УДК 004.75

Algorithms for Assessing Quality Indicators Functioning of a Distributed System Storing Confidential Data

Alexander G. Basyrov[1] — Doctor of Technical Sciences, Professor; professor of the Department of the A. F. Mozhaysky Military Space Academy. Research interests: parallel computing, information systems. E-mail: alexandrerbas@mail.ru

Igor N. Koshel[2] — Candidat of Technical Sciences. Head of the Department of the A. F. Mozhaysky Military Space Academy. Research interests: parallel information processing planning, information systems. E-mail: kin1470@mail.ru

Valery V. Abramenkov[1] — Head of FSAU "Central Department of Housing and Social Infrastructure (Complex)". Research interests: information systems, data storage systems. E-mail: info@fgaul.ru

[1] A. F. Mozhaysky Military Space Academy, Saint Petersburg, Russia

[2] FSAU "Central Department of Housing and Social Infrastructure (Complex)", Moscow, Russia

For citation: Basyrov A. G., Koshel I. N., Abramenkov V. V. Algorithms for Assessing Quality Indicators Functioning of a Distributed System Storing Confidential Data // Intelligent technologies on transport. 2024. No. 2 (38). P. 13–19. (In Russian). DOI: 10.20295/2413-2527-2024-238-13-19

Abstract. *The article presents algorithms for assessing the quality indicators of distributed data storage systems used to accumulate and process data from automated information systems for various purposes, serving a large number of geographically distributed clients. Mathematical expressions for indicators of confidentiality, availability and cost of data storage are given. The proposed algorithms make it possible to analyze the distribution of information resources among the elements of the data storage system in order to select rational solutions for organizing the storage of confidential information.*

Keywords: *distributed data storage system, data confidentiality.*

Introduction

Modern automated information systems that serve a large number of users contain data storage systems (DSS), the quality of which is subject to high requirements [1–3].

In order to analyze the values of quality indicators of the functioning of storage systems, a methodological apparatus is needed for assessing the following indicators of storing data located in storage systems: confidentiality, availability and cost.

A feature of the proposed approach is the use of quantitative estimates of the indicators under consideration.

The article discusses mathematical expressions that make it possible to quantify the values of the listed indicators and proposes algorithms for their calculation.

Quality indicators of a distributed data storage system

A distributed information system (RIS), including client automated workstations (AWS) and data processing centers (DPCs), united by a global computer network (fig. 1), allows solving applied problems based on the collection, storage, processing and transmission of target information [4–7].

We will assume that the operating technology of such a RIS involves collecting information from client workstations and processing it in one or more data centers. In this case, information is accumulated in a database (DB) and stored in a data center.

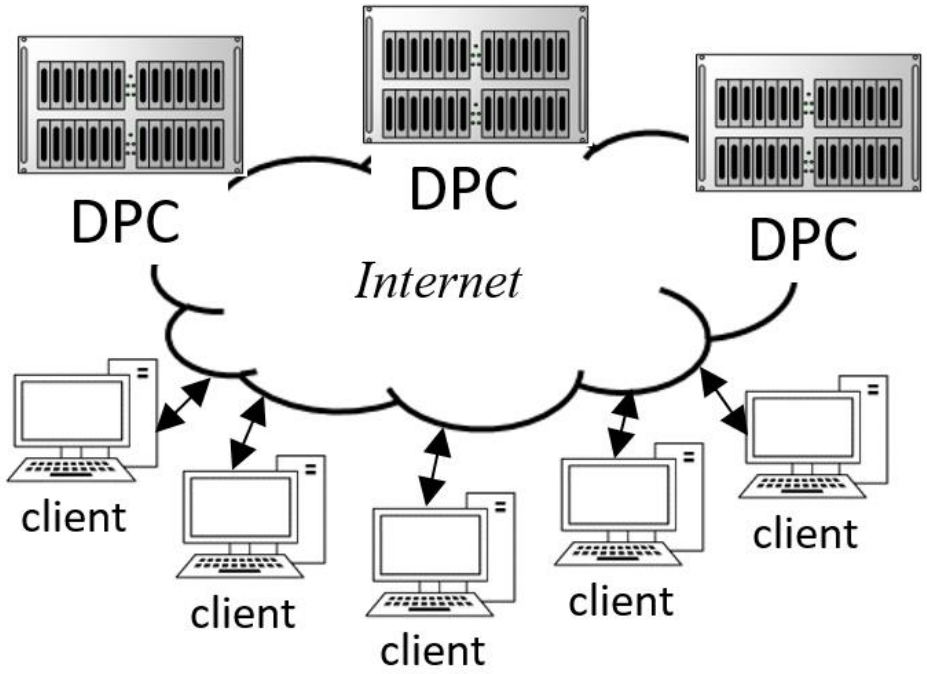

Fig. 1. Distributed information system

Information received from clients is stored in relational databases with a table structure (fig. 2). The database consists of n records or their blocks, each of which contains service information from one RIS client.

Records	Recording storage location			
	α_1	α_2	...	α_m
r_1	0	1		0
r_2	1	0		0
...				
r_n	1	0		1

Fig. 2. Structure of stored information in the database

In order to increase reliability, database records can be stored on several elements of a data storage system, which can have a different architecture and represent a data storage of various capacities, accessed by users via the global Internet. At the same time, storing data on each storage system element has a certain cost, depending on the volume of stored information and the quality of the services provided for its storage (reliability, confidentiality, communication channel capacity, etc.).

In normal mode, RIS uses a database located in the data center, and the remaining data centers are used for backup information storage.

Each data center can be subject to destructive effects, both external (cyber-attack, terrorist attack, natural disaster, etc.) and internal (breakdown, failure, unauthorized access to data), as a result of which the integrity, availability and confidentiality of stored information can be violated (or parts thereof), which leads to denial of service to the RIS.

In order to meet the requirements for the quality of information storage, database records can be distributed among storage elements (data center and workstation) in such a way as to ensure restoration of access to data in the event of a data center service failure within a given time, subject to restrictions on the cost and confidentiality of their storage.

Data storage confidentiality indicator

Let a rectangular matrix be given $H_{[n,m]}$, each element h_{ij} of which is equal to 1 if the i-th record, $1 \le i \le n$, is placed on the j-th element of the storage system, $1 \le j \le m$ and is equal to 0 otherwise.

Let also be given a vector $C_{\langle m \rangle} = \langle c_1, c_2, ..., c_m \rangle$, each element c_j of which is the probability (*guarantee level*)

of ensuring confidentiality of data storage on the *j*-th element of the storage system.

Let's consider two options for ensuring the confidentiality of data storage, each of which is advisable to use in appropriate conditions [8–12].

Option 1. The confidentiality requirement applies to each record separately.

Data confidentiality indicator — the minimum level of confidentiality *L* for all database records will be

$$L = \min_i \left(\prod_{j=1}^{m} \beta\left(h_{ij}, c_j\right) \right), i \in [1, n],$$

where $\beta(x,y) = \begin{cases} 1 \text{ at } x = 0; \\ 1 \text{ at } x = 1 \end{cases}$

Option 2. The confidentiality requirement applies to the entirety of the records.

Data confidentiality indicator — average value of information storage confidentiality:

$$L = \frac{1}{2^m} \sum_{k=1}^{2^m} \sum_{i=1}^{n} \prod_{j=1}^{m} \beta\left(h_{ij} \cdot comb_{jk}, c_j\right),$$

where $comb_{jk}$ is the value of the *j*-t*h* digit (0 or 1) in the binary representation (combination) of the number *k*.

Data storage cost indicator

The cost of data storage is determined by the total cost of storing it on all storage elements. If the cost of storage is calculated through the price $Z_{\langle m \rangle} = \langle z_1, z_2, ..., z_m \rangle$ for storing a unit of data volume on the *j*-th storage element, then the total cost of storing all records on all storage systems will be:

$$S = \sum_{j=1}^{m} \left(z_j \sum_{i=1}^{n} h_{ij} \cdot v_i \right),$$

where z_j is the cost of storing a unit of volume (GB, TB) of data on the *j*-th storage system;

v_j — vector component;

$V_n = v_1, ..., v_n$ — volume of the *i*-th record.

Data availability indicator

The availability of data depends on the readiness of the storage system elements to provide the databases stored in it. This indicator can be estimated based on the availability factors *g* of the corresponding storage elements [13–15]:

$$g = \frac{t_p}{t_p + t_в},$$

t_p — the time of regular operation of the RIS, during which the requested database records must be provided with the established efficiency;

$t_в$ — time during which the requested database records are unavailable.

If the availability coefficient of the *j*-th storage element is g_j, then the minimum level of availability *G* for all data records will be

$$G = \min_i \left(1 - \prod_{j=1}^{m} \beta\left(h_{ij}, 1 - g_j\right) \right), i \in [1, n].$$

Obviously, this indicator depends on the placement of database records, the readiness and state of storage elements (serviceable/faulty).

To increase data availability, storage elements are combined into clusters, duplicating information on each element of such a cluster. However, duplication of information increases the risk of unauthorized access to it, which leads to a decrease in the confidentiality of data storage.

Thus, the optimization problem arises of placing information on storage elements in order to ensure its maximum confidentiality under restrictions on the availability and cost of data storage.

To ensure the adequacy of the solution to this problem, algorithms are proposed for assessing the confidentiality, availability and cost of data storage, the use of which will make it possible to select a rational configuration for placing information on storage elements.

The following describes the step-by-step operation of algorithms for assessing quality indicators of storage systems, common to which are the following notations: matrix $H_{[n,m]}$ — matrix for placing records on storage systems; *n* — number of database records (blocks); *m* — number of storage elements.

Algorithm for assessing information availability in a distributed data storage system

In the presented algorithm g_i, the availability coefficient of the i-th storage element.

Algorithm 1.

Step 1. Start.

Step 2. i :=1 , G :=1.

Step 3. j :=1, d :=1.

Step 4. If $h_{ij} = 1$, then $d := d \cdot (1 - g_i)$.

Step 5. $j := j + 1$.

Step 6. If $j > N$, then go to step 7, otherwise — to step 4.

Step 7. $G := min(G,1 - d)$.

Step 8. $i := i + 1$.

Step 9. If $i > m$, then go to step 10, otherwise — to step 3.

Step 10. End.

As a result of the algorithm's operation, the variable G will have the value of the minimum level of data recording availability. By comparing this value with the required one, it is possible to assess the suitability of data distribution among storage elements from the point of view of information availability requirements. The computational complexity of the algorithm is $O(mn)$.

Algorithm for estimating the cost of information storage in a distributed data storage system

In the presented algorithm z_j, is the price of storing a unit of data volume on the j-th storage element, v_i and is the volume of the i-th data block.

Algorithm 2.

Step 1. Start.

Step 2. $i := 0$, $S := 0$.

Step 3. $j := 1$.

Step 4. If $h_{ij} = 1$, then $S := S + z_j \cdot v_i$.

Step 5. $j := j + 1$.

Step 6. If $j > N$, then go to step 7, otherwise — to step 4.

Step 7. $i := i + 1$.

Step 8. If $i > m$, then go to step 10, otherwise — to step 3.

Step 9. Finish.

At the end of the algorithm, the variable will have the value of the cost of data storage. By comparing this value with the required one, it is possible to assess the suitability of data distribution among storage elements from the point of view of requirements for the cost of information storage. The computational complexity of the algorithm is $O(mn)$.

Algorithms for assessing information confidentiality in a distributed data storage system

In the algorithms under consideration c_i, the level of confidentiality of information storage on the i-th storage element.

As a result of the operation of the algorithms, the variable C will have a value of a certain level of data storage confidentiality. Comparing this value with the required one, it is possible to evaluate the value of the confidentiality of data distributed among storage elements in terms of information confidentiality requirements.

Algorithm 3. Estimation of the minimum level of data confidentiality.

Step 1. Start.

Step 2. $i := 1$, $C := 1$.

Step 3. $j := 1$, $d := 1$.

Step 4. If $h_{ij} = 1$, then $d := d \cdot c_i$.

Step 5. $j := j + 1$.

Step 6. If $j > N$, then go to step 7, otherwise — to step 4.

Step 7. $C := min(C,d)$.

Step 8. $i := i + 1$.

Step 9. If $i > m$, then go to step 10, otherwise — to step 3.

Step 10. Finish.

Thus, the variable C will have the value *of the minimum* level of confidentiality of data storage. The computational complexity of the algorithm is $O(m^2)$.

Algorithm 4. Estimating the average level of information confidentiality (brute force).

Step 1. Start.

Step 2. $i := 1$, $C := 0$.

Step 3. $j := 1$, $u := 0$.

Step 4. $k := 1$, $d := 1$.

Step 5. If $h_{ik} = 1 \cap comb(i,k - 1) = 1$, then $d := d \cdot c_i$.

Step 6. $k := k + 1$.

Step 7. If $k > m$, then go to step 8, otherwise — to step 5.

Step 8. $u := u + d$.

Step 9. $j := j + 1$.

Step 10. If $j > N$, then go to step 11, otherwise — to step 4.

Step 11. $C := C + u$.

Step 12. $i := i + 1$.

Step 13. If $i > 2^m - 1$, then go to step 14, otherwise — to step 3.

Step 14. If C := $C / \left(n\left(2^m - 1\right)\right)$.

Step 15. End.

In the presented algorithm, the value of the function β is equal to the value β of the digit $\beta \in \{0,1\}$, in the binary code of the integer α, $0 \le \alpha \le 2^m - 1$.

As a result of the algorithm's operation, the variable *C* will have the value *of the average* level of confidentiality of data storage.

It should be noted that the algorithm has high computational complexity $O(2^m)$ and is applicable for relatively small values of the number of storage elements.

Since in practice there is a need to assess the confidentiality of data storage on a significant number of storage elements, it is advisable to use an approximate approach to assessment based on "greedy algorithms". The algorithm proposed below, which belongs to this type, implements a strategy of "greedy" data collection on storage elements with the goal of unauthorized collection of the maximum amount of information at minimal cost.

Algorithm 5. Evaluating the confidentiality of data storage (greedy algorithm).

Step 1. Start.

Step 2. Define the vector $F = f_1, f_2, ..., f_m, f_i \in \{0,1\}$.

Step 3. $F := 0$, $C := 0$, $k := 1$.

Step 4. $i := 1$, $d := 0$.

Step 5. If $f_i = 1$, then go to step 13, otherwise — to step 6.

Step 6. $j := 1$, $u := 0$.

Step 7. $u := u + H(j,i)$.

Step 8. $j := j + 1$.

Step 9. If $j > N$, then go to step 10, otherwise — to step 7.

Step 10. If $u \cdot (1 - c_i) \ge d$, then go to step 11, otherwise — to step 13.

Step 11. $d := u \cdot (1 - c_i)$, $l := i$.

Step 12. $f_l := 1$, $c := c + d$.

Step 13. $i := i + 1$.

Step 14. If i > m, then go to step 15, otherwise — to step 5.

Step 15. If $h_{jl} = 1$, then $h_{jr} = 0$ $\forall r = 1,...,m$.

Step 16. $k := k + 1$.

Step 17. *If* $k > m$, then go to step 18, otherwise — to step 4.

Step 18. End.

As a result of the algorithm's operation, the variable *C* will have the value of *the pessimistic (minimum)* level of data storage confidentiality. The computational complexity of the algorithm is $O(m^2)$.

Thus, the presented algorithms provide estimates of data storage quality indicators.

Conclusion

The given algorithms have a practical focus on studying the effectiveness of using storage systems for collecting, storing and processing confidential information.

The main problem of using algorithms is obtaining initial data, namely, estimates of the values of indicators of the availability of storage elements and the level of their provision of confidentiality of stored information. Availability estimates can be obtained based on software agents that operate on each storage element and keep track of the active location of the corresponding storage element on the network. Obtaining confidentiality estimates can be found using fuzzy data analysis models [16], taking into account many parameters of their storage on storage elements.

REFERENCES

1. Suhoroslov O. V. Novye tekhnologii raspredelennogo hraneniya i obrabotki bol'shih massivov dannyh. M.: Institut sistemnogo analiza RAN, 2021. 40 s. (In Russian)

2. Radchenko G. I. Raspredelennye vychislitel'nye sistemy. Chelyabinsk: Fotohudozhnik, 2012. 184 s. (In Russian)

3. Dokuchaev V. A., Kal'fa A. A., Maklachkova V. V. Arhitektura centrov obrabotki dannyh. M.: Goryachaya liniya — Telekom, 2023. 240 s. (In Russian)

4. Cil'ker B. Ya., Orlov S. A. Organizaciya EVM i sistem: uchebnik dlya vuzov. 2-e izd. SPb.: Piter, 2011. 668 s. (In Russian)

5. Melekhin V. F., Pavlovskij E. G. Vychislitel'nye mashiny, sistemy i seti. M.: Akademiya, 2007. 560 s. (In Russian)

6. Horoshevskij V. G. Arhitektura vychislitel'nyh sistem: ucheb. posobie. 2-e izd. M.: MGTU im. N. E. Baumana, 2008. 520 s. (In Russian)

7. Kosyakov M. S. Vvedenie v raspredelennye vychisleniya. SPb.: NIU ITMO, 2014. 115 s. (In Russian)

8. Amelin R. V. Informacionnaya bezopasnost'. M.: Yurist", 2010. S. 121. (In Russian)

9. Danilov A. D. Cennost' informacii. Tekhnologii informacionnogo obshchestva. 2011. № 10. S. 137–140. (In Russian)

10. Stepanov E. A., Korneev I. K. Informacionnaya bezopasnost' i zashchita informacii: uchebnoe posobie. M.: Infra-M. 2010. 336 s. (In Russian)

11. Fat'yanov A. A. Problemy zashchity konfidencial'noj informacii, ne sostavlyayushchej gosudarstvennuyu tajnu. Informacionnoe obshchestvo. 2010. № 5. S. 49–56. (In Russian)

12. Informacionnaya bezopasnost'. M.: Oruzhie i tekhnologii, 2011. S. 35–44. (In Russian)

13. Rahman P. A. Koefficient gotovnosti sistemy obrabotki dannyh s osnovnym i rezervnym uzlami. Mezhdunarodnyj zhurnal prikladnyh i fundamental'nyh issledovanij. 2015. № 9. S. 608–611. (In Russian)

14. Polovko A. M., Gurov S. V. Osnovy teorii nadezhnosti. 2-e izd., pererab. i dop. SPb.: BHV-Peterburg, 2006. 704 s. (In Russian)

15. Shishmarev V. Yu. Nadezhnost' tekhnicheskih sistem: uchebnik dlya stud. vyssh. ucheb. zavedenij. M.: Izdatel'skij centr “Akademiya”, 2010. 304 s. (In Russian)

16. Bronevich A. G., Lepskij A. E. Nechetkie modeli analiza dannyh i prinyatiya reshenij: uchebnoe posobie. Nac. issled. un-t “Vysshaya shkola ekonomiki”. M.: Izd. dom Vysshej shkoly ekonomiki, 2022. 264 s. (In Russian)

Received: 23.04.2024
Accepted: 20.05.2024

Алгоритмы оценивания показателей качества функционирования распределенной системы хранения конфиденциальных данных

Басыров Александр Геннадьевич[1] — доктор технических наук, профессор. Профессор кафедры Военно-космической академии имени А. Ф. Можайского. E-mail: Vka_24kaf@mil.ru

Кошель Игорь Николаевич[2] — канд. техн. наук. Начальник факультета Военно-космической академии имени А. Ф. Можайского. E-mail: kin1470@mail.ru

Абраменков Валерий Валерьевич[1] — начальник ФГАУ «Центральное управление жилищно-социальной инфраструктуры (комплекса)». E-mail: info@fgaul.ru

[1] Военно-космическая академия имени А. Ф. Можайского, Россия, Санкт-Петербург

[2] ФГАУ «Центральное управление жилищно-социальной инфраструктуры», Россия, Москва

Для цитирования: Басыров А. Г., Кошель И. Н., Абраменков В. В. Алгоритмы оценивания показателей качества функционирования распределенной системы хранения конфиденциальных данных // Интеллектуальные технологии на транспорте. 2024. № 2 (38). С. 13–19. DOI: 10.20295/2413-2527-2024-238-13-19

Аннотация. *В статье представлены алгоритмы оценивания показателей качества распределенных систем хранения данных, применяемых для накопления и обработки данных автоматизированных информационных систем различного назначения, обслуживающих большое количество территориально распределенных клиентов. Приведены математические выражения показателей конфиденциальности, доступности и стоимости хранения данных. Предложенные алгоритмы позволяют проанализировать распределение информационных ресурсов по элементам системы хранения данных для выбора рациональных решений по организации хранения конфиденциальной информации.*

Ключевые слова: *распределенная система хранения данных, конфиденциальность данных.*

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Сухорослов О. В. Новые технологии распределенного хранения и обработки больших массивов данных. М.: Институт системного анализа РАН, 2021. 40 с.

2. Радченко Г. И. Распределенные вычислительные системы. Челябинск: Фотохудожник, 2012. 184 с.

3. Докучаев В. А., Кальфа А. А., Маклачкова В. В. Архитектура центров обработки данных. М.: Горячая линия — Телеком, 2023. 240 с.

4. Цилькер Б. Я., Орлов С. А. Организация ЭВМ и систем: учебник для вузов. 2-е изд. СПб.: Питер, 2011. 668 с.

5. Мелехин В. Ф., Павловский Е. Г. Вычислительные машины, системы и сети. М.: Академия, 2007. 560 с.

6. Хорошевский В. Г. Архитектура вычислительных систем: учеб. пособие. 2-е изд. М.: МГТУ им. Н. Э. Баумана, 2008. 520 с.

7. Косяков М. С. Введение в распределенные вычисления. СПб.: НИУ ИТМО, 2014. 115 с.

8. Амелин Р. В. Информационная безопасность. М.: Юристъ, 2010. С. 121.

9. Данилов А. Д. Ценность информации. Технологии информационного общества. 2011. № 10. С. 137–140.

10. Степанов Е. А., Корнеев И. К. Информационная безопасность и защита информации: учебное пособие. М.: Инфра-М. 2010. 336 с.

11. Фатьянов А. А. Проблемы защиты конфиденциальной информации, не составляющей государственную тайну. Информационное общество. 2010. № 5. С. 49–56.

12. Информационная безопасность. М.: Оружие и технологии, 2011. С. 35–44.

13. Рахман П. А. Коэффициент готовности системы обработки данных с основным и резервным узлами. Международный журнал прикладных и фундаментальных исследований. 2015. № 9. С. 608–611.

14. Половко А. М., Гуров С. В. Основы теории надежности. 2-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2006. 704 с.

15. Шишмарев В. Ю. Надежность технических систем: учебник для студ. высш. учеб. заведений. М.: Издательский центр «Академия», 2010. 304 с.

16. Броневич А. Г., Лепский А. Е. Нечеткие модели анализа данных и принятия решений: учебное пособие. Нац. исслед. ун-т «Высшая школа экономики». М.: Изд. дом Высшей школы экономики, 2022. 264 с.

Дата поступления: 23.04.2024
Решение о публикации: 20.05.2024