XIV ГЛОБАЛЬНЫЕ НАУКИ И ИННОВАЦИИ 2021: ЦЕНТРАЛЬНАЯ АЗИЯ

МЕЖДУНАРОДНЫЙ НАУЧНО-ПРАКТИЧЕСКИЙ ЖУРНАЛ

22-27 октября 2021, Нур-Султан (Астана), Казахстан

CONSOLIDATION OF LEGAL ENTITIES IN THE FORM OF AN ASSOCIATION «NATIONAL MOVEMENT «BOBEK» CONGRESS OF SCIENTISTS OF KAZAKHSTAN

ISSN 2664-2271

«GLOBAL SCIENCE AND INNOVATIONS 2021: CENTRAL ASIA»

No. 3(14). OCTOBER 2021
SERIES "TECHNICAL SCIENCES"
The journal was founded in 2018.

II VOLUME

BIG DATA, CHARACTERISTICS OF BIG DATA. USE OF BIG DATA IN CONTROL OF COMPLEX PROCESSES

[1]A.R.Akhatov, [2]A.E.Rashidov, [3]L.Y.Xurramov
[1,2,3]Samarkand State University

Annotation. *The features of the Big Data-type information and the level of study of this type of data were systematically analyzed. The problems found in large data and the approaches to solving these problems and the methods of detection of complex processes in these processes and methods of detection through the use of Big Data technology.*

Keywords. *Flow of data, Big Data, 3V characteristics, various sources of data, data reliability, variability, visualization, structural and unstructural data.*

By the end of the twentieth century, the digitization of a lot of data and the support of this process by many countries and organizations around the world have led to the emergence of sources that generate a very large flow of information from the beginning of the XXI century. As a result, today a large flow of data is generated from these sources, which is difficult to store and process using modern information and communication technologies. Such rapid growth of data has led to the emergence of a new term in the field of science for the storage and processing of this data - "Big Data". The first definition of the term was introduced in 1997 by two NASA researchers, Michael Cox and David Ellswort, and in 1998 by Silicon Graphics Inc (SGI) researcher J.R.Mashey used this concept, and a year later Bryson and other researchers published an article on Big Data in communications at the Association of Computing Machines (ACM) (1999).

There is no clear official definition of the concept of “Big Data”, most of the initial definitions are based on the reason for the origin of this concept such as the volume of data. Some of these definitions are given below.

Microsoft researchers Daniel Fisher, Rob De Line, Mary Czervinski, and Steven Drucker describe their research as "Big data is a large set of data that cannot be stored, transmitted, or processed in a simple way."

McKinsey Global Institute researchers James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Bers described Big Data in their research as in the following:

Big Data are data sets that can not be retrieved, stored, managed, and analyzed by conventional database software in terms of size [2].

Another group of definitions given to Big Data focuses not only on its size but also on other features. One such definition is that of foreign scholars A.D. Mauro, M. Greco and M. Grimaldi states in one of their researches: Big Data is an information assets characterized by high volume, speed, and diversity that requires specific technology and analytical methods to convert it into value [3]. In this definition, the authors emphasize that Big Data has three properties (3 V), Volume, Velocity, and Variety.

Based on the results of systematic scientific articles and research, many scientists have stated in their research that Big Data has exactly the three characteristics (dimensions) mentioned above, and the data flow that combines these features is the basis for calling Big Data. But some scientists are adding a few more features of Big Data in addition to these “3 V” features. For example:

- value (Jean-Pierre Dijcks and, researchers of Fujitsu, I. Mitchell, M. Locke, M. Wilson, A. Fuller [4, 5]);
- veracity (researchers of IBM, S. Miele and R. Shockley [6]);
- veracity and variability (Russian researchers I.A. Radchenko and I.N.Nikolaev add [7]);
- value, veracity, variability and complexity (A. Gandomi and M. Haider [8]);

- veracity, value, variability, visualization, valuation and viability (Vishnu Pendyala the properties of [9]);
- volatility, validity, and vulnerability, veracity, value, variability and visualization (V.A.Raznichenko [10]).

What these features of Big Data mean is shown below:

Volume. Represents the unwanted flow of data and the amount of data stored. The size of the data volume is an important indicator that they can be considered Big Data. Foreign scientists P. Elmasri, B. Navathe in their research states that in today's context, the range of data volumes considered to be Big Data varies from terabytes or petabytes to exabytes [11]. But most foreign scientists have not come up with scientific opinions or conclusions about putting a strict limit on this range of Big Data volumes. One reason for this is that while data from any active source will one day produce large volumes of data, the second reason is that the time limit set for the size of Big Data data will change over time. About this Fujitsu scientists I. Mitchell, M. Locke, M. Wilson, A. Fuller says in their research: "Yesterday's big is today's normal. Some of those over the age of 40 who have read this book will probably remember how they thought about filling up the 1 kilobyte of memory on the Sinclair ZX81. Big Data simply means a larger amount of data than today's norm." [5]

Foreign scholars Paul Zikopoulos and others, in their research on Big Data and its size feature, say: "The term Big Data is a bit misleading because the data already available is somewhat small or the only difficulty is not in its size. In short, the term Big Data refers to data that cannot be processed or analyzed using traditional processes or tools "[12].

Velocity. Big Data Rate refers to the speed at which data is generated (growth rate) and processed for analysis. In other words, speed refers not only to the speed at which data enters a database, but also the speed at which important data is retrieved from that database. Typically, Big Data requires real-time high-speed data analysis. To do this, Big Data data processing technology must be able to analyze them before entering the database.

Velocity is one of the most important features of Big Data in terms of practical use. Fujitsu scientists I. Mitchell, M. Locke, M. Wilson, A. Fuller also supports this idea in his research through the following two ideas [5]:

1) "Petabytes (1 million gigabytes) seem big to people right now, but tomorrow this size will become normal and over time it will also become smaller in our eyes. So forget the size - Big Data is about speed of making decision. In business terms, Big Data is a quick response''.

2) "Reuters information agency has been dealing with Big Data for 100 years because its business model is based on getting relevant news from multiple sources quickly and delivering it to the right people as quickly as possible".

Ukrainian scientist V.A. Reznichenko in his research argues that the Big Data data category includes a large proportion of data streams above 100 Gbyte per day [10]. In conclusion, according to Reznichenko, a group of data that generates a data flow at an average speed of more than 1.19 Mbyte per second during the day can be called Big Data.

Variety is the ability to receive, store, and process different data from different sources. That is, diversity means:

- Availability of data from various sources. Examples of such sources are: social networks, mobile networks, audio and video surveillance devices, metrological data sources, various signaling devices, GPS, people recording and storing information about them, automatic data collectors, and so on.
- Variety of data receiving and storing formats: text, audio, video, images.
- Semantic diversity. That is, the same information is presented in different ways. For example, the presentation of information that a student has received five or three grades in terms such as excellent or satisfactory.

- Different structure of data. According to the structure of the data are divided into 3 groups: 1) structured; 2) unstructured; 3) semi-structured. Structured data refers to data presented in a rigid form used in traditional relational databases, where the meanings have clearly defined attributes (e.g. if there are name, surname, address attributes for the students entity, all lines in the corresponding table will have no fields other than first, last name, address fields). Therefore, they will be relatively easy to search, categorize, sort. In some cases, data is collected in a special way before it is known how it will be stored and managed. This data may have a specific structure, but not all of the data collected will have the same structure. Some attributes can be distributed between different objects, but other attributes can only exist in a few instances. In addition, some objects can be added additional attributes at any time and do not have a pre-defined schema. This type of data is called semi-structured data. Unstructured or unstructured data is data that does not have a predefined form. Typically, such information includes text documents, web pages, Twitter twits, video, audio, and image information that contain some information.

Vareity is also one of the key features of Big Data, and this is emphasized in most of the definitions given to Big Data by many foreign scholars. One such definition is given by foreign scholars R. Kazi, A. The Lions stated in their researchs: “Big Data is a very large amount of structured and unstructured data”[13].

Veracity is a feature that characterizes the accuracy and reliability of data. In his research, Vishnu Pendyala writes, “Reliability is a crucial aspect of using Big Data and getting results from it’’, and emphasizes the importance of the verasity feature [9]. Big Data access technology does not imply that carefully selected and validated data enter the database. The initial data can be "raw", i.e they can come to the database without any initial processing, random and with a lot of "noise". While Big Data offers great opportunities for analysis and decision making, its value largely depends on the quality of the original data. Big Data technology takes this feature into account and allows you to work reliably with such data.

Value - This explains its importance in terms of practical problems. Ishvarappa and J.Anuradha state their opinions on the importance of value as an important feature of Big Data in his research: “The potential value of Big Data is so great that value is the most important aspect of big data” [14]. Azerbaijani researchers Z.T. Magerramov, V.G. Abdullaev, A.Z. Magerramova argues in her research that value is a key feature that allows Big Data to be distinguished as a new phenomenon [15].

Variability refers to any change in the data in motion over time, including changes in flow rate, format, or content.

Volatility - refers to how long the data is valid and how long it should be stored. Relevance depends on the size of the Big Data (how much data needs to be stored and how much does it cost to store?), value (how long does the data remain up to date, and does it create value?), and the need to process the data.

Validity. This feature is closely related to reliability and describes how accurate and inaccurate the available data is in terms of its intended use.

Visualization. The sheer size and complexity of the data is so immense that many organizations struggle to gain any competitive advantage by analyzing them. Big Data should be presented in a way that is easy to understand for different users. Visualization is the process of making a very large amount of information understandable and readable.

Viability is the process of carefully selecting features and factors that are most likely to predict the most important outcomes for an enterprise. Vishnu Pendyala describes the nature of vitality as follows: “Vitality means that not everything about Big Data is useful. It is only necessary to assess the viability of existing attributes to select properties, which helps to determine the value”[9].

Valuation is the process of extracting value from Big Data. The valuation feature should not be confused with the value feature. Because the first is a process and the second is a result [9].

From the above descriptions of Big Data and its properties, we can conclude that when a data group appears in Big Data, its size, speed, and diversity properties are manifested. The remaining features of Big Data are usually manifested not in determining its existence, but in its use i.e. in the process of processing, analyzing, obtaining the result and determining its importance. These features themselves show that the study of Big Data and its technologies is one of the most pressing issues today. The urgency of studying Big Data, which has already entered our lives today, is also reflected in two aspects, such as its features:

1) The urgency of implementing large-scale, fast and diverse data storage technologies, which are growing day by day;

2) The urgency of discovering interesting and useful aspects of big data, which are not yet known to us, through the introduction of technologies for rapid processing of data collected and entered into the database in real time.

Today, the effective use of Big Data-type data provides great opportunities in all areas, as well as in the control of complex processes and the detection of small changes in these processes. But this is a bit complicated process and requires the use of special methods and algorithms. One of the main reasons for the complexity is the large number of data types and the size of the data volume. The largest volumes of data in these processes are usually image and video format data. The source of this data is surveillance cameras and similar sensors. Typically, this data falls into the category of large unstructured data. Sometimes there is a need for real-time analysis of this video data. In this case, the use of Big Data technology is difficult. That is, before storing data, it is necessary to use processing technologies (Figure 1).

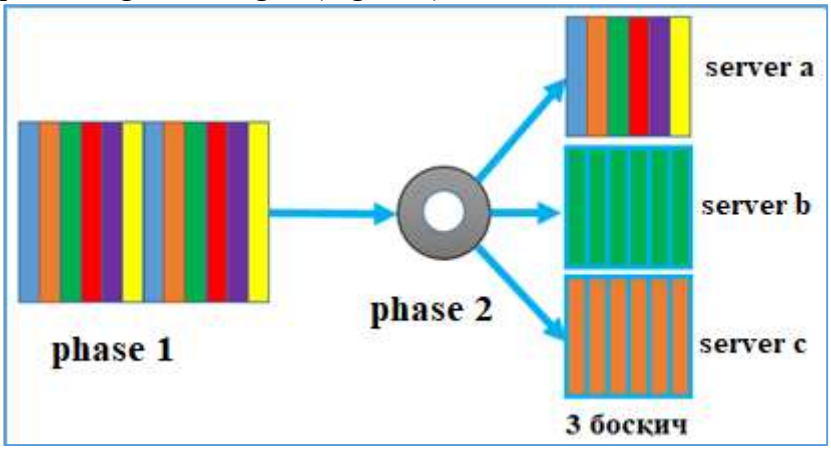

Figure 1. The process of separating structured data from video data

In phase 1 of the process described in Figure 1, unstructured video data is obtained from the surveillance camera. In phase 2, the required data is extracted from the video data obtained in phase 1 in real time. In phase 3, the separated data and primary data are allocated to separate servers (e.g. the data is placed on the server b and the surveillance camera and the time-determining data are placed on the server c). Even today, data on a server is used to control complex processes and to analyze data, and is usually deleted after a certain period of time due to the inability to store it for a long time. Because the data on servers b and c do not occupy a large volume and are structured, they can be used in real-time or in the analysis of the status of processes or similar data over a period of time.

Conclusion

Today, Big Data is becoming increasingly important for a variety of industries. There is a growing urgency today to collect, store, and process large amounts of data to gain valuable insights, make important decisions, and justify support for future actions. Research has shown that

although Big Data and its applications in the field are being studied extensively, Big Data has not yet been formally defined. However, two-thirds of studies on Big Data have given different definitions of Big Data [10]. Officially, the absence of a single tariff means that research on Big Data is still ongoing. Based on most of the existing definitions, Big Data refers to a large amount of data flow that refers to a variety of data sets that cannot be received, managed, and processed in a short period of time by traditional information systems and software and hardware.

Big Data, unlike traditional data, uses a large number of sometimes unrelated data sources to identify previously unrelated relationships. This provides new optimal solutions for controlling complex processes, as in all areas, and for detecting small changes in these processes.

REFERENCES:

1. D. Fisher, R. DeLine, M. Czerwinski, S. Drucker, “Interactions with Big Data Analytics”, Interactions 2012.

2. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, Ch. Roxburgh, A. H. Byers “Big Data: The next Frontier for Innovation, Competition, and Productivity”, McKinsey Global Institute, May 2011

3. A.D. Mauro, M. Greco, and M. Grimaldi, “What is Big Data? A Consensual Definition and a Review of Key Research Topics” International Conference on Integrated Information, 2015.

4. J.P. Dijcks, “Big Data for the Enterprise”, Oracle report (2012).

5. I. Mitchell, M. Locke, M. Wilson, A. Fuller, “Big Data. The definitive guide to the revolution in business analytics”, THE WHITE BOOK OF... 2012.

6. S. Miele, R. Shockley, “Analytics: The real-world use of big data”, IBM report, April 2013.

7. И.А. Радченко, И.Н. Николаев “Технологии и инфраструктура Big Data” - СПб: Университет ИТМО, 2018. - 52 с.

8. A. Gandomi, M. Haider “Beyond the hype: Big data concepts, methods, and analytics”, International Journal of Information Management 35 (2015) 137-144

9. V. Pendyala, “Veracity of Big Data”, San Jose, California, USA, 2018

10. В.А. Резниченко, “Что такое Big Data”, Проблеми програмування 2019. № 3, 86-100 с.

11. R. Elmasri, B. Navathe, “Fundamentals of Database Systems”, Pearson 2016.

12. Ch. Eaton, D. deRoos, T. Deutsch, G. Lapis, P. Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, The McGraw-Hill Companies, 2012.

13. R.U. Rehman Qazi, A. Sher “Big Data Applications in Businesses: An Overview”, The International Technology Management Review, Vol. 6 (2016), № 2, 50-63

14. Ishwarappa, J. Anuradha “A Brief Introduction on Big Data 5 Vs Characteristics and Hadoop Technology”, Procedia Computer Science 48 (2015) 319 – 324

15. З.Т. Магеррамов, В.Г. Абдуллаев, А.З. Магеррамова “BIG DATA: проблемы, методы анализа, алгоритмы”, РИ, 2017, №3