INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT

UDC 004.414.23 https://doi.org/10.23947/1992-5980-2020-20-2-196-200

Artificial intelligence in data storage systems

V. V. Zhilin[1], O. A. Safar'yan[2]
[1] Budenny Military Academy of Communication (St. Petersburg, Russian Federation)
[2] Don State Technical University (Rostov-on-Don, Russian Federation)

Introduction. The artificial intelligence (AI) performance in data storage systems is considered. When working with data, the advantage of its use both in economic terms and for security is determined. The work objective is the introduction of artificial intelligence in data storage systems. The key tasks involve the description of methods for data separation, organization of its storage and counteraction to security threats.
Materials and Methods. The data that should be fed into the drives is divided into parts so that it can be restored without one of the parts. This is necessary to be able to access and recover information in the event of a software or hardware failure.
Results. The AI performance under detecting security threats is considered. Since the model implies the interaction of users with data, it was found out how the data access control is carried out and the keys are stored.
Discussion and Conclusions. The use of AI in organizing a data warehouse will speed up the system. Artificial intelligence with built-in machine-learning algorithms will provide responding to a situation that affects the state of the system. Analysis of the state of the drives will avoid a possible hardware or software failure. Minimization of the human factor in the system operation contributes to the improvement of its work.

Keywords: artificial intelligence, threshold separation, threat, machine learning, storage organization, dynamic change, key, encryption, backup, attack, hash.

For citation: V. V. Zhilin, O. A. Safar'yan . Artificial intelligence in data storage systems. Vestnik of DSTU, 2020, vol. 20, no. 2, pp. 196–200. https://doi.org/10.23947/1992-5980-2020-20-2-196-200

Introduction. At the present stage of development of information and telecommunication technologies, a person comes across various kinds of information. The data storage in a safe place is an urgent issue. The science that deals with the storage of information in an encrypted form is cryptography. It studies how to hide data and provide its confidentiality[1].

The information is static in all data storage implementations. That is, the data will be where it was for the first time. Information can get to another disk or another sector only after removing it from its original location. However, this fact may be regarded as weakness of existing algorithms.

To increase the protection effectiveness, it is recommended to share data [1] in order to store their parts. This provides organizing fault-tolerant storages for information recovery algorithms in the event of any fault, failure of one of the drives, as well as with loss of one or more parts of the data.

If an attacker has access to the disk, he will be able to obtain the necessary information. In case of static storage, if an illegitimate user gains access to several such disks, he can recreate information. Separated data storage increases information security. In addition, in large storages, as a rule, hard disk drives are used, whose running speed (reading and writing operations) is significantly lower than that of solid state disks (SSD) and flash drives[2]. In this regard, you have to choose between the volume used and the operation speed.

The work objective is to describe the performance of artificial intelligence (AI) in data storage systems. The task is to describe the AI operation algorithms when performing the following operations:

• data recording on drives;

[1] Ryabko BYa, Fionov AN. Kriptografiya v informatsionnom mire [Cryptography in the Information World]: Goryachaya liniya-Telekom; 2018. 305 p URL: https://www.techbook.ru/book.php?id_book=1001 (accessed 04.05.2020).
[2] Sravnenie SSD i HDD diskov v real'nykh usloviyakh ispol'zovaniya [Comparison of SSD and HDD drives under real-time conditions of usage]. Habr. URL: https://habr.com/ru/post /394135/ (accessed 04.05.2020).

• referral from users;
• data warehouse analysis;
• in case of information security threats.

Artificial intelligence impact on the storage system quality. In modern devices, such as smartphones and computers, developers give special consideration to the implementation of artificial intelligence. In these technologies, machine learning algorithms are implemented, which increases the operation speed and reduces the response time to conducting frequently repeated information. It should be considered how the introduction of artificial intelligence and machine learning into storage systems can significantly improve the quality of their operation [2].

Currently, large companies and organizations are starting to implement AI in their data warehouses. A widespread adoption of machine learning and artificial intelligence technologies helps to improve the operating quality at the management level. This facilitates the work of network and data warehouse administrators through constantly diagnosing the causes of traffic overload and reduction. This will allow them to identify potentially vulnerable segments of the model beforehand.

Artificial intelligence involves the use of integrated deep learning algorithms that can predict the state of the entire system and respond quickly to possible changes. This will significantly reduce the cost of eliminating the consequences caused by equipment failure. In addition, the introduction of artificial intelligence in the organization of fault-tolerant storages will allow for their automation [3]. This implies an analysis of the system state and processing of incoming data in dynamic mode.

Artificial intelligence and machine learning provide minimizing the likelihood of data loss. Together with redundant arrays of independent disks, such a system increases the availability and speed of overcoming forced downtime thanks to intelligent data recovery, a backup strategy and the transfer of necessary data [4].

Data distribution by input parameters. To store data in a shared form, threshold separation methods can be applied. In classical algorithms, the in parameters are static values, which is a major disadvantage. Gaining access to one data through selecting input parameters threatens safety of all other data in the system.

Thus, the greatest preference in terms of information security is given to the algorithm that uses various input parameters for threshold sharing. These parameters can be randomly generated according to a certain algorithm or depend directly on the input data, which will be analyzed accordingly to select the most preferred parameters. The generation of such parameters will be handled directly by artificial intelligence. At the same time, it should consider the number of users to whom this information is available. Threshold sharing parameters and the location of the first share are stored in the database. This database is encrypted on the keys of specific users, which corresponds to the application of an asymmetric cryptoalgorithm.

Parameters of the proposed algorithm for data reliability. When distributing data among drives, artificial intelligence uses the following parameters: drive speed, availability, free volume and reliability index. At the same time, AI uses the data value parameters, size and frequency of accesses to it for calculation.

The algorithm for the distribution of data shares across drives is based on the calculation of drive coefficients. To determine the speed of the drive *S*, it is required to find the arithmetic average of the writing speed $s_{\text{зап}}$ and read rate $s_{\text{чт}}$:

$$S = \frac{s_{\text{зап}} + s_{\text{чт}}}{2}.$$

Drive accessibility A has 3 levels: 0 — unavailable, 1 — frequent situation of drive unavailability, 2 - rare situation of drive unavailability, 3 — anytime available. The exponential coefficient of the drive can be expressed by the formula:

$$K_{\text{н}} = S \cdot A \cdot V \cdot R,$$

where *V* is disk capacity; *R* is reliability, it is ranked from 1 to 10 points; *A* is the number of file accesses in a given period of time.

The significance factor of the file $K_{\text{ф}}$ depends on the value level (1 — low, 2 — medium, 3 — high). Based on the known attributes, this parameter can be calculated from the formula:

$$K_{\text{ф}} = S \cdot V \cdot A.$$

Consider the technique of selecting a storage drive for the remaining files. To determine the priority of the drive, we use the file number n_1 in the sorted table in descending order of the parameter $K_{\text{ф}}$. Then the disk number N_1 participating in the sample for storing the first share of the data is found from the formula:

$$N_1 = round\left(\frac{d \cdot n_1}{f}\right) mod\ n_2,$$

where d is the number of disks used for storage; f is the number of files in the system; n_2 is the number of drives; *round* is rounding to the nearest integer.

In the same way, drive N_2 is determined, which may be selected as a drive for storing the first share:

$$K_{first} = K_{\text{ф}} \cdot K_{d.max},$$

where $K_{\text{ф}}$ is the split file ratio; $K_{d.max}$ is maximum drive coefficient.

Next, the drive, whose difference in the coefficients K_{first} and $K_{\text{н}}$ is the smallest, is selected as N_2. To store the first share, one of the found disks N_1 and N_2 is randomly selected.

Advantages of the considered algorithm of data distribution among storage devices. Classically, disks combined into one data array (*RAID*-array) operate according to a static algorithm. That is, when the first share of data is detected, a search procedure for the second share can be performed. Each time, upon detection of the next share, the probability of determining a specific distribution algorithm increases. In the proposed algorithm, the distribution is based on information data and drives. In other words, the task of finding all the data shares is *NP*-complete, that is, it cannot be solved in polynomial time. It follows that the use of this algorithm increases the reliability of the information storage system.

However, users do not create a key; it is generated automatically and stored on the devices of the users themselves. Thus, when trying to decrypt data, a specific key must be used, otherwise this operation fails. The location of the keys changes dynamically over time. Such changes can occur at certain hours, or at specific time intervals [5]. Artificial intelligence is responsible for the key location with account of the availability of all devices connected to the system. This complicates the work of an attacker whose goal is to access information stored on drives [6].

Since the information in the repository is operated by users, it should be determined who of them is allowed to conduct operations with data, in other words, how to differentiate the data access.

Access comparison is performed according to the hash-sum table[1]. If a user has access to data, then they are restored. If there is no entry in the database on granting access rights, the entity requesting access to the object is denied.

In addition, an attempt to gain access will be recorded in the log file, and the subject who owns the information stored in the system will be informed about the access attempt by an illegal user or subject. In this case, artificial intelligence analyzes the actions of each user to make decision in the event of a specific situation. For example, it determines whether the query is invalid, or it has any purpose [2].

Fig. 1 shows an example of granting access to data recovery and access denial with recording the event in a log file and the user notice.

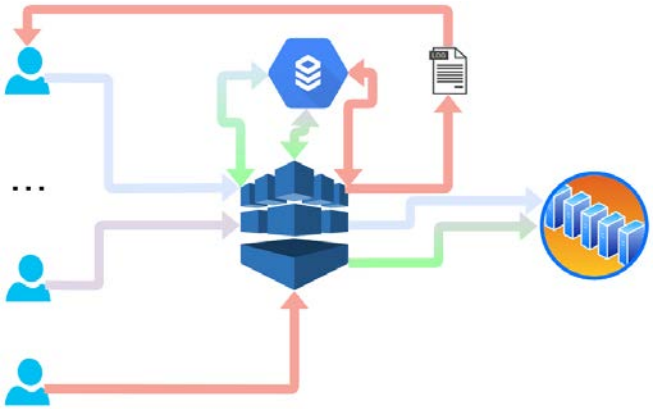

Fig. 1. Granting and denial of access to data recovery

Consider the access control function. Since the primary subjects in the system are users, the issue of providing access to data is particularly topical. Full control over user operations is very important. The decision to provide access to data is made by a distribution device with support of artificial intelligence. In this case, the use of AI will increase the

[1] Legkii sposob poddelki kontrol'noi summy s pomoshch'yu kollizii [Easy way to fake checksum using collisions]. xakep.ru. Available from: https://xakep.ru /2012/11/22/light-fake-checksum/ (accessed 04.05.2020).

user's data handling speed. This is achieved through transforming the most commonly used information, which avoids waiting for the response time to the sent data recovery request.

We consider dynamic change of storage. In certain cases, artificial intelligence controls the transfer of data and is responsible for data backup and restore including the time-based ones. In this case, the data migrates from one disk to another, and the information in the database also changes. This minimizes the likelihood of predicting the system state at a specific time. Since all keys are generated on the ground of previous sequences using hash sums, the task of finding a key designed to decrypt a database is an *NP*-complete problem [7]. In this case, artificial intelligence analyzes the state of drives, as well as the system as a whole. When a malfunction is detected, data is transferred from those drives on which failures are noticed. This minimizes the likelihood of information loss.

When a threat is detected, artificial intelligence analyzes the attack to determine its target. If the final goal is the storage object, then the data from it is transferred to another drive. Thus, in the event of a successful attack, AI will take effective actions to hide information [8].

As users work with data, the human factor plays an important role. Therefore, the system is designed so that only users with access can work with the data. There is an analogy with the mandatory access control model. An exception is that the distribution of access is controlled by the user who has created the information in the system and is its owner.

Conclusion. Thus, the use of artificial intelligence increases the system speed under organizing a data warehouse. Restricting user access to the algorithm of the system operation improves information security. Artificial intelligence with built-in machine learning algorithms provides a quick response to any situation affecting the state of the system. Analysis of the state of the drives avoids a possible hardware or software failure. Minimization of the human factor in the system operation helps to improve its functioning and a deeper analysis of user queries. In addition, the collection of information on possible attacks enables maintaining the system security at a proper level.

References

1. Mogilevskaya NS, Kul'bikayan RV, Zhuravlev LA. Porogovoe razdelenie failov na osnove bitovykh masok: ideya i vozmozhnoe primenenie [Threshold file sharing based on bit masks: concept and possible use]. Vestnik of DSTU. 2011;11(10):1749–1755. URL: https://vestnik.donstu.ru/jour/article/ view/912/907 (accessed 04.04.2020). (In Russ.)

2. Nikolenko SI, Kadurin AA, Arkhangel'skaya EV. Glubokoe obuchenie. Pogruzhenie v mir neironnykh setei [Deep learning. Immersion in the world of neural networks]. St. Petersburg: Piter; 2018. 481 p. (In Russ.)

3. Dubrova E. Fault-Tolerant Design. Springer; 2013. 185 p.

4. Flakh P. Mashinnoe obuchenie [Machine Learning]. Moscow: DMK Press; 2015. 400 p. (In Russ.)

5. Zhilin VV, Drozdova II, Cherkesova LV, et al. Trekhmernaya model' bezopasnosti komp'yuternykh sistem [Three-dimensional model of computer systems security]. Young Researcher of the Don. 2018;5:30–37. URL: http://mid-journal. ru/upload/iblock/f81/6_620_ZHilin_30_37.pdf (accessed 04.05.2020). (In Russ.)

6. Parloff R. Why Deep Learning Is Suddenly Changing Your Life. Fortune. 2016. Retrieved 13 April, 2018.

7. Cormen Th, Leiserson Ch, Rivest R. Algoritmy: postroenie i analiz [Algorithms: construction and analysis]. Moscow: Vil'yams; 2006. 1296 p. (In Russ.)

8. Hutson M. Missing data hinder replication of artificial intelligence studies. Science. 15 February, 2018. doi:10.1126/science.aat3298.

Submitted 09.04.2020
Scheduled in the issue 12.05.2020

About the authors:

Zhilin, Viktor V., student of the Cybersecurity of IT Systems Department, Military Academy of Communication of S.M. Budenny (3, Tikhoretsky Av., K-64, St. Petersburg, 194064, RF), ORCID: https://orcid.org/0000-0001-6277-3795, zhilin95@inbox.ru

Safar'yan, Olga A., associate professor of the Cybersecurity of IT Systems Department, Don State Technical University (1, Gagarin sq., Rostov-on-Don, 344000, RF), Cand.Sci. (Eng.), associate professor, ScopusID 57210832767, ORCID: https://orcid.org/0000-0002-7508-913X, safari_2006@mail.ru

Claimed contributorship

V. V. Zhilin: collection and analysis of literature data; determination of research techniques; task setting. O. A. Safar'yan: academic advising; basic concept and the paper structure formulation.

All authors have read and approved the final manuscript.