ALGORITHM FOR SOLVING PROBLEM SYNTHESIS THE OPTIMAL LOGICAL STRUCTURE DISTRIBUTED DATA IN ARCHITECTURE OF GRID SERVICE

Nurmatova E.V.[1], Gusev V.V.[2]

[1] *MIREA — Russian Technological University, 20, Stromynka, Moscow, 107996, Russia*

[2] *National Research Center "Kurchatov Institute" - Institute for High Energy Physics, 1, Science sq, Protvino, 142281, Russia*

E-mail: [a] nurmatova@mirea.ru

The questions of constructing optimal logical structure of a distributed database (DDB) are considered. Solving these issues will make it possible to increase the speed of processing requests in DDB in comparison with a traditional database. In particular, such tasks arise for the organization of systems for processing huge amounts of information from the Large Hadron Collider. In these systems various DDB are used to store information about: the system of triggers of data collection from physical experimental installations, the geometry and the operating conditions of the detector while collecting experimental data. Two interrelated stages in the synthesis algorithm are proposed. At the first stage, the problem of distribution of database clusters between the server and clients, followed by the problem of optimal distribution of data groups of each node by types of logical records are addressed. At the second stage the problem of database localization on the nodes of the computer network is solved, in addition to the results of the first stage, the characteristics of the DDB are taken into account. Optimal logical structure of DDB will ensure the efficiency of the information system on computational resources. As a result of its solution, the local network of the DDB is decomposed into a number of clusters that have minimal information connectivity with each other. Solving the problem of synthesis of the optimal logical structure is also of great practical importance for the automated design of logical structures, for the automated formation of query specifications and adjustments of the DDB.

Keywords: data warehouse, optimal logical data structure, applications, large data volume, synthesis algorithm.

Elena Nurmatova, Victor Gusev

1. Previous work

For a grid architecture with a large number of requests, users, and large amounts of data, it is advisable to formulate the problem of synthesizing the optimal logical structure of a DDB based on the criterion of the minimum total time for implementing a set of user requests. Indeed, along with the necessary information that is really required by the user, redundant information, which arises as a result of the localization of information elements that are not required by the user in one record, is transmitted from the database server. Excessive information "clogs up" the communication channels, which in the future will require an increase in network bandwidth due to the power of hardware and software.

The variety of options for alternative solutions for the choice of not fully defined evaluation criteria, is a rather weak side of the problem of determining the optimality of the developed logical data structure in a distributed architecture.

When working with *quantitative* criteria, which include request response time, update cost, memory cost, time to create, reorganization cost, the contradiction of the criteria to each other can cause the difficulty [1].

There are optimality criteria, which are immeasurable properties, poorly represented in quantitative terms or in the form of an objective function. The *qualitative* criteria for evaluating a DB include flexibility, adaptability, availability for new users, compatibility with other systems, the ability to convert for use on another computing platform, the possibility of recovery, the possibility of fragmentation and expansion of the structure.

It is expedient to consider the synthesis of the logical structure of the DDB as a sequential solution of three particular problems, the results of which are the determination of:

1) optimal localization of data groups, providing a minimum of total traffic in the system and satisfying the specified restrictions;
2) the structure of the optimal distribution of groups by types of logical records, providing a minimum of the total time of local processing of information on the servers of network nodes under the given restrictions;
3) the structure of the database localization by the system nodes, providing the minimum value of the total time of access to the localization and DB processing nodes.

The solution of the problem of data processing in real time requires a special organization of the logical and physical structure of the DDB to ensure the response time of the system on the order of 1-2 seconds or less (the minimum implementation time for operational requests).

2. Stages of an approximate algorithm for synthesizing the optimal logical structure of the DDB

Consider an approximate algorithm for solving the problem of synthesizing the optimal logical structure of the DDB and the structure of localization of the database according to the criterion of the minimum total time for the implementation of a set of user requests , consisting of a sequence of stages (figure 1).

Stage 1. At this stage, the localization of data groups in the computing system is determined by the criterion of the minimum total traffic. To solve this problem, an approximate algorithm for distributing DDB clusters between the server and clients of the local network is used. At the *first step* of the stage, the graph of the canonical structure of the DDB is reduced to a disconnected graph with the calculation of the "weight" of each data group. The weight of each group consists of the weight of the data group itself and the weight of the arcs, taking into account the requirements of users:

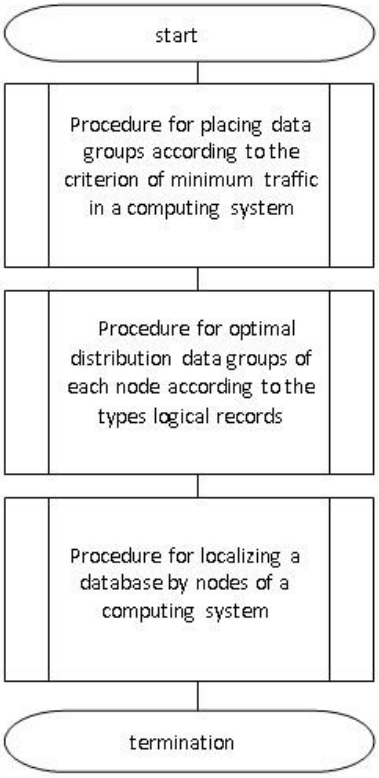

Figure 1. The main stages synthesis algorithm

$$V_i = V_i^{\text{гр}} + V_{ii'}^{\text{св}}$$

where, $V_i^{\text{гр}}$ — the total weight of the data group; $V_{ii'}^{\text{св}}$— the weight of the arcs of the graph of the canonical structure of the DDB.

$$V_i^{\text{гр}} = \sum_{k=1}^{k_0} \sum_{p=1}^{p_0} \xi_{kp}^{э} \varphi_{kp}^{э} \omega_{pi}$$

$$V_{ii'}^{\text{св}} = \sum_{k=1}^{k_0} \sum_{p=1}^{p_0} \xi_{kp}^{э} \varphi_{kp}^{э} \omega_{pi} \sum_{i' \neq i}^{I} \omega_{pi'} a_{ii'}^{\Gamma}$$

Then the weight of the i-th group is

$$V_i = \sum_{k=1}^{k_0} \sum_{p=1}^{p_0} \xi_{kp}^{э} \varphi_{kp}^{э} \omega_{pi} (1 + \sum_{i' \neq i}^{I} \omega_{pi'} a_{ii'}^{\Gamma})$$

where, $\xi_{kp}^{э}$— the frequency of usage of queries by users; $\varphi_{kp}^{э}$— the elements of the matrix for using queries by DDB users; ω_{pi}— the matrix for using data groups when executing queries; $a_{ii'}^{\Gamma}$— the semantic contiguity matrix of data groups.

At the *second* step of the stage, the computer network graph is transformed to a disconnected graph with the calculation of the "weight" of each node:

$$V_r = t_r + \sum_{r' \neq r}^{R_0} t_{rr'}$$

where t_r — the total average duration of data processing in the r-th node, consisting of the time of decomposition of the query into subqueries, route selection and connection establishment, etc.;

$t_{rr'}$ — the average duration of data transmission between nodes, determined based on the matrix of logical distances between the servers of the nodes of the computer network.

At the *third* step of the first stage, the matrix $V = \|v_{ir}\|$ is formed, whose elements are equal: $v_{ir} = V_i \times V_r$ for $i = \overline{1,I}; r = \overline{1,R_0}$.

At the *fourth* step of the first stage, the problem

$$\min_{\{x_{ir}\}} \sum_{i=1}^{I} \sum_{r=1}^{R_0} v_{ir} x_{ir}$$

is solved under the constraints:

— by the number of data groups, the localization of which is possible on one node

$$\sum_{i=1}^{I} x_{ir} \leq N_r, r = \overline{1,r_0}$$

— on the admissible duplication of groups by network nodes $\sum_{r=1}^{r_0} x_{ir} \leq M_i$,

$$\sum_{r=1}^{r_0} x_{ir} \leq M_i, i = \overline{1,I}$$

— on the amount of available external memory of the network servers for storing data

$$\sum_{i=1}^{I} x_{ir} \rho_i \pi_i \leq \eta_r^{\text{ВЗУ}}$$

where, ρ_i — the vector of group lengths in bytes; π_i — the vector of number of instances in groups; $\eta_r^{\text{ВЗУ}}$ — the amount of available memory on the server of the r-th host; $x_{ir} = 1$, if the i-th data group is included in the r-th network node; $x_{ir} = 0$ – otherwise.

This is a linear integer programming problem. Its solution makes it possible to determine the optimal localization of data groups by network nodes.

Stage 2. At this stage, the problems of optimal distribution of data groups of each node according to the types of logical records are solved by the criterion of the minimum total time of local data processing in each network node. The number of synthesis tasks for this stage is determined by the number of network nodes.

The initial data are subgraphs of the graph of the canonical structure of the DDB, as well as the temporal and volume characteristics of the subgraphs of the canonical structure of the DDB, the set of requests from users and network nodes [2].

The synthesis problem for this stage is solved using exact or approximate algorithms [3]. The following restrictions are used: restrictions on the number of groups in a logical record, on the one-time inclusion of groups in records, on the cost of storing information, on the required level of information security of the system, on the time of performing operational transactions on DB servers, on the total time of servicing operational requests on servers. As a result, we determine the logical structures of the DB for each node in the network.

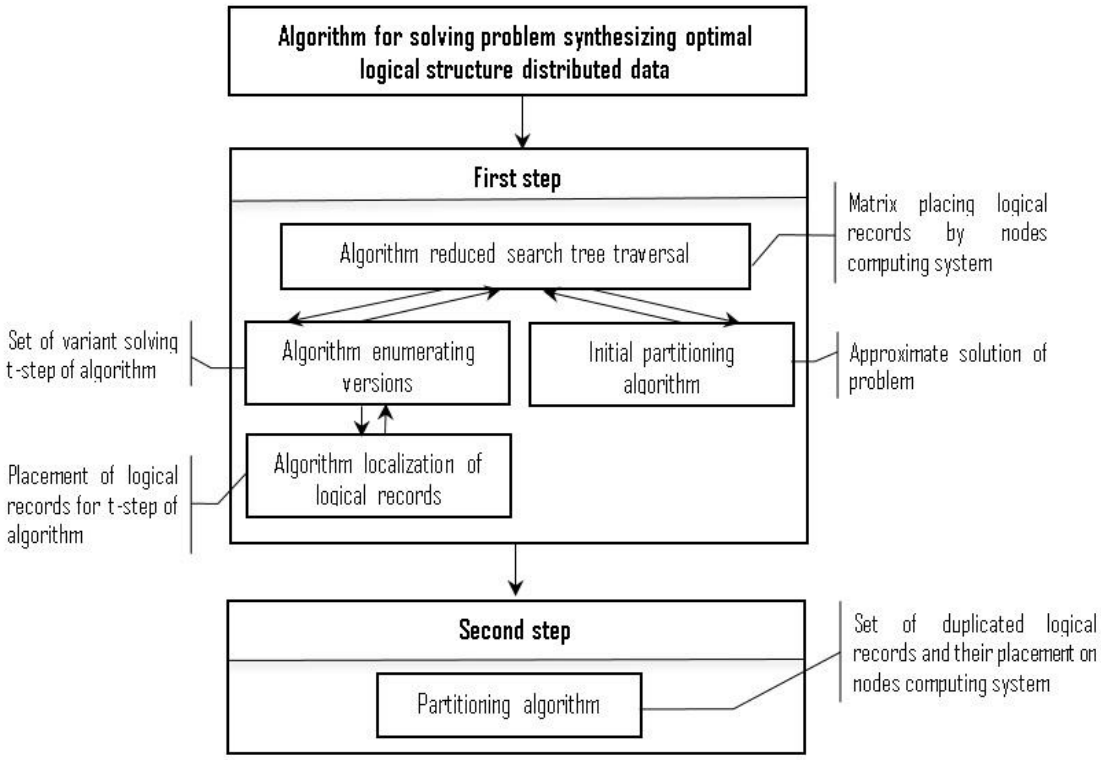

Figure 2. Summary diagram stages solving synthesis problem

Stage 3. Localization of the DB by network nodes. Initial data of the stage: results of the previous stages and characteristics of the DDB. Restrictions: on the total number of synthesized logical records located on the server of the r-th node of the computer network; on the amount of available external memory of the network servers for storing the database; on the number of copies of logical records placed on the network.

As a result of the proposed algorithm (figure 2), localization matrices of the set of data groups by the types of logical records are formed (result of stage 1) and then groups of records by network nodes (result of stage 2, table 1) are formed. The timing of the algorithms is also evaluated.

Table 1 is an example of a matrix for localizing records by network server nodes. It is indexed row by row by t-numbers of logical record groups, and by columns by r-numbers of network nodes.

Table 1. Matrix of localization of logical records by network nodes-servers

		Network node numbers, t		
		1	**2**	**3**
Logical record group numbers, r	**1**	1	0	0
	2	0	1	0
	3	0	0	1
	4	0	1	0
	5	0	1	0

3. Similar solutions

As alternative solutions for comparing the results of synthesizing the data structure according to various criteria, we analyzed the analogue that solves the NP-hard nonlinear integer discrete optimization problem from the DDB domain [2] and implements 3 neural network algorithms for synthesizing the optimal logical structure DDB according to the criterion of the minimum total time of

sequential processing of a set of user requests:

1. NS-GA-algorithm (HNN) – the evolutionary optimization algorithm based on artificial Hopfield neural networks and genetic algorithms;
2. TM-algorithm (TM) – the neural network optimization algorithm based on modified taboo search;
3. RTM algorithm (DTM) – the distributed neural network optimization algorithm based on modified taboo search.

Another analogue considers the formulation of the problem of synthesizing the optimal logical structure of the DDB in fuzzy conditions according to the criterion of the minimum total loading time of the DDB [1].

4. Conclusion and Future plans

Further work within the framework of this topic is the development of software that implements the search algorithm for a variant of the logical structure of the DDB, which ensures the optimal value of the specified criterion for the efficiency of the functioning of the grid system and satisfies the main system, network, and structural constraints.

References

[1] Kosterin E.V., Minin Yu.V., Ivanova O.G., Al-Matari N.A.Kh. Statement of the problem of synthesizing the optimal logical structure of a network database in fuzzy conditions.– Information and security, Voronezh, 2014. – 574 – 579 p.

[2] Nurmatova E.V., Gusev V.V., Kotliar V.V. Analysis of the features of the optimal logical structure of distributed databases// Collection of works the 8th International Conference “Distributed Computing and Grid-technologies in Science and Education”.— Dubna, 2018.— 167 p.

[3] Amaru L.G. New Data Structures and Algorithms for Logic Synthesis and Verification.- Springer, 2016. — 262 p. — ISBN: 9783319431734, EISBN: 9783319431741