Ekoloji 28(108): 855-859 (2019)

LETTER TO THE EDITOR

Construction of Multi-scale Ecological Environment Database Management Model Based on Hadoop

Juxiang Ren*
Information Faculty, Business College of Shanxi University, Taiyuan 030031, China
*Email: 18636180797@163.com

In order to further deepen the research on the ecological environment database management model, a method of building a multi-scale ecological environment database management model based on Hadoop is proposed. Taking the multi-scale land boundary as the research area, the trapezoidal grid and rectangular grid are divided according to the 1:100,000, 1:10000, and 1:2000 topographic maps, and the multi-scale ecological environment data integration analysis is performed for the data of different elements. Based on this, the multi-scale ecological environment database management model is built based on Hadoop platform. The experimental results show that the model has short running time and strong practicability in both stand-alone and distributed environments, which show that the proposed method is practical.
Hadoop; Multi-scale grid; Ecological environment database; Model.

1 INTRODUCTION

With the progress of major scientific programs such as global change, the relationship and mechanism of the Earth's circle, human-land relations and environmental effects, it is increasingly needed to support scientific data resources that have been deposited and accumulated for many years. From a broad perspective, ecological resources refer to all the materials, energy and information that can maintain the ecological functions of the natural environment. It includes natural ecological resources, economic ecological resources and social ecological resources. In a narrow sense, ecological resources are also called natural resources, including biological resources on the surface of the earth, water resources, land resources, and atmospheric resources surrounding the earth. At present, there are still problems in the construction of China's ecological environment database, such as weak integration of 3S technology and database, poor data quality, poor data standardization, and difficulty in network data sharing. It is urgent to integrate various data resources systematically with modern computer technology, to establish a spatial database, and need to scientifically manage and use these data. Therefore, using Hadoop's big data processing platform to build a multi-scale ecological environment database management model has became imminent, it can better provide useful information and services for government macro decision-making, land and resources planning, ecological environment improvement, monitoring, governance and evaluation (Ingarao et al. 2018, Mahmoud et al. 2018, Tan and Bi 2018).

Qing Zhu published an article in the journal Ekoloji's 2019 Issue 107, entitled "Data Acquisition Technology of Air Pollution Sources in Ecological Monitoring Database", This article shows that the acquisition of atmospheric pollution source data is a key link in the establishment of an ecological monitoring database. A new air pollution

source data acquisition technology is studied. Using the location algorithm based on the sparse system identification, the problem of the location of the atmospheric pollution source is modeled as the problem of the parameter identification of the sparse system. Based on the incomplete random sampling data, the sparse signal is reconstructed by the norm regularization minimum mean square error optimization method to obtain the air pollution source. The ZigBee sensor network is used to collect the air pollution source location results. While describing the working principle of the sensor and concentrator node hardware, the key functions of data acquisition in the ZigBee sensor network are designed. The collected air pollution source data is stored by a storage table and a writing method based on the HBASE design. The results show that the leak rate of air pollution source data collected by the research technology is between 0.1% and 1.0%. The time for using different air pollution source materials is longer. The time for using different air pollution sources is about 50ms, which is a high-performance data acquisition technology for air pollution sources. Reference (Wang Chen and Wang 2017) explains how to use Hadoop and HTML5 technology to solve the problem of visualization efficiency in the construction of WebGIS system under the environment of environmental protection big data. Aiming at the problem of long load delay and slow response in image rendering of massive data, building a Hadoop-based Big Data-Processing Model and the Architecture of WebGIS. The efficiency of the WebGIS system is improved through a plurality of links such as a database, a spatial data service, a WebGIS client implementation and the like. Aiming at the problem that relational database and single node processing are difficult to meet the storage and processing of massive meteorological and hydrological data, Reference (Li Wang and Ye 2018) proposed a large-scale meteorological and hydrological data concurrent processing model (CPHDH) based on Hadoop. The model combines cloud computing technology, realizes distributed storage of massive meteorological hydrological data by computer cluster and HDFS file system, and implements efficient parallel processing of massive meteorological hydrological data based on MapReduce programming framework. The above methods have obtained good research results, but there are still some shortcomings need further analysis. To this end, this paper proposes a multi-scale ecological environment database management model build method based on Hadoop.

2 IDEA DESCRIPTIO

2.1 Ecological environment database multi-scale grid construction

This paper establishes an ecological environment database on three different scales at the global, national and Beijing levels, and uses grid theory to divide the terrestrial grounds of the world, the whole country, and Beijing. The global, national, and Beijing multi-scale land boundaries are used as research areas. According to the 1:100,000, 1:10000, and 1:2000 topographic maps, trapezoidal grids and rectangular grids are divided. According to the international numbering rules, 103000, 72500 and 16300 multi-scale topographic coverage maps are formed respectively (Niu 2018). The spatial data is divided into multiple grid data blocks to achieve seamless connection of the block spatial data. Each grid map contains homogeneous and continuous ecological environment attribute data, such as topographical factors and forest vegetation coverage factors, limited meteorological factors, land and soil factors, and atmospheric pollution factors, which can not only directly read the attribute information in each grid data block, realize the visual expression of various types of data, which facilitates the subsequent processing and analysis of data (Zheng Hao and Huang 2017). The multi-scale grid of the constructed ecological environment database is shown in Table 1:

Table 1 Multi-scale grid construction of ecological environment database

size	scale	Grid type	Number of grids / 10,000
global	1：100 000	Ladder grid	10.3

National	1:10 000	Ladder grid	7.25
Beijing	1:2 000	Rectangular grid	1.63

2.2 Multi-scale ecological environment data integration analysis

Data integration is the unified processing of data space, time and attributes, but due to the discretization of computer data representation, the way people think things and the static features of existing geospatial data, time is often used as a constant or in data integration. The parameters are treated, and the result is that data integration at different spatial scales becomes the most frequent form of data integration. The reasons for spatial multi-scale data integration that require spatial data from ecological environments is: The phenomena or processes of the same geological environment show different properties on different spatial scales, and the data of each scale need to be used to fully reflect a physical process. In the multi-factor analysis, the integration of multi-scale data is also involved when data on one scale is used to use other element data on another scale. Here we mainly analyze the spatial multi-scale data integration of different elements:

2.2.1 Comparable time scales of different elements

When the spatial scale is comparable, the purpose of data integration mainly includes spatial correlation analysis and new data generation. The former refers to the use of certain or some data to perform operations such as quality inspection, data synthesis, data refinement, and auxiliary derivation. The latter is to use the correlation between different data elements to generate new data from some data, such as high-precision pollution source distribution map and regional small-scale topographic map to generate pollution status map.

2.2.2 When the spatial scales of different elements are incomparable

When the scales of the two datasets are very different, the correlation analysis between them is carried out, otherwise it is of little significance to use their correlation to generate new datasets. At this time, the data integration is presented as background reference analysis and element weight analysis using different data sets to give full play to the function of multi-source data

2.3 Hadoop-based multi-scale ecological environment database management model

As an open source software platform, Hadoop makes it easier to write and run applications that process massive amounts of data. The core design in the Hadoop framework is Map Reduce and HDFS. HDFS is an acronym for Hadoop Distributed File System, the Hadoop Distributed File System, which provides underlying support for distributed computing storage. HDFS has much in common with existing distributed file systems. But at the same time, the difference between it and other distributed file systems is also obvious. HDFS is a highly fault-tolerant system suitable for deployment on inexpensive machines (Zhou Wang 2017). HDFS provides high-throughput data access, making it ideal for large-scale dataset applications and for deployment on inexpensive machines. Map Reduce is a simplified distributed programming model that allows programs to be automatically distributed to a large cluster of ordinary machines for concurrent execution.

Based on Hadoop's multi-scale ecological environment database management model construction process, the ecological environment data obtained by 2.2 integrated processing is preprocessed, and the files are processed into the sequential files defined in Hadoop, which is convenient for reading. The first step is word segmentation; The second step treats each line into the form of <document id+" \t" + mark, document content>; The third step is to perform tfidf vectorization preprocessing. The vector corresponding to each document draws on the Vector Writable data type in the mahout open source project, which greatly facilitates processing and saving memory. According to the above steps, the construction of the multi-scale ecological environment database management model can be completed.

3 RESULTS

The experimental analysis is carried out under the Hadoop platform to verify the feasibility of the research method. Two sets of data sets A and B are selected, and each set of data sets contained 1000 ecological environment data as experimental data.

Experiments are performed in both stand-alone and distributed environments. The training data is the same, 90% of which is extracted as the training set, and the other 10% is the test set. The experiment under the Hadoop platform increased from one node to five nodes one by one. In the stand-alone algorithm, the running time of the database management model construction of these two data sets is shown in Table 2:

Table 2 Running time in a single machine environment

Data set	Number of experiments/s	Paper method /s
A	50	12
	100	13
B	50	12
	100	14

After the database management model is built on the Hadoop platform, the running time of the two data sets on the five nodes is shown in Table 3:

Table 3 Runtime in a distributed environment

Data set	Number of experiments/s	Paper method /s
A	50	10
	100	13
B	50	11
	100	12

Analysis of Tables 2 and 3, we can see that in the stand-alone environment and the distributed environment, the application of this method has less time-consuming operation, which indicates that the multi-scale ecological environment database management model constructed in this paper has certain effectiveness.

4 DISCUSSION

According to the experimental analysis results, the multi-scale ecological environment database management model constructed in this paper has better performance. In both the stand-alone environment and the distributed environment, the running time is short. The main reason is that in the construction of this model, the multi-scale grid of ecological environment database is constructed at first, and the model is constructed in the multi-scale grid, which greatly saves the running time and improves the efficiency of the construction of this model.

5 CONCLUSIONS

This paper builds a multi-scale ecological environment database management model, and verifies the validity of the model through experiments. The results of this study can provide data support and reference for scientific research, teaching and production practice, and provide ideas and services for related research.

ACKNOWLEDGEMENTS

The work was supported by Shanxi Province 1331 Engineering Service Industry Discipline Group Innovation Project: Soil Pollution Ecological Restoration.

REFERENCES

Ingarao G, Priarone PC, Deng Y, Paraskevas D (2018) Environmental modelling of aluminium based components manufacturing routes: additive manufacturing versus machining versus forming. Journal of Cleaner Production 176:261-275.

Li H, Wang J, Ye M (2018) Concurrent processing model of massive meteorological hydrological data based on Hadoop. Journal of Computer Applications 38 (z2):62-68.

Mahmoud ME, El-Khatib AM, Badawi MS, Rashad AR, El-Sharkawy RM, Thabet AA (2018) Recycled high-density polyethylene plastics added with lead oxide nanoparticles as sustainable radiation shielding materials. Journal of Cleaner Production 176:276-287.

Niu Q, Yu L, Jie Q, et al. (2018) An urban eco-environmental sensitive areas assessment method based on variable weights combination. Environment, Development and Sustainability 28 (2):1-17.

Tan F, Bi J (2018) An inquiry into water transfer network of the yangtze river economic belt in china. Journal of Cleaner Production 176:288-297.

Wang YF, Chen G, Wang D (2017) Optimum Design and Implementation of Environmental Protection WebGIS System Architecture Based on Hadoop and HTML5. China Market Marketing 45 (9):29-30.

Zheng J, Hao YU, Huang S (2017) Evaluation and obstacle factors study on eco-environmental carrying capacity in Fujian Province based on DPSIR-TOPSIS model. Acta Scientiae Circumstantiate 32 (17):122-128.

Zhou X, Wang MG (2017) Research on Data Storage and Processing Technology of Multi-source Heterogeneous Power Distribution Based on Hadoop. Automation & Instrumentation 35 (12):223-224.