МРНТИ 20.23.17
УДК 004.45

https://doi.org/10.51889/2021-3.1728-7901.09

BIG DATA PROCESSING IN THE DIGITALIZATION OF ENTERPRISE ACTIVITIES

Balakayeva G.T. [1], *Darkenbayev D.K.* [1], *Turdaliyev M.*[1]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
*e-mail: dauren.kadyrovich@gmail.com

Abstract

The growth rate of these enterprises has increased significantly in the last decade. Research has shown that over the past two decades, the amount of data has increased approximately tenfold every two years - this exceeded Moore's Law, which doubles the power of processors. About thirty thousand gigabytes of data are accumulated every second, and their processing requires an increase in the efficiency of data processing. Uploading videos, photos and letters from users on social networks leads to the accumulation of a large amount of data, including unstructured ones. This leads to the need for enterprises to work with big data of different formats, which must be prepared in a certain way for further work in order to obtain the results of modeling and calculations. In connection with the above, the research carried out in the article on processing and storing large data of an enterprise, developing a model and algorithms, as well as using new technologies is relevant. Undoubtedly, every year the information flows of enterprises will increase and in this regard, it is important to solve the issues of storing and processing large amounts of data. The relevance of the article is due to the growing digitalization, the increasing transition to professional activities online in many areas of modern society. The article provides a detailed analysis and research of these new technologies.

Keywords: MongoDB, data, technology, processing, analysis.

Аннотация
Г.Т. Балакаева [1]*, Д.К. Даркенбаев* [1] *, М. Турдалиев* [1]
[1]*Казахский Национальный университет им. аль-Фараби, г. Алматы, Казахстан*

ОБРАБОТКА БОЛЬШИХ ДАННЫХ ПРИ ЦИФРОВИЗАЦИИ ДЕЯТЕЛЬНОСТИ ПРЕДПРИЯТИЯ

Темпы роста данных предприятий значительно возросли в последнее десятилетие. Исследования показали, что за последние два десятилетия объем данных увеличивается примерно в десять раз каждые два года – это превысило закон Мура, который удваивает мощность процессоров. Каждую секунду накапливается около тридцати тысяч гигабайт данных, и их обработка требует увеличения эффективности обработки данных. Загрузка видео, фотографий и писем пользователей в социальных сетях приводит к накоплению большого объема данных, в том числе неструктурированных. Это приводит к необходимости работать предприятиям с большими данными разных форматов, которые должны быть определенным образом подготовлены к дальнейшей работе в целью получения результатов моделирования и вычислений. В связи с вышеизложенным проведенные в статье исследования обработки и хранения больших данных предприятия, разработка модели и алгоритмов, а также использование новых технологий является актуальным. Несомненно, с каждым годом информационные потоки предприятий будут нарастать и в связи с этим важно решать вопросы хранения и обработки больших объемов данных. Актуальность статьи обусловлена растущей цифровизацией, возрастающим переходом на профессиональную деятельность в режиме онлайн во многих сферах жизни современного общества. В статье проводится подробный анализ и исследование этих новых технологий.

Ключевые слова: MongoDB, большие данные, технологии, обработка, хранение, анализ.

Аңдатпа
Г.Т. Балақаева [1] *, Д.Қ. Даркенбаев* [1]*, М. Тұрдалиев* [1]
[1]*әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы қ., Қазақстан*

КӘСІПОРЫНДАРДЫҢ ҚЫЗМЕТТЕРІН ЦИФРЛАНДЫРУДА ҮЛКЕН КӨЛЕМДІ ДЕРЕКТЕРДІ ӨНДЕУ

Соңғы онжылдықта кәсіпорындардың деректерінің өсу қарқыны айтарлықтай біліне бастады. Зерттеулер көрсеткендей, соңғы екі онжылдықта мәліметтер саны әр екі жыл сайын он есеге көбейіп, процессорлардың қуатын екі есе арттыратын Мур заңынан да асып түсті. Әр секунд сайын шамамен отыз мың гигабайт деректер жинақталады және оларды өңдеу тиімділігін арттыруды қажеттілігі туындады. Әлеуметтік желілерде қолданушылардан бейнелерді, фотосуреттер мен хаттарды жүктеу көптеген деректердің, соның ішінде құрылымданбаған деректердің жиналуына әкеледі. Бұл кәсіпорындардың әр түрлі форматтағы үлкен деректермен жұмыс жасау қажеттілігіне әкеледі, оларды модельдеу мен есептеу нәтижелерін алу үшін одан әрі жұмыс істеу үшін белгілі бір жолмен дайындау керек. Жоғарыда айтылғандарға байланысты мақалада ірі кәсіпорын деректерін өңдеу және сақтау, модель мен алгоритмдер құру, сондай-ақ жаңа технологияларды қолдану бойынша жүргізілген зерттеулердің өзектілігі зерттелген. Жыл сайын кәсіпорындардың ақпараттық ағындары арта түсетіні сөзсіз, осыған байланысты үлкен көлемдегі деректерді сақтау және өңдеу мәселелерін шешу маңызды. Мақаланың өзектілігі цифрландырудың өсуіне, қазіргі қоғамның көптеген салалары кәсіби қызметтердің онлайн режимге ауысуына байланысты болып отыр. Мақалада осы жаңа технологияларға кеңінен зерттеулер жүргізілген.

Түйін сөздер: MongoDB, деректер, технологиялар, өңдеу, талдау.

1. INTRODUCTION

Big data is defined as aggregated datasets that include structured and unstructured data that are important in size and vary in structure. Usually, big data refers to the process of constant accumulation of various unstructured data [1]. The rapidly growing volume of data presents us with new and challenging storage and processing challenges. Large amounts of data pose complex challenges for traditional systems. According to IDC, the digital universe in 2020 will be 40 zeta bytes, and since 2010 the amount of data has increased 50 times [2].

Large-scale data processing challenges the development of business processes and the transformation of information flows into intelligent digital resources. The concept of big data does not require very strict definitions. This concept describes an exponentially growing set of data that is large, raw, and unstructured for analysis using relational database techniques. It is important to understand not the size of the increase, such as terabytes or pet bytes, but how to handle it. Big data was introduced on September 3, 2008 by Clifford Lynch, editor of the journal Nature. In his article "The Impact of Large-Scale Data Processing Technologies on the Future of Science", he showed that data is growing rapidly, and their processing is one of the most pressing problems in the future [3].

The informational value of big data is clear. The tasks that are solved when analyzing the information flow of big data are as follows:

- Traffic forecasting based on traffic analysis from call centers, technical support services and the site;
- Creation of forecast models;
- Fraud detection in real time;
- Risk analysis;
- Operational analytical processing, etc. [4].

The move from analog to digital has led to an increase in the volume of business data on a daily basis. There were 1 trillion gigabytes of data recorded in 2010, according to IDC. This data is driving the use of billions of phones, tens of billions of social media posts, an increase in the number of sensors connected to the Internet in cars, the growth of POS terminals, and more. can be directly related to the increase in the number of devices [5]. In many cases, data analysis is perceived as large-scale data processing. In fact, if you are faced with different unstructured data, you don't need to pay much attention to their size. Today, there is no one-size-fits-all method or algorithm for large-scale data processing that is suitable for every situation. Every time we get the necessary knowledge from raw data, it becomes necessary to create a specific method and algorithms for each task. Experts from all over the world are studying the development of large-scale data processing methods and how this will affect the prospects of enterprises. Long-term storage of data in a warehouse can not only cause various difficulties, but can also be beneficial if properly stored and processed. By 2020, data is projected to reach 5,200 GB per capita, including seniors and babies, of which only 15% will be written to the cloud.

The volume of data is expected to double each year. When IDC analyzes data for 2020, 33% believe it is valuable data. A person studying big data should pay special attention to 3 questions for data processing. These are issues of collecting, processing and storing data. Thus, we can say that Big Data is a complex system. Each system has its own function and can be easily integrated with other systems.

NoSQL is a generic description that differs significantly from traditional models of accessing information through the SQL language, aimed at implementing database management mechanisms. NoSQL has flexible state that can change over time and is available for every request. NoSQL database can be used for all information models - text, graphics, documents using a key value pair. There are different databases for the term NoSQL, but they all have the same characteristics. Depending on the nature of working with data, you can choose databases and work with them at your discretion.

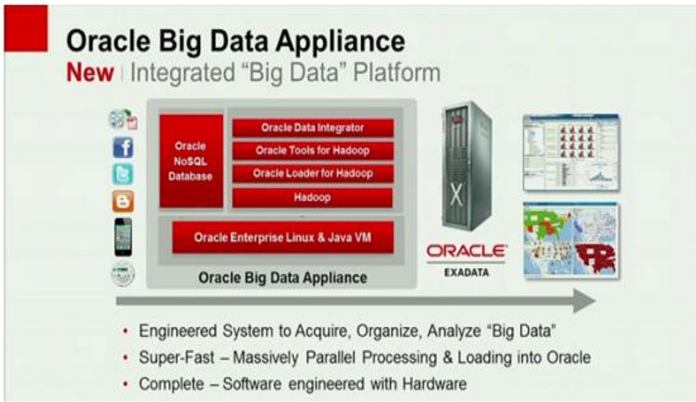
A document database associates each key with a complex data structure called a document. A document key is a collection of value pairs. MongoDB is a sample document storage database. A MongoDB document group is called a collection. This is the equivalent of a DBMS schedule.

Graphical storages are used to store information about data transmission networks such as social networks. Graphic stores include Neo4J and Giraph.

Database key-value store stores each individual item in the database as a key along with its value. Examples of key-value storage include Riak and Berkeley DB. Some key and value repositories, such as Redis, allow each value to be in integer form, which adds functionality.

2. EXISTING PRODUCTS FOR STORING AND PROCESSING LARGE AMOUNTS OF DATA

Oracle Big Data Appliance is used to solve problems of processing, collecting and structuring data. It is a tool that Oracle NoSQL Database can integrate with the pre-installed Hadoop cluster and other data stores. Oracle Big Data Appliance is based on the processing and storage of unstructured or slightly unstructured data. This indicates that these tasks perform well in the Hadoop database [6].



Oracle Big Data Appliance

New | Integrated “Big Data” Platform

- Engineered System to Acquire, Organize, Analyze “Big Data”
- Super-Fast – Massively Parallel Processing & Loading into Oracle
- Complete – Software engineered with Hardware

Figure 1. Example of data storage in Oracle Big Data Appliance [7]

Gartner, an IT research and analysis company, has written three key characteristics of large data in its articles and labeled them as "three V's" [8, 9]:

- volume (translated from English "volume") - the physical volume of stored data;

- speed (translated from English "velocity") - the rate of change of data and subsequent analysis of these changes;

- Versatility (translated from English "variety") - a variety of processed data, structured and unstructured data.

The work of the model with large data is becoming more and more popular. It was implemented by the Apache Software Foundation in the Apache Hadoop project. Apache Hadoop consists of 2 components: the Hadoop Distributed File System (HDFS) distributed cluster system and the Map Reduce software interface.

Hadoop is a software platform for creating distributed applications for mass processing of parallel (MPP) data. The Hadoop platform can be divided into two main components:

Hadoop Distribution File System (HDFS) - a distributed file system that provides high-speed access to application data;

MapReduce is a software platform for distributing and processing large amounts of data in a computing cluster [10].

A special type of NoSQL database was created to solve the problem of large-scale data processing. A comparison of the relational database and NoSQL properties is given in Table 1.

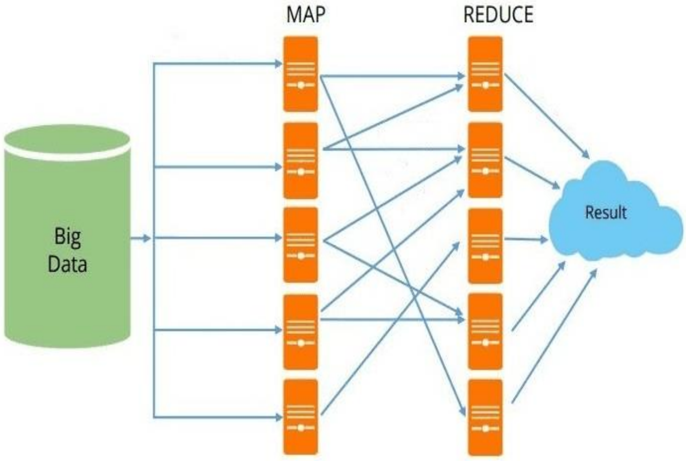


Figure 2. Map-Reduce technology map [10]

The comparison of relational database and NoSQL properties is tabulated:

Table 1. Comparison of database properties

Relational data fund	*NoSQL data fund*
Complex data relationships	*Very simple relationship*
Centralization of drawings	*Optional scheme:Unstructured data*
Scalability	*Allocated edits*
Static memory	*Memory is scaled with computing resources*
Universal functions and properties	*The system is aimed at applications and developers*

NoSQL technology (for example, Cassandra) is not designed to replace a relational database, but rather to help solve problems when the amount of data is too large. NoSQL often uses clusters of inexpensive standard servers. This solution allows you to reduce the cost of one gigabyte per second several times [11].

3. ANALYTICAL TECHNOLOGIES FOR LARGE-SCALE DATA PROCESSING

Today, many organizations have a large amount of data. Depending on the task at hand, different methods and techniques must be used to process this data. There is a wide range of data storage equipment, some of which use special storage technologies for the convenience of their further processing and analysis: multiprocessor systems [12].

Multiprocessing (Multiprocessing, Multiprocessing, Eng. Multiprocessing) - the use of a pair or more physical processors in one computer system. It also means the system's ability to support multiple processors or distribute tasks between them. There are many variations of this concept, and the definition of multiprocessing can vary depending on the context, mainly how the processors are defined.

During the years of development of computer technology, multiprocessor systems have gone through a number of stages of development. Historically, SIMD technology was the first to be mastered. However, now there is a constant interest in MIMD architecture. This interest is mainly determined by two factors:

SISD processing means: one processor sequentially processes instructions on a computer with a single instruction stream and a single data stream; one data element is processed on each machine.

The MIMD architecture provides great flexibility: with the appropriate hardware and software, MIMD can work for a single user, processing high-performance data from a single application, as a multi-tasking machine that performs multiple tasks in parallel, and as some combination of these features. The MIMD architecture is able to take full advantage of modern microprocessor technologies based on strict cost and performance requirements. Virtually all modern multiprocessor systems are built on the same microprocessors that are available on personal computers, workstations and small single-processor servers [13]. When working with a SIMD on a computer with multiple streams and multiple streams at the same time, a single processor processes a stream of instructions, each of which can perform parallel calculations on the data set. This processing is widely used in computer modeling, but it is rarely used for business processes and is not recommended. In order to take advantage of all the features of the architecture, programs must be written and defined individually for each task. The third existing multiprocessor architecture is MISD. It is recommended that a processor with multiple instruction streams and a single data stream often maintain the advantage, as multiprogramming modules perform the same task on the same data, reducing the chances of getting a result if one of the modules fails. The MISD architecture allows you to compare the results of calculations to determine the defect.

Working with large amounts of data within databases is a topical issue. As the amount of data increases, our hard drives may experience some problems and, most importantly, have time to access the required data. You can use cache, but that doesn't help in the end. You can divide the database by placing information of each class in its own database. As the amount of data increases, the speed of the system decreases significantly. One way to reduce data access time is to place the database in random access memory. This method allows you to increase the speed up to 100 times. Memory Databases - IMDB is a database that uses a computer's main memory to store data. Random access memory is the main data warehouse in such systems. As the cost of memory decreases day by day, its use as storage can be effective in increasing the speed of data processing. There are new types of databases for working with large data, such as analytical databases. Today, this concept is used in almost all databases [14]. However, the developers of Terradata were the first to perform an embedded analysis in the database [15]. In addition, one type of database is column data storage. In recent years, a number of column databases have emerged, including MonetDB and C-Store. The developers of these systems claim that this approach increases the results of analytical workloads, which are in high demand for reading data similar to certain loads, especially applications in the data warehouse [16]. Unlike a set of Data Mining algorithms and a database, analytics platforms are initially focused on data analysis and are designed to create ready-made analytical solutions.

Analytical platform is an information-analytical system, as well as a specialized software solution that includes all the tools for obtaining templates from "raw" data, the process of obtaining some templates from the entire data array: means of consolidating information in one source, data acquisition, conversion, storage, information extraction algorithms, visualization, switching of simple and complex methods and models [17]. There is no single method or algorithm for properly processing large amounts of data. Accordingly, each task has its own execution algorithm, data analysis algorithm. There are many algorithms for processing large data, but new algorithms need to be developed to achieve certain goals.

4. DOCUMENT PROCESSING FOR THE ENTERPRISE DIGITALIZATION SYSTEM BASED ON MONGODB

Documents related to one documentation system have general rules for working with them - preparation, agreement, approval, registration, execution. The general rules for working with documents included in one documentation system are determined by typical management functions. Combining documents into documentation systems reduces the number of local regulations governing the work with documents.

When collecting and sorting documents for the administrative module, a number of differences and rules were taken into account. Collected and put in order various types of documents in the administrative and personnel modules. Particular attention was paid to literacy: there were no spelling, punctuation and other errors. When writing all types of letters, the standard rules of business correspondence were used, the official business style of speech was applied, which did not include colloquial expressions in the text.

The system under development uses database technology to organize data catalogs. To implement the data catalog management complex, the document-oriented MongoDB database was chosen.

A study of methods of interaction with documents stored in the MongoDB was carried out. Typical queries to the MongoDB database are considered.

To add an object to the database in MongoDB, use the insert () function. This function can only be called on a collection, for example: db.collection.insert (). Where, db is the name of the database containing the required collection; collection is the name of the collection into which you want to insert the document. If documents are added to a collection that does not exist, then such a collection is created with the documents inserted into it. When adding a new document to a collection, MongoDB automatically adds an _id field to it with a random Objected (...) value, unless this field is explicitly specified when adding it. Objected () is a function that creates a 12 byte id object. As a rule, a specially generated string of 12 characters is fed into it. All collections are located in one database called rmsDB (Researches Management Systems Database). The following are methods for adding objects to each collection using the standard 25 mongo shell, with a description of the correct sequence for adding and the reasons for this sequence of adding objects [18].

To remove documents in MongoDB, there is the db.collection.remove () function. But, in the presence of a large number of documents and many links between them, such an operation can be quite complicated and, in case of an error, can lead to the formation of false links to already deleted documents. In addition, there is always the possibility that the document was deleted by mistake, and no one is ever safe from this. To change, add and remove fields and arrays in documents, in MongoDB, the db.collection.update () function was used. This function has many operators to provide the desired operations.

5. USING MONGODB FOR BIG DATA STORAGE

MongoDB is the leading enterprise NoSQL technology database written in the C ++ programming language. Integrates with any software application written in Python. Let's take a look at the difference between SQL and NoSQL. SQL databases use Structured Query Language (SQL) to identify and manipulate data. When using SQL, we need a relational database management system (DBMS) such as SQL Server, MySQL Server, or MS Access. In a DBMS, data is stored in database objects called tables [19].

A table is a collection of related data records made up of columns and rows.

NoSQL database has a dynamic schema for unstructured data. In NoSQL, data is stored in several ways: it can be based on a column, document, graph, or a key and value store. NoSQL database has the following advantages:

- documents can be created without prior definition of the structure;
- each document can have its own unique structure;
- Differences in database syntax from other databases
 may be;
- large-scale structured, semi-structured and
 can store unstructured data;
- object-oriented programming is simple and compact to use;
- horizontal scaling.

To create a database in MongoDB, we use the MongoClient instance, and after specifying the name of the database, MongoDB creates and works with the database:

db = client [datacampdb]

It should be noted that databases and collections are created slowly in MongoDB. This means that collections and databases are created only when the first document is entered. MongoDB data is presented and stored using JSON-style documents. At PyMongo, we use dictionaries to present documents. Here is an example of a PyMongo document:

article = {“authors”: “Darkenbayev D., Balakayeva G”,
“About”: “MongoDB and Python”; “Tags”: [“mongodb”, “python”, “pymongo”]}

When managing a MongoDB database, you can name the collection, add additional information, and delete it. We can fully manage and use our large amount of unstructured data in the non-relational database MongoDB, which also covers security issues.

CONCLUSION

This article examines the topical issues of large-scale data processing of enterprises and data storage and processing technologies of giant companies in the field of information technology. The works of leading foreign and domestic scientists were reviewed, the topicality of the issue was revealed. Along with the growth of the world's population, the volume of data is growing rapidly. Obviously, because of the different formats of data, it is difficult to process them. NoSQL technology for storing and processing unstructured data has shown positive results in solving non-relational data processing. According to Chris Phillips, a professor at the University of Newcastle (UK), there is no set of rules for processing large data, so there is a lot of debate about how to process it. He expressed his opinion that good results can be achieved by using machine learning algorithms to solve the problem. In future articles we plan to publish an article on the integration of research and technology in DataMining and NoSQL technologies.

References:

*1 Big Data. Explanatory dictionary on Academician. 2014. URL: https: //dic.academic.ru/dic.nsf /ruwiki/1422719. (*Date of the application*: 18.04.2019).*

2 Rubanov V.A. Between management standards and the information element // Technological Forecast. – 2010. – No. 3.

3 Tom White Hadoop: The Definitive Guide, 3rd Edition. O'Reilly Media, 2012, – 688 p.

*4 Artemov C. Big Data: New Opportunities for Growing Business // Jet Infosystems. URL: http://www.pcweek.ru. (*Date of the application*: 20.08.2018).*

5 Doug L. 3D Data Management: Controlling Data Volume, Velocity and Variety // Meta Delta. – 2001. – P. 949-951.

*6 Pettey C., Goasduff L. Gartner Says Solving Big Data Challenge Involves More Than Just Managing Volumes of Data. URL: http://www.gartner.com. (*Date of the application*: 11/21/2019).*

7 Doug L. 3D Data Management: Controlling Data Volume, Velocity and Variety // Meta Delta. 2001. P. 949-951.

*8 Pettey C., Goasduff L. Gartner Says Solving Big Data Challenge Involves More Than Just Managing Volumes of Data. URL: http://www.gartner.com. (*Date of the application*: 06/27/2020).*

*9 HDFS Architecture Guide. URL: https: //hadoop.apache.org/docs/r1.2.1 /hdfs_design.html. (*Date of the application*: 15.04.2019).*

*10 Apache Cassandra. URL: http://cassandra.apache.org. (*Date of the application*: 15.04.2019).*

*11 Semenov Yu.A. Large amounts of data (big data). URL: http://book.itep.ru. (*Date of the application*: 21.04.2020).*

12 Jacobs A. The pathologies of big data // Communications of the ACM. 2009. Vol. 52. Iss. 8. R. 36-44.

13 Tsvetkov V.Ya., Lobanov A.A. Big Data as Information Barrier // European researcher. Series A. 2014. Vol. 78. – Iss. 7-1. – R. 1237-1242.

*14 Lockwood G.K. Conceptual Overview of Map / Reduce and Hadoop. URL: http://www.glennklockwood.com. (*Date of the application*: 06/28/2020).*

15 Anshina M. Methods of working with big data and their effectiveness // Big Data Conference: Opportunity or Necessity, March 26. – Moscow, 2013 . – 312 p.

16 Abadi D.J., Madden S., Hachem N. ColumnStores vs. RowStores: How Different Are They Really? // Proceedings of the ACM SIGMOD International Conference on Management of Data. – Vancouver. – 2008 . – 475 p.

17 Fernández A, Río S, López V, Bawakid A, del Jesus M, Benítez J, Herrera F. Big data with cloud computing: an insight on the computing environment, MapReduce and programming framework. WIREs Data Min KnowlDiscov 4 (5). – P. 380-409.

18 Balakayeva G.T., Phillips C., Darkenbayev D.K., Turdaliyev M. Using NoSQL for processing unstructured Big Data. News of the National Academy of Sciences of the Republic of Kazakhstan. ISSN 2224-5278. – Volume 6, Number 438, 2019 . – P.12-21. URL: https://doi.org/10.32014/2019.2518-170X.151

19 Balakayeva G.T., Darkenbayev D.K., Chris Phillips. Investigation of technologies of processing of Big Data. International Journal of Mathematics and Physics 8, – No. 2, (13) 2017. – P.13-18. URL: https://doi.org/10.26577/ijmph.v8i2