RESEARCH AND DESIGN OF INSPECTION CLOUD PLATFORM FRAMEWORK FOR SURVEYING AND MAPPING PRODUCTS

Zhang Libo[1] Chen Hui[2] Li Zheng[2]
(1.National Quality Inspection and Testing Center for Surveying and Mapping Products, Beijing 100830, China; 2. Information Center, Ministry of natural Resources, Beijing 100830, China)

KEY WORDS: Quality Inspection; Inspection Cloud Platform; Platform Framework; Massive Spatial Data; Cloud Computing

ABSTRACT:

With the continuous improvement of modern surveying and mapping technology and with the plentiful of achievements, traditional quality inspection software for single machine, single task and single data type, difficult to massive multi-source isomerization achievements, difficult to meet the requirement of rapid, accurate and efficient quality inspection. With the development of IT technology such as cloud computing, big data and artificial intelligence, the quality inspection software needs to combine cloud computing technology with quality inspection business, refactoring software framework. Facing to the storage and spatial query requirement of inspection for surveying and mapping products, the paper researches and designs the spatial data distributed storage and the spatial data distributed index in cloud platform. The Management of inspection rule is the core in cloud platform. Inspection rule is the minimum operating independent unit, which becomes inspection item by parameterization, the paper builds full run-time operating mechanism in cloud platform for inspection rule. Finally, Combining the inspection requirement for surveying and mapping products and business, the paper researches and design the cloud framework for surveying and mapping products.

1. Introduction

Geographic Information Quality Inspection Software for Surveying and Mapping has experienced tool development, system construction and platform evolution, it has become an important tool and means for the quality control of Surveying and Mapping Geographic Information Products. With the arrival of the big data, new requirements for the software are put forward in data utilization, response speed, data security and energy consumption. Data reuse, mutual reference, rapid response of quality inspection efficiency, timeliness and security of data transmission are facing new challenges. Developing cloud platform for surveying and mapping geographic information quality inspection has become an important way to solve these problems. At present, IT Domain has provided a relatively perfect solution and framework technology for cloud platform construction. Hadoop, Storm, Spark and other computing engines provide a new distributed computing framework for large-scale data analysis and processing. With scalability, memory-based computing characteristics, and the ability to read and write data in any format directly on Hadoop, Spark is more convenient and efficient in large-scale data processing, and it has widespread recognition and support. But, mature solution in the field of surveying and mapping geographic information is not exist, and the same situation was happened in the field of surveying and mapping geographic information quality inspection. Based on

the analysis of the operation problems of current surveying and mapping geographic information quality inspection software, this paper combines the quality inspection business requirements with cloud computing, researches and designs a cloud platform for surveying and mapping geographic information quality inspection based on Hadoop framework, Spark computing engine and distributed database MongoDB.

2. Problem Analysis

2.1 Spatial data storage and management

Storage and management of geospatial information data has experienced File-RDBMS, RDBMS, ORDB, OODB.(Chen GuoPing, 2013) In the era of huge data with increasing demand for geographic information data, the above four methods can't solve the problem of storage and management of massive spatial data well. Distributed database and distributed file system came into being. Massive spatial data was decomposed into smaller data blocks and distributed to each node for storage, which greatly shortened the time of data storage process.

2.2 Spatial data indexing

Spatial data indexing arises at the historic moment in spatial database technology. Its main goal is to speed up the system's retrieval of spatial data. It can improve the retrieval speed by reducing irrelevant data or clustering related data. Dozens of spatial indexing methods have been developed, including B-tree-based index, binary-tree-based index, Hibert-based spatial filling curve index and hashing-based index. With the emergence of distributed computing and distributed storage technology, distributed spatial data index emerges as the times require. Distributed spatial index is a spatial index established for nodes in a cluster of distributed storage systems. Each node's data block holds the spatial index of the corresponding data block. Research on distributed spatial index still faces some difficulties, such as process scheduling of name nodes, uneven distribution of spatial data, data load balancing, inefficient data operation, etc.

3. Related technologies

3.1 Hadoop

Hadoop is a distributed system infrastructure developed by the Apache foundation. Users can develop distributed programs without knowing the underlying details of the distribution. Hadoop makes full use of the power of clustering for high-speed computing and storage, which is reliable, efficient and scalable. The core design of Hadoop is HDFS(Hadoop Distributed File System) and MapReduce. HDFS provides storage for massive amounts of data, MapReduce provides computation for massive amounts of data. HDFS is designed for deployment on inexpensive hardware and provides high throughput to access application data.

3.2 Spark

Spark memory computing framework was born in 2009 at AMPLab laboratory at the university of California, Berkeley. The target of the framework design and development is one stack to rule them all, which is completing a variety of big data analysis tasks in a set of software stack. Spark is not part of the Hadoop ecosystem strictly, it is an open source cluster computing environment similar to Hadoop, but actually recognized as a complement to Hadoop. The main feature of Spark is to provide a clustered distributed memory abstraction RDD(Resilient Distributed Dataset). RDD is an immutable collection of partitioned records, which is the programming model of Spark. The model provides operations on two types of RDDs: transformation and action. Transformation is used to define a new RDD, including map, flatMap, filter, union, sample, join, groupByKey, cogroup, ReduceByKey, Cros, sortByKey, mapValues etcs. Action is a result for return, including collect, re-duce, count, save, lookupKey.(Apache Spark, 2015)

3.3 MongoDB

MongoDB is a distributed file system-based NoSQL open source database project that uses loose storage similar to JSON to store more complex data types, with indexing, sharding, load balancing, aggregation, and more(Zhou Yao, 2018). For spatial data, MongoDB provides a series of indexing and query mechanisms based on 2dsphere and 2d for geospatial data, enabling indexing, storage, and fragmentation of large-scale geospatial data.

Spark's advantage lies in its efficient memory computing power. MongoDB is mainly used for distributed storage and query of big data. In order to give full play to their respective advantages, the middleware MongoDB Connector for Spark is used to realize the seamless connection between the computing engine and the database. MongoDB is used instead of HDFS and its advantages in aggregation and auxiliary indexing are fully utilized to provide Spark computing engine. When performing data analysis, Spark first extracts and filters data through MongoDB, and only reads the data required by Spark operation, improves memory usage efficiency while eliminating data redundancy, and improves Spark memory computing power.

4. Key technology

4.1 Massive vector data storage technology

In order to solve the problem of uneven distribution of geospatial data and low correlation of adjacent geographic entities, the spatial data division method based on STR tree is used to divide the spatial position of vector data, and the divided data store into database parallel by MongoDB Connect for Spark, each database is organized by layer. Since MongoDB uses loose storage similar to JSON, data conversion and organization of raw data is required before storage.

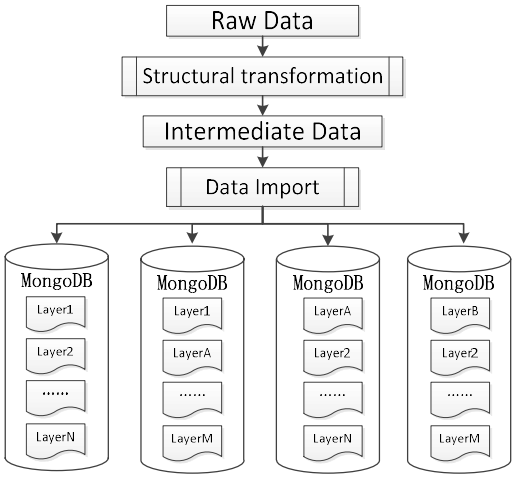

Figure 1. The process for vector data store into database

4.2 Distributed spatial index construction technology

After storing into database, each node builds a spatial index in parallel according to the layer, and the spatial index file is organized according to the layer and is independent of the database. To meet a variety of spatial data inspection needs, highlight the differences between spatial representations and spatial indexes of various object geometry types, improve inspection efficiency, the framework builds double indexes. If the geometry type is point, the framework builds R tree index and GeoHash+ index; if the geometry type is line, the framework builds R tree index and GeoHash+ index for start point and end point in line; if the geometry type is polygon, the framework builds R tree index. GeoHash+ technology is a point matching technology. It is based on the GeoHash technology and overcomes the defects of the latter in the applicability of coordinate types, inconsistent grid length and width, and low coding similarity of adjacent points.(Zhang Libo, 2016)

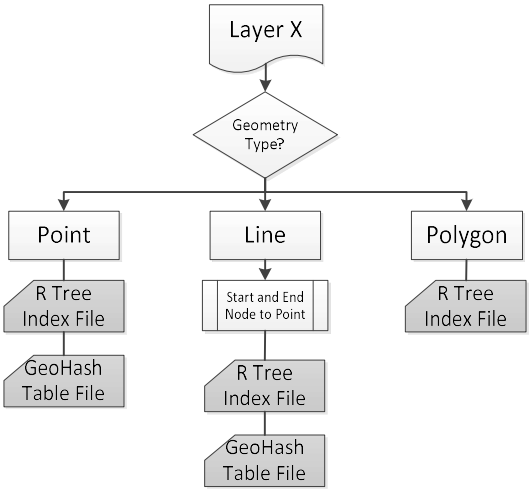

Figure 2. Spatial index for different geometry data

4.3 Construction technology of quality inspection rule library based on spark

The quality inspection rule is an abstraction of possible data defects. It has universal applicability, is the smallest unit of the quality inspection model, and is the most dynamic quality inspection element. It is the entity that carries the quality inspection function. In the quality inspection process, the quality inspection function of the software is the embodiment of the quality inspection rules. The quality inspection rules are not static. They may increase due to the increase in the type of data, or may be adjusted due to the adjustment of data standards.

The quality inspection rule is instantiated by parameter activation to form an inspection item. The inspection item is a rule that is used to check whether a data has a specific error, and is the result of the rule instantiation. A complex inspection scheme can be completed by the logical combination of inspection items.

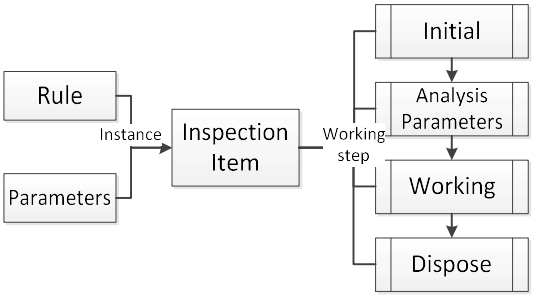

Figure 3. Instantiation and working step for inspection item

Spark-based quality inspection rule library, is based on the quality inspection rule library. By using Spark's ability in memory computing and multi-iterative batch processing and combining with the distributed spatial query and geometric operations of the spatial data, the library extended the resilient distributed dataset(RDD), and achieved full run-time management of inspection items, including the analysis of inspection items, the instantiation of rules, the operation of rules, the destruction of rules, etc.

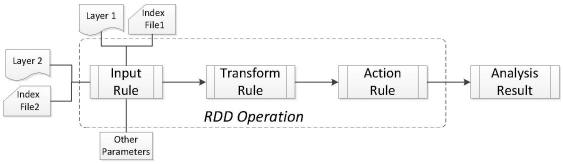

Figure 4. The process of spatial inspection analysis

The input rule reads the data into the RDD space, mainly including Parallelize, TextFile and other operations.

The transform rule completes the intermediate process of the job, but not trigger the submission of the job, mainly includes operations such as filter and map.

The action rule triggers Spark to submit the job, mainly including reduce, collect, count, etc.

5. Framework design

The inspection cloud platform framework for surveying and mapping products is divided into 3 layers.

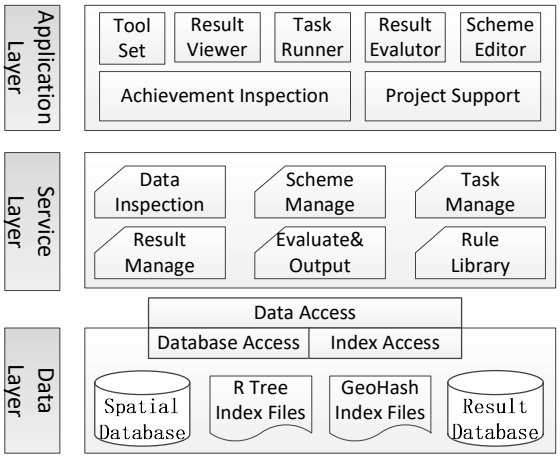

Figure 5. The structure of the inspection cloud platform framework

The data layer provides data storage and access. The data contains four types, including spatial data, table data, index file, and inspection result.

The service layer provides inspection services and manage services, including full run-time management of inspection items, data scheme management, inspection scheme management, evaluation scheme management, inspection task management, inspection result management, etc.

The application layer provides interface performance that interacts directly with the user. The framework has different interactive interfaces according to project needs, including system tools and project support. System tools are tools that are not related to specific needs, including scheme editor, inspection task runner, inspection result viewer and other tool set. Project support is a complete quality inspection automation system based on specific data production technical regulations, data inspection technical regulations, project acceptance technical regulations.

References

[1]Chen, GuoPing. 2013. Spatial database technology application. Wuhan University Press

[2]Zhou, Yao. 2018. Design and Implementation of Application and Analysis System for Big Spatial Data based on Spark and MongoDB. Geomatics & Spatial Informatioin Technology Vol 41, No 9, Sep,2018:71-74

[3]Zhang, LiBo. 2016. Research and Application of point matching technology based on mass data. Bulletin of Surveying and Mapping Vol 11, 2016:122-125