ISSN: 2687-8097
Scientific Journal of Research and Reviews
DOI: 10.33552/SJRR.2021.03.000554
Iris Publishers

Review Article

Management and Visualization of Geoscience Big Data based on Hadoop Platform

Le Gao, Yongjie Huang, Zhiqiang Zeng*,Shu Zhou and Wenqi Li

Wuyi University, Jiangmen, China

***Corresponding author:** Zhiqiang Zeng, Wuyi University, Jiangmen, China

Received Date: April 10, 2021

Published Date: April 21, 2021

Abstract

This study attempts to use big data technology to manage and analyze the geoscience data with large amount, wide sources and various formats accumulated for many years. The traditional data processing methods are mostly based on pure database and single machine, which is slow and difficult to achieve good results of data service. Hadoop is a distributed, semi-structured data management platform, which can be used to access large data quickly and conveniently. It has been proved by practice that the Hadoop platform can not only effectively complete all kinds of geoscience data processing, but also greatly improve the processing speed. It has good practicability and versatility, and provides technical support for further analysis and application of geoscience big data.

Keywords: Big Data; Hadoop; Data Management; Visualization

Foreword

Since the 21st century, with the rapid development of cloud computing, Internet and mobile communication technology, the amount of information data generated has increased at an exponential rate, and human beings have entered the era of information explosion, that is, the era of big data. Big data not only describes the large scale of data, but also indicates that the data is diversified, fast changing, and true or false. Big data technology is a method of processing data, not the scale of the data itself [1-3]. In recent years, big data has changed people's way of thinking in dealing with affairs, and has penetrated into all fields of human society and scientific research [4-9]. Based on the background of geoscience big data, this paper constructs a distributed and parallel platform for geoscience big data, and conducts research on data management and visualization. After years of accumulation, geoscience data has accumulated a large number of structured, unstructured and semi-structured data such as word document, PDF document; excel document, picture, video, audio, etc., including topographic map, geological map, mineral map, profile map, fault, remote sensing, geophysical exploration, geochemical exploration, drilling, trenching and other geoscience data. Because of the multi-source heterogeneity of these data, although after finishing, it is difficult to unify the standards. It makes the analysis and processing of geoscience data slow and inefficient. Therefore, distributed parallel processing is needed for all types of geoscience data. The high-performance parallel computing platform based on geoscience big data can complete data storage and analysis with high efficiency, high scalability, high fault tolerance and high reliability. This research is based on Hadoop geoscience big data management platform, which can process geoscience big data efficiently and distributed, and effectively solve the bottleneck problems encountered in the past.

Geoscience big data Hadoop platform

The Hadoop platform of geoscience big data is needed to solve the data storage and call of a large number of historically accumulated, multi-source and heterogeneous geoscience big data, and carry out data management and visualization analysis research

on this basis. Hadoop software is an open source project designed by AAPCHE for big data analysis. It has the characteristics of high scalability, high efficiency, high reliability, and other functions such as big data storage and distributed computing of big data. At present, there are mainly three Hadoop platform construction modes [10].

i. Mode 1: stand-alone mode for debugging MapReduce programs;

ii. Mode 2: Pseudo distributed mode, adding HDFS distributed storage and code debugging function on the basis of stand-alone mode;

iii. Pattern 3: Fully distributed, with clustered distributed storage and high performance computing.

The big data platform in this study includes data resource layer, data processing layer, data application layer and data presentation layer from the bottom to the top. The capabilities provided by each level are shown in Figure 1. The big data platform supports the storage and management of billions of pieces of data. The multi-source data is collected and stored uniformly in the data resource layer. High-performance parallel computation is carried out through the MapReduce function of the data processing layer, and then distributed storage is completed through HDFS. HDFS uses master-slave structure for data operation, in which multiple Data Node control Name Node, and then Name Node distribute data uniformly, and finally map all kinds of data back to Data Node. This layer can achieve the file new, open, modify, delete and other work. Data application layer mainly provides geoscience GIS service, image visualization and other functions. The data presentation layer can provide the browsing and downloading service of the data achievements.

Big data algorithm analysis

The parallel processing method of geo-big data on Hadoop platform is divided into four steps (see Figure 2) :

I. The geo-data is distributed processed under the Hadoop platform HDFS. The client issues data writing requirements to the Name Node, and then distributes the decomposed pieces to the Data Node in turn.

II. In each Data Node, the client writes the Map function, each map function processes the data of the Data Node, and the data input and output are stored in the file system;

III. Map function processes the input geoscience data (Key, value), generates new (Key, value) , and then transmits them to Reduce for processing, and outputs the processing results;

IV. The client sends a request to the Name Node to read the data, and the Name Node returns the stored data information to the client.

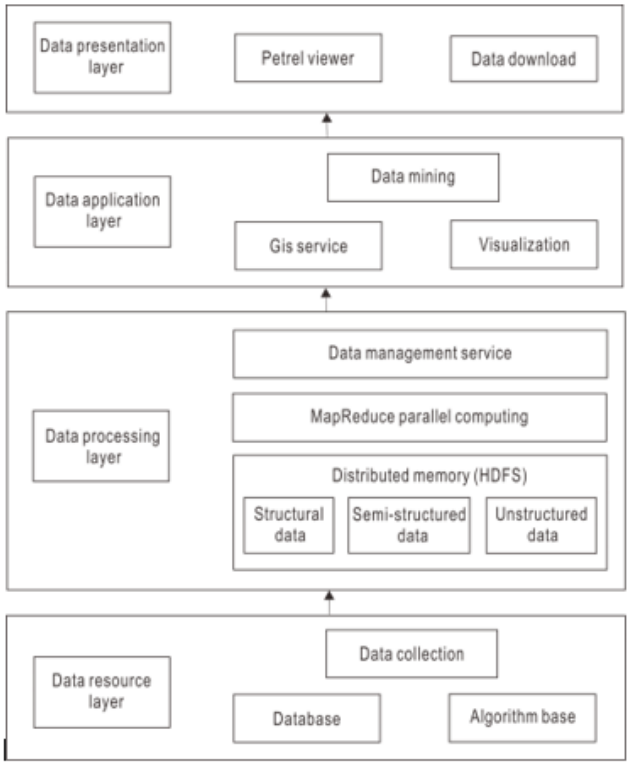

Figure 1: Function Diagram of Geo-Big Data.

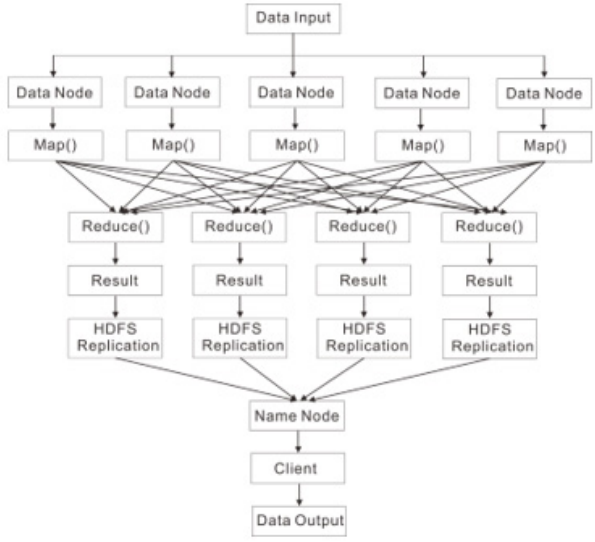

Figure 2: Flowchart of Geoscience Big Data Algorithm.

The geochemical big data have a wide range of sources and various formats. The geochemical data in this paper are mainly studied by factor analysis algorithm. The factor analysis method is used to obtain the common factor from the data of the sampling points in the research area. The common factor retains the original information to the maximum extent, and then the pan-kriging method is used to obtain the anomaly, which is obtained by adopting the regional variable X(a)=Y(a)+Z(a). Assuming that X(a) only has a function of variation, and the estimation of Y(a) can be carried out under the condition of a constant difference, the increment of Y(a) has non-stationary first and second moments, and Z(a) can be obtained under optimal conditions and unbiased conditions, namely, the anomaly.

On the Hadoop big data platform, the original geoscience data A is set, and the map function is used to identify the small pieces of data in multiple Data Nodes. The data is stored in A data document and transmitted to reduce for standardized processing, making it become dimensionless data with uniform scale. The processed result data is stored in a temporary file TempFile1.0.

Data Processing of Geoscience Big Data Platform

In order to reflect the data processing effect of the geoscience big data Hadoop platform, data processing research was carried out in the big data center of Wuyi University according to the platform design framework. In order to excavate the correlation between data, 20 nodes Hadoop and MapReduce framework are used to realize distributed computing mining. The research group used 160,000 stream sediment geochemical data and 50G geological map data from Pangxidong area, west Guangdong in China as experimental data.

The geochemical data of Ag, As, Au, B, Bi, Cu, F, Mn, Mo, Pb, Sb, Sn, W and Zn are used for factor analysis, and the results can effectively reflect the regional geological characteristics of the study area. Through big data platform calculation, the cumulative contribution of the first 5 factors of elements was 81.38%>75% (qualified if the cumulative contribution reached 75%). From the results, it was concluded that there were 5 element combinations in the study area: F1: As-Au-B-Cu-Sb; F2: Ag Pb - zinc; F3: Bi - Mo - Sn - W; F4: Mn; F5: F. the geochemical factor scores of the five main factors were compiled (Figure 3). The spatial distribution information of these geochemical element anomalies reflected the aggregation process of different ore-forming elements. F1 is an element with strong migration ability, which reflects the enrichment process of hydro thermal ore-forming elements at middle and low temperature. F2 is an element that is not easy to migrate and represents a very complex metallogenic geological background. F3 reflects the enrichment of high-temperature ore-forming elements; F4 indicates the enrichment of deposits of Mn and other sedimentary types. The element F of F5 usually indicates the enrichment of Sn, W, and Mo deposits. The acceleration and comparison of the data analysis time between the geo-big data Hadoop platform and the single serial computer shows: with the increase of data volume, the processing speed of the big data platform is faster than the serial processing speed, and it shows an exponential increase. When the data reaches 160,000 data, the processing speed of Hadoop big data platform is 1798 times that of stand-alone serial, as shown in Figure 4.

In order to realize the effective management of unstructured data, testing and data mining are carried out on the data of the study area including images and borehole profiles. The average image load time is 2400 ms in the standalone client test environment, and 15 ms in the big data platform environment. Fig. 5 is a three-dimensional map of the study area obtained by processing multiple image data. Spatial anomaly map of ore-

forming elements is obtained by analyzing geochemical, image and borehole data (Figure 6). The comprehensive prediction of metallogenic target area (Figure 7) was delineated based on factor analysis characteristics, three-dimensional map and spatial map of metallogenic element anomalies, so as to provide technical support and visual display for further analysis and application of geoscience big data. The results show that in terms of the processing speed of unstructured geospatial big data, the effect of big data platform is significantly better than that of single-machine serial processing speed (see Figure 4).

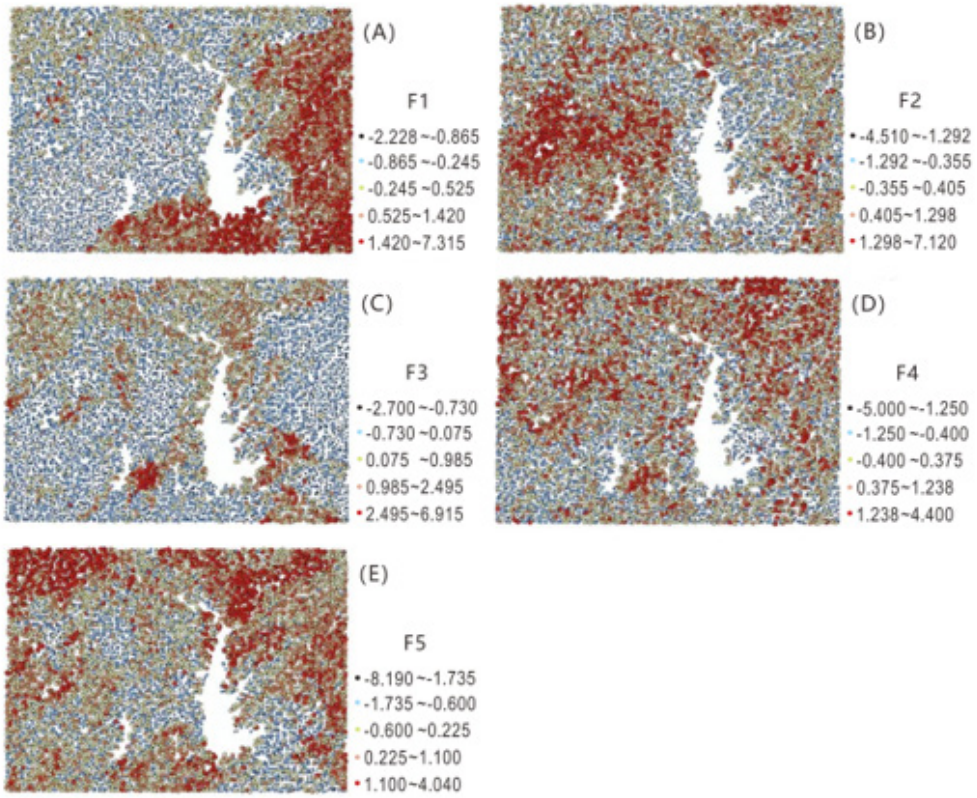

Figure 3: Factor Analysis Score Chart.

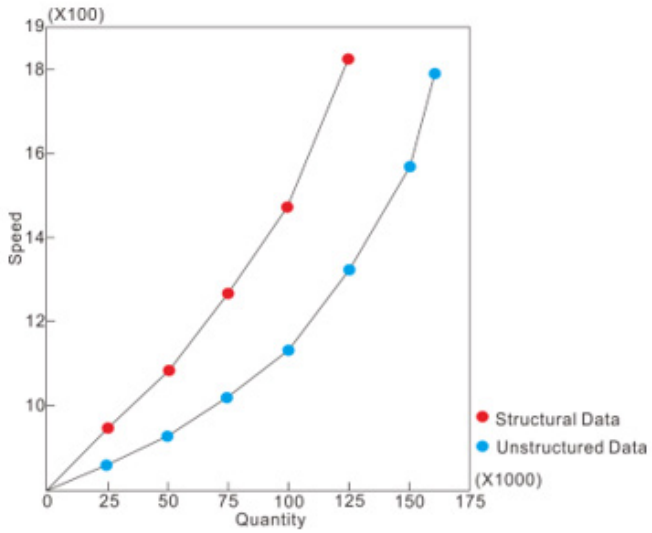

Figure 4: Acceleration Ratio between Big Data Platform and Single Machine.

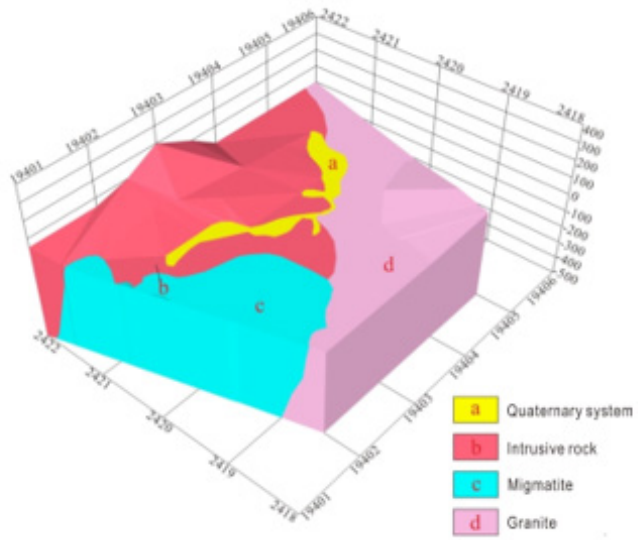

Figure 5: Three-dimensional Stereogram of the Study Area.

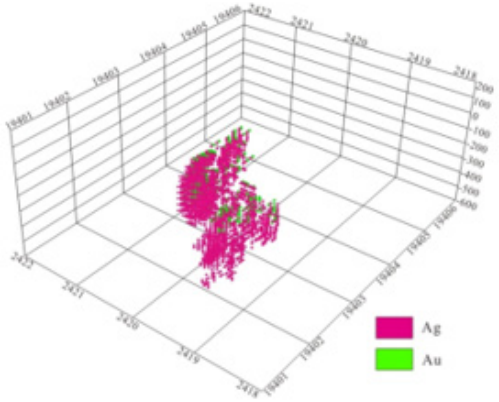

Figure 6: Spatial Anomaly Diagram of Ore-forming Elements.

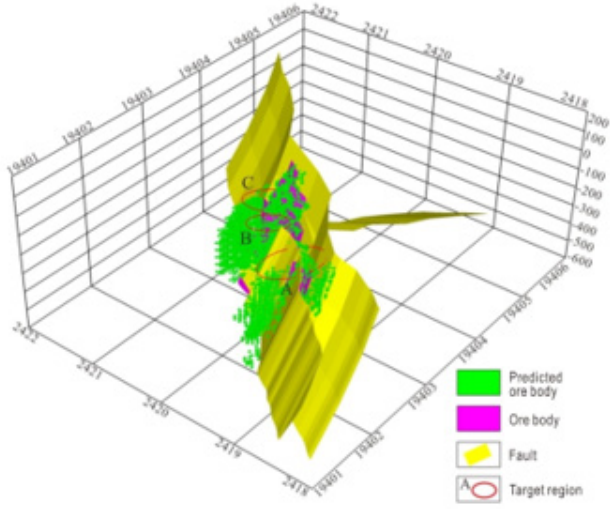

Figure 7: Comprehensive Prediction of Metallogenic Target Area.

Conclusion

This paper designs a distributed parallel Hadoop platform for geoscience big data based on geoscience structured, semi-structured and non-structured multi-source heterogeneous data, and draws the following conclusions:

I. The distributed and parallel platform of geoscience big data was designed on the big data center Hadoop, and the parallel transformation of the geoscience data processing algorithm was carried out on the platform, which reached the international advanced level;

II. Based on the Hadoop geoscience big data platform, different components are used for structured, semi-structured and unstructured data to realize the management and presentation of multi-source big data;

III. Hadoop geoscience big data platform can conduct reasonable analysis on geoscience data, and the comprehensive element anomalies extracted by factor analysis method can effectively indicate regional geological phenomena, which is consistent with the actual situation;

IV. The parallel data processing of Hadoop geoscience big data platform is faster than the single serial data processing and the larger the data, the higher the efficiency.

Acknowledgement

This project is supported by "Natural Science Foundation of Guangdong Province:18zxxt52", "Wuyi University Hong Kong Macao Joint Research and Development Fund: 2019WGALH23" and "Wuyi University Youth Research Group Fund:2019td10".

Conflict of Interest

No conflict of interest.

References

1. Yongzhang Zhou, Qianlong Zhang, Yongjian Huang, Wei Yang, Fan Xiao (2021) Construction of knowledge graph of porphyry copper deposit from Qinzhou Bay- Hangzhou Bay and insight into knowledge graph based mineral resource prediction and evaluation. Earth Science Frontiers.
2. Zhimei Lei, Yandan Chen, Ming K. Lim (2021) Modelling and analysis of big data platform group adoption behavior based on social network analysis. Technology in Society 65:101570.
3. Edward Harshany, Ryan Benton, David Bourrie and William Glisson (2020) Big Data Forensics: Hadoop 3.2.0 Reconstruction. Forensic Science International: Digital Investigation 32: 300909.
4. Le Gao, Yutong Lu, Pengpeng Yu, Fan Xiao (2017) Three-dimensional visualization and quantitative prediction for mine: A case study in Xiayuandong Pb-Zn ore deposits, Pangxidong region, southern part of Qin-Hang metallogenic belt, China. Acta Petrologica Sinica 33(3): 767-778.
5. Yan Zhang, Yongzhang Zhou, Zhenghai Wang, Rui Huang, Wenchao Lv (2011)The Recognition and Extraction of Geochemical Composite Anomalies: A Case Study of Pangxidong Area. Acta Geoscientica Sinica 32(5): 533-540.
6. Jinli Miao, Wu Shang, Youhua Wei, Zhixin Gao and Zhe Xu (2015) Construction and Practice of Geological Big Data Management Platform Based on Hybrid Architecture. Scientific and Technological Management of Land and Resources 32(2): 114-119.
7. Le Gao, Yongzhang Zhou, Kun Wang, Zhiqiang Zeng and Yutong Lu (2020) Application of Partial Least Square Modeling in Stream Sediment Geochemical Survey of Stratabound Lead-zinc Mineralization in Western Guangdong Province. Geotectonica et Metallogenia 44(2):258-266.
8. Kalia Khushboo, Gupta Neeraj (2021) Analysis of hadoop MapReduce scheduling in heterogeneous environment. Jin Shams Engineering Journal 12(1): 1101-1110.
9. Rogerio Luis de C.Costa, Jose Moreira, Paulo Pintor, Sergio Lifschitz (2021) A Survey on Data-driven Performance Tuning for Big Data Analytics Platforms. Big Data Research 25: 100206.
10. Feilong Qin, Heping Cheng, Yali Cheng, Xinyue Zhou, Hanjin Hu (2019) Research on the Parallel Processing of Big Data of Deep Mineral Resources Based on the Hadoop Platform. Journal of Chengdu Technological University 22(4): 50-54.