Views on Big Data technology information security

A.Kubigenova, Al.Aktayeva, A.Sharipbay, A.Beissekov, G.Muradilova, A.Seraliyeva

Abstract. The article discusses the security issues of the Big Data technology directly related to the architecture of the system, as well as the implementation of its following main components: data collection, data storage, normalization, analysis, and visualization of data.

BigData technology anticipates certain security hazards bearing adverse effects on business processes of safety-critical systems. In case of massive computer attacks, (MCA) BigData (BD) has unpredictable behavior dynamic, while the incidents are merely recognized rare.

The significance and novelty of the study is substantiated by the positive outcome of assessment of the Monte-Carlo method for evaluation the survivability of MCA-exposed recoverable BDs possessing time redundancy (recoverability). The statement of the problem, the modeling algorithm, and examples of solutions are given, based on which the Monte-Carlo method is recommendable for solving the problem.

The results obtained might be promising in a way of improving the development of solutions to ensure the stability of the BD function under the conditions of object-orientated aggressive actions.

Keywords: BigData, Monte-Carlo method, massive computer attack, survivability, security issues, information sensitivity.

I.INTRODUCTION

These days, BigData is crucial to the success of any business in a data-driven world. However, BigData (BD) is interrelated to various security threats that can negatively impact organizations. Advanced solutions in innovative BigData technology enable for data flow enhancing efficiency and shifting into real-time operation and better decision making. A business can reap some benefits from the use of BigData technology, like:

• BigData is a good way of improvement the products quality level and creation of customized marketing by getting a comprehensive overview of the behavior and motivation of the customers.

• It may be efficiently used in tracking fraudulent activity in real time, identifying unusual patterns and behavior by means of predictive analytics.

• Larger volumes of data provide more chances to explore the untapped area, as well as to conduct deeper and more detailed analysis for the benefit of all concerned parties.

• BigData is used to train machine-learning models to identify patterns and make informed decisions with little or no human intervention.

• Data analysts use various types of data primarily to make better business decisions by understanding the behavior of participants. [3], [4]

Manuscript received December, 2022.*The study is part of a doctoral thesis titled "Models and methods of post-quantum cryptographic security of big data."*

1. A.Kubigenova, doctoral student of KazATU named after S. Seifullin, Astana, Kazakhstan, phone: 87024584790; e-mail: akku_kubigenova@mail.ru

2. A.Aktayeva , Associate Professor of the Department of Information Systems and Informatics, Kokshetau University named after A. Myrzakhmetov, Kokshetau, Kazakhstan. e-mail: aktaewa@list.ru

3. A.Sharipbay, Doctor of Technical Sciences, Professor of the Department of Artificial Intelligence Technology of the L.N. Gumilyov, Astana, Kazakhstan. e-mail: sharalt@mail.ru

4. A.Beissekov, Lecturer, Department of Information Systems, Sh.Ualikhanov Kokshetau University, Kokshetau, Kazakhstan. e-mail: b.akilbek@mail.ru

5. .G.Muradilova, Lecturer of the Department of Information Systems, Sh.Ualikhanov Kokshetau University, Kokshetau, Kazakhstan. e-mail: mgs_kz@mail.ru

6. A.Seraliyeva, lecturer of the Department of Information Systems and Informatics of Kokshetau University named after A.Myrzakhmetov, Kokshetau, Kazakhstan. e-mail: seralieva_a_a@mail.ru

II.MATERIALS AND METHODS

Data mining, machine learning, and predictive analytic's are just a few of the recently developed techniques being used to gain new insights into untapped areas of BigData sources. [1]

BigData represent large, diversified datasets coming from multiple channels of unlimited choice, like social media platforms, websites, e-registries, sensors, grocery shopping, call logs. BigData has three unique parameters:

1) Volume, which means extra-large amount of data,

2) Velocity, which stands for very high data transfer rate,

3) Variety refers to weak data structure, which is understood primarily as the irregularity of the data structure and the difficulty of extracting homogeneous data from the flow and identifying correlations.

The development of BigData technologies is fortified by other criteria, such as veracity (reliability), variability, value, visibility (Fig. 1).

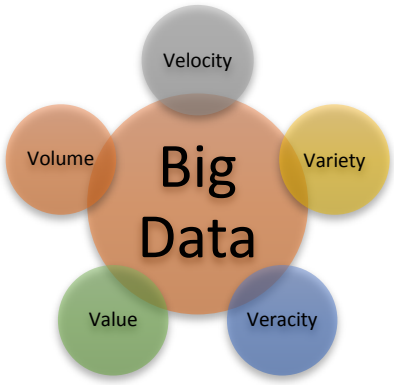

Fig. 1. BigData basic parameters

The ever-growing data flows present both opportunities and challenges. On the one hand, the prospect of better analysis of information resource data flow allows for more informed decisions, on the other hand, such drawbacks as security issues, can put companies in a difficult position when dealing with sensitive information.

The more sophisticated the digital technologies become and more complicated the information infrastructure is the larger is the number of cybercrimes [1], [2].

The cybercrime is feasible through the implementation of computer attacks, or massive computer attacks. The issues of testing and monitoring the computer attacks security of information systems were contemplated in the works of V.V. Kulba, S.I. Makarenko, I.I. Livshits, A.S. Markov and et al. Separate research results are provided in [3], [4], [5], [6], [7].

The scope of the present article assumes a massive computer attack (MCA) as a targeted unauthorized impact on the information resources of an automated information system or gaining unauthorized access to them using software or firmware [8]. At the same time, the systems must have reliable, scientifically substantiated methods to facilitate the assessment of the stability of the BigData functioning under *massive computer attacks* (MCA).

III. RESEARCH OBJECTIVE

Ensuring the security and confidentiality of business data together with sensitive customer information are major security concerns related to BigData use. In this regard, it is highly desirable to increase the level of system stability against cyberattacks, set up tools for automatic data cleaning, data masking and documents protection, and establish mandatory authorization for employees, accompanied by uninterrupted system status monitoring.

One of the key security concerns associated with the adoption of BigData technologies is the complexity of data in terms of its structure, source, storage location, format, device type, etc. Variety of procedures conducted with BigData, especially storage, cleaning, masking, and so on, makes managing information resource data a complicated task. Lack of security measures when storing and processing BigData can lead to data leakage.

The opportunity of better analysis, despite allowing for more informed decisions, has certain disadvantages, such as security issues which bring the businesses into difficulties when dealing with sensitive information. Classification of the main security problems when using BigData technologies is presented in Figure 2.

Data storage. Enterprises tend to adopt cloud storage to expedite data movement and speed up business operations. However, the associated risks are exponential with security concerns. Even the slightest mistake in data access control can lead to anyone getting a lot of sensitive data. As a result, large tech companies are using both on-premises and cloud storage for security and flexibility.

Critical information may be stored in local databases, whereas less important data is stored in the cloud for ease of use. However, to enforce security policies on local databases, companies need cybersecurity experts. This obviously increases the cost of managing data in on-premises databases, and still, companies shouldn't take the security risks for granted by storing all data in the cloud.

Fake data. Fake data generation poses a serious threat to business since it takes up the time that could otherwise be spent identifying or resolving other pressing issues. There is more scope for using inaccurate information on a very large scale, as soon as estimating individual data points can be challenging for companies.

False flags for fake data can also lead to unnecessary actions that could potentially reduce performance or slow down other essential business processes. Criticism of data to work with is one of the means to avoid the abovementioned and improve business processes. The optimal approach consists in validating the data sources by periodically evaluating and analyzing machine learning models using different test datasets to search for abnormalities.

Data sensitivity. Data sensitivity maintenance is a huge task in today's digital world. It aims to protect personal or confidential information from cyberattacks, hacks, and intentional or unintentional data loss. Companies should closely adhere to data privacy guidelines presumed by cloud access control services, including very strict privacy enforcement to enhance data protection. It is best to follow several rules along with the implementation of one or more data protection technologies. General rules include data knowledge, vaster control over the data storage and backup, protecting the network from unauthorized access, conducting frequent risk assessments, and regularly training users in data privacy and security principles.

Data control. A security breach can be devastating to a business, including exposing critical business information to a fully compromised database. Deploying highly secure databases is vital to ensuring the all-level data security. An excellent database management system comes with various access controls. While it is recommended to follow strong and rigorous physical security practices, it is even more important to follow extensive software security measures to protect data storage. Example of some methods to effectively achieve this goal: data encryption, data segmentation and separation, protection on the way, and implementation of a trusted server. In addition, some security tools can integrate with databases to automatically track data exchanges and notify users of data compromises.

Identity Access Management. Through control over the data available for viewing or editing by the users the companies ensure the data consistency and confidentiality. Yet, implementation of access control is not an easy thing to do, especially taking large companies with thousands of employees. Still, shifting from local solutions to cloud services streamlines the Identity Access Management (IAM) handling. IAM controls the data flow by means of identification, authentication and authorization. Another good benchmark for the company to follow IAM best experience is meeting the relevant ISO standards.

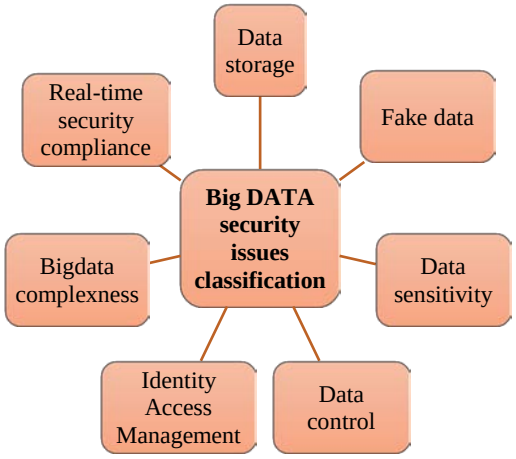

Fig. 2. BigData basic security issues classification

BigData complexness. One of the biggest threats to BigData is the diversity of datasets in use at any given time. Such information can be structured as well as unstructured and can come from mobile devices, servers, email files, cloud applications, and more. The more complex the data becomes the harder it is to secure, so it's important to use a proven Extract, Transform, Load (ETL) service to increase data compatibility.

Real-time security compliance. A perfect BigData store should include real-time tools that ensure compliance with security requirements. These resources are dedicated to constant processes monitoring to ensure that the company is taking the proper steps to protect data. However, as they run, these tools generate their own large amounts of data, which need further processing. Breach of this data might be as detrimental as hackers gaining access to other databases.

The above mentioned real-time compliance tools, despite they create a lot of extra data that needs to be protected, are nevertheless essential. It is necessary to explore all the options and make good choice. Better to choose the tools designed to minimize potential false positives. One false violation warning can lead to a waste of resources and the possibility of overlooking actual violations [6].

Data Mining Challenges. Data mining is a powerful tool to ensure better understanding and usage the company-owned data. Even though this process is handled by professionals, it can still create BigData security issues that should not be overlooked.

Control of information that IT-specialists gain access to when doing data mining is of almost importance. Access to age and geographic demographics can help reveal invaluable patterns, however, there's usually no real reason to provide access to credit card information, social security numbers, or other sensitive data.

Theft of employees' sensitive data. An advanced data culture has allowed every employee to have some level of critical business information. Such data democratization at the same time presumes the risk that an employee will intentionally or unintentionally cause the leakage of sensitive information.

Theft of confidential employee data is typical for large technology companies, as well as upstarts. To avoid employee data theft, companies should implement legal policies along with securing the network with a virtual private network.

Additionally, companies can use Desktop as a Service (DaaS) to eliminate the functionality of data stored on local drives.[2]

IV.RESULTS AND DISCUSSION

Considering the models of scenarios of massive computer attacks on the example of the process of BigData managing access control. It is of practical interest to use the Monte-Carlo method for modeling the processes of functioning of a MCA-exposed BigData, taking into account the importance of the object of a computer attack (CA) as part of a computer network (CN), for which a block for specifying various attack scenarios was introduced into the model.

Models of MCA-exposed BD can be divided into 3 subgroups:

– Accurate mathematical model;
– Approximate analytical model;
– Statistical model based on the Monte-Carlo method

Verbal statement of the problem:

Given:

a) the pole BigData configuration of (0 - *N*), where *N* is the number of elements representing the terminal vertex, 0 is the index of the control element, means, the vertex of the network; elements from 1 to *N* are the end vertices;

b) BD availability requirements;

c) a subset of controlled elements that are included in the technological chain (peculiar of the attack period) and belong in BigData elements.

It takes: to ensure the exchange of technological information between subjects and objects of control (end vertices BD).

Hacker's objective is to inflict maximum damage on the manageability of the BigData (minimize the number of end vertices of the BigData that have connections with the subject).

Problem statement:

Initial data:

a) *n* – number of attacks;

b) configuration (1-*N*), 1 – decision maker pole, *N* – controlled objects (CO) poles;

c) the threshold number of controlled objects N_{tr}, having at least one connection with the subject, at which the controllability of the process control system is ensured;

d) BigData communication network configuration – $S_{BD}=(V,E)$, where:

– $V = \{v_i\}$ – set of BD nodes, v_i= 1, if the element is in operating condition, v_i=0 in another case, i=0,1,2,…,K,K – set of BD nodes, i=0, pole index, i= 1 up to K – BD nodes indices;

– $E = \{e_{ij}\}$ – set of connection lines between node elements V, $e_{ij} = 1$, if the connection between nodes i-m and j-m is envisaged, $e_{ij}=0$, if the connection between nodes is not envisaged, $i=1,2,\ldots,K-1, j = 2, 3, \ldots, K$, K – number of V-set elements, including the control element and N – controlled elements;

– $V^* = \{v^*\}$ V – the subset of controlled objects that are under the control of the subject.

MCA scenarios:

1. The attacker is not aware whether the elements of the computer network belong to the BigData.

2. The MCA initiator has no knowledge of the BigData - S_{BD} network configuration, but there is information about affiliation with the BigData.

3. The MCA initiator knows the S_{BD}, and is capable of assessing the structural importance of the S_{BD} elements, and plans the first and subsequent CA considering the structural importance of the S_{BD} elements. A computer attack is planned according to the criterion of decreasing w_i, where w_i is the coefficient of structural importance of the element, which is calculated according to [18].

4. The MCA initiator is aware of the result of the previous attack, which allows him to correct the initial attack scenario. For example, the attack was unsuccessful, and the target of the attack remained undamaged (operative), consequently, this object can be included in the attack plan again.

Limitations:

1. The probability of striking the S_{BD} elements, meaning subjective probability, is set by means of expert assessments [11], [12] and [13] or is found using a statistical model.

2. S_{BD} elements change their conditions, shifting from operative to damaged and back at time t_i, where $i=0+\Delta t$, Δt is the discretization interval of statistical model events.

3. The cost of a successful attack on a vertex element is incommensurably high compared to successful attacks on S_{BD} elements.

4. Controlled objects, means elements of the set V^*, are equivalent.

Experimentally [18]:

1. The following initial data is entered:

a) About the security of the object of attack;

b) About the capabilities of the attacker.

2. The results of simulation implementations are generalized.

3. BigData operative availability is estimated.

4. An interpretation of the result obtained is carried out under the conditions of decision-making under many criteria [18] and in definitions.

Thus, the study outlines the solution of the actual scientific problem of assessing stability under the MCA exposure conditions, which is of practical importance in determining the information risks of BigData technologies.

V.CONCLUSION

The conclusions obtained are of interest for planning further scientific research towards developing advanced solutions to ensure the stability of the BigData functioning under the massive computer attacks (MCA):

a) random attacks on the considered computer network configuration are recognized the least effective;

b) knowing the configuration of the computer network, the hacker can inflict CA, based on the importance of the elements. This method of implementing the attack is more effective compared to random attacks;

c) knowing the configuration of the network and having an attack plan taking into account the structural importance of the elements, the hacker can cause serious damage to the computer network at a lower cost;

d) security of the confidentiality of the of the BigData communication network configuration is a reasonable measure to ensure its MCA-survivability.

So far, insufficient attention has been paid to the security issues of BigData systems as soon as the vast majority of projects are implemented regardless the information security, which sooner or later will lead to a significant increase in the time and cost of implementing the security systems.

The results obtained might be promising in a way of improving the development of solutions to ensure the stability of the BigData function under the conditions of object-orientated aggressive actions.

REFERENCES

[1]. 2020 Network Security and Accessibility Report, https://habr.com/ru/company/qrator/blog/548314/ (date of access 16.02.2023)

[2]. 2019 Data Breach Investigations Report, https://enterprise.verizon.com/resources /reports/2019-data breach-investigations-report.pdf (date of access 16.02.2023)

[3]. Makarenko S. I. Audit of information security: main stages, conceptual framework, classification of measures // *Control Systems, Communications and Security*. 2018. vol 1. pp. 1–29.

[4]. Livshits II Modern practice of information security audit // *Quality management.* 2011. vol. 7. pp. 34–41.

[5]. Kulba V. V., Shelkov A. B., Gladkov Yu. M., Paveliev S. V. *Monitoring and audit of information security of automated system*s. Moscow: IPU im. V. A. Trapeznikova RAN, 2009. 94 p.

[6]. Markov A. S., Tsirlov V. L., Barabanov A. V. *Methods for assessing the inconsistency of information security tool*s / ed. A. S. Markova. M.: Radio and communication, 2012. 192 p.

[7]. Voevodin V. A. *Methods of audit and monitoring of the information security management system in terms of ensuring the protection of information in web application*s // Modern research in the field of social, economic and technical sciences: monograph. N. Novgorod: NOO "Professional Science", 2021. pp. 8–44.

[8]. GOST R 51275–2006. *Data protection. Informatization object. Factors affecting information. General provisions.* M.: Standardization, 2018.

[9]. Voevodin V. A., Kovalev I. S., Folomeev L. A. Insurance of information risks as a tool for managing information protection. *In. International CONFERENCE, 2019. Ser.: Scientific. conf. dedicated to Radio Day (issue LXXV).* M.: Publishing House of Moscow. NTO radio engineering, electronics and communications them. A. S. Popova, 2020. pp. 152–155.

[10]. GOST R 59516-2021. Information Technology. Information security management. Information security risk insurance rules. Moscow: Standartinform, 2021. 20 p.

[11]. Statyev V.Yu. Information Security in the Big Data Space, DOI: 10.36724/2072-8735-2022-16-4-21-28

[12]. Big Data Security Concerns https://www.integrate.io/blog/big-

data-security-concerns/ (date of access 16.02.2023)

[13]. Top 7 Big Data Security Issues and Their Solutions https://hevodata.com/learn/big-data-security/ (date of access 16.02.2023)

[14]. Security of the information space in the context of big data. Kurbatsky V.A. *In the collection: Comprehensive Protection of Information. Materials of the XXII scientific-practical conference. 2017*. pp.250-253.

[15]. Popova A.V., Voronova Yu.S. Information security of personal data of citizens in the conditions of technologies "big data" ("big data"). *Alley of Science*. 2020. Vol. 1. pp. 683-686.

[16]. Problems of ensuring information security in big data management systems. Poltavtseva M.A. In the collection: XIII All-Russian meeting on the management of VSPU-2019. *Proceedings of the XIII All-Russian Conference on Management Problems of the VSPU-2019*. Institute of Management Problems. V.A. Trapeznikov RAS. 2019. pp. 2606-2611

[17]. Technologies of Big Data (Big Data) in the field of information security. Petrenko A.S., Petrenko S.A. In *the collection: The 2018 Symposium on Cybersecurity of the Digital Economy (CDE'18).* Second international scientific and technical conference, 2018, pp.248-255.

[18]. Korobov V. B. Teoriia i praktika ekspertnykh metodov: monografiia [Theory and practice of expert methods: monograph]. Moscow, INFRA-M Publ., 2019.

[19]. Big Data Facts 2012 Promotion of mobile applications (alakris.ru) (date of access 26.02.2023)