Hydraulic and Civil Engineering Technology VI
M. Yang et al. (Eds.)

doi:10.3233/ATDE210214

Design and Implementation of a Hybrid Architecture Big Data Platform for Catchment Water Resource Spatial Temporal Management and Control

Yuhong LI [a,1], Jiajun LU [a], Qiongfeng JIANG [a] and Zhiyuan ZENG [a]
[a] *School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, P.R. China*

Abstract. The management and protection of catchment water resource are effective measures to promote the harmonious coexistence of human and nature, and accelerate the construction of ecological civilization. Effective storage, management, and retrieval of large spatial temporal data in catchment water resource are facing enormous challenges. At the same time, higher requirements are put forward for data concurrent access support capability and security reliability. Therefore, it is urgent to carry out research on intelligent management and control of large spatial temporal data in catchment water resource. This paper develops a hybrid architecture storage and retrieval system for large spatial temporal data of catchment water resource, which solves the problems of efficient storage and retrieval of massive multi-source heterogeneous data and concurrent access support. Combined with the technical specifications of water resources and geographic information related countries and industries, the existing water-related management system is migrated and integrated by using the " one-source-one-repository " model, avoiding repeated collection and storage of observation data, improving data consistency, and facilitating data sharing among various subsystems. HBase-based tile pyramid storage is used to implement fast visual display and query of data. Metadata model based on MongoDB document model is used to simplify metadata description. At the same time, the Elasticsearch search engine is used to build metadata full-text index, which provides multiple matching methods such as exact matching, fuzzy search, and range query. Spatial vector feature data storage model is established based on GeoJSON and MongoDB, build spatial index, design auxiliary index to accelerate data query and filtering, design sharing strategy in shared replication cluster, balance the contradiction between data distribution and query efficiency.

Keywords. Catchment water resource management and control, hybrid architecture big data platform, spatial temporal data storage and retrieval, HBase-based tile pyramid storage, MongoDB document model

1. Introduction

In recent years, water conservancy combined with major information construction projects, special work of resource investigation and general survey and daily work has

[1] Yuhong Li, School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, P.R. China; E-mail: liyuhong@hust.edu.cn.

generated and accumulated a large amount of water conservancy data, and various forms of river and lake management systems have emerged. However, the above system is only customized and developed for the business needs of the management department, and only realizes a single control goal through limited special data. The Guiding Opinions on Promoting the Development of Water Resources Big Data points out that it is necessary to strengthen the integration and sharing of water conservancy data resources across businesses and across vertical levels, and establish a scientific and standardized data sharing application mechanism. In this paper, through the research of data storage and retrieval strategy, the application of big data platform technology, to build a reliable, extensible distributed, mixed architecture storage and retrieval management system [1-3]. To realize the control of multi-source heterogeneous data in Catchment Water Resource [4], provide unified data transmission and retrieval interface, and strengthen the interconnection between data. At the same time, provide the system concurrent access support, ensure the safety and reliability of data [5-9]. The relevant research results have great theoretical research value and engineering application value in promoting the great protection of the Yangtze River, promoting the overall effect of the river and lake chief system, and realizing the fine management of Catchment Water Resource.

2. State of art on Catchment Water Resource Management and Control

With the passage of time, the multi-dimensional water conservancy system with the "nature-society" dual water cycle and its associated water ecology, water environment, water economic and social processes become more and more complex, and the existing water space management and control system are faced with more difficulties:

(1) The lack of unified technical standards and effective system guarantee makes the data resources such as basic water conservancy hydrological information data and professional database have the situation of repeated construction, and it is difficult to share resources. All kinds of information resources are scattered in different business departments, and the phenomenon of data fragmentation is serious [10];

(2) The volume of data is large and storage is difficult. In recent years, with the continuous development of satellite remote sensing, Internet of Things, intelligent sensor monitoring and other technologies, the volume of catchment water resource data has exploded, which has increased the difficulty of data storage and management;

(3) Water conservancy data is complex and diverse. In addition to traditional structured data, water conservancy data also includes unstructured data such as satellite images [11], digital pictures, and shape data. There is no platform that can integrate and merge water conservancy data.

3. Architecture Design of Catchment Water Resource Management and Control Big Data Platform

3.1. Big Data Storage Strategy for Catchment Water Resource Management and Control

3.1.1. Structured Data Storage Strategy

Structured data is a basic data type with a fixed table structure model, and it has high requirements on consistency constraint and data integrity, so it is suitable to be processed by structured query language.

(1) Structured data " one-source-one-repository " storage specification

Now the data between the system of each management department is isolated from each other, and the structure of the database table may be inconsistent. In order to promote data sharing and facilitate unified management, the system, in accordance with the principle of " one-source-one-repository " and in combination with relevant technical specifications of the Ministry of Water Resources, formulates unified database table structure specifications to store the data of subsystems of various departments in a unified and standardized manner.

The classification model of " one-source-one-repository " table of catchment water resource structured data is shown in the figure 1:

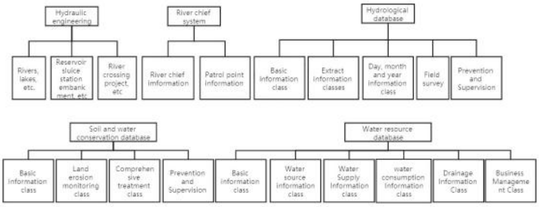

Hydraulic engineering
Rivers, lakes, etc.
Reservoir sluice station embankment, etc
River crossing project, etc
River chief system
River chief imformation
Patrol point information
Hydrological database
Basic information class
Extract information classes
Day, month and year information class
Field survey
Prevention and Supervision
Soil and water conservation database
Basic information class
Land erosion monitoring class
Comprehensive treatment class
Prevention and Supervision
Water resource database
Basic information class
Water source information class
Water Supply Information class
water consumption Information class
Drainage Information Class
Business Management Class

Figure 1. The classification model of " one-source-one-repository " table of catchment water resource structured data.

(2) Classified and mixed storage strategy

1) Scalar data without spatial attributes: This type of data needs to maintain the original data structure, standardize, and improve the existing database tables in accordance with relevant standards, and then transfer to the new system for unified management and use MySQL relational database storage.

2) Infrastructure and construction data with spatial attributes: In addition to the standard structured storage, the spatial characteristics are processed. Combined with the spatial index characteristics of MongoDB document database, the structured two-dimensional table data is converted into a fixed document format, and MongoDB is used for storage.

Mixed storage process of catchment water resource structured data migration and classification is shown in the figure 2:

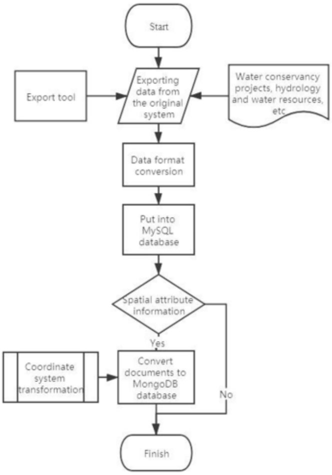

Start
Export tool
Exporting data from the original system
Water conservancy projects, hydrology and water resources, etc
Data format conversion
Put into MySQL database
Spatial attribute information
Yes
No
Coordinate system transformation
Convert documents to MongoDB database
Finish

Figure 2. Migration and import process of catchment water resource structured data.

3.1.2. Unstructured Data Storage Strategy.

(1) Storage strategy of raster image data

1) HBase-based tile pyramid storage: In this system, tile storage is mainly used for rapid visual display of hot image data, which can be transformed into the closest level in tile map. The storage model of HBase is a huge sparse table structure, and the number of tiles varies greatly among image data with different resolutions. Therefore, a fixed tile sequence family is adopted to store tile data by dynamically adding qualifiers [12].

2) MongoDB based metadata store: This system uses MongoDB based document mode to store remote sensing image data metadata. On the one hand, it has good compatibility for different metadata formats. On the other hand, it can transform the traditional metadata from XML files with tree hierarchy into linear K-V structure, which is convenient for storage and management. At the same time, MongoDB can easily realize horizontal expansion of database, expand system capacity through sharded replication cluster, and provide support for high concurrent access to data [13].

(2) Storage strategy of space vector data

This system uses the OGR data model and GeoJSON data format to describe the spatial vector element data, and uses MongoDB for storage [14-15]. This approach has the following advantages:

1) The OGR vector factor model is highly consistent with the MongoDB document data model;

2) The GeoJSON data format can be easily converted to WKT/WKB and other description formats;

3) MongoDB database has built-in support of two-dimensional spatial index, which can easily and quickly establish spatial index on GeoJSON object fields, and support spatial query and analysis operations.

The flow of space vector data storage is shown in the figure 3:

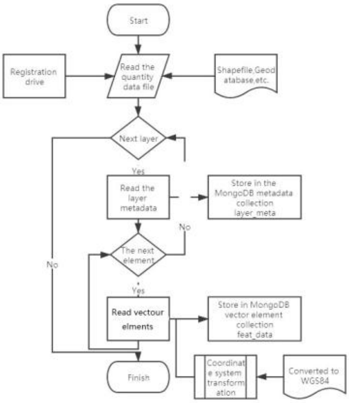

Start
Registration drive
Read the quantity data file
Shapefile,Geodatabase,etc.
Next layer
Yes
Read the layer metadata
Store in the MongoDB metadata collection layer_meta
No
The next element
No
Yes
Read vectour elments
Store in MongoDB vector element collection feat_data
Finish
Coordinate system transformation
Converted to WGS84

Figure 3. The flow of space vector data storage.

To sum up, the catchment water resource big data storage process is shown in the figure 4:

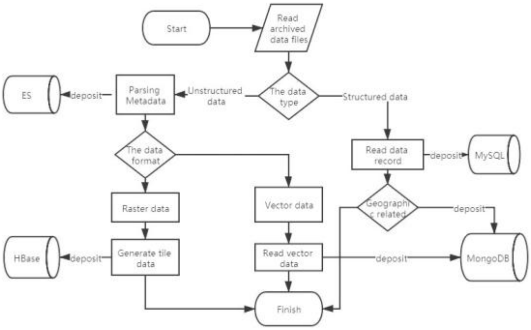


Figure 4. Integral storage structure of catchment water resource big data.

3.2. Catchment Water Resource Big Data Retrieval Strategy

3.2.1. Elasticsearch Engine

This system uses Elasticsearch engine as the secondary index of HBase. HBase only supports row key indexes and has no secondary index support, which means you cannot locate data by column families and qualifiers. The row key is associated with the column family and qualifier and stored in the ES index as a document record. By using the search function of ES, the associated row key can be retrieved according to the column family and qualifier, and then the data in HBase can be retrieved. Another use of ES is to retrieve the original data in HDFS. The storage path is associated with the original data, stored in the ES index, and the HDFS storage path is retrieved according to the metadata properties of the original data.

3.2.2. Index Design and Retrieval of Vector Elements

This system uses MongoDB for spatial index and 2dSphere spatial index to create index for vector element data set, so that all vector elements within the specified coordinate range can be queried. At the same time, in order to improve the efficiency of data retrieval, auxiliary indexes such as attribute filtering can be created according to different requirements and application scenarios. The commonly used indexes are single field index and combined index.

3.3. Big Data Hybrid Architecture Design for Catchment Water Resource Management and Control

According to the needs of big data in catchment water resource management and control, this paper establishes a hybrid storage and retrieval system for massive multi-source heterogeneous big data in catchment water resource, constructs a catchment water resource big data platform, integrates data resources of various departments, and provides data storage support, data retrieval and management functions for all departments [16]. The system adopts layered design to reduce the coupling between layers and enhance the scalability of the system. The system mainly includes interface layer, processing layer and storage layer, as shown in the figure 5:

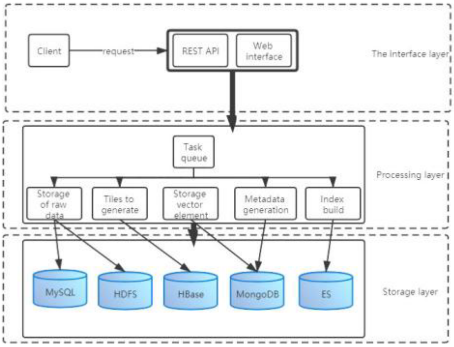


Figure 5. Overall system architecture.

3.3.1. The Interface Layer

The interface layer provides the user interaction ability, including data uploading, retrieval, management, download and other functions.

(1) Web page: Provide user interaction interface, supporting single or batch upload and download of files; Provide fuzzy query function, users only need to input keywords to carry out fuzzy matching query; Provide administrator control panel, easy to manage data, monitor system running status and other information.

(2) Rest API interface: Rest API interface is provided for developers to use, with higher customizability, developers can flexibly filter and filter data according to different application requirements.

3.3.2. Processing Layer

The processing layer is the core of the whole system. The implementation of business logic depends on the storage foundation of the lower layer and improves the processing capacity of the upper interface.

(1) Task queue: it is mainly used to realize the asynchronous processing of tasks and improve the system's concurrent ability. After the user submits the request, the interface layer encapsulates the task into different categories according to the request type, which is implemented in the form of thread pool and blocking queue. The task is encapsulated into an abstract class, and the specific task type corresponds to different logical implementations, which are scheduled uniformly by the thread pool.

(2) Data warehousing: storing data in different databases.

(3) Metadata generation: There are two main sources of metadata: one is the original file data set itself, and the other is the attached XML file. The system will extract the metadata respectively and convert them into JSON linear format. After the extraction and merger, the metadata will be stored in the MongoDB database.

(4) Index build: The system uses the ElasticSearch search engine to store the fields that need to be retrieved from the metadata into ElasticSearch, constructs the full-text search, and carries out fuzzy matching according to the keyword information. In addition, spatial index and auxiliary index are added in MongoDB to speed up filtering and filtering operations and improve spatial query capability.

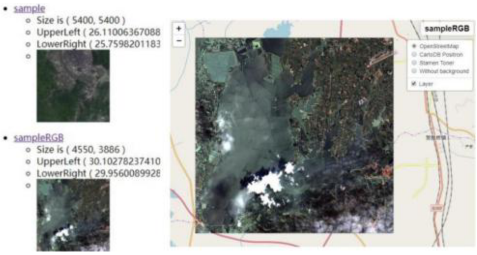
3.3.3. Storage Layer

The storage layer is the foundation of the whole system, which is used to store massive data, provide data redundancy, ensure data integrity, and prevent data loss due to node downtime. The raw data is stored in the HDFS distributed file system. Water data is stored in MySQL relational database. The image pyramid tile data is stored in the HBase key value database. The data of space vector elements are converted into documents and stored in MongoDB document database. The metadata is stored in the MongoDB document database. The index build data is stored into the ES search engine.

4. System Application

4.1. Visualization of Raster Data

The remote sensing image is imported into the system and viewed. As shown in the figure 6, the left side is the list of remote sensing image data information, which mainly contains layer name, image resolution size, geographical range of image, thumbnail preview image and other information. Click the corresponding layer, and the right side is the visual display of the corresponding remote sensing image tile, and different base image and zoom level can be selected.



- sample
  - Size is (5400, 5400)
  - UpperLeft (26.11006367088
  - LowerRight (25.7598201183
- sampleRGB
  - Size is (4550, 3886)
  - UpperLeft (30.10278237410
  - LowerRight (29.9560089928

Figure 6. Remote sensing image tile data visualization.

4.2. Spatial Query Performance Testing

PostGIS database and the MongoDB vector element storage mode adopted by this system are compared and tested for range query and distance query. The test data was a vector data set provided by OpenStreetMap in the form of Shapefile. The elements of a single data set ranged from tens of thousands of items to nearly six million items, and were calculated by averaging multiple queries.

The results of the rectangular range query and the specified point distance query are shown in the figures 7-8:

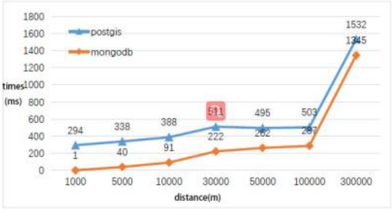


Figure 7. Rectangular range query

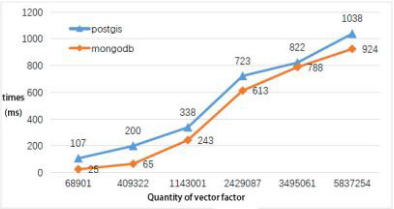


Figure 8. Specify a point distance query.

As can be seen from the above figure, the spatial query time of both increases with the increase of data volume. The spatial query performance of MongoDB vector element storage model is slightly better than PostGIS spatial database, especially when the query scope and data volume are not very large, the performance advantage of MongoDB is obvious.

Acknowledgment

This project was supported by the Major projects of technological innovation in Hubei Province (No. 2019ACA155). Thanks are due to colleagues in our laboratory for their valuable comments and other helps. The authors are also grateful to the reviewers in the Program Committee.

References

[1] Munoz M, Gil JD, Roca L, Rodriguez F, Berenguel M. An IoT architecture for water resource management in agroindustrial environments: A case study in Almeria (Spain). Sensors-Basel. 2020 Feb; 20(3): 596-616.

[2] Lu X, Cheng C, Gong J, Guan L. Review of data storage and management technologies for massive remote sensing data. Sci China Technol Sc. 2011 Dec; 54(12): 3220-32.

[3] He G, Zhou Y. The research of multidimensional analysis based on multi-source heterogeneous real estate data. In: 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA); 2018 Apr 20-22; Chengdu, Sichuan. IEEE; p. 285-9.

[4] Ai P, Yue Z. A framework for processing water resources big data and application. Applied Mechanics and Materials; 2014 Feb;519-520: 3-8.

[5] Hajjaji Y, Boulila W, Farah IR, Romdhani I, Hussain A. Big data and IoT-based applications in smart environments: A systematic review. Computer Science Review. 2021 Feb; 39: 100318-34.

[6] Zhang Y, Luo W, Yu F. Construction of Chinese smart water conservancy platform based on the blockchain: Technology integration and innovation application. Sustainability-Basel. 2020 Oct;12(20): 1-16.

[7] Yang CT, Huang GL, Huang FH, Ho TW, Yang JM. SOA-based platform for water resource information exchanging. In: 2009 17th International Conference on Geoinformatics; 2009 Aug 12-14; Herndon VA, IEEE; p. 1-4.

[8] Liang J, Xie J, Zhang X, Wang X. Study on the Construction of big data and valorization services of intelligent water. In 2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC); 2021 Jun 18-20. Beijing, Beijing, IEEE; P.145-149.

[9] Wang XC, Sun Z. The design of water resources and hydropower cloud GIS platform based on big data. Geo-Informatics in Resource Management and Sustainable Ecosystem. 2013;312-322.

[10] Li Y, Song B, Lv P. Thinking on the integration and sharing of water conservancy information resources in Henan province. Henan Water Conservancy and South-to-North Water Diversion. 2019 Oct; 48(10): 82-3. (In Chinese)

[11] Liu J, Xue Y, Ren K, Song J, Windmill C, Merritt P. High-performance time-series quantitative retrieval from satellite images on a GPU cluster. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2019 Aug; 12(8): 2810-21.

[12] Wang L, Chen B, Liu Y. Distributed Storage and Index of Vector Spatial Data Based on HBase. In: Hu S, Ye X, editors. 2013 21st International Conference on Geoinformatics; 2013 Jun 20-22; Kaifeng, Henan: IEEE: p.1-5.

[13] Fan J, Yan J, Ma Y, Wang L. Big data integration in remote sensing across a distributed metadata-based spatial infrastructure. Remote Sens-Basel. 2018 Jan;10(1): 7-26.

[14] Zhang D, Wang Y, Liu Z, Dai S. Improving NoSQL storage schema based on Z-Curve for spatial vector data. IEEE Access. 2019; 7:78817-29.

[15] Huang K, Li G, Wang J. Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding. Remote Sens Lett. 2019 Nov;10(2): 1070-8.

[16] Fazio M, Celesti A, Puliafito A, Villari M. Big data storage in the cloud for smart environment monitoring. In: Shakshuki E, editor Procedia Computer Science; 2015. p. 500-6.