Ensuring availability of digital education platforms’ databases

I N Selyutin¹, A M Kadnova², O I Bokova³, N S Khokhlov² and R V Belyaev⁴

¹ Voronezh Institute of Government Communications (branch) of the Academy of the Federal security service of Russia, 2, Minskaya str., Voronezh, 394042, Russian Federation
² Voronezh Institute of the Ministry of Internal Affairs of the Russian Federation, 53, Patriots ave., Voronezh, 394065, Russian Federation
³ Ltd Kaskad, 2, Ferganskaya str., Moscow, 109444, Russian Federation
⁴ Air Force Academy named after Professor N.E. Zhukovsky and Yu. A. Gagarin, 53 a, Starykh Bolshevikov str., Voronezh, 394064, Russian Federation

E-mail: aizhana_kadnova@mail.ru

Abstract. The article offers a procedural model for optimization of making queries against the digital education platform under the conditions of negative information impacts. The model is developed on the basis of the formalized description of the process of database query processing. The model allows you to study the influence of negative information impacts and of the structure of a digital education platform on the availability indicators of its database. A model for determining plans of making queries against a digital education platform database is offered. On the basis of the genetic algorithm, a model for optimizing the allocation of database areas of a digital education platform is constructed. The effectiveness of the proposed models is confirmed by the results of a computational experiment, during which a fragment of a digital education platform was deployed. The probabilistic and temporal characteristics of database queries processing are presented. Assessment of modeling accuracy was carried out by comparing the data obtained experimentally with the data obtained during direct-analysis simulation. The developed models for optimizing the allocation of pieces of data and database query execution plans can be used to improve not only existing database of digital education platforms but also the ones under development.

1. Introduction
At the present stage of information technology development, the distributed processing of information is becoming more widespread [1].

To ensure the operational reliability, the digital education platforms (DEP) must have means of protection against the negative information impacts (NII) and of availability control of DEP databases (DB) [2-4,5-7].

The DB availability depends on the performance of the information system on which the given DB is deployed, therefore, the incorrect choice of elements of the information system and the misallocation of DBs in it can also lead to a violation of its availability.

When building distributed information systems, the optimization of the number and location of data storage and processing centers, as well as of data storage system components is required. This task reduces itself to the problem of linear integer programming with Booleans, and for its solution it is
possible to use the branch and bound method [8]. In order to determine a balanced spending level for the formation of a data storage and allocation system, methods of the theory of welfare (Pareto optimality principle) and of the probability theory are applied [9, 10].

2. Materials and methods
With increase of the dimension of problems, the time of solution by exact methods of discrete optimization increases exponentially, therefore, when optimizing the structure of distributed information systems, the NaturalComputing method is used in work, which is based on the principle of using natural decision-making mechanisms and proposes the use of genetic algorithms [11-17]. For application, a choice of genetic operators and of a coding scheme for the parameters of a distributed information system in «chromosomes» [18] is necessary.

For optimization in large systems with strict time limits for decision-making, the genetic algorithms ensure the reduction of time costs in comparison with the branch and bound method.

3. Results and arguments
The initial data for solving the problem of ensuring the DEP DB availability are:

- the set of management tasks specified for DEP solution, and the corresponding set of queries and their characteristics;
- characteristics of negative information impacts.

Based on the source data, upon the restrictions imposed on the structure of the information system, it is required to maximize the effectiveness function of making queries against the DEP DB:

\[ P(t_{imp}, J, t_{add}) \overset{\otimes}{\text{max}} \]

where \( P(t_{imp}, J, t_{add}) \) is the probability that the request will be implemented in time \( t_{imp} \), not exceeding the specified \( t_{add} \) under the given restrictions:

\[ V_i \leq V_{i,\text{max}}, i = \overline{1, I}, m_j \geq N_j, j = \overline{1, J} \]

where \( V_i \) is the total amount of pieces of data stored on the \( i \)-th data storage node; \( m_j \) the number of copies of pieces of data of the \( j \) type; \( N_j \) is the replication coefficient for the \( j \)-type data pieces.

To study the DB availability, the DEP DB functional model was used, that takes into account the negative information impacts [19].

Considering the determinacy of the structure of the information system, based on which the DEP is deployed, the model takes into account the time required to make a query of data search in the DB:

\[ T_{\text{que}} = Q_i + \max(2R_i + T_i) \]

where \( Q_i \) is the time of subquery transit through the information system during the request shipping to the collector node; \( T_i \) is implementation time of the \( i \) subquery as a part of the request, \( i = \overline{1, I} \); \( I \) is the number of subqueries into which the source query was divided; \( R_i \) is the data transfer time between the storage nodes.

The processes of implementation of the queries to the DEP DBs can be represented as servicing the flow of applications within a stochastic queuing network (QN).

As a model of request processing the model M/G/1 describing network queuing systems (NQS) with a Poisson stream of requests and general-independent service time was chosen. For this type of NQS, the time spent in the queue on the \( i \)-th switching element is defined by the Pollaczek–Khinchine formula:
\[ Q_i = \frac{\lambda_i \cdot t_{\text{serv}}^2}{2(1 - \rho)} \]

where \( \lambda_i \) is the flow rate of the \( i \) queries; \( t_{\text{serv}} \) is the average service time; \( \rho \) is the bandwidth of the switching element.

To study the model presented as a queuing network, it is possible to apply a simulation or analytical research method.

The proposed model for optimizing query execution plans (figure 1) reduces itself to setting the optimal DB query execution plan, which is characterized by a high degree of data fragment (DF) duplication, and assumes the use of the developed DEP DB functional model [19].

![Figure 1. A procedural model for determining the DB queries execution plans.](image)

The criteria for choosing the query execution plan are the maximum probability of timely query execution and the minimum average time of its execution.

The essence of the proposed approach is as follows. The structure of the information system is known and described by the graph \( G = (A,W) \), for which \( A \) is the set of nodes, \( W = \{w_{ij}\} \) is the set of arcs \((i,j \in A,i \neq j)\). Elements of the sets \( S \) – data storage nodes, \( G \) – data transfer path between nodes, \( I \) – query sources correspond to graph nodes. The arcs correspond to the connections between nodes. Also the DB query flow \( Z = \{K,W_w,W_K\} \), is known, as well as the option of database structure, it is described by the \( M_{\text{pod}} \) matrix, where \( W_w \) is the subset of DB queries for data retrieval; \( W_K \) is a subset of DB queries for data correction; \( K \) is a set of categories of urgency of DB queries; \( M_{\text{pod}} \) is the matrix for DB DF s allocation within the nodes.
The query flows generated by the sources are distributed according to the Poisson law with intensities $l_i, l_{OI}, k_{OW} = W_k W_{OB}$. Any request can be performed according to one of the $N_K$ schemes. These execution schemes are defined by the oriented graphs $S_{K_n}(M_{K_n}, O_K)$, where $M_{K_n}, n = 1, \overline{N_K}$ is a subset of the $M_{POD}$ set, used by the $k$-th query within the $n$-th execution scheme, $O_K$ is the matrix of the data volumes transferred between the nodes during execution of $k$ requests.

All DFs used by the $k$-th request are included in the set $M_K = M_{K_1} M_{K_2} M_{K_3} \ldots M_{K_n}$, while the number of schemes for this request implementation $N_K$ is defined as the product of the number of copies of each of the DFs.

To solve the problem, an analytical model of the DEP DB functioning is used [19]. The model is described in terms of queuing theory, it allows us to calculate the average implementation time of the $k$-th request with the $n$-th execution scheme $T_{K_n}(n = 1, N_K)$, as well as the probability of timely execution of the request.

The problem variables determine the query execution scheme and are determined by the matrix $X_K = \|x_{ij}\|$, where $x_{ij} = 1$, if the $i$-th DF on the $j$-th data storage node is involved into the implementation of the $k$-th request, $i O M_{K_j} j O S$, in the other case $x_{ij} = 0$.

The problem is solved as follows. For the analytical model, the variant of DB allocation in the information system is defined. Using the DB DF allocation matrix ($M_{POD}$) for the $k$-th query, $k_{OW}$, and the directed graphs $S_{K_n}(M_{K_n}, O_K)$, $n = 1, \overline{N_K}$ the set $M_K$ is calculated, as well as the option of the $X_K$ execution scheme for this request is generated. Then, by using the model the query is executed according to the previously generated scheme, the current values of $T_{K_n}$ and $P_K(T_{K_n}, J_{don})$ are calculated. Then, the next implementation scheme for the $k$-th request is defined, and then the calculation is performed again by using the analytical model.

It is necessary for each $k$-th query to calculate such a $X_K$ matrix that would ensure the objective function to reach its maximum:

$$P = \sum_{k \in \Omega k} \sum_{i \in I} P_K(T_{K_n} \leq T_{don}) \lambda_{ik} I \sum_{k \in \Omega k} \sum_{i \in I} \lambda_{ik} \rightarrow \text{max}.$$

The analysis of the dependence of the query execution time has shown that minimizing the information exchange between each pair of nodes of information storage relative to the pieces of data located on them ensures a reduction of the execution time of DB queries:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} V_{ij} \rightarrow \text{min}$$

where $V_{ij}$ is the amount of data transmitted from the $i$-th node of information storage to the $j$-th one.

The choice of the optimal variant of FD allocation is a nondeterministic polynomial time complete problem; therefore, the use of a genetic algorithm (GA) in work is proposed in order to search its solution (figure 2).

The initial data for the FDs allocation algorithm is the initial allocation of the FDs on the information storage nodes, the restrictions to the minimum number of copies of each FD and the maximum data amount on each node.
For the GA function as part of the algorithm, a set of FDs on each node is represented as a chromosome. After selecting a set of parental chromosomes on the basis of ranking by the maximum number of FDs in the chromosome, they are subjected to the N-point crossing-over and mutation procedures, while for each generation, a verification is carried out in accordance with the restrictions, and a model of query execution plans determination is applied to a given generation. If the generation is successful, the above-described procedures are repeated towards it, if not, a new selection of the parental pair from the previous generation is made.

The proposed models were tested as a part of software package for ensuring the DEP DB availability.

4. Conclusion
Procedural models of defining the DEP DB query execution plans and allocating the DB pieces of data have been developed, that allow providing the necessary analytical information to the DEP administrator to make a reasonable decision on ensuring the functional availability.

A computational experiment was carried out using the developed software package. The application of the developed procedural models allowed us to reduce the query execution time and to increase availability by 18-23%.

The developed models of defining the DEP DB query execution plans and allocating the DB pieces of data mentioned in the article can be further used to develop proposals for improving both existing and required DEP DBs in order to reduce the query execution time. The results can be used, for example, when designing software for the following tasks: big data analytics for information security risk estimation for cloud infrastructure [22].

References
[1] Foster I and Kesselman C 2004 The Grid 2: Blueprint for a New Computing Infrastructure (USA: Morgan Kaufmann Publishers) p 748
[2] Foster I, Kesselman C, Tsudik G and Tuecke S 1998 A security architecture for computational grids Proc. of ACM Conf. on Computers and Security (ACM) 83-91
[3] Selyutin I N, Drovnikova I G and Korobkin D I 2015 About the requirements to performance
indicators of information security tools of automated systems Technosphere safety technologies 4 299-305

[4] Kotenko I V, Saenko I B and Polubelova O V 2013 Advanced data storage systems for information security monitoring and control SPIIRAS Proceedings 25 113-34

[5] Atanasov I, Pencheva E, Nametkov A and Trifonov V 2019 On Functionality of Policy Control at the Network Edge International Journal on Information Technologies and Security 3(11) 3-24

[6] Tulloch M 2003 Microsoft Encyclopedia of Security (Washington: Microsoft Press) p 414

[7] Mishra N 2014 Security issues in grid computing International Journal on Computational Sciences & Applications 4 53-64

[8] Yesikov D O 2015 Tasks of ensuring the functional stability of distributed information systems Software products and systems 4 133-41

[9] Yesikov D O, Ivutin A N, Larkin E V and Kotov V V 2017 Multi-agent approach for distributed information systems reliability prediction Procedia Computer Science 103 416-20

[10] Yesikov D O, Akinshin R N, Abramov P I and Lutina L E Mathematical models for designing a subsystem of information safety ensuring in distributed information systems Scientific Bulletin of MSTU GA 20 161-70

[11] Ahmad I 2002 Evolutionary algorithms for allocating data in distributed database systems Distributed and Parallel Databases 11 5-32

[12] Zemlyanskaya S Yu Formation of the optimal configuration of a distributed information system of railway transport using evolutionary algorithms Collection of scientific papers DONIZHT 44 29-41

[13] Yanyushkin V V 2010 Models and algorithms of data allocation optimization News of Higher Educational Institutions North Caucasian Region 2 10-16

[14] Ivutin A N and Yesikov D O 2015 Evaluation of the effectiveness of the adaptive reproduction scheme in the island genetic algorithm for solving the problems of ensuring the functional stability of distributed information systems Izvestiya TulGU 9 119-28

[15] Ivutin A N and Yesikov D O 2016 Experimental evaluation of the parameters of the island genetic algorithm for solving the problems of ensuring the functional stability of distributed information systems IzvestiyaTulGU 2 99–104

[16] Khomonenko A D and Yakovlev E L 2015 Neural network approximation of the characteristics of multichannel non-Markov queue systems SPIIRAS Proceedings 4 81-93

[17] Shalyto A A, Tsarev F N and Egorov K V 2010 The joint use of genetic programming and verification for the generation of finite state machines for controlling systems with complex behavior SPIIRAS Proceedings 4 123-35

[18] Skobtsov Yu 2008 Fundamentals of evolutionary computing (Donetsk: DonNTU) p 326

[19] Selyutin I N, Meshcheryakova T V and Nikulina E A 2016 Modeling the elements interaction process of the security subsystem information infrastructure of distributed information and computing systems Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia 4 132-7

[20] Tsaregorodtsev A V, Kravets O Ja, Choropov O N and Zelenina A N 2018 Information Security Risk Estimation for Cloud Infrastructure International Journal on Information Technologies and Security 4(10) 67-76