International Journal of Innovative Research and Reviews, 5(1) 6-9

International Journal of Innovative Research and Reviews
ISSN: 2636-8919
Website: www.injirr.com
doi:
Research paper, Short communication, Review, Technical paper

REVIEW ARTICLE

Overview of Methods of Collection and Processing of Geographic Data

Nazila Ali RAGIMOVA[1], Vugar Hacimahmud ABDULLAYEV [1, *], Jala JAMALOVA[1]

[1] *Department "Computer Engineering", Azerbaijan State Oil and Industry University*

**Corresponding author E-mail: abdulvugar@mail.ru*

HIGHLIGHTS

- The use of modern methods in the collection and processing of geographic data is crucial in that it makes exhaustive and long works easy and quick as well as more reliable.
- This study outlines such methods and their use in processing of geographic data.

ARTICLE INFO

Received : 10.02.2020
Accepted : 02.08.2021
Published : 07.15.2021

Keywords:
Geoinformatics,
Geographic data,
Big Data,
Internet of Things,
MapReduce

ABSTRACT

On the eve of Industry 4.0, there are global processes of digitalization and intellectualization of many scientific and economic areas. This article examines changes in geography under the influence of advanced information technologies. These technologies are the Internet of Things (for geographic data collection), Cloud Computing (for data storage), Big Data (for data processing), and Cyber Physical Systems (for physical and digital process management required to operate on geographic data). These technologies turn geographies into geoinformatics, which contributes to the further development of this science. The most useful technologies for geographic data collection are the Internet of Things (including things, people, data and processes) and remote sensing of the Earth (for remote geographic data collection). The most useful technologies for processing geographical data are On-Line Analytical Processing (for analytical processing of multidimensional data), Data Mining (for finding patterns in the obtained geographical data), Machine Learning (for deep analysis of geographical data) and MapReduce (for parallel processing of a large amount of data). Using methods, a geographical data processing algorithm develop, which consists of three stages. The first step is to implement the server necessary to form the foundation for data storage and processing. In order for a server to support the operational processing of big data, it must have a distributed file system. The second stage is the design of the database used for the organization and storage of geographical data. The last step is the basic processing and analysis of available geographical data. A paradigm MapReduce uses as an example of data processing.

Contents

Cite this article Ragimova NA, Abdullayev VH, Jamalova J. Overview of Methods of Collection and Processing of Geographic Data. *International Journal of Innovative Research and Reviews (INJIRR)* (2021) 5(1) 6-9
Link to this article: http://www.injirr.com/article/view/57

1. Introduction

With the approach of the Information Age, economics and science began to change. Many interdisciplinary sciences have appeared, derived from informatics and other science. For example, social computing is a joint scientific discipline aimed at studying social research using information methods.

This article will examine the impact of information technology on geography. Geoinformatics formed under the influence of these processes. Geoinformatics is an interdisciplinary science that uses methods and tools of Information Technologies to collect, process, store and present geographic data and its metadata. Geographic data can be seen as a collection of seismological, ecological, metrological, hydrological, geological, spatial and other data that is geographical [1].

During the information era, geoinformatics uses the following advances in information technology: the Internet of Things, Cloud Computing, Big Data, Cyber-Physical Systems, and other technologies.

The Internet of Things (IoT) is a set of smart gauges and sensors connected in a single network, for exchanging information and for interacting with each other and with the outside world [2].

The concept of a Cloud Computing, according to which programs run and output results to a window of a standard web browser on a local PC, while all applications and their data necessary for work are located on a remote server on the Internet. Computers performing Cloud Computing are called the "computing cloud." Additionally, the load between computers included in the "computing cloud" is automatically distributed [3].

Big data is a collection of technologies that are designed to perform three operations. First, to handle large amounts of data compared to "standard" scenarios. Secondly, to be able to work fast with very large volumes of data. That is, there is not just a lot of data, but also they constantly become larger and larger. Third, to be able to work with structured and poorly structured data in parallel in different aspects. Big data suggests that algorithms receive a stream of information that is not always structured at the input and that more than one idea can be extracted from it [4].

Cyber-Physical Systems (CPS) are systems that consist of various natural objects, artificial subsystems and controllers that allow representing such an entity as a whole. The CPS provides close communication and coordination between computing resources and physical resources. Computers monitor and control physical processes using a feedback loop where what happens in physical systems affects computing and vice versa [5].

In order for any information system to function, it must have reliable means of collection and processing of initial data.

2. Materials and Methods

2.1. Collection of geographic data

Geographic data is in a natural environment. Various smart sensors are used to obtain this data, which can be represented as Internet of Things, Internet of Everything and Remote Sensing.

The concept of the Internet of Everything (IoE) is based on the Internet of Things. The main place in the IoT is occupied by things, while the main elements in the IoE, in addition to things, are people, data and processes. The concept of IoE is aimed at connecting all devices to a single network [6].

Remote Sensing refers to the non-contact study of the Earth (also other planets, or satellites), its surface and subsoil, individual objects and other phenomena by recording and analyzing their own or reflected electromagnetic radiation.

The following types of radiation is recorded:

- natural radiation determined by the natural illumination of the Earth's surface by the Sun;
- thermal radiation - the Earth's own radiation;
- artificial radiation, which is generated when the area is irradiated with a source located on the recording device carrier [7].

Recording can be carried out using technical means installed on aerospace aircraft, as well as on the earth's surface. The image may be in the form of a two-dimensional analog recording, for example photographic, or digital recording on magnetic storage devices. If the devices integrate into IoE, then Remote Sensing is a method using which the basic geographical data receive.

2.2. Tools of processing of geographic information

Once the data were received, it is necessary to organize method of the geographic data storage and processing. Given the heterogeneity and diversity of data collected, tools of conventional data processing are not suitable for this role. Therefore, big data processing means should be used here. Common big data methods are On-Line Analytical Processing (OLAP), Data Mining, MapReduce, and Machine Learning.

OLAP is a class of applications and technologies designed for online analytical processing of multidimensional data (collection, storage, analysis) to analyze the activities of the studied object and predict the future state in order to support decision-making. The OLAP technology is used to simplify the work with multi-purpose accumulated data on the activities of the studied object in the past and not to mess with their large volume, as well as to turn a set of quantitative indicators into qualitative ones, using quick, uniform, rapid access to various forms of information presentation. Such forms, obtained from the primary data, allow the user to form a full idea of the object being studied [8].

Data Mining is the process of detecting in "raw" data previously unknown, non-trivial, practically useful, and accessible interpretations of knowledge necessary for making decisions in various areas of human activity. Data Mining is one of the steps of Knowledge Discovery in Databases. The information found in the process of applying Data Mining methods must be non-trivial and previously unknown. Knowledge should describe new relationships between properties; predict the values of some characteristics based on others, etc. The knowledge found should also be applied to new data with some degree of

confidence. The usefulness is that this knowledge can provide some benefit in its application [9].

MapReduce is a model of distributed computing developed by Google, used to reliably perform parallel big data computing on large clusters up to several terabytes in size. The main advantage of MapReduce technology is the ease of scalability of data processing on several clusters. Each of these clusters can consist of several computers; if the number of computers changes, there is necessary simply change the configuration. The MapReduce paradigm consists of "sending a computer to where the data is located," that is, processing big data do on the same cluster where it is stored [10].

Programs that use MapReduce automatically parallels and executed on distributed nodes of the cluster, while the executive system itself takes care of the details of the implementation (splitting the input data into parts, separating tasks into cluster nodes, processing failures and message between distributed computers). Thanks to this, programmers can efficiently use the resources of distributed Big Data systems. [11]

The technology is almost universal: it can use to count words in a large file, create a list of all addresses with the necessary data and other tasks of processing huge arrays of distributed information. In addition, the areas of application of MapReduce include distributed data search and sorting, referencing the graph of web links, processing statistics of network logs, and building inverted indexes, clustering of documents, machine learning and statistical machine translation. The MapReduce is also adapted for multiprocessor systems, voluntary computing, dynamic cloud and mobile environments. [11]

The MapReduce paradigm, like any other technology in addition to the advantages, has disadvantage. Usually, disadvantages mean in view of the situation when this model is undesirable to use.

Below is a list of these cases:

- Real-time processing.
- It is not always very easy to implement everything and everything in the form of the MapReduce application.
- When processing requires many data to shuffle across the network.
- When to process streaming data. MapReduce is best suited for batch processing of huge amounts of data.
- When you can get the desired result using a standalone system. Obviously, setting up and managing an autonomous system is less painful than a distributed system.

When there are OLTP needs. MapReduce is not suitable for a large number of short online transactions. [12]

The most common tool for big data processing is Apache Hadoop, implemented based on MapReduce. Apache Hadoop is a set of libraries and utilities that allows to develop and execute distributed programs running on clusters that are capable of consisting of a large number of nodes (up to several thousand) [13].

Machine learning (ML) is a branch of artificial intelligence. Specifically, this is a technique for analyzing data that allows a machine, robot, or analytical system to be independently trained by solving an array of similar problems. To simplify, ML technology is the search for patterns in the array of presented information and the choice of the best solution without human participation [14].

One of common tool that implements ML is Apache Spark. Spark is a framework of Big Data for distributed batch and streaming of unstructured and weakly structured data, which is part of the ecosystem of Hadoop project [15].

Thanks to this variety of interactive tools of data analytics, Spark actively used in IoT systems, as well as in various business applications, including for ML, for example, for predicting customer outflows and assessing financial risks. Spark can operate both in a Hadoop cluster environment running YARN and without Hadoop core components; in particular, it based on the Mesos cluster management system. Spark maintains several popular distributed storage systems: HDFS, OpenStack Swift, Cassandra and Amazon S3. Spark also guarantees APIs for programming languages often used in the Big Data area: Java, Scala, Python and R.

2.3. Algorithm of Processing of Geographic Data

Given the above, the geographic algorithm of data processing may look like Fig. 1. As can be seen from Fig. 1, the algorithm can be divided into three stages for efficient processing of geographic data: implementation of server, design of database, and main processing of geographic information.

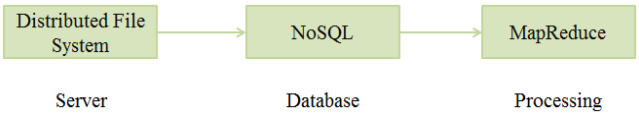

Figure 1 Algorithm of Processing of Geographic Data

Before start processing of data, it needs to implement a server that can guarantee real-time access to data and efficient processing of big data. To guarantee the above, a distributed file system must be implemented on the server. A distributed file system is a file system where clusters of a single file system can reside on different machines, that is, a file system can include several computers. An example of distributed file systems is HDFS, OpenStack Swift, Cassandra, and Amazon S3 [16].

After implementing the server, there is necessary design the database. Given the characteristics of geographical data, relational databases are not suitable for their processing. Instead, there should apply NoSQL technology, which is based on a non-relational database. NoSQL is a scalable storage with an elastic data model. This term is also used to refer to approaches implementing databases that are different from relational database. Unlike traditional database, NoSQL has the following properties: basic availability, flexible state, consistency in the end [17].

There begins the stage of main processing of large geographical data after the design of the database. In general, the MapReduce paradigm consists of three stages: Map, Shuffle and Reduce.

There is received the collected geographic data from the designed database at the input of the Map stage. The main task of this stage is to distribute geographical data by factors. The factor refers to the geographical phenomenon required to sort the data. For example, the geological phenomenon is presented as a geological factor, the same with meteorological, hydrological, ecological and other phenomena. In other words, a combination of factor-geographic data pairs will be obtained at the output.

In the intermediate Shuffle stage, the data obtained at the output of the Map stage sorted by a predetermined factor. Sorting will result in a collection of data that will act as a "value" at the output of this stage, and a GIS will act as a "key." Simply put a pair of values "GIS - list of data" is formed at the output.

In the last Reduce stage, there takes place the main processing of geographical data. The output values of the Shuffle stage are used to map to predetermined geographic factors. Combining these maps for one geographic area produces a smart map. The result is a collection of smart maps for small areas of one region, and then this region can be divided into smaller areas for more effective geographical monitoring. As a result, there is a set of smart maps for each geographic area. This can be represented as a collection of pairs of values "geographical area - smart map." This pair allows to get a Smart Geographic Area.

3. Conclusions

As a result of this study, it can be concluded that geographical data has the character of big data. This means that big data tools should be used to process this data. To do this, there were considered the stages of MapReduce, its input and output data.

It is also important to organize the storage of geographical data. To do this, distributed file systems and NoSQL technology were considered. In addition, geographic data collection tools were considered.

Also, this article pays attention to the advanced achievements of information technologies that can develop geography.

References

[1] Sinha, A., Malik, Z., Rezgui, A., Fox, D., Barnes, C., Lin, K., Heiken, G., Thomas, W., Gundersen, L., Raskin, R., Jackson, I., Fox, P., McGuinness, D., Seber, D., Zimmerman, H. (2010). Geoinformatics: Transforming Data to Knowledge for Geosciences. GSA Today 20(12) 4-10.

[2] Malche, T., Maheshwary, P. (2015). Harnessing the Internet of Things (IoT): A Review. International Journal of Advanced Research in Computer Science and Software Engineering 5(8) 320-323.

[3] Birje, M., Challagidad, P., Goudar, R., Tapale, M. (2017). Cloud Computing Review: Concepts, Technology, Challenges and security. International Journal of Cloud Computing 6(1) 32-57.

[4] Taylor-Sakyi, K. (2016). Big Data: Understanding Big Data. https://www.researchgate.net/publication/291229189_Big_Data_Understanding_Big_Data

[5] Sanislav, T., Miclea, L. (2012). Cyber-Physical Systems – Concept, Challenges and Research Areas. Control Engineering and Applied Informatics 14(2) 28-33.

[6] Internet of Everything vs Internet of Things: What is the difference? https://www.itransition.com/blog/internet-of-everything-vs-internet-of-things

[7] Waghmare, B., Suryawanshi, M. (2017). A Review – Remote Sensing. International of Engineering Research and Application 7(6) 52-54.

[8] Maliappis, M., Kremmydas, D. (2015). An Online Analytical Processing (OLAP) Database for Agricultural Policy Data: A Greek Case Study. HAICTA 2015.

[9] Mostafa, A. (2016). Review of Data Mining Concept and its Techniques. https://www.researchgate.net/publication/301297991_Review_of_Data_Mining_Concept_and_its_Techniques?channel=doi&linkId=5710ef8a08aeff315b9f6deb&showFulltext=true

[10] Khezr, S., Navimipour, N. (2017). MapReduce and its Applications, Challenges and Architecture: A Comprehensive Review and Directions for Future Research. Journal of Grid Computing 15(3) 1-27.

[11] MapReduce, https://www.bigdataschool.ru/wiki/mapreduce

[12] What are the disadvantages of MapReduce?, https://stackoverflow.com/questions/18585839/what-are-the-disadvantages-of-mapreduce

[13] Polato, I., Re, R., Goldman, A., Kon, F. (2014). A Comprehensive View of Hadoop research – A Systematic Literature Review. Journal of Network and Computer Applications 46 1-25.

[14] Alzubi, J., Nayyar, A., Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. Journal of Physics Conference Series 1142 1-15.

[15] Parwej, F., Akhtar, N., Perwej, Y. (2018). A Close-Up View About Spark in Big Data Jurisdiction. International Journal of Engineering Research and Application 8(1) 26-41.

[16] Unver, M., Erguzen, A. (2016). A Study on Distributed File Systems: An Example of NFS, Ceph, Hadoop. ICENS 2016 1-5.

[17] Gupta, A., Tyagi, S., Panwar, N., Sachdeva, S., Saxena, U. (2017). NoSQL Databases: Critical Analysis and Comparison. 2017 International Conference on Computing and Communication Technologies for Smart Nation 293-299