Chemical Technology, Control and Management

Volume 2020 | Issue 4 Article 13

8-29-2020

A STUDY OF SECURITY PROBLEMS IN BIG DATA AND THEIR SOLUTIONS

Nozima Akhmedova
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi Address: 108, Amir Temur st., 100200, Tashkent city, Republic of Uzbekistan E-mail: ANFscience@gmail.com, Phone:+998-93-508-31-23, anfscience@gmail.com

Recommended Citation
Akhmedova, Nozima (2020) "A STUDY OF SECURITY PROBLEMS IN BIG DATA AND THEIR SOLUTIONS," *Chemical Technology, Control and Management*: Vol. 2020 : Iss. 4 , Article 13.
DOI: https://doi.org/10.34920/2020.4.81-85
Available at: https://uzjournals.edu.uz/ijctcm/vol2020/iss4/13

ISSN 1815-4840, E-ISSN 2181-1105
Himičeskaâ tehnologiâ. Kontrolʹ i upravlenie
CHEMICAL TECHNOLOGY. CONTROL AND MANAGEMENT
2020, №4 (94) pp.81-85. https://doi.org/10.34920/2020.4.81-85
International scientific and technical journal
journal homepage: https://uzjournals.edu.uz/ijctcm/
Since 2005

UDC 004.056

A STUDY OF SECURITY PROBLEMS IN BIG DATA AND THEIR SOLUTIONS

Akhmedova Nozima

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Address: 108, Amir Temur st., 100200, Tashkent city, Republic of Uzbekistan
E-mail: ANFscience@gmail.com, Phone:+998-93-508-31-23;

Abstract: *Statistical data on information security that concerns Big Data and is the most important for enterprises are provided. Based on this data, we studied problems such as the lack of big data practices and protection, the lack of techniques for protecting big data, the lack of standards for protecting big data, the lack of regulation of big data and ecosystems, security problems in Big Data, and proposed several proposals to improve the security of systems that use this technology.*

Keywords: *Big Data, Hadoop, cyber security, information technologies, information security, security systems, methodology of security, identification, classification, data passport, access.*

Аннотация: *Корхона учун муҳим ҳисобланган ва Big Data технологиясига тегишли бўлган ахборот хавфсизлиги бўйича стастистик маълумотлар келтирилган. Бу маълумотлар асосида Big Data билан ишлаш амалиёти ва ҳимоясини йўқлиги, Big Data ҳимояси учун услубларнинг йўқлиги, Big Data ҳимояси учун стандартларни йўқлиги, Big Data ва экотизимлар назоратининг йўқлиги, Big Data технологиясида хавфсизликни таъминлаш муаммолари ўрганилган ва бу технологиялардан фойдаланадиган хавфсизлик тизимларини яхшилаш учун бир нечта таклифлар ва ечимлар таклиф этилган.*

Таянч сўзлар: *Big Data, Hadoop, кибер хавфсизлик, ахборот технологиялари, ахборот хавфсизлиги, хавфсизлик тизимлари, хавфсизликни таъминлаш услубияти, идентификация, таснифлаш, маълумотлар паспорти, рухсат.*

Аннотация: *Приведены статистические данные по информационной безопасности в Big Data и являющемся наиболее важным для предприятия. Были изучены такие проблемы как отсутствие практики по работе с Big Data и защиты, отсутствие методик и стандартов для защиты Big Data, отсутствие регулирования Big Data и экосистем, проблемы обеспечения безопасности в технологии Big Data, предложено несколько путей улучшения безопасности этих систем.*

Ключевые слова: *Big Data, Hadoop, кибербезопасность, информационные технологии, информационная безопасность, системы безопасности, методология обеспечения безопасности, идентификация, классификация, паспорт данных, доступ.*

Introduction

New technologies have created huge amounts of data and the ability to process them. The emergence Big Data has become the embodiment of a long-standing business dream — to learn everything about customers, competitors and market trends. In 2030, according to Forrester researchers, almost all companies that use data Analytics for decision-making will also implement Big Data processing. [1].

Among the main advantages of Big Data for business, according to information obtained because of a survey by the research company The Economist Intelligence Unit and the consulting company Accenture, there are:

- search for new sources of income (56 %);
- improving the customer experience (51%);
- new products and services (50%);

- attracting new customers and maintaining the loyalty of old ones (47%).

According to experts of the service, cybersecurity, the security of processing, storage and transmission is the most important aspect of studying Big Data: information has a price, and its leaks can cause significant damage to business. So, according to the European Commission, published in the 2016 report The EU Data Protection Reform and Big Data: Factsheet, the personal data of European citizens (450 million people) will be estimated at one trillion euros by 2020. As the research shows, customers ' trust in the company directly depends on the reliability of their data protection. With a leak, businesses can lose customers ' trust and money in addition, get into trouble with regulators. These include fines, suspensions, and legal proceedings.[2].

Over the past three years, the top companies affected by information leaks included Yahoo (data leaks — more than 500 million customers), the Home Depot (50million plastic cardholders), Target (70 million credit and debit cardholders).

There are several distributions of Hadoop: Hortoworks, Cloudera, MapR, IBM BigInsights, etc. Hadoop is very popular, including with such IT giants as Facebook, Alibaba, Amazon, Linkedin, and eBay. The reason lies primarily in the ability of Hadoop to accept and analyze huge amounts of data of different structures from multiple sources without preparation, as well as in its performance and availability. In addition, Hadoop includes the HDFS file system, which allows you to significantly reduce the cost of terabyte of data storage. According to the magazine, Readwrite, the cost of storing a terabyte in Hadoop is 2.5 times lower than in Oracle databases. Calculations by experts show that the total cost of owning a terabyte of information in Hadoop is ten times lower than that of commercial database manufacturers [3].

Problems in security of Big Data and solutions

In the process cyber security of organizing a big data protection system, specialists of the cybersecurity service identify a number of problems. They are related, first, to the features of Hadoop, which, in fact, is not an ordinary classical database, but a file system organized in a so-called "data lake", where data from various sources is stored. However, the information in such a lake is physically distributed across the server cluster and is accessible via various interfaces (APIs) or application layers, each of which must be protected. Secondly, with the lack of regulation of big data in General. In addition, third, with individual processes for processing and providing access to big data.

Problem # 1. Lack of practice on working with Big Data and protecting

Big Data is a new paradigm for data storage and processing. IT services may lack the competence to support and service new technologies, as there are not enough ready-made specialists on the market. There are no courses or textbooks for studying Big Data technology. To get the necessary knowledge, you need one or two years of daily work with the technology, which is incompatible with the current activities of specialists. It is also not always easy for his services to protect new technologies. They do not always understand what exactly is going on inside the big data cluster, what the threats and vulnerabilities of new technologies are. The IP protection methodologies of the classic three-link architecture are not applicable to new technologies. There is a need to create and train a new class of IT and IS specialists to work with Big Data, which in itself is quite expensive and resource-intensive process.

Experts on cyber security recommend:

* allocate separate divisions in IT and Information Security (IS) services that will deal with Big Data technologies on a permanent basis;

* involve specialists at all levels of IT and is services from the first day of creation of Big Data systems: developers, administrators, IT specialists, testers, etc., so that the experience gradually accumulates with the growth of the system;

* send staff to appropriate courses at least once every two years.

Problem # 2. The lack of methodologies for the protection of Big Data

There is no single industry-accepted methodology for ensuring big data security that could help develop and implement a Big Data security management system yet. Various organizations publish their own methodologies and recommendations, but none of them has yet reached the ISO level. Cyber Security experts recommend paying attention to the following:

- IBM: Top tips for Big Data Security;
- Oracle: Enterprise Security for Big Data Environments;
- Forrester: Big Data Security Strategies For Hadoop Enterprise Data Lakes;
- ENISA: Big Data Security: Good Practices and Recommendations on the Security of Big Data Systems;
- Cloud Security Alliance: Big Data Security and Privacy Handbook;
- Securosis: Securing Hadoop: Security Recommendations for Hadoop Environment
- Cloudera: Cloudera Security.

All of these methodologies have their drawbacks. There are no generally accepted criteria for selecting a methodology, because each organization has its own individual data storage and processing processes, and the methodology describes, among other things, the security of data processing processes, and so on.

Problem #3. The lack of standards for the protection of Big Data

In addition to security methodologies, there are no standards that describe a complete list of rules and regulations on Big Data security, which is considered normal practice in the cybersecurity industry. Currently, several working groups are working on creating standards, for example, WG9 under the auspices of the ISO JTC 1 Committee and the Big Data Working group from the Cloud Security Alliance community. In the United States, in addition to participating in international communities, there is a working group on Big Data security — NIST SP1500-4: Big Data Security and Privacy [4].

However, none of them has yet published a single standard.

In their research, the working groups concluded that security and privacy measures should be embedded in the design of Big Data systems, and not appear as they develop. There is no description of the measures themselves yet [5].

Problem #4: the Big Data ecosystem

The reason for the lack of standards lies in the huge size of the big data ecosystem and the speed of development of this direction. In other words, the Big Data eco system is developing too rapidly and growing too fast, which makes it more difficult to standardize it.

Problem #5. Lack of regulation of Big Data

There are laws on the protection of personal data, on Bank secrecy, on state secrets, on commercial secrets, and so on, but there is no government regulation in the field of Big Data protection.

Due to existing problems, companies are forced to independently develop approaches to ensuring the security of Big Data.

Having studied all possible solutions available on the market, the cyber security Bank's cybersecurity specialists developed their own methods and approaches to solving this problem [6]. Therefore, it makes sense to divide big data security into two phases (Fig. 1):

* identification and classification of information (what to protect).

The tasks that need to be solved during this phase are to identify, classify security objects, and assign data privacy labels. The company should develop criteria for data confidentiality independently. For example, a password or plastic card data can be considered confidential data and deleted from the cluster. The more sensitive data is placed in the data lake, the more difficult it will be to differentiate access to it;

* security (how to protect).

The task that is solved in this phase is to apply security measures to the objects of protection. For example, administrative, physical, and technical security measures. Security requirements can be found in various standard collections, such as ISO 27001.

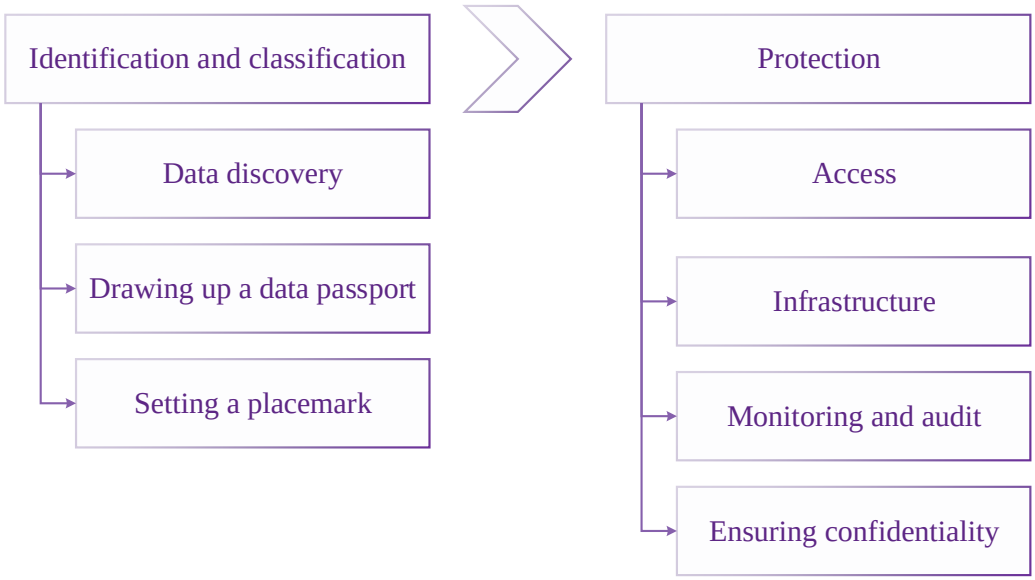

Figure 1. ***Methodology of Security of Big Data.***

Drawing up a "data passport". After identifying the data, they should be classified, that is, a certain "data passport" should be compiled, which will also include a privacy label (Fig. 2).

Data sheet	
Name of the asset	Name of a database, dataset, file, table, or any trusted source
Confidential	Privacy label
Asset type	Oracle, SAS, DB2, MSSQL, XML, CSV, JSON, Hive, XLS, or flat-file
The data type and its length, dimension	i.e. number, char, date, char(20), and so on.
Data owner	full name or division in the company that owns the data in the source
Usage context description	Clients, projects, and role descriptions that use data
Content of the secret	Personal data, PCI DSS, etc.
Jurisdiction data	RU, UZ, EN, etc.
Update frequency	Frequency of data compilation or data Analytics

Figure 2. ***Data passport.***

All data must have a privacy label. The more sensitive data gets into the data lake, the more security measures you need to apply to ensure access to it. Highly sensitive data, such as passwords, should be deleted from the data pool or kept out of it. It is precisely in order to restore order in the data and discipline the staff, the "data passport" is used.

Reaction to privacy tags:

· high privacy.

You should prevent high-privacy data from entering the data lake and delete those that have already entered it;

· average privacy.

You should control access to data. Each company determines a specific set of measures based on its objectives;

· no privacy policy.

You do not have to control access.

Access. One of the main security rules is to restrict access to a level that is sufficient to perform your work tasks. Access control means that a specific user gets access to specific data at a specific time. This requires implementing authentication mechanisms and conducting periodic checks of employee privileges. For example, Hadoop supports a special Kerberos Protocol that controls access to Hadoop resources. However, Kerberos does not work by default, and it will take time and money to implement it. You can also connect other software products that implement role-based access functionality — such as Sentry, Apache Accumulo, and others. If necessary, the "accuracy" or granularity of data access can be limited to the level of a column or even a cell.

When working in a data lake, you should create a "secure workplace" for the data analyst in order to:

1) exclude the possibility of copying data from the data lake. In this case it is possible to apply virtual automated systems with the appropriate settings;

2) logging actions with data that the analyst performed. To do this, you can integrate the analyst's virtual arm with event logging tools.

Monitoring and auditing. An audit implies that any activity that occurs in Hadoop is logged. To ensure data security, you need to log certain events: traffic, user activity, and so on, so that you can restore a picture of the incident based on these events. You cannot protect yourself from an attack if you can't see it, so you should monitor it centrally, for example, in a SIEM system, so that you can get visibility of how apps are working and the traffic pattern.

Eliminating the value of data. If the data is devalued, then its attractiveness will disappear — it will become uninteresting to steal it. To "devalue" data, various methods of abstraction are used, that is, encryption, tokenization, masking data, and even deleting them. Methods and recommendations for eliminating the value of data while preserving its useful properties are described in various methodologies, for example, in ISO 29100 Privacy Framework [7].

Conclusion

The problem of ensuring the security of big data storage and processing lies precisely in the huge amounts of unstructured, disparate data.

Companies need to develop a process-based approach to data analysis and processing, as well as automate processes related to ensuring the security of big data within established practices. Automation can also include elements of machine learning (artificial intelligence, AI). Using AI, it is possible to extract signs of "privacy" from data added to the cluster, detect patterns that are not typical for normal data processing, create user profiles, and record deviations in users 'work from their normal behavior profile, i.e., identify users' motives when working with data.

References

1. SAS Analytics. URL: http://www.sas.com/ru_ru/software/analytics.html (date accessed: 05.01.2017).
2. Free software for statistics, SAS University Edition SAS. URL: http:// www.sas.com/ru_ru/software/university-edition.html (accessed: 05.01.2017).
3. I.BudzkoV, A.V.Schmid, N.A.Ivanov, S.P.Sadovsky, “Security and confidentiality of the centralized system of disease prevention” URL: https://yadi.sk/d/F-CaJt3J3DgeS4 (date accessed: 19.03.2017).
4. “Bol'shie danny'e” (Big Data Data: changing the future of humanity. URL: http://www.kitaichina.com/se/txt/2013-03/28/content_530849.htm (date accessed: 05.01.2017).
5. V.Mayer-Schoenberger, KukierTo, “Big data. A revolution that will change the way we live, work and think” Moscow, 2014.
6. A.G.Ostapenko, E.V.Ermilov, A.O.Kalashnikov “Risks of damage, chances of utility and viability of components of automated systems under the influence of information threats on them” *Information and security*, vol. 2, no. 2, 2013.
7. “Tehnologii analiza danny'h kompanii”, BaseGroup Labsdata analysis technologies SAS Analytics. URL: https://basegroup.ru/ (accessed: 05.01.2017).