





Review Article

Big Data Management for Cloud-Enabled Geological Information Services

Yueqin Zhu ^{1,2}, Yongjie Tan ^{1,2}, Xiong Luo ^{3,4} and Zhijie He ^{3,4}

¹Development and Research Center, China Geological Survey, Beijing 100037, China

²Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing 100037, China

³School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing 100083, China

⁴Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

Correspondence should be addressed to Yueqin Zhu; yueqin_zhu@126.com and Xiong Luo; xlueo@ustb.edu.cn

Received 20 October 2017; Revised 10 December 2017; Accepted 31 December 2017; Published 29 January 2018

Academic Editor: Anfeng Liu

Copyright © 2018 Yueqin Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cloud computing as a powerful technology of performing massive-scale and complex computing plays an important role in implementing geological information services. In the era of big data, data are being collected at an unprecedented scale. Therefore, to ensure successful data processing and analysis in cloud-enabled geological information services (CEGIS), we must address the challenging and time-demanding task of big data processing. This review starts by elaborating the system architecture and the requirements for big data management. This is followed by the analysis of the application requirements and technical challenges of big data management for CEGIS in China. This review also presents the application development opportunities and technical trends of big data management in CEGIS, including collection and preprocessing, storage and management, analysis and mining, parallel computing based cloud platform, and technology applications.

1. Introduction

In the era of big data, the data-driven modeling method enables us to exploit the potential of massive amount of geological data easily [1–3]. In particular, by mining the data scientifically, one can offer new services that bring higher values to customers. Furthermore, it is now possible to implement the transition from digital geology to intelligent geology by integrating multiple systems in geological research through the use of big data and other technologies [4].

The application of geological cloud makes it possible to fully utilize structured and unstructured data, including geology, minerals, geophysics, geochemistry, remote sensing, terrain, topography, vegetation, architecture, hydrology, disasters, and other digitally geological data distributed in every place on the surface of the earth [4, 5]. Moreover, the geological cloud will enable the integration of data collection, resource integration, data transmission, information extraction, and knowledge mining, which will pave the way for the transition from data to information, from

information to knowledge, and from knowledge to wisdom. In addition, it provides data analysis, mining, organization, and management services for the scientific management of land resources, prospecting breakthrough strategic action and social services, while conducting multilevel, multiangle, and multiobjective demonstration applications on geological data for government decision-making, scientific research, and public services [5].

Big data technologies are bringing unprecedented opportunities and challenges to various application areas, especially to geological information processing [2, 6, 7]. Under these circumstances, there are some advancements achieved in the development of this area [8, 9]. Furthermore, from various disciplines of science and engineering, there has been a growing interest in this research field related to geological data generated in the geological information services (GIS). We analyze the number of those documents indexed in “Web of Science” [10]. In Figures 1 and 2, we can easily find that, in the past ten years, the number of those documents in which “geological data” is in the “Title” and in the “Topic” are both increasing, respectively. Hence, the analysis for geological

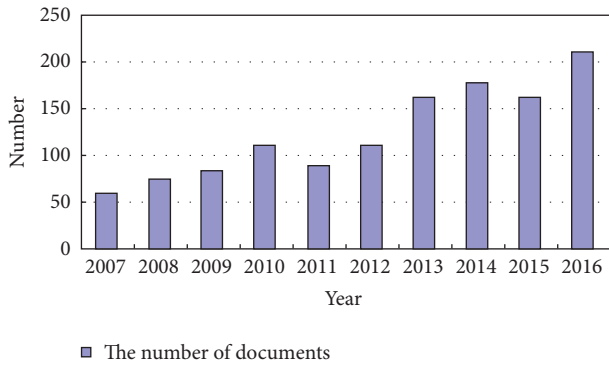


FIGURE 1: The trend of the number of documents in which “geological data” is in the “Title” from 2007 to 2016.

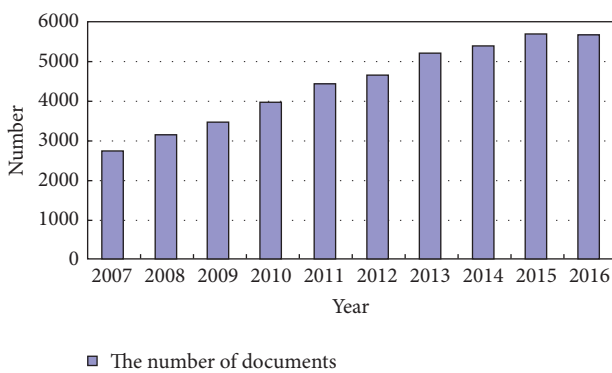


FIGURE 2: The trend of the number of documents in which “geological data” is in the “Topic” from 2007 to 2016.

data in GIS is an interesting and important research topic currently.

Considering the development status of cloud-enabled geological information services (CEGIS) and the application requirements of big data management analysis, this article describes the significant impact and revolution on GIS brought by the advancement of big data technologies. Furthermore, this article outlines the future application development and technology development trend of big data management analysis in CEGIS.

The remainder of this article is organized as follows. In Section 2, we provide a review on CEGIS, with an emphasis on the descriptions for the system architecture and those requirements from big data management. Then, the challenges for big data management in CEGIS are presented in Section 3. Furthermore, the key technologies and trends on big data management in CEGIS are analyzed in Section 4. Finally, conclusion is drawn in Section 5.

2. Review on Cloud-Enabled Geological Information Services

The construction of geological cloud differs from the current big data analysis based on Internet and Internet of Things (IoT). Having a deep understanding of data characteristics is necessary to collect, process, analyze, and interpret data in

different fields, because the nature and types of data vary in different fields and in different problems. Geology is a data intensive science and geological data are characterized with multisource heterogeneity, spatiotemporal variation, correlation, uncertainty, fuzziness, and nonlinearity. Therefore, the geological cloud has a certain degree of confidentiality and it is highly domain-specific; meanwhile, it is developed on the basis of a large amount of geological data accumulated over a long period of time [5, 11]. There are many real-time data generated from some issues like geological disasters and geological environment. The geological cloud includes core basic data, which can be divided into three parts: existing structured database, some unstructured data, and public application data. Therefore, it is important to take good advantage of the existing traditional structured data, use the big data technologies to deal with the relevant unstructured data, and also consider the peripheral public data.

Geological big data are multidimensional, and they consist of both structured and unstructured data [12]. The technical methods of big data analysis differ greatly from those of professional databases. Long-term geological survey, geological study, and years of geological information accumulation have formed a rich and professional database, which is an important fundamental assurance for land and resources science management, geological survey, and geological information public service [13]. This “professional cloud” objectively requires the technology research and development, such as construction of professional local area network, data sharing platform, and geological big data visualization services. Hence, the construction of geological cloud is closely related to land resource management, deployment decision, and the application demand of public service. The key technologies of research and development include the following: unstructured data extraction and mining analysis, structured and unstructured data mixed storage and management, big data sharing platform, data transmission, and visualization [11].

Generally, the construction of geological cloud is a long-term systematic project. This means that it is required to follow the basic principles of “standing on the reality, focusing on the future” and “focusing on the long-term and overall situation, embarking on the current and local situation,” in order to achieve the analysis and application of geological cloud public data and core data gradually in accordance with the technical route of big data analysis; thus the construction of geological cloud will be implemented eventually. For the earth, the land and resources management should cover many respects, including human behavior, climate change, development and utilization of various resources, natural disasters, environmental pollution, and the ecosystem cycle. Then, the introduction of big data technologies can integrate this type of resource information to provide the ability of uniformly dealing with the problems related to the entire earth information resources, which has a significant effect on the strategic planning of land and resources management [3].

The geological cloud is an important part in the science system for geological data research. The ultimate goal for developing geological cloud is to better describe and understand the complex earth system and geological framework,

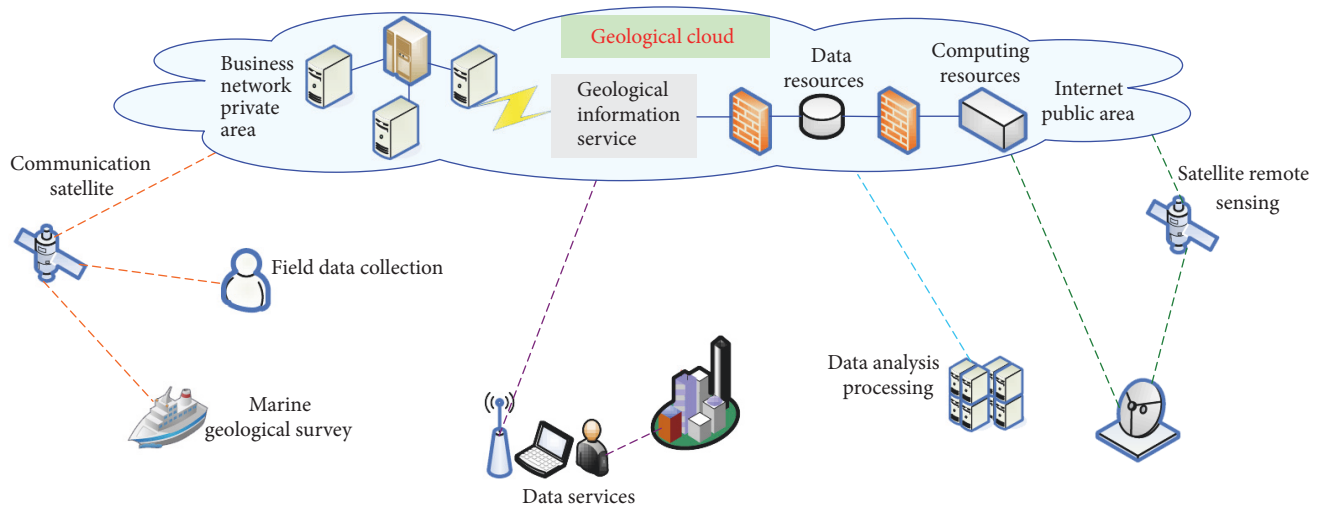


FIGURE 3: The system architecture of geological cloud.

provide the scientific basis for the description of the land surface and the biodiversity characteristics of the earth, and improve the ability to deal with complex social problems.

2.1. System Architecture. Because the business service functions of each country are different, the system architecture of the geological cloud would vary. In the following, we present a system architecture in Figure 3 [14], using China as an example.

The geological cloud combines the geological survey Intranet and the geological survey Extranet. It enables the sharing and management of computing resources, storage resources, network resources, software resources, and geological data resources [15].

Geological cloud can be summarized with the following characteristics [14].

- (i) *“One Platform: The Geological Cloud Management Platform.”* It uniformly manages computing resources, storage resources, network resources, software resources, and geological data resources.
- (ii) *“Two Networks: The Geological Survey Intranet and the Geological Survey Extranet.”* Here, the Intranet is constructed by creating a network that is physically isolated from the Internet. The Intranet is developed on the basis of the existing geological survey network and each node is linked through a dedicated line or bare fiber. All of the internal business management systems, software systems, and data are deployed on the Internet, providing services to 28 local units and those users of more than 350 geological survey projects. Facilitated by the public geological survey network, the geological survey business management system, geological data information service system, and public geological data can be deployed on the Extranet accessed by the general public. The communication between the Intranet and the Extranet,

including data exchange and audit, can be carried out by single-directional light gate.

- (iii) *“One Main Node and Three Domain-Specific Nodes.”* One main node is constructed in China Geological Survey Development Research Center. In addition, three domain-specific nodes—namely, marine node, geological environment node, and aviation geophysical exploration and remote sensing node—are constructed, respectively. Each node is configured with the corresponding servers, storage equipment, network equipment, management platform, large-scale specialized data processing software system and various customized applications. Each node would store huge amounts of geological data and conform to current data security standards. The master node and the domain-specific nodes are linked via optical fibers. The master node will consist of 200 computing nodes with 3 PB storage capacity and will be equipped with some geological data processing software system. The master node will be hosted in a medium-sized supercomputing center and it will provide support for the three-dimensional seismic exploration data processing and other large-scale computing. The three domain-specific nodes are to maintain their scale in the near future to facilitate reasonable scheduling and efficient utilization of information resources and data resources.

On the Extranet, it deploys a system for geological survey business management and auxiliary decision-making. The system provides a real-time tracking and management function for geological survey projects and various resources.

Main users of the geological cloud include institutional users, geological survey project users, and the general public users. The institutional users can store the current geological database and newly collected data in the geological cloud through the geological survey business network and can

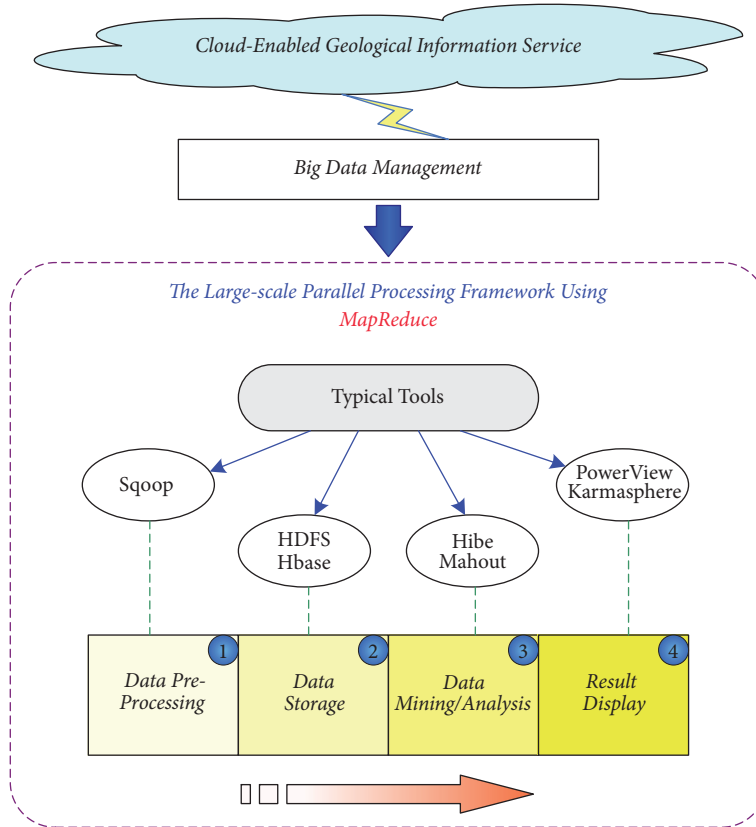


FIGURE 4: Schematic diagram of big data analysis.

obtain the geological data of other institutions from the cloud as needed. The geological survey project users can access the cloud geological background data through 4G or satellite lines and can collect data through data the collection system.

2.2. Requirement from Big Data Management. The construction of geological cloud must meet customer demand. Big data technologies are then used as the means to implement the geological cloud.

The types and quantity of geological data have been continuously growing over the years. Geological data include all kinds of electronic documents, structured, semistructured, and unstructured data, such as various databases (map database, spatial database, and attribute database), pictures, tables, video, and audio. Generally speaking, those important data may be buried in the massive data without the guidance for requirements. Hence, the first step is to understand the user requirements and then gain the capability of large-scale data processing. This is followed by data mining, algorithm, and analysis, which will ultimately generate value. Big data technologies in the field of geography must meet different needs from people at different levels, including the public demand of the geologic data services and professional data demand for geological research institutions, as well as related enterprises and government departments [16].

On the basis of big data analysis technologies, a complete data link is formed connecting data, information, knowledge,

and service, through the use of advanced cloud computing system, IoT, and big data processing flow. It is shown in Figure 4 [5].

3. Challenges for Big Data Management in Cloud-Enabled Geological Information Services

Geological big data are generated regarding various layers of the earth, the history of the conformation and evolution of the earth, and the material composition of the earth and its changes. It also involves the exploration and utilization of mineral resources. In the current geological work, the collection, mining, processing, analysis, and utilization of various complex type data are closely related to those general big data. The “4V” characteristics of big data—namely, Volume, Velocity, Variety, and Veracity—also apply to geological big data.

3.1. Volume. Currently, there is no consensus on the size of geological data. Geological big data are a collection of data, including geology, minerals, remote sensing, geophysical exploration, geochemical exploration, surveying, and mapping, which are interconnected and integrated. In terms of the number of mines, there are at least 70000 in China, and some official documents and popular science books indicate that there are more than 200000 deposits and minerals that have been found. The information is huge and cannot be processed

using conventional tools. For example, an Excel spreadsheet cannot contain all the information of 70000 mining areas. Then, it is difficult to classify and rank the 200000 mines, so it is necessary to rely on the concepts and technologies of big data [17].

Especially in recent years, images, video, and other types of data have emerged on a large scale. With the application of 3D scanning and other devices, the data volume has been increasing exponentially. The ability to describe the data is more and more powerful, and the data are gradually approximated to the real world. In addition, the large amount of data is also reflected in the aspect that the methods and ideas used by people to deal with data have undergone a fundamental change. In the early days, people used the sampling method to process and analyze data in order to approximate the objective with a small number of subsample data. With the development of technologies, the number of samples gradually approaches the overall data. Using all the data can lead to a higher accuracy, which can explain things in more detail [18].

Recently, the China geological survey system has built databases including regional geological database (covering the 1:2500000, 1:1000000, 1:500000, 1:250000, and 1:200000 regional geological map; the national 1:200000 natural sand; the isotope geological dating; and the lithostratigraphic unit database), basic geological database (covering the national rock property database and national geological working degree database), mineral resources database (covering the national mineral resources, the national mineral resources utilization survey mining resources reserves verification results, the national survey of large and medium-sized mines, the prospect of mineral resources, the survey of the resources potential of major solid mineral resources in China, and the geological and mineral resources database), oil and gas energy database (covering the oil and gas basins in China, the geological survey results of the national oil and gas resources, the national petroleum and geophysical exploration, national shale gas, national coal bed methane, national natural gas hydrate, and other databases), geophysical database (covering 1:1 million, 1:500000, 1:250000, 1:200000, and 1:50000 gravity, national regional gravity, national aeromagnetism, national ground magnetism, national electrical survey, seismic survey, national aviation radioactivity, and national logging database), geochemical database (covering the databases of national 1:250000 and 1:20 geochemical exploration, national multiobjective geochemical and national land quality evaluation results), remote sensing survey database (covering national aeronautical remote sensing image, China resources satellite data, space remote sensing image, national mine environmental remote sensing monitoring, national high score satellite, and other databases), drilling database (covering the national geological borehole information, the national important geological borehole, the Chinese mainland scientific drilling core scanning image library, and so on), hydraulic cycle hazards database, data literature database, special subject database (covering the national mineral resources potential evaluation database, the important mineral “three-rate” investigation and evaluation database), work management

database (covering the national exploration right, mining right, mining right verification, geological information meta-data database, and many others) [17].

For those databases, they are still expanding and accumulating, and their practical values have not yet been fully reflected. However, the vast majority of researchers are virtually impossible to have all of the above data, at most, using their own accumulated data. Anyway, even if their accumulated data, both on the quantity and on type, is incomparable by 10 years and 20 years ago, they are, in fact, in the era of the “relatively big data.” From 1999 to 2004, for example, in “the Chinese mineralization system and regional metallogenic evaluation” project, although there are 202 national academic experts that participated in it, they only master data of 4500 properties (all kinds of minerals). From 2006 to 2013, the study of “national important mineral and regional mineralization laws” was conducted; meanwhile, the mining area covered only by the mineral resources research institute was 30600. Therefore, the increase of information and the amount of data are unprecedented in the last ten years.

3.2. Variety. From the formal point of view, the geological big data have many characteristics, including multidimensionality, multiscale, and multitenses. And they contain structured, semistructured, and unstructured data and usually are stored in forms of text, graphics, images, databases (including image database, spatial database, and attribute database), tables, videos, and audios in a fragmented state. For example, a large number of field outcrop description data, borehole core description data, and all kinds of geological survey, exploration report, and a large number of geological maps, drawings, and photos were stored and managed in the form of paper for a long time; even the numerous relational databases and spatial databases were primarily used to store and manage structured data that are tabulated and vectorized, while the text descriptions, records, and summaries were directly stored. Very few standardized processing and structural transformations were performed. Furthermore, there is no tool available to effectively integrate storage and manage structured, semistructured, and unstructured data.

3.3. Velocity. The increase of geological data is very fast, especially in remote sensing geology, aviation geophysical exploration, regional geochemical exploration, and other fields, due to the introduction of new technologies and new methods. Meanwhile, high speed processing is also a characteristic of big data. In addition to the need of analyzing data in real time, people also need to describe the results of data mining and processing through the use of several data processing techniques, such as image and video, while requiring effective and efficient handling skills. For example, the detection of the deep earth information not only needs to obtain parameters of the seismic wave reflection and refraction but also needs to conduct quick processing, so as to timely predict whether earthquake will occur and forecast the time, location, and intensity. In this way, we can avoid the disaster effectively. When applying a variety of data to a particular mountain, one should learn which ones have

spatial limitations and which are not related to spatiality, so that one deduces the metallogenic law and guides the prospecting better [17].

3.4. Veracity. For the understanding of the value of big data, most people consider it low value density. It means that the real useful information in the vast amount of data is very little. Taking video as an example, the useful data may be only a second or two in the continuous monitoring process. While big data is high value, it does not need to be invested too much; just collecting information from the Internet can bring business value. Therefore, big data has the characteristics of low value density and high business value. The same is true for geological big data. So far, there has been a lot of information about geophysical prospecting, but only a few have been confirmed, and the discovered mines were less. But once a breakthrough was made, its socioeconomic value was enormous, such as the lithium polymetallic deposit in Tibet and the newly discovered Jima copper polymetallic deposit in the outskirts of Sichuan [17].

In addition, the spatial attribute and temporal attribute of geological data also bring a big challenge to data accuracy. Any geological data have spatial attribute, and their values are reflected in the spatial law of distribution of mineral resources. For this reason, in the process of establishing the metallogenic series, exploring the metallogenic law, and constructing the mathematical model, the spatial attribute of the metallogenic model should be considered. Obviously, every metallogenic series has its spatial attribute. Geological data also has the time attribute, which is very different from physical, chemical, and other natural sciences. One of the fundamental pillars of geology is the geological time scale. The rocks, strata, and deposits of different geological periods have different distribution characteristics and regularity, so those data have their own time attribute.

It is obvious that those characteristics of geological big data mentioned above impose very challenging obstacles to the data management in CEGIS. The challenges related to geological big data management can be summarized as follows:

- (i) It is quite difficult to describe and model geological big data, since there are few effective characteristics description mechanisms and object modeling approaches under the cloud computing environment.
- (ii) There remain many technical issues that must be addressed to fully manage, mine, analyze, integrate, and share those geological big data, in consideration of those complex characteristics, including multi-source heterogeneous data, highly spatiotemporal variation, high-volume and high-correlation data, and many others.
- (iii) Many issues appear in achieving decision support, such as data incompleteness, data uncertainty, and high-dimensionality of data.

The broad range of challenges described here make good topics for research within the field of big data management in CEGIS. They are analyzed in the next section.

4. Key Technologies and Trends on Big Data Management in Cloud-Enabled Geological Information Service (CEGIS)

With the rapid advancement of big data technologies, some key technologies are accordingly developed for big data management in CEGIS. Specifically, a schematic diagram of those key technologies is shown in Figure 5. Then, in this section we present an analysis on those key technologies. Meanwhile, the trends along this direction are also discussed.

4.1. Geological Big Data Collection and Preprocessing. Geological big data collection and preprocessing aim to categorize those geological big data obtained from geological data, geological information, and geological literature.

4.1.1. Geological Data Collection Access. In addition to the traditional collection ways, it is also required to carry out large-scale network information access and provide real-time, high concurrency, and fast web content acquisition, combining with the application characteristics in the cloud environment. Currently, considering that the growth rate of geological information generated from the network is very fast, the big data analysis system should obtain relevant data quickly.

4.1.2. Quality and Usability Characteristics of Geological Data. It needs to distinguish and identify valuable information through intelligent discovery and management technologies. Because the information value density contained in different data sources differs from each other, filtering out the useless or low-value data source can effectively reduce the data storage and processing costs. Then, it can also further improve the efficiency and accuracy of analysis.

4.1.3. Geological Data Entity Recognition Model. According to the subject domain of geology, the distributed data are extracted to form a data warehouse, after conducting the operation of processing and integration. When extracting data in the field of geology, it needs to use entity modeling method to abstract entities from the vast numbers of data, so as to find out the relationship between those entities. This approach ensures that the data used in warehouse data can be consistent and relevant in accordance with the data model [19]. These recognized data are directly input into the system, stored as metadata, which could be used for data management and analysis.

4.1.4. Aggregation of Geological Big Data. Generally, different data sources and even the same data source may generate data with different formats. As mentioned above, because these structural, semistructured, and unstructured multimodal geological big data are integrated together, the data heterogeneity is obvious in big data analysis. Then, data aggregation as the key technology in achieving data extraction and transformation [20] enables data sharing and data fusion between heterogeneous data sources. Through

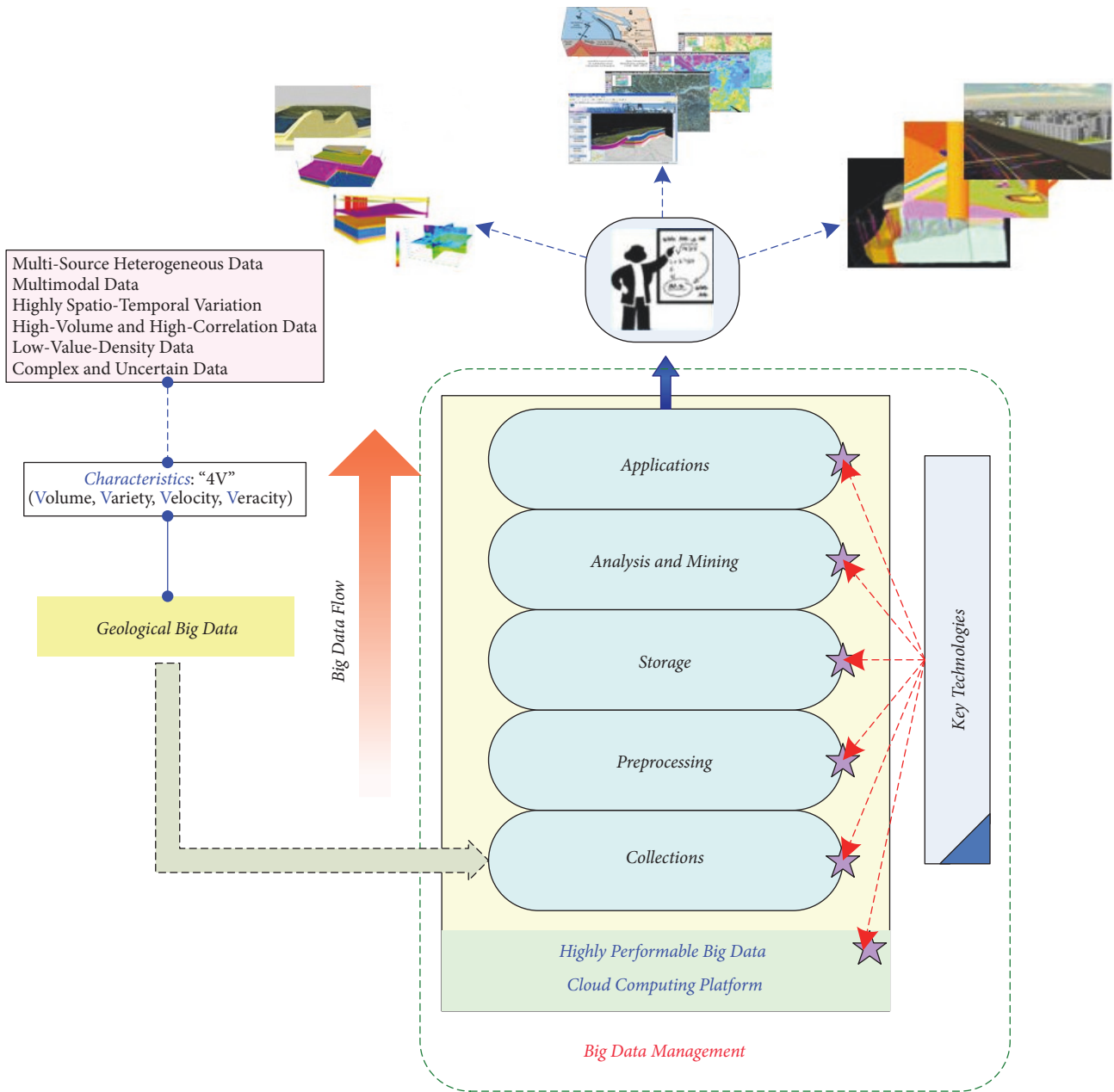


FIGURE 5: Schematic diagram of key technologies for big data management in CEGIS.

the use of heterogeneous information aggregation technologies, the unified data retrieval and data presentation could be achieved. On the basis of it, after aggregating those distributed heterogeneous data sources, they are extracted and converted to achieve the functions of automatically constructing subject domain database and data warehouse [21].

4.1.5. Management of Geological Big Data Evolution Tracking Records. In order to effectively utilize geological big data, it needs to track the evolution of big data during the whole life cycle of GIS, with the purpose of achieving the traceable big data management.

Here, we provide an example of aggregating and collecting geological big data in CEGIS. Figure 6 illustrates this process. While developing CEGIS, all kinds of geological data should be processed. Through the use of geological cloud, big data are collected, and then they are aggregated to achieve some key functions in geological information service platform, including catalog sharing, intelligent searching, data products release, and collaborative service.

4.2. Geological Big Data Storage and Management. From the data collection perspective, geological data can be divided into field survey data, drilling and engineering exploration data, remote detection data, analytical test data, and

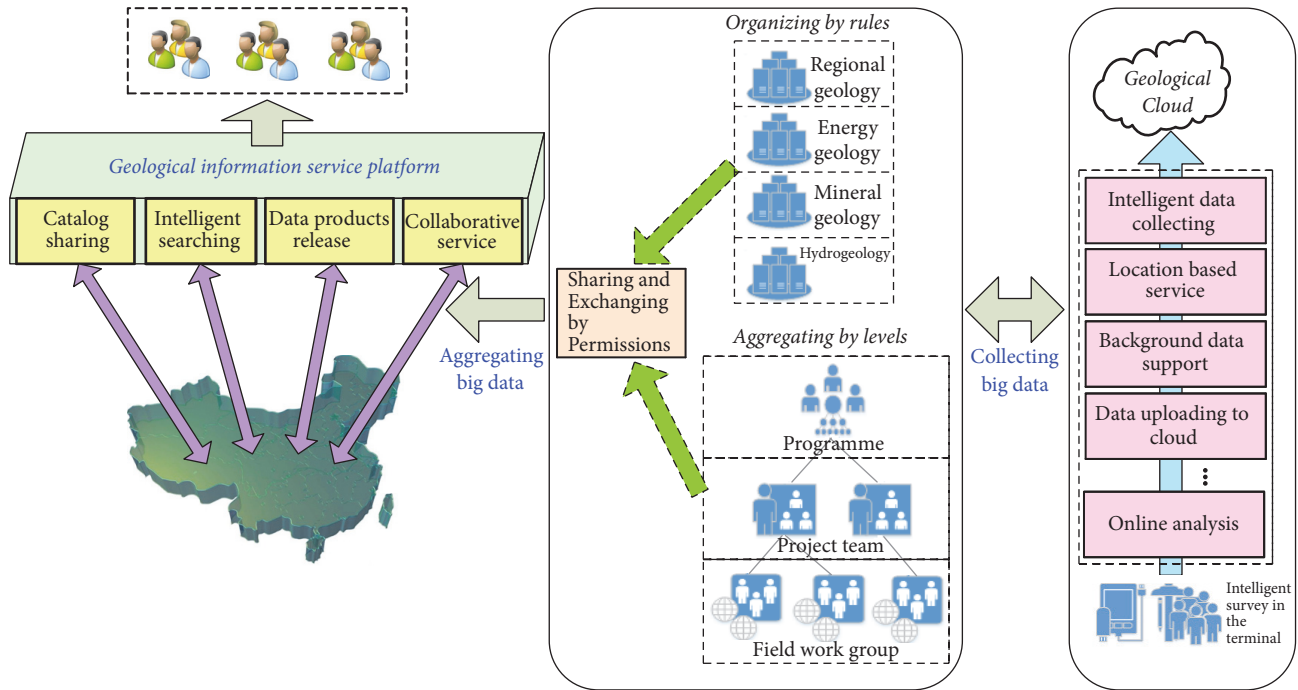


FIGURE 6: An example of aggregating and collecting geological big data in CEGIS.

comprehensive study data. From the angle of comprehensive application fields, they can be also divided into regionally geological survey data, energy and mineral resources evaluation and exploration survey results data, geological disaster monitoring and early warning data, geological environment survey and evaluation results data, and marine geological survey and evaluation data. From the data formality point of view, they can be divided into picture data, text report data, tabular data, and image data. These data are collected by different units.

Facing these complex geological big data mentioned above, the traditional relational database will be difficult to handle them, while the distributed storage system can be used to store such huge amounts of data and manage them. Then, the data system places the massive data in many machines, which avoids such limitation of storage capacity, and also brings many problems that have not occurred before in stand-alone systems. Hence, some distributed data storage solutions have accordingly emerged, including Hadoop, Spark, and other nonrelational database systems (like HBase, MongoDB, and many others) [22]. These different solutions satisfy the specific requirements from different applications. When applying to the analysis of big data, different solutions can be employed according to the specific needs of different intelligence analysis. Furthermore, different solutions can be combined to meet specific needs. Actually, there have been some attempts to develop combination strategies for distributed storage model, varying in the big data management performance requirement, and the complexity of collected big data that are supported by the distributed storage system [23]. Hence, there is still a room for improvement and optimization of geological big data storage, while designing

a hybrid distributed storage model through the use of cloud advantages of flexibly scalable deployment, to meet the users' requirement for geological big data resource management with satisfactory data durability and high availability [23].

Here, the hot research topics include the following:

- (i) For geological applications, the load optimization storage should be implemented to achieve the coupling for data storage and application and the coupling for distributed file system and the new storage system.
- (ii) Based on the application characteristics of distributed databases, more studies could be conducted on the application of new databases NoSQL and NewSQL in geological survey work.

With the development of big data technologies, more and more mature distributed data storage solutions will emerge and will be applied to big data analysis [24, 25].

Specifically, in the management of geological big data, the implementation of data query—for example, spatial query—has been a long-term focus. Generally, considering those advantages with unified modeling language (UML) and computer-aided software engineering (CASE) methodology, the spatial database could be accordingly designed and implemented to characterize and realize the object-oriented spatial vector big data firstly [26]. And then, in the developed spatial database, the function of self-generating codes would be achieved to realize two-way spatial query between graphic-objects and property data [26]. Moreover, in consideration of the complex characteristics of geological big data, the spatial query is achieved finally through the use of Flex technology in ArcGIS Server software platform [27]. Practically speaking,

in this technology, the spatial query could be implemented through two functions, including “Query” and “Find” query methods [28].

4.3. Geological Big Data Analysis and Mining. In terms of geological data analysis and mining, it needs to combine geological data, geological information, and geological literatures, through the analysis of geological application demand of real-time mining, to explore geological big data environment analysis and mining algorithm, in an effort to fully achieve the goal of intelligent mining for geological big data.

Figure 7 shows a schematic diagram of discovering geological knowledge through analyzing and mining geological big data. It can be easily found that geological big data analysis and mining play an important role in achieving the final goal. More relevant research work related to it mainly involves the following aspects.

4.3.1. Geological Big Data Analysis. Considering the special applications, geological big data technologies would apply big data concepts to analyze the metallogenic rules by making full use of various data related to ore, to recognize deposit metallogenic series, to summarize the metallogenic regularities and express in an appropriate way (like voice, image, and many others), and to establish the scientifically mathematical model. The model then uses new exploration data to predict future data and to guide geological prospecting.

In addition, it is necessary to pay special attention to the analysis of new geological big data information collected from social medium and networks [29]. These include the geological text information flow data from microblog web sites, the geological multimedia data from media sharing web sites, the geology-related user interaction data on social networking web sites, and many others [30]. These multisource data complement traditional big data. Specifically, such data should be addressed with the help of multilingual information processing, multilingual machine translation, and social network cross-language retrieval [31]. Big data analysis of such data is a key to deep use of geological data in a broader dimension. With the maturity of big data analysis technologies, it becomes possible to analyze and extract valuable information from these data [32] and to provide effective solutions for geological big data applications.

4.3.2. Geological Big Data Mining. Data mining is to extract the unknown and useful knowledge and information from the massive multilevel spatiotemporal data and attribute data, using statistics, pattern recognition, artificial intelligence, set theory, fuzzy mathematics, cloud computing, machine learning, visualization, and relevant techniques and methods. Data mining could reveal the relationship and evolution trend behind the geological big data, achieve the automatic or semiautomatic acquisition of the new knowledge, and provide the decision basis for resource prediction, prospecting, environmental assessment, and disaster prevention and mitigation [33]. Therefore, the knowledge is obtained directly from known geological data to provide relevant decision support [34]. In consideration of the amount of data, it may

deal with terabytes or even petabytes of data, as well as multidimensional data, all kinds of noise data and dynamic data. Because data mining algorithms will directly influence the outcome of the discovered knowledge, selecting the most appropriate algorithms and parallel computing strategy is the key to data mining.

Effective data mining also could reduce manual intervention during information processing and make use of methods and tools of big data intelligent analysis [35, 36]. Recently, there has been a growing interest in the geological big data mining through the use of some novel computational intelligent methods—for example, rough set [37] and fuzzy aggregation [38]. Moreover, with the development of those neural network based machine learning algorithms in recent years, some popular methods, including extreme learning machine [39, 40], approximate dynamic programming [41], and kernel learning [42], could be used to further improve mining effectiveness for geological big data in the future.

4.4. Highly Performable Big Data Cloud Computing Platform. Highly performable big data cloud computing platform is the foundation for big data analysis. It enables parallel computing for large-scale incremental real-time data and large-scale heterogeneous data [43–46].

With the advent of massive data storage solution, many big data distributed computing frameworks have been proposed. Among them, Hadoop, MapReduce, Spark, and Storm are the most important distributed computing frameworks. These frameworks have different characteristics and solve different problems in applications [47–50]. The Hadoop/MapReduce is often used for offline complex big data processing, the Spark is often employed in offline fast big data processing, and the Storm is often available for real-time online big data processing. Different computing frameworks have their different advantages and disadvantages. Hadoop/MapReduce is easy to program, and it is with satisfactory scalability and fault tolerance. In addition, it is suitable for offline processing of massive data with petabyte level, but it does not support real-time computation and flow calculation. Spark is a memory-based iterative computing framework. By placing intermediate data in memory, Spark can achieve higher iterative calculation performance. The programming model of Spark is more flexible than that of Hadoop/MapReduce, but Spark is not suitable for those applications in which the fine-grained updates are conducted asynchronously. Hence, Spark may be unavailable for those application models that require incremental changes. Storm is suitable for stream data processing. It can be used to handle a stream of incoming messages and can write the processed result to a specified storage device. Another major application of Storm is real-time data processing where data are not necessary to be written into storage devices, which usually results in low time delay. Hence, Storm is particularly suitable for scenarios where real-time online analysis is required to obtain results for big data analysis.

An application example is geological big data aggregation mining framework based on Hadoop [16]. Geological big data aggregation mining platform research is based on the

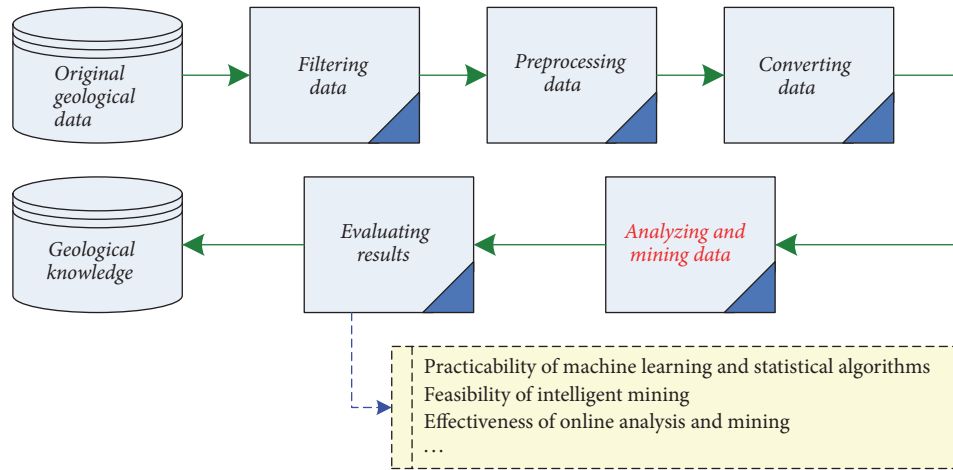


FIGURE 7: Schematic diagram of discovering geological knowledge through analyzing and mining geological big data.

China geological survey data network, and it uses the Hadoop technology to improve and modify existing platform, to make it suitable for big data applications, and to provide a platform for the pilot applications. The geological survey grid platform can be updated in three layers—that is, the virtual layer, the computing layer, and the terminal application layer. The virtual layer is the virtualization of computer resources based on Hadoop distributed file system (HDFS) virtualization technology, which is the foundation of cloud computing and cloud services. The computing layer mainly uses MapReduce method to implement the analysis algorithms for geological big data. Currently, the geological big data technologies mainly use the block calculation strategy to achieve parallel analysis through the utilization of the characteristics of Hadoop, in an effort to speed up the analysis and processing of geological data. The terminal application layer is designed to display the results and receive user feedback to improve system availability.

MapReduce has been used to perform morphological correlation analysis, which involves the analysis of geochemical data processing and the study of the correlation between multi-elements. Figure 8 shows the pattern correlation between elements. It can be seen from Figure 8 that the elements of Mn, Co, and Be are similar in the distribution of morphology. Therefore, from a qualitative point of view, the correlation is relatively high. Moreover, after testing, the proposed prototype system is running three times more quickly than the existing common computing platform, showing that the geological big data is applicable to the Hadoop platform. Furthermore, some applications of using MapReduce could be found in [51].

4.5. Applications of Geological Big Data Technologies

4.5.1. Exploration of Metallogenic Law. The metallogenic law is the human regular knowledge of the temporal and spatial distribution of mineral resources, and its cognitive level, ability, and scope are all related to the size of data, the type of data, and the way of data processing. Therefore, to deduce

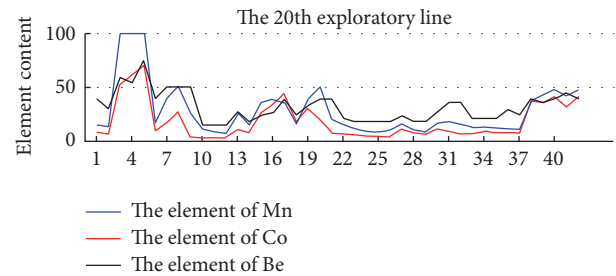


FIGURE 8: Correlation among three element morphologies.

the metallogenic law, it is necessary to fully understand the massive data about spatial distribution, reserves and production in mineral origin, the geological structure of the mineral origin, and related geological survey data. Then, it is to conduct the regular speculation and objectivity expression of these geological big data, so that one can identify the essential reasons for the distribution of mineral origin. Actually, using geological big data technologies could help to translate data into new understanding or knowledge and help to guide the future geological prospecting work.

4.5.2. Smart Prospecting. The types of deposits vary, and the formation of them is related to certain geological backgrounds and geological effects, respectively. The geological backgrounds include tectonic unit and stratigraphic unit, deep upper mantle and lithosphere conditions, and paleogeography and palaeoclimate environment on the surface of the earth. Geological effects include tectonism, magmatism, sedimentation, metamorphism, and weathering. These geological backgrounds and effects in the wide range of space, and in the long geologic history, are a dynamic change and repeated stack, and large deposits can be formed only in a variety of favourable conditions. Long-term scientific research and experience accumulation formed mineral deposit and mineralization prediction subject. Professionals

are guided by certain theories and methods to adopt quantitative or qualitative methods to predict prospecting with the existing knowledge and experience.

However, in view of the difficulties of geological data sharing and the limitations of calculation tools and calculation methods, most of the known deposits in the past are independent of each other. In the future, we can use geological data to connect several adjacent deposit exploration data, conduct unified analysis and specialized processing, determine the “digital” characteristics of the distribution of metallogenic materials, find out metallogenic potential, delineate the abnormal area and prospective area, and promote geological prospecting. Furthermore, geological data informatization and standardization could be improved [52].

4.5.3. Service of People’s Livelihood Geology. After entering the 21st century, geological work is more closely related to economic development, and geological work plays an important role in every aspect of social and economic life. Agricultural geology, urban geology, environmental geology, tourism geology, disaster geology, and other works have been strengthened, and the service area has also been expanded [53]. Meanwhile, the public demand for geological information is increasingly urgent [54].

In order to meet the social demand for geological data, China Geological Survey carried out the construction of geological cloud, which built cluster geologic data service system with the National Geological Information Center and the Provincial Geological Information Center as the backbone nodes, conducted the integration of data resources, and applied the GIS cloud technology, in order to obtain large-scale computing ability and solve those key problems, such as the distributed storage, processing, query, interoperability, and virtualization of massive spatial data [5, 13]. Recently, in China, Shandong Provincial Bureau of Geology and Mineral Resources also carried out the construction of “the application system of geological business based on e-government cloud platform.” It mainly relies on the public service cloud platform of the e-government in Shandong province and constructs the government external network service system and Internet service system to achieve the unified management and information service of the mineral resource. Using technical methods of spatial analysis, big data mining, and three-dimensional geological model, it develops a basic system framework for geological mining services, featured by “a (cloud) platform, a (data) center, and many application systems,” to improve the ability of the people’s livelihood geological service, promote interaction with the public, realize socialization services, and promote the clustering and industrialization of the mineral resources information services.

4.5.4. Application of Knowledge Visualization Service. With the continuous development of web technology, human beings have experienced the “Web 1.0” era, which is characterized by document interconnection, and “Web 2.0”, which is characterized by data interconnection, and are moving towards the new “Web 3.0” era based on the interconnected

knowledge of the entity. Due to the continuous release of user-generated content and linking open data on the Internet, people need to explore knowledge interconnection methods which both conform to the development of the network information resources and meet users’ requirements from a new perspective according to the knowledge organization principles in the large data environment, to reveal human cognition on a deeper level [55].

In this context, knowledge graph (KG) was formally put forward by Google in May 2012, and its goal is to improve the search results and describe the various entities and concepts that exist in the real world and the relationship between these entities and concepts. KG is a great choice to select the essence and discard the dross, as well as the sublimation of the present semantic web technology. In recent years, the applications of KG have been increasing rapidly, and there is now a mature method used to draw a KG and conduct intelligent searching research based on KG [34]. However, the function of KG has not been fully implemented at present, especially for the specific object of geological big data; the application aspect still needs to be further strengthened. Along this direction, the visualization service for geological data in the web-based system is attracting more and more attention [56, 57].

5. Conclusion

Big data technologies make it possible to process massive amount of unstructured and semistructured geological data. And the geological cloud enables us to explore the application of demand-driven geological core data and to extract new information from unstructured data, while supporting the decision-making in land resources management. Thus, the geological cloud could effectively organize and use geological big data, to mine the data scientifically, with the purpose of producing higher value and achieving the corresponding service.

In the architecture of geological cloud, this article describes the application background of CEGIS and the demands from big data management. Furthermore, we elaborate the application requirements and challenges faced in big data management technologies. Then, more analyses are provided from four aspects, including data size, data type, data processing speed, and data processing accuracy, respectively. In addition, this article outlines the research status and technology development opportunities of big data related in CEGIS, from the perspectives of big data acquisition and pre-processing, big data storage and management, big data analysis and mining, highly performable big data cloud computing platform, and big data technology applications. With the continuous development of big data technologies in addressing those challenges related to geological big data, such as the difficulties of describing and modeling geological big data with some complex characteristics, CEGIS will move towards a more mature and more intelligent direction in the future.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by the Key Laboratory of Geological Information Technology of Ministry of Land and Resources under Grant 2017320, the National Key Technologies R&D Program of China under Grant 2015BAK38B01, and the National Key R&D Program of China under Grant 2016YFC0600510.

References

- [1] P. Vermeesch and E. Garzanti, "Making geological sense of 'big data' in sedimentary provenance analysis," *Chemical Geology*, vol. 409, pp. 20–27, 2015.
- [2] J. Chen, J. Xiang, Q. Hu et al., "Quantitative geoscience and geological big data development: a review," *Acta Geologica Sinica*, vol. 90, no. 4, pp. 1490–1515, 2016.
- [3] Y. Zhu, Y. Tan, R. Li, and X. Luo, "Cyber-Physical-Social-Thinking modeling and computing for geological information service system," *International Journal of Distributed Sensor Networks*, vol. 12, no. 11, 2016.
- [4] M. M. Song, Z. Li, B. Zhou, and C. L. Li, "Cloud computing model for big geological data processing," *Applied Mechanics and Materials*, vol. 475–476, pp. 306–311, 2014.
- [5] J. P. Chen, J. Li, N. Cui, and P. P. Yu, "The construction and application of geological cloud under the big data background," *Geological Bulletin of China*, vol. 34, no. 7, pp. 1260–1265, 2015.
- [6] C. Li, "The technical infrastructure of geological survey information grid," in *Proceedings of the 18th International Conference on Geoinformatics*, pp. 1–6, 2010.
- [7] L. Wu, L. Xue, C. Li et al., "A geospatial information grid framework for geological survey," *PLoS ONE*, vol. 10, no. 12, Article ID e0145312, 2015.
- [8] K. Evangelidis, K. Ntoursos, S. Makridis, and C. Papatheodorou, "Geospatial services in the Cloud," *Computers and Geosciences*, vol. 63, no. 2, pp. 116–122, 2014.
- [9] M. Huang, A. Liu, T. Wang, and C. Huang, "Green data gathering under delay differentiated services constraint for internet of things," *Wireless Communications and Mobile Computing*, 2018, <http://downloads.hindawi.com/journals/wcmc/aip/9715428.pdf>.
- [10] <https://www.webofknowledge.com/>.
- [11] C. Yang, M. Yu, F. Hu, Y. Jiang, and Y. Li, "Utilizing Cloud Computing to address big geospatial data challenges," *Computers, Environment and Urban Systems*, vol. 61, pp. 120–128, 2017.
- [12] L. Wu, L. Xue, C. Li et al., "A knowledge-driven geospatially enabled framework for geological big data," *ISPRS International Journal of Geo-Information*, vol. 6, no. 6, article no. 166, 2017.
- [13] Y. Tan, "Architecture and key issues of geological big data and information service project," *Geomatics World*, vol. 23, no. 1, pp. 1–9, 2016.
- [14] Y. Tan, "Architecture investigation of the construction of geological big data system," *Geological Survey of China*, vol. 3, no. 3, pp. 1–6, 2016.
- [15] W. He and Y. Wang, "Prototype system of geological cloud computing," *Progress in Geophysics*, vol. 29, no. 6, pp. 2886–2896, 2014.
- [16] Y. Zhu, Y. Tan, J. Zhang, B. Mao, J. Shen, and C. Ji, "A framework of hadoop based geology big data fusion and mining technologies," *Acta Geodaetica et Cartographica Sinica*, vol. 44, no. S0, pp. 152–159, 2015.
- [17] D. Wang, X. Liu, and L. Liu, "Characteristics of big geodata and its application to study of minerogenetic regularity and minerogenetic series," *Mineral Deposits*, vol. 34, no. 6, pp. 1143–1154, 2015.
- [18] B. Pan and R. Yang, "Management and utilization of big data for geology," *Surveying and Mapping of Geology and Mineral Resources*, vol. 33, no. 1, pp. 1–3, 2017.
- [19] P. Yang and L. J. Lu, "The research on encoding methodology of the character of geological entity based on mass geological data," *Advanced Materials Research*, vol. 962–965, pp. 208–212, 2014.
- [20] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 85–96, 2016.
- [21] C.-L. Kuo and J.-H. Hong, "Interoperable cross-domain semantic and geospatial framework for automatic change detection," *Computers & Geosciences*, vol. 86, pp. 109–119, 2016.
- [22] Y.-J. Wang, W.-D. Sun, S. Zhou, X.-Q. Pei, and X.-Y. Li, "Key technologies of distributed storage for cloud computing," *Journal of Software*, vol. 23, no. 4, pp. 962–986, 2012.
- [23] M. Armbrust, A. Fox, R. Griffith et al., "Above the clouds: a berkeley view of cloud computing," Tech. Rep. UCB/EECS-2009-28, University of California at Berkeley, California, Calif, USA, 2009.
- [24] J. Xia, Z. Bai, B. Wang, J. Chang, and Y. Wu, "Design and implementation of comprehensive management platform for geological data informatization," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 50, no. 2, pp. 295–300, 2014.
- [25] W. Hua, J. Liu, and X. Liu, "Data management of object type geological features on control dictionary," *Earth Science - Journal of China University of Geosciences*, vol. 40, no. 3, pp. 425–430, 2015.
- [26] B. Jia, C. Wang, C. Liu, and W. W. Sun, "Design and implementation of object-oriented spatial database of coalfield geological hazards -Based on object-oriented data model," in *Proceedings of the 2010 International Conference on Computer Application and System Modeling, ICCASM '10*, pp. V1282–V1286, Taiyuan, China, IEEE, October 2010.
- [27] <http://server.arcgis.com/>.
- [28] X. Zhou, X. Li, A. Chen et al., "Design and implementation of the service system of spatial data for geological data," *Journal of Geomatics*, vol. 38, no. 4, pp. 57–60, 2013 (Chinese).
- [29] H. Huang, Z. Chao, and C. Feng, "Opportunities and challenges of big data intelligence analysis," *CAAI Transactions on Intelligent Systems*, vol. 11, no. 6, pp. 719–727, 2016.
- [30] S. Jin, W. Lin, H. Yin, S. Yang, A. Li, and B. Deng, "Community structure mining in big data social media networks with MapReduce," *Cluster Computing*, vol. 18, no. 3, pp. 999–1010, 2015.
- [31] C. C. Yang, C.-P. Wei, and L.-F. Chien, "Managing and mining multilingual documents: Introduction to the special topic issue of information processing management," *Information Processing & Management*, vol. 47, no. 5, pp. 633–634, 2011.
- [32] X. Luo, J. Deng, W. Wang, J.-H. Wang, and W. Zhao, "A quantized kernel learning algorithm using a minimum kernel risk-sensitive loss criterion and bilateral gradient technique," *Entropy*, vol. 19, no. 7, article 365, 2017.
- [33] C. H. Tse, Y. L. Li, and E. Y. Lam, "Geological applications of machine learning in hyperspectral remote sensing data," in *Proceedings of Conference on Image Processing - Machine Vision Applications VIII*, 2015.

- [34] Y. Zhu, W. Zhou, Y. Xu, J. Liu, and Y. Tan, "Intelligent learning for knowledge graph towards geological data," *Scientific Programming*, vol. 2017, Article ID 5072427, 13 pages, 2017.
- [35] H. X. Vo and L. J. Durlofsky, "Data assimilation and uncertainty assessment for complex geological models using a new PCA-based parameterization," *Computational Geosciences*, vol. 19, no. 4, pp. 747–767, 2015.
- [36] A. Gasmi, C. Gomez, H. Zouari, A. Masse, and D. Ducrot, "PCA and SVM as geo-computational methods for geological mapping in the southern of Tunisia, using ASTER remote sensing data set," *Arabian Journal of Geosciences*, vol. 9, no. 20, article 753, 2016.
- [37] Z.-S. Luo and Y.-T. Wei, "Research on Rough set applied in the geological measure data prediction model," *Advanced Materials Research*, vol. 457–458, pp. 792–798, 2012.
- [38] M. Farzamian, A. K. Rouhani, A. Yarmohammadi, H. Shahi, H. A. F. Sabokbar, and M. Ziaie, "A weighted fuzzy aggregation GIS model in the integration of geophysical data with geochemical and geological data for Pb–Zn exploration in Takab area, NW Iran," *Arabian Journal of Geosciences*, vol. 9, no. 2, article no. 104, pp. 1–17, 2016.
- [39] Y. Xu, X. Luo, W. Wang, and W. Zhao, "Efficient DV-HOP localization for wireless cyber-physical social sensing system: a correntropy-based neural network learning scheme," *Sensors*, vol. 17, no. 1, article 135, 2017.
- [40] X. Luo, Y. Xu, W. Wang et al., "Towards enhancing stacked extreme learning machine with sparse autoencoder by correntropy," *Journal of The Franklin Institute*, 2017.
- [41] X. Luo, H. Luo, and X. Chang, "Online optimization of collaborative web service qos prediction based on approximate dynamic programming," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 452492, 9 pages, 2015.
- [42] X. Luo, J. Deng, J. Liu, W. Wang, X. Ban, and J. Wang, "A quantized kernel least mean square scheme with entropy-guided learning for intelligent data analysis," *China Communications*, vol. 14, no. 7, pp. 127–136, 2017.
- [43] J. Passmore, J. Laxton, and M. Sen, "EarthServer for geological applications opening up access to big data using OGC web services," *Advances in Soil Mechanics and Geotechnical Engineering*, vol. 3, pp. 123–129, 2014.
- [44] C. Li, M. Song, L. Xia, X. Luo, and J. Li, "The spatial data sharing mechanisms of geological survey information grid in P2P mixed network systems network architecture model," in *Proceedings of the 9th International Conference on Grid and Cloud Computing, GCC '10*, pp. 258–263, November 2010.
- [45] S. A. B. Cruz, A. M. V. Monteiro, and R. Santos, "Automated geospatial Web Services composition based on geodata quality requirements," *Computers & Geosciences*, vol. 47, pp. 60–74, 2012.
- [46] J. Xia, C. Yang, K. Liu, Z. Li, M. Sun, and M. Yu, "Forming a global monitoring mechanism and a spatiotemporal performance model for geospatial services," *International Journal of Geographical Information Science*, vol. 29, no. 3, pp. 375–396, 2015.
- [47] S. Ibrahim, H. Jin, L. Lu, L. Qi, S. Wu, and X. Shi, "Evaluating MapReduce on virtual machines: the hadoop case," in *Proceedings of the First International Conference on Cloud Computing*, pp. 519–528, 2009.
- [48] M. H. Iqbal and T. R. Soomro, "Big data analysis: apache Storm perspective," *International Journal of Computer Trends and Technology*, vol. 19, no. 1, pp. 9–14, 2015.
- [49] J. L. Reyes-Ortiz, L. Oneto, and D. Anguita, "Big data analytics in the cloud: spark on Hadoop vs MPI/OpenMP on Beowulf," *Procedia Computer Science*, vol. 53, no. 1, pp. 121–130, 2015.
- [50] X. Meng, J. Bradley, B. Yavuz et al., "MLlib: machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, 2016.
- [51] R. Giachetta, "A framework for processing large scale geospatial and remote sensing data in MapReduce environment," *Computers and Graphics*, vol. 49, pp. 37–46, 2015.
- [52] S. Huang and X. Liu, "Geological data informatization and standardization based on geological big data," *Coal Geology of China*, vol. 28, no. 7, pp. 74–78, 2016.
- [53] K. J. A. Kouame, F. Jiang, Y. Feng, and S. Zhu, "The strengthening of geological infrastructure, research and data acquisition - using gis in ivory coast gold mines," *MATEC Web of Conferences*, vol. 95, p. 18001, 2017.
- [54] C. S. J. Karlsson, S. Miliutenko, A. Björklund, U. Mörtberg, B. Olofsson, and S. Toller, "Life cycle assessment in road infrastructure planning using spatial geological data," *The International Journal of Life Cycle Assessment*, vol. 22, no. 8, pp. 1302–1317, 2017.
- [55] K. Stock, T. Stojanovic, F. Reitsma et al., "To ontologise or not to ontologise: an information model for a geospatial knowledge infrastructure," *Computers & Geosciences*, vol. 45, pp. 98–108, 2012.
- [56] J. Hunter, C. Brooking, L. Reading, and S. Vink, "A Web-based system enabling the integration, analysis, and 3D sub-surface visualization of groundwater monitoring data and geological models," *International Journal of Digital Earth*, vol. 9, no. 2, pp. 197–214, 2016.
- [57] R. D. Müller, X. Qin, D. T. Sandwell et al., "The GPlates portal: Cloud-based interactive 3D visualization of global geophysical and geological data in a web browser," *PLoS ONE*, vol. 11, no. 3, Article ID e0150883, 2016.

