Lubko D.V.
Candidate of Technical Sciences,
Associate professor of the department of computer science
Dmytro Motornyi Tavria State Agrotechnological University

OVERVIEW OF DISTRIBUTED PROCESSING TECHNOLOGIES OF BIG DATA VOLUMES AND ENSURING THEIR SECURITY

Abstract. *The application of Big Data in global systems plays an important role in many areas of life, including business, science, medicine, the public sector and many others. Big Data is large, diverse and complex data that originates from various sources and is processed using specialized tools and technologies. Processing big data is becoming an increasingly important task for many organizations. Building an effective Big Data application requires careful planning and consideration of various aspects. This article discusses the key steps in building a Big Data application and the main aspects to consider.*

Key words: *distributed data processing, data security, technology, big data, data protection, Big Data.*

Лубко Д.В. Огляд технологій розподіленої обробки великих обсягів даних та забезпечення їх безпеки. *Дослідження має на меті зробити огляд технологій розподіленої обробки великих обсягів даних та забезпечення їх безпеки та дати практичні рекомендації з даної проблемної області. Однією з головних тенденцій є зростання обсягів даних, що генеруються щодня, що вимагає потужних інструментів для їх обробки та аналізу. Технології, такі як Apache Hadoop та Apache Spark, дозволяють здійснювати розподілену обробку даних та обчислення великих обсягів даних. Машинне навчання та аналітика даних стали необхідними для вибуття корисної інформації з великих обсягів даних. Бібліотеки, такі як Scikit-learn та TensorFlow, дозволяють розробникам та дослідникам створювати моделі машинного навчання та проводити аналітику даних. Забезпечення безпеки та конфіденційності даних важливо для обробки великих обсягів даних, особливо в чутливих сферах, таких як охорона здоров'я та фінанси. Важливо дотримуватися відповідних правових норм і регулювань, таких як регуляція GDPR. Висновуючи, обробка великих обсягів даних є*

важливою для сучасного суспільства та бізнесу. Зростання обсягів даних вимагає від розробників та дослідників використовувати нові технології та методи для витягування користі з цих даних. Розвиток цього напрямку залишається актуальним та перспективним завданням для майбутніх досліджень та інновацій.

Ключові слова: *розподілена обробка даних, безпека даних, технології, великі обсяги даних, захист даних, Big Data.*

Introduction. As shown by the analysis of recent research and publications in this problem area, namely the issue of working with and processing large amounts of data, their existing technologies and ensuring their security and protection, it was determined that many scientists and scholars have been actively working and are working for this. Namely, these are such specialists and scholars as: Chakraborty S. [1], Sharov S. [1], Kenneth Cukier [2], Viktor Mayer-Schönberger [2], Matthias Niehoff [3], Marin Ivan, Shukla Ankit [4], Steven Pinker [5], Steven Pinker [5], Rajkumar Buyya [6], Edd Wilder-James [7], Bernard Marr [9], Thomas H. Davenport [10]. But, despite the large number of works and research in this area, aspects of big data mining and issues of ensuring their security and data protection have not yet been sufficiently covered. That is why the topic of the work is very relevant.

The purpose of the study – to review technologies for distributed processing of large amounts of data and ensuring their security, and to provide practical recommendations in this problem area.

Presentation of the main material. The first step in creating various processing applicationslarge amounts of data- Big Data, is the development of its architecture. To effectively process large amounts of data, it is important to develop a distributed system. This means that the application must be divided into independent components that can run in parallel.

To process large amounts of data, you need to consider how that data will be collected and stored. Using data collection systems such as Apache Kafka or Apache Flume, you can efficiently move data from various sources to a central repository.

This data store needs to be well organized. Using distributed databases such as Apache HBase or Apache Cassandra will help ensure fast data access and the ability to scale depending on the volume. An example of the Hbase Performance Monitoring dialog box is in Figure 1.

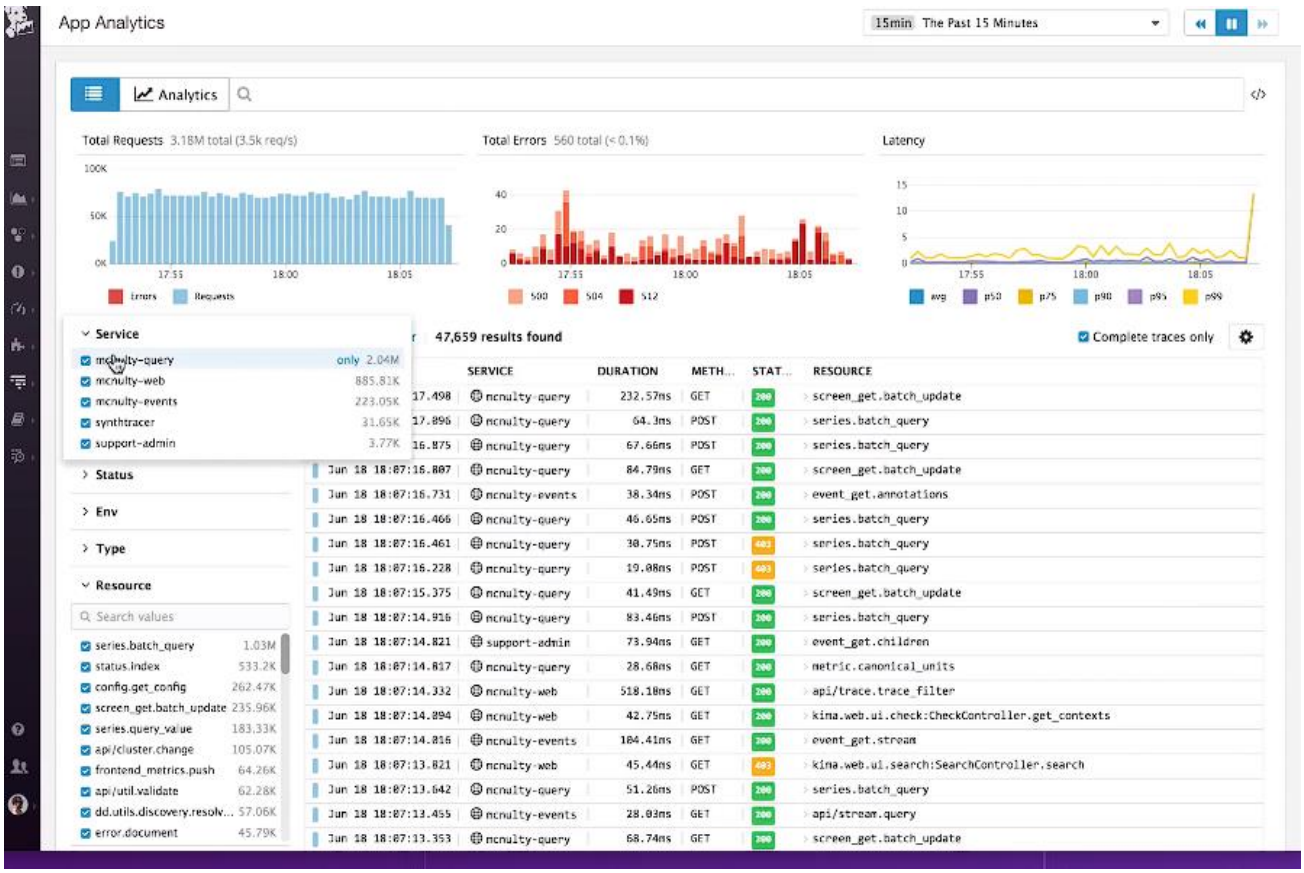

Figure 1. Example of "Hbase Performance Monitoring"

Various tools and technologies are used to process and analyze large amounts of data. Apache Hadoop and Apache Spark are popular frameworks for data processing and analysis. They provide tools for distributed data processing and the ability to perform complex operations on large amounts of information.

A Big Data application often needs to interact with other systems and data sources [2; 4]. For this, monitoring and logging systems are used to track and analyze the application’s interaction with other systems. Security and authentication play an important role in ensuring the security of interaction with other systems. Authentication and encryption mechanisms are used. APIs and interfaces make it easier for other developers or teams to integrate.

Ensuring data security and confidentiality in a Big Data application is an extremely important task, as large amounts of data can be valuable and vulnerable to various threats.

Below we will look at steps and strategies to ensure data security:

1. Encryption to protect data at rest and in transit. Use TLS/SSL encryption to ensure data is secure when it is transmitted over the network. Encrypt data in databases and at rest.

2. Establish strict data access controls, restricting access rights to specific roles and users. Use authentication and authorization to verify user identity and authority.

3. Maintain activity logs to track events and unauthorized access attempts. Use monitoring systems and log analysis to detect anomalies.

4. Store passwords in hash form, not in plain text. Use password hashing technologies to reduce the risk of using tables to attack password hash functions.

5. Regularly applying updates and patches to software and dependencies to fix known vulnerabilities.

6. Protect physical access to servers and data infrastructure. Apply security measures to premises and server centers.

7. Measures are used to prevent attacks such as SQL injections, code injection, and other vulnerabilities. To this end, code audits are conducted and data filtering mechanisms are implemented.

8. Ensuring that staff understand the importance of safety and are able to identify potential threats and act accordingly.

9. Regularly back up your data and verify its recovery capabilities. This is important to protect against data loss due to accidental failures or attacks.

10. The organization must be subject to certain regulatory requirements, such as GDPR, HIPAA, or others.

Data visualization integration is extremely important for a Big Data application because it helps users better understand large amounts of information and find visual patterns. Interactive graphs and charts allow users to create interactive visualizations to display different parameters and dependencies in the data. The ability to create personalized reports allows users to choose the data they want to display and analyze. Some data is better displayed in a three-dimensional form. The ability to create 3D graphs and models is required to analyze spatial data and volumetric data. Animation and time visualization will allow users to create animations and visualize dynamic changes over time. This is especially useful for tracking and analyzing data dynamics.

If the data includes networks or relationships between objects, graph and network visualization capabilities are needed. Users can display the structure and interactions between nodes. Filtering and highlighting capabilities allow users to filter data and highlight important parts of the visualization for better understanding and analysis. Scaling will provide the ability to scale

visualizations to work with large amounts of data without losing performance. Support for different data types allows you to work with different types of data such as numbers, text, images, audio, graphs, etc. Support for special effects and visualizations will allow you to create impressive and expressive graphical effects to best display the data.

After developing and launching an app, it is important to engage users and listen to their suggestions and feedback. This will help you understand their needs and wants and improve the app with subsequent updates and improvements. It is important to establish feedback collection mechanisms such as surveys, comments, feedback loops, and support.

Increasing application requirements and data volume may require changes and extensions. The application must remain flexible and open to change. It is important to verify that the application can handle even the largest data volumes and complies with all legal and regulatory requirements. Processing large amounts of data can be subject to various legal regulations, especially in areas where data privacy and security are critical, such as healthcare or finance. It is important to examine all relevant regulations and standards and ensure that the application complies with these requirements. This may include the use of encryption mechanisms, access controls and other means to ensure data security.

Machine learning can be a powerful tool for analyzing large amounts of data. The application of machine learning algorithms allows for automated analysis, pattern recognition, and prediction based on data. It is important to explore different machine learning methods and choose the ones that are best suited to the context. Cloud platforms such as Amazon Web Services, Microsoft Azure, Google Cloud, and others can be great helpers in processing Big Data. They provide secure storage and scaling of data and provide access to powerful tools for data processing and analysis.

By considering all these aspects, you can successfully build a big data application that meets the highest security standards, analyzes data effectively, and scales for future growth. Building such an application is a challenging task, but with the right planning, tools, and knowledge, you can succeed in big data processing. In today's world, big data processing has become an essential component for many fields, from business and science to healthcare and public policy.

Conclusions. One of the main trends is the growth of data volumes generated every day, which requires powerful tools to process and analyze

them. Technologies such as Apache Hadoop and Apache Spark enable distributed data processing and big data computing. Machine learning and data analytics have become essential to extract useful information from large volumes of data. Libraries such as Scikit-learn and TensorFlow allow developers and researchers to create machine learning models and perform data analytics. Ensuring data security and privacy is important for processing large volumes of data, especially in sensitive areas such as healthcare and finance. It is important to comply with relevant legal norms and regulations such as the GDPR regulation. In conclusion, processing large volumes of data is essential for modern society and business. The growth of data volumes requires developers and researchers to use new technologies and methods to extract benefits from this data.

The development of this direction remains a relevant and promising task for future research and innovation.

References

1. Chakraborty S., Sharov S. Indexing-based approach in document-centric big data. *In Secure Big-data Analytics for Emerging Healthcare in 5G and Beyond: Concepts, paradigms, and solutions*. 2024. Chapter 11. Pp. 229–247.
2. Viktor Mayer-Schönberger, Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Publishing: John Murray General Publishing Division, 2013.
3. Matthias Niehoff, Dr. Daniel Pape. *Big Data, Fast Data*. Publishing: Bookwire, 2000.
4. Marin Ivan, Shukla Ankit. *Big Data Analysis with Python*. Publishing: Packt Publishing, 2019.
5. Seth Stephens-Davidowitz, Steven Pinker. Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are. Published: Dey Street Books, 2017.
6. Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi. *Big Data: Principles and Paradigms*. Publisher: Morgan Kaufmann, 2016.
7. Edd Wilder-James. *Planning for Big Data*. Published: O'Reilly Media, 2012.
8. Hadoop Big Data Interview Questions You'll Most Likely Be Asked (Job Interview Questions Series). Publishing: Vibrant Publishers, 2017.
9. Bernard Marr. *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Published: Wiley, 2016.
10. Thomas H. Davenport. *Big data@work: dispelling the myths, uncovering the opportunities*. Published: Harvard Business Review Press, 2014.