УДК 004.056.523:[003.26:004.62
DOI: 10.31673/2409-7292.2024.010011

**Havrylova A., Korol O., Voropay N.,
Sevriukova Y., Bondarenko K.**

ANALYSIS OF CRYPTOGRAPHIC AUTHENTICATION AND MANIPULATION DETECTION METHODS FOR BIG DATA

The emergence of Big Data has made it possible to obtain more information about the customer and competitive base, as well as about market trends. Therefore, the desire of criminals to gain unauthorized access to this data also increases in direct proportion. However, not enough attention is paid to security issues of Big Data systems and their creation does not take into account the information security component. The purpose of the article is to develop approaches to the cyber protection of Big Data, which are stored and transmitted by telecommunication communication channels when responding to a request, namely their integrity and authentication. The main problems when working with Big Data are analyzed. Based on the results of the analysis of Big Data protection problems, a cryptographic transformation of data indexes when they are stored in the database and when transmitted in the form of hash codes is proposed to increase the efficiency of search by user requests. The expediency of using crypto-code constructions on elliptic curves of various types of modifications to form a pseudo-random substrate in order to increase the crypto-resistance of transmitted messages is proven. When forming hash codes, it is advisable to use modified elliptical codes, and under stricter conditions, the level of cryptographic resistance of authentication codes can be increased due to hybrid crypto-code constructions. A pseudo-random substrate can be represented by varieties that must equally ensure the necessary transformations and preservation of universality properties by the UMAC algorithm.

Keywords: Big Data, crypto-code constructions, hybrid crypto-code constructions, elliptic curves, pseudo-random layer, damaged codes.

Introduction

New technologies make it possible to create huge arrays of data and the ability to process it. The emergence of Big Data has made it possible to obtain more information about the customer and competitive base, as well as about market trends. Thus, from 2020 to 2025, the amount of data generated by mankind will grow from 51 zettabytes to 175 zettabytes [1]. And the global Big Data analytics market, which was estimated at $41.85 billion in 2019, will grow to $115.13 billion, with an average growth rate of 11.9% by 2028 [2]. Among the main advantages of Big Data for business, according to information obtained as a result of a survey by the research company The Economist Intelligence Unit and the consulting company Accenture, the following can be highlighted: search for new sources of income (56%); improving customer experience (51%); new products and services (50%); influx of new customers and retention of loyalty of old ones [3].

Therefore, the desire of criminals to gain unauthorized access to this data also increases in direct proportion. However, not enough attention is paid to security issues of Big Data systems and their creation does not take into account the information security component.

Therefore, the purpose of the article is to develop approaches to the cyber protection of Big Data, which are stored and transmitted by telecommunication communication channels when responding to a request, namely their integrity and authentication.

Analysis of problems when working with Big Data

Big Data is characterized not only by volume, velocity and variety, but also by reliability and potential value. In order to maintain the required values according to these characteristics, the following problems are now identified when working with Big Data [3]:

1) lack of practice in working with Big Data and its protection;
2) lack of Big Data protection methodologies;
3) lack of Big Data protection standards;
4) a large ecosystem of Big Data;
5) lack of Big Data regulation.

Among the leading standardization institutes, special attention is paid to the recommendations of the US National Institute of Standards and Technology (NIST), because its experts advise first to focus on ensuring the security and confidentiality of data at all technological levels of their processing, covering five main interfaces of interaction with data [3], namely:

interaction interface between data providers and application providers;
the interaction interface between application providers and data consumers;
the interaction interface of the application provider and the platform for working with Big Data;
data protection during the internal interaction of various Big Data technologies and platforms;
protection of Big Data system management tools.

Among Big Data protection tools, can also be singled out cryptographic means of information protection, which can be effectively used both when storing data and when transmitting it.

Using multilayer hashing functions using the UMAC algorithm

In targeted attacks, even a file downloaded from a trusted source can become a loophole for an attacker. And if quantum computers are connected, then new threats to encryption appear, as they will be able to radically change the time of selection of cipher keys. Therefore, new opportunities for information protection should appear.

Thus, NIST specialists point to the developments of the Cloud Security Alliance (CSA) [4] and recommend focusing on four areas of protection:

1) infrastructure security;
2) data confidentiality;
3) data management;
4) integrity and response procedures.

In this work 2) and 4) are considered.

The implementation of the data privacy direction is considered through the impact of social data on security and privacy when making requests to Big Data. Moreover, data protection must be ensured regardless of where it is stored or used, and ensuring the confidentiality and manageability of Big Data must be considered through data inventory and classification, the use of data masking technologies, the formation of management policies and data access rules [5].

Maintaining the integrity and conducting the response procedure is carried out through Big Data analysis to detect malicious activity and understand the state of Big Data processing systems, detect security events and respond to identified threats, detect, analyze and investigate incidents, as well as security of the analytics results.

To implement these directions, it is proposed to use the approach of converting indexes in the database into hash codes, which will reduce the capacity of the database itself and increase the efficiency of responses to requests to it. Also, when the results of requests are transmitted to the database via telecommunication communication channels, hashing of transmitted messages is used, which is usually carried out using tamper detection codes (to control data integrity) and message authentication codes (to confirm the authenticity of data).

Therefore, the use of hashing as database indexes and data transmission on request through telecommunication communication channels can be represented in the form of the following tuple [6–10]:

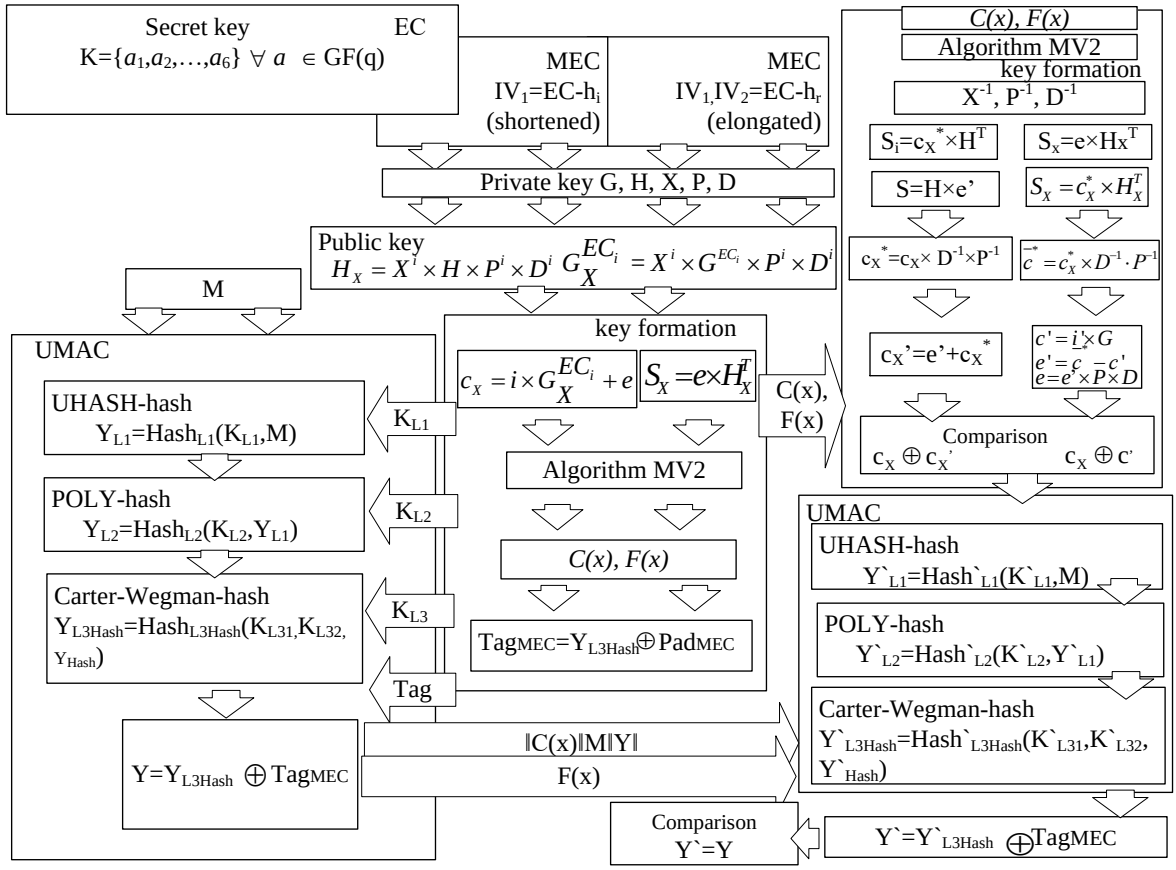
$$\{C_{Index/mess}; Hash_{UMAC}; Pad\}, \quad (1)$$

where $C_{Index/mess}$ – index / message codegram; $Hash_{UMAC}$ – index/message hash code; Pad – pseudorandom layer.

When using a reliable hash function, it is computationally difficult to create a fake message with the same hash code (MAC code - message authentication code) value as the real message. However, these threats can be implemented due to the weakness of specific hashing algorithms, signature or errors in their implementation [7, 11].

The method of building multilayer hashing functions using the example of the UMAC algorithm (Fig. 1) is expedient.

It is based on a combination of multi-stage universal key hashing and the use of a block symmetric cipher.



Pic 1. Structural diagram of the implementation of the modified UMAC algorithm on McEliece and Niederreiter CCC, on EC, on MEC and on DC

This algorithm uses many universal hash functions and provides proven security when forming an authentication code [8, 12]:

1st layer – the value of the universal hash function (UHASH-hash) of the first level of hashing:

$$Y_{L1} = Hash_{L1}\left(K_{L1}, M_{index/mess}\right), \tag{2}$$

where $M_{index/mess}$ – index / message; $Hash_{L1}$ – universal key hashing function $M_{index/mess}$ using a secret key of the first hashing level K_{L1};

2nd layer – the value of the universal hash function (UHASH-hash) of the second hashing level:

$$Y_{L2} = Hash_{L2}\left(K_{L2}, M_{index/mess}\right), \tag{3}$$

where $Hash_{L2}$ – universal key hashing function Y_{L1} using a secret key of the second hashing level K_{L2};

3rd layer – the value of the universal hash function (UHASH-hash) of the third hashing level:

$$Y_{L3} = \left(\left(\left(\sum_{i=1}^{m} Y_{L2_i} K_{L3_{1i}}\right) \bmod(prime(36))\right) \bmod(2^{32})\right), \tag{4}$$
$$xor(K_{L3_2})$$

where *xor* – "exclusive OR" operation on previous and next values.

Additional crypto-resistance of the code is provided by the use of a pseudo-random lining on the last layer *Pad* [9, 13–15]:

$$Y = Hash_{UMAC} \oplus Pad\ . \tag{5}$$

Thus, universal hashing in a multi-layer UMAC design allows for the same probability of forming hash images for a variety of used key data. This property ensures the security of the encryption algorithm [10, 16].

Pseudo-random lining *Pad* strengthens the crypto-resistance of the MAC authentication code. Formation of a pseudo-random lining *Pad* is presented in the form of hybrid McEliece crypto-code constructions on modified elliptic curves (MEC) with various types of modifications (on extended and shortened, as well as damaged codes) [10].

To form *Pad* y in the form of crypto-code structures on modified elliptical codes with lengthening and shortening and in the form of hybrid crypto-code structures on modified elliptical codes with damage (DC) [8]: text formation with modifications (lengthening/shortening, damage):

$$C_j^* = C_j - C_{k-h_j}, E_{K_{MV2}}, \tag{6}$$

$$C_j^* = C_{h_r}, E_{K_{MV2}}\ . \tag{7}$$

where C_j – index/message codegram; C_{k-h_j} – modified index/message code when shortened; C_{h_r} – modified index/message code when extended; $E_{K_{MV2}}$ – damage based on the use of the key K^i_{MV2} and algorithm *MV*2.

Thus, as a mechanism for the formation of a pseudorandom lining *Pad* for the third layer of the cascade UMAC hashing algorithm, it is proposed to use crypto-code constructions on elliptic curves and their modifications.

To assess the stability of crypto-algorithms for formation Pad we will use the package NIST STS 822 [17]. The results of the research are shown in Table 1. According to the given data, it was concluded in [18] that the statistical characteristics of the proposed crypto-code constructions are not inferior to the traditional asymmetric crypto-code systems of McEliece. All cryptosystems passed 100% of the tests and following results were obtained: 155 out of 189 tests passed at the level of 0.99 (hybrid McEliece crypto-code constructions on shortened MEC), which is 82% of the total number of tests; 149 out of 189 tests passed at 0.99 (traditional asymmetric McEliece crypto-code).

Table 1

Statistical security studies results [18]

Cryptosystems	Number of tests in which more than 99% of the sequences passed the test	Number of tests in which more than 96% of the sequences passed the test	Number of tests where less than 96% of the sequences passed the test
McEliece's ACCS	**149 (78,83%)**	**189 (100%)**	**0 (0%)**
McEliece's MACCS on shorter MECs	151 (79,89%)	189 (100%)	0 (0%)
McEliece's MACCS on extended MECs	152 (80,42%)	189 (100%)	0 (0%)
HCCCDC on extended MECs	153 (80,95%)	189 (100%)	0 (0%)
HCCCDC on short MECs	**155 (82 %)**	**189 (100%)**	**0 (0%)**

From the above, it can be seen that *Pad*, which is formed on damaged codes (DC) has the greatest stability. *Pad* formed on traditional elliptical codes (EC) are characterized by the least

stability. Based on the conducted studies of the properties of McEliece code cryptosystems, stability estimates are given in works [6 – 11]. This also extends to quantum cryptanalysis. It is also shown that the relative speed of data transmission for the most important cases of applied nature, the McEliece cryptosystem allows to increase the relative speed of information transmission by 30-40% [19]. The results of the study of the effectiveness of the formation of crypto-code transformations using the McEliece cryptosystem are shown in Table 2.

Table 2

Effectiveness of performing transformations in Niederreitter and McEliece cryptosystems on the EC

Type of cryptosystem	Information vector length, symbols			Crypto conversion rate, bit/s	
	10	100	1000	block size, bit	
				256	128
McEliece	0,55	1,53	4	357,534	365,551

In the proposed method of forming data integrity and authenticity control codes, the first layers of the transformation are proposed to be implemented using high-speed, but cryptographically weak universal hashing schemes traditional for the UMAC algorithm, the last layer is proposed to be implemented using a developed secure (cryptographically strong) strictly universal hashing scheme based on algebraic geometric codes and crypto-code constructions [8-10, 20 – 22].

This algorithm for generating a hash code is cryptographically resistant to cracking in the post-quantum period, since the technical characteristics of quantum computers will only slightly reduce the number of operations required to crack this algorithm.

Conclusions

Based on the results of the research, it can be concluded that there is no single way to prevent leakage and distortion of information during its transmission. Therefore, a comprehensive approach to the organization of authenticated access and data hashing is important. When forming hash codes, it is advisable to use modified elliptical codes, and under stricter conditions, the level of cryptographic resistance of authentication codes can be increased due to hybrid crypto-code constructions. A pseudo-random substrate can be represented by varieties that must equally ensure the necessary transformations and preservation of universality properties by the UMAC algorithm.

The results of research on crypto-resistance and efficiency confirmed the feasibility of using the modified UMAC algorithm to transform data indexes in Big Data into more stable crypto-code structures, and their use when transmitting data at the request of authorized users through telecommunication channels.

Перелік посилань

1. Hordienko, N. Protect big data and minimize the risk of information loss, 2022, URL: https://www.ukrlogos.in.ua /10.11232-2663-4139.04.32.html.

2. Iluk, A. Risks associated with the protection of personal data in context Big Data. 2017, vol. 42 (592). URL: https://yur-gazeta.com /publications/practice /inshe/riziki-povyazani -iz-zahistom-personalnih-danih-v-konteksti-big-data.html.

3. NIST Special Publication 1500-1. NIST Big Data Interoperability Framework, 2020, URL: https://bigdatawg.nist.gov/_upload files/NIST.SP.1500-1.pdf.

4. Big Data Taxonomy, Cloud Security Alliance, 2021, URL: https:// downloads. cloudsecurityalliance. org/initiatives/bdwg /Big_Data_Taxonomy.pdf.

5. Big Data Security and Privacy Handbook: 100 Best Practices in Big Data Security and Privacy. Cloud Security Alliance, URL: https://downloads. cloudsecurityalliance.org/ assets/research/big-data/BigData_Security_ and_Privacy_Handbook. pdf.

6. Havrylova, A. A., Korol, O. H., Milevskyi, S. V., Bakirova, L. R. Mathematical model of authentication of a transmitted message based on a McEliece scheme on shorted and extended modified elliptic codes using UMAC modified algorithm, Кібербезпека: освіта, наука, техніка, 2019, No 1(5), pp. 40 – 51.

7. Yevseiev, S., Havrylova, A., Korol, O., Dmitriiev, O., Nesmiian, O. [and etc.]. Research of collision properties of the modified UMAC algorithm on crypto-code constructions, EUREKA: Physics and Engineering, Talli, Osauhing "Scientific Route", 2022, Number 1 (38), pp. 34 – 43.

8. Korol, O., Havrylova, A. Mathematical models of hybrid crypto-code constructions in the UMAC algorithm Przetwarzanie, transmisja i bezpieczenstwo informacji, Bielsko-Biala, Wydawnictwo naukowe Akademii Techniczno-Humanistycznej w Bielsku-Bialej, 2020, Vol. 12, pp. 125 – 134.

9. Yevseiev, S., Havrylova, A. Improved UMAC algorithm with crypto-code McEliece's scheme, Modern Problems Of Computer Science And IT-Education : collective monograph / [editorial board K. Melnyk, O. Shmatko], Vienna, Premier Publishing s.r.o., 2020, pp. 79 – 92.

10. Korol, O., Havrylova, A., Yevseiev, S. Practical UMAC algorithms based on crypto code designs, Przetwarzanie, transmisja I bezpieczenstwo informacji, Bielsko-Biala, Wydawnictwo naukowe Akademii Techniczno-Humanistycznej w Bielsku-Bialej, 2019, Tom 2, pp. 221-232.

11. Evseev, S., Kotz, H., Korol, O. Analysis of the legal framework for the information security management system of the nsmep, Eastern-European Journal of Enterprise Technologies, 2015, 5(3), pp. 48–59.

12. Evseev, S., Abdullayev, V. Monitoring algorithm of two-factor authentication method based on passwindow system. Eastern-European Journal of Enterprise Technologies, 2015, 2(2), pp. 9–16.

13. Yevseiev, S., Tsyhanenko, O., Ivanchenko, S., Milov, O., Shmatko, O. Practical implementation of the Niederreiter modified crypto-code system on truncated elliptic codes, Eastern-European Journal of Enterprise Technologies, 2018, 6(4-96), pp. 24–31.

14. Yevseiev, S., Kots, H., Liekariev, Y. Developing of multi-factor authentication method based on Niederreiter Mc-Eliece modified crypto-code system. Eastern-European Journal of Enterprise Technologies, 2016, 6(4), pp. 11–23.

15. Yevseiev, S., Korol, O., Kots, H. Construction of hybrid security systems based on the cryptocode structures and flawed codes, Eastern-European Journal of Enterprise Technologies, 2017, 4(9-88), pp. 4–21.

16. Milov, O., Yevseiev, S., Ivanchenko, Y., Tiurin, V., Yarovyi, A. Development of the model of the antagonistic agents behavior under a cyber conflict. Eastern-European Journal of Enterprise Technologies, 2019, 4(9-100), pp. 6–10.

17. Rukhin, J. Soto. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. NIST Special Publication 800-22, 2000.

18. Yevseiev, S., Ponomarenko, V., Laptiev, O., Milov, O. Synergy of building cybersecurity systems: monograph, Kharkiv: PC Technology Center, 2021, 188 p.

19. Yevseiev, S., Tsyhaneko, O., Gavrilova, A., Guzhva, V., Milov, O., Moskalenko, V., Opirskyy, I., Roma, O., Tomashevsky, B., Shmatko, O. Development of Niederreiter hybrid crypto-code structure on flawed codes, Eastern-European Journal of Enterprise Technologies, 2019, № 1/9 (97), pp. 27 –38.

20. Gavrilova, A., Volkov, I., Kozhedub, Yu., Korolev, R., Lezik, O., Medvediev, V., Milov, O., Tomashevsky, B., Trystan, A., Chekunova, O. Development of a modified UMAC Algorithm based on crypto-code constructions, Eastern-European Journal of Enterprise Technologies, 2020, № 4/9 (106), pp. 45 –63.

21. Havrylova, A., Tkachov, A., Shmatko, A. Development of a pseudo-random substrate for the UMAC algorithm on crypto-code constructions, Information Protection and information systems security 2021, november 11–12, 2021, Lviv, Ukraine: Materials of the VIII Intern. sci.-tech. conf., Lviv: Education of the Lviv Polytechnic, 2021, pp. 49 - 50. URL: https://drive.google.com/drive/folders/18xwh1 x6hp2ggE14Znhf4Cknl1164FRyi?usp=sharing.

22. Yevseiev, S., Milevskyi, S., Bortnik, L., Voropay, A., Bondarenko, K., and Pohasii, S.,"Socio-Cyber-Physical Systems Security Concept", 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 09-11 June 2022, Ankara, Turkey DOI: 10.1109/HORA55278.2022.9799957.

23. Laptiev, O., Tkachev, V., Maystrov, O., Krasikov, O., Open'ko, P., Khoroshko, V., Parkhuts, L. The method of spectral analysis of the determination of random digital signals. International Journal of Communication Networks and Information Security (IJCNIS). Vol 13, No 2, August 2021 P.271-277. ISSN: 2073-607X (Online). DOI : 10.54039/ijcnis.v13i2.5008 https://www.ijcnis.org/index.php/ijcnis/article/view/5008 .

24. Kyrychok, R., Laptiev, O., Lisnevsky, R., Kozlovsky, V., Klobukov, V. Development of a method for checking vulnerabilities of a corporate network using bernstein transformations. Eastern-European journal of enterprise technologies. Vol.1№9 (115), 2022 P. 93–101. ISSN (print)1729 - 3774. ISSN (on-line) 1729-4061. DOI: 10.15587/1729-4061.2022.253530.

25. Petrivskyi, V., Shevchenko, V., Yevseiev, S., Milov, O., Laptiev, O., Bychkov, O., Fedoriienko, V., Tkachenko, M., Kurchenko, O., Opirsky, I. Development of a modification of the method for constructing energy-efficient sensor networks using static and dynamic sensors. Eastern-European journal of enterprise technologies. Vol.1№9 (115), 2022 pp. 15–23. ISSN (print) 1729 - 3774. ISSN (on-line) 1729-4061. DOI: 10.15587/1729-4061.2022.252988.

Надійшла 07.02.2024