, network load, disk load, etc.) LB cluster is typically used for high user access to database and Web server load.

Through the design of parallel algorithms for specific HPC cluster, part to break large tasks into subtasks, assign to calculate multiple compute nodes at the same time, shorten calculation time needed to solve the whole problem. Unlike LB cluster tasks of HPC cluster correlation between child tasks, are not entirely independent. HPC can break through the performance limits of a single computer, to solve the bottleneck of a single high-end computer system.

Data storage is the basic technology of information systems. In recent years, along with the rapid development of Internet technology, rapid data accumulation and rapid data growth problem her promote the development of data storage technology. Common data storage technologies including disk array technology (Redundant Array of Independent Disks, RAID), direct-attached storage technologies (Direct Attached Storage, DAS), network storage (Network Attached Storage, NAS), storage area network (Storage Area Network, SAN), as well as distributed storage technology.

By via RAID Controller (Hardware, Software) , combining RAIDs into a single large-capacity hard drives, data mirroring, data striping technology, spreading the data on the disks in the RAID. RAID breaks through single disk volume limitation, implementing a virtual high-capacity disk. DAS connects directly with the server. I/O requests are sent directly to the storage device. This way is used for alone or two small clusters of servers. NAS uses file access protocols (for example, NFS, CIFS) to access the data based on server through the TCP/IP network. SAN uses Fibre Channel Protocol (Fiber Channel, referred to as FC), solves the underlying transport protocol-layer protocol is still using the SCSI protocol.

Distributed storage is a P2P based distributed storage, and the other is cluster storage. Distributed Google File System storage solutions include distributed file systems, and Google BigTable a distributed database system.

Web service of high performance mainly solved the performance bottleneck in web service, and achieved web service of high performance through extending the storage, transfer and computing resources. Cooperative work of multiple cluster nodes to improve the whole performance of system, and heart rate monitoring technology to monitor the operation of the cluster nodes, and the failure takeover technology to implement the failure node storage and the transfer of computing, ensured that the services still ran well under the case of the node failure. Cloud computing realized the virtualization of the server through virtualization technology, which can provide vast amounts of data storage, computing and service resources.

Geospatial Information Matching Engine

The framework of address library system is divided into three layers, namely, application layer, service layer and data layer. Application layer mainly provides for specific business application systems of related industries address, including portal system and address application management template sets, and the service interface for all kinds of business application based on the system. Service layer is responsibility to realize the "high cohesion and loose coupling" of the system, using coarse-grained remote interface to minimize communications between presentation layer and business layer, including service sets, tool sets and core functional components. Among them, tool sets contain data management and data mining tools and operational management system; the address data management and collection subsets constitute the core functional components, and each subset is made up a number of components. Data layer includes data storage and data services, including the two entities of standard address geospatial database, address resources database. As shown in Figure 3.

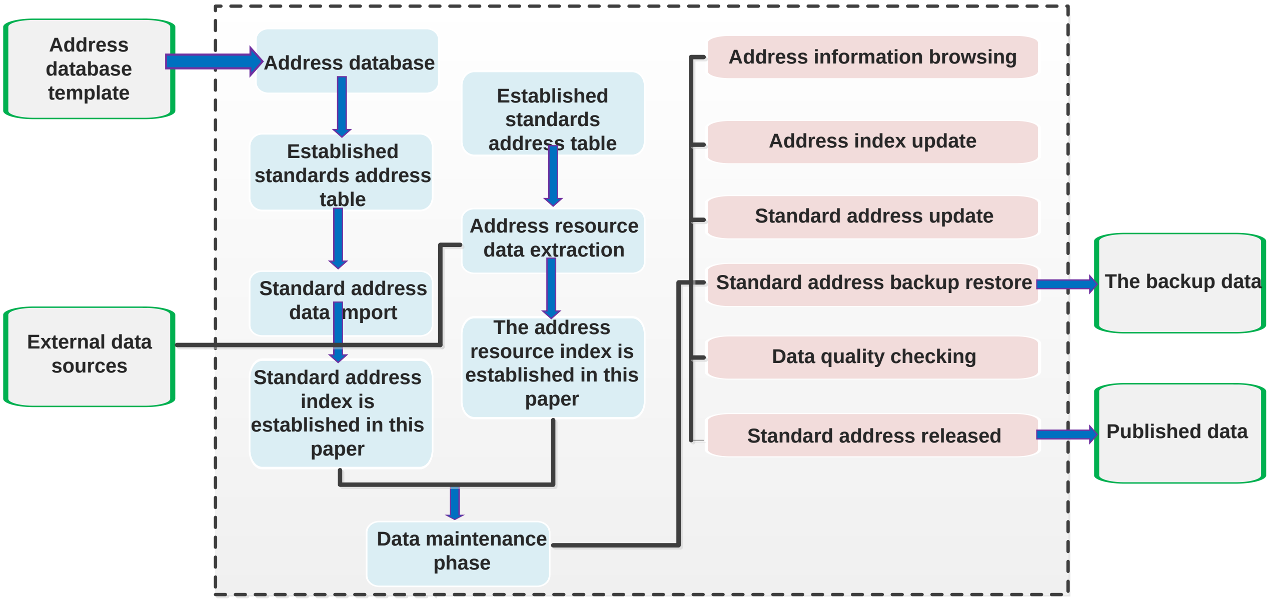

Address database template
Address database
Established standards address table
Standard address data import
External data sources
Standard address index is established in this paper
Established standards address table
Address resource data extraction
The address resource index is established in this paper
Data maintenance phase
Address information browsing
Address index update
Standard address update
Standard address backup restore
Data quality checking
Standard address released
The backup data
Published data

Fig.3. Geospatial information matching engine technology flow chart

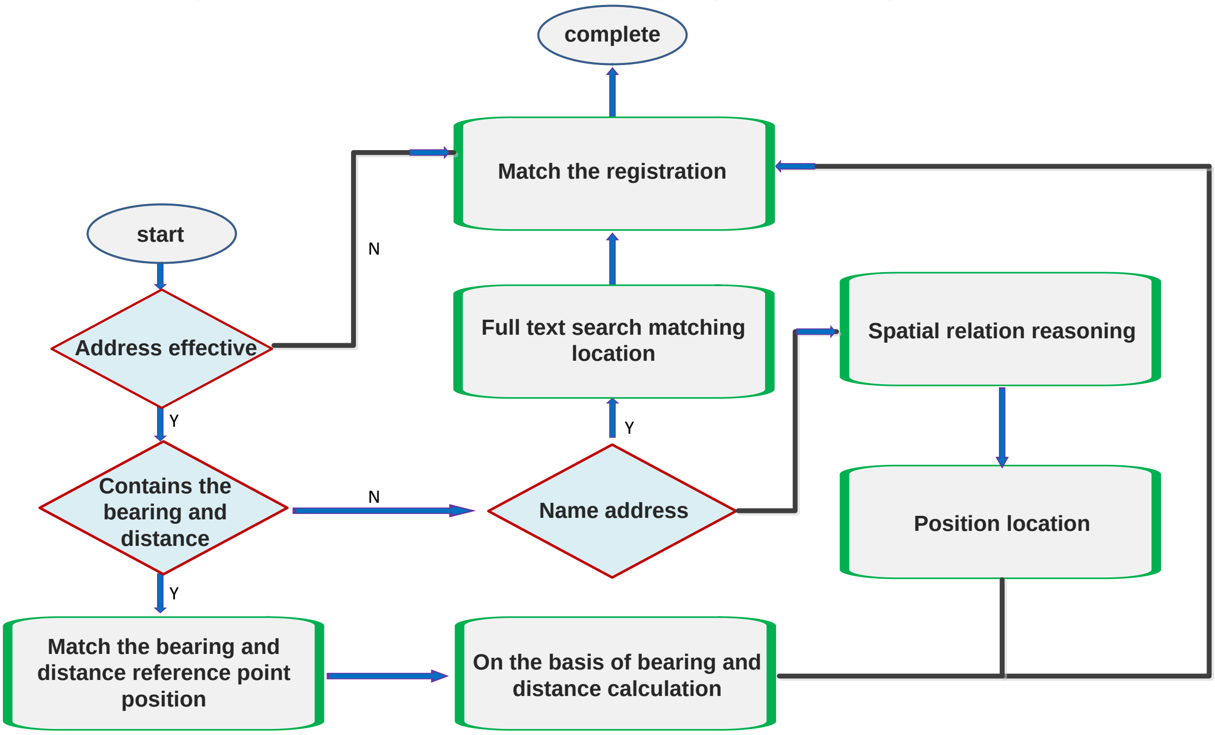
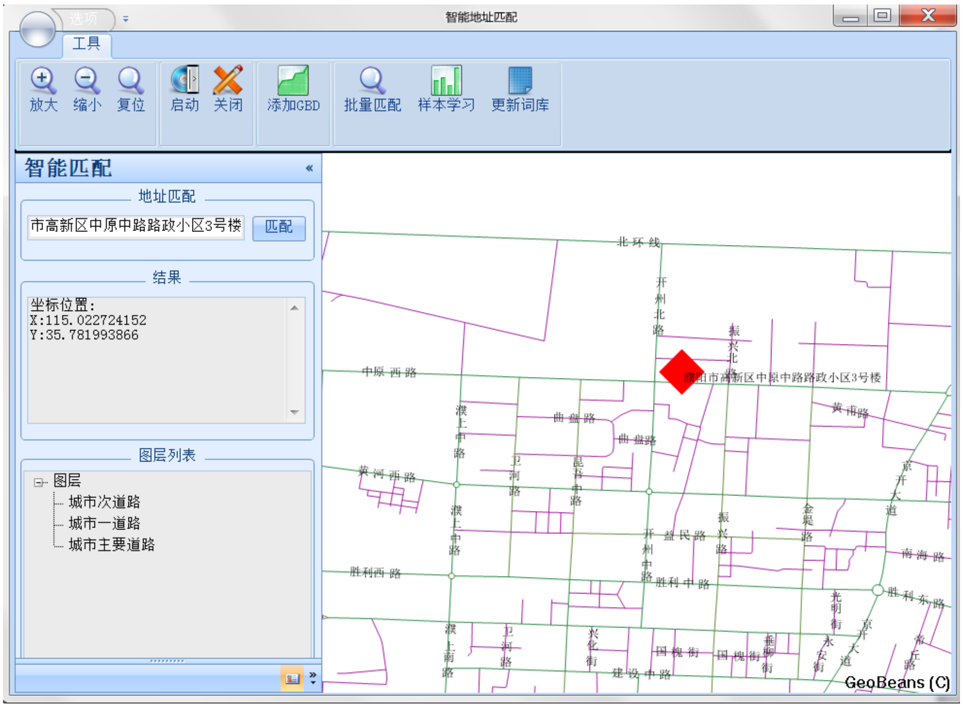

complete
Match the registration
start
N
Address effective
Full text search matching location
Spatial relation reasoning
Y
Y
Contains the bearing and distance
N
Name address
Position location
Y
Match the bearing and distance reference point position
On the basis of bearing and distance calculation

Fig.4. Geospatial information matching algorithm process

Address matching engine of the service layer was developed using natural language reasoning algorithm with heuristic strategy, and knowledge (address elements relationship) was stored using the implicit memory to solve the problem with implicit graph search method. In order to reduce the size of the search, backward inference strategy was used. The core of intelligent matching is reasoning based on the knowledge and finding the optimal solution in knowledge base. The knowledge base of the algorithm refers to the address database. The final result of knowledge reasoning process using depth first strategy is a path from a root node to a leaf node. Because the address database records the pointer (the parent node id) from child nodes to parent nodes, the complexity of the depth first search is significantly reduced, which only need to find the local path of the current child node as the starting point and the parent node as the end. If the path exists, then it shows the transfer from the parent node to the current node is successful, and deep search can continue; but if the path does not exist, exception handling needs to do, as shown in Figure 4.



市高新区中原中路路政小区3号楼

坐标位置:
X:115.022724152
Y:35.781993866

Fig.5. Visualization matching of geospatial information

Table 1 Spatial information matching ability testing results schedule

number	Matching adress	Data quantity	Handware index	performance
1	Haidian, Beijing	34663	Common desktop computer(2 GB memory，CPU Intel E6550，hard-disk7200rps）	Matching rate 98.7%、Accurate law 93.8%,avgerage every hands address matching consuming time 0.25 seconds.
2	Lhasa Tibet	20551		Matching rate 99.1%、Accurate law 94.6%,avgerage every hands address matching consuming time 0.22 seconds.
3	Puyang, Henan	15336		Matching rate 97.9%、Accurate law 93.5%,avgerage every hands address matching consuming time0.20 seconds.

Prototype system mainly includes the function modules of management of the place name dictionary library, address sample training, and the visual matching. The management module of place name dictionary library is used to manage place name dictionary, solving the editing problems of the address resolution thesaurus; the module of address sample training for training model parameters, resolves the training of the address type annotation model (HMM); the visualization module of matching is used to the interactive match, and retains the corresponding geographic coordinates of the address using the address matching algorithm based on natural language understanding, and visually displays them in the map, as shown in Figure 5. By the address matching performance tests of Beijing, Lhasa and Puyang, the results are shown in Table 1.

Conclusions

The traditional geospatial information service architecture consists of single-node application server and database. To solve the limitations of system performance, stability, expansibility, etc. of the traditional single node architecture of geospatial information service with mass data storage, multi-user concurrent access, through cloud computing of the fusion of the collection computer cluster technology, data storage technology and high-performance network services, the paper achieves association and search services of massive geospatial information, and provides the basis for establishing a geospatial information service platform with the global coverage, high resolution, multi-subject and adapts to the global strategy formulation and implementation.

Acknowledgement

In this paper, the research was sponsored by the National Science and Technology Support Plan (project No. 2015BAJ08B02) and Industry Database Management Software and Remote Sensing Data Comprehensive Service Platform Construction Project (Project No. 06-Y30B04-9002-13/15).

References

[1] Information on http://www.EECS.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf.

[2] Wang Jianzong, Wan Jiguang, Liu Zhuo, et al. Data Mining of Mass Storage Based on Cloud Computing [C]. International Conference on Grid and Cooperative Computing [D], 2010:426-431.

[3] Wu Yu, She Kun, Zhu Williams, et al. A Web Text Filter Based on Rough set Weighted Bayesian[C]. 8th IEEE International Symposium on Dependable [D]. Autonomic and Secure Computing, 2009: 241-245.

[4] Hillol Kargupta, Jiawei Han, Philip S. Yu, et al. Proceedings of Next Generation Data Mining [J]. Taylor and Francis, 2008, 218-236.

[5] Chu Wang, Qian De-pei. Pattern Oriented Software Development for Software Reuse [J]. Acta Electronica Sinica, 2005, 12A:2357-2359.

[6] Xiao Han. Research on Software Development Approach based on Reusable Component [J]. Microelectronics, 2007, 24(1):176-179.