УДК 004.9 **doi: 10.26906/SUNZ.2022.4.100**

Nina Kuchuk, Anna Shyman

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

A METHOD FOR DISTRIBUTING TRANSACTIONS TO HYBRID CLOUD DATA STORAGE

Abstract. The article discusses the current task of optimal resource allocation in CLOUD systems that support hybrid cloud data storage. The purpose of this article is to develop an optimal distribution method of several heterogeneous transactions to a hybrid cloud data storage, connected by a common bandwidth limitation. The optimization criterion will be the minimum cost. Information flows have certain points of departure and destination, are heterogeneous flows with common restrictions on the bandwidth of the communication channels used. **The results obtained.** The method is based on the construction of network graphs. The optimization problem is reduced to the distribution of flows in the network in such a way that, if the requirements of the cloud hybrid data storage are met, the cost of transmitting the flow in the network is minimal. To solve it, an iterative algorithm for constructing the maximum admissible flow is proposed. At each iteration, the simplex table of network graphs is modified. The direction of further research is the development of a method for optimizing the structure of cloud data storage.

Keywords: cloud technologies, data storage, hybridity, CLOUD system, network graph.

Introduction

In recent years, "cloud technologies" have gained enormous popularity in the field of information technologies [1, 2]. The modern IT market offers a large number of software products that support this technology (hereinafter referred to as CLOUD systems) and ensure the functioning of cloud data storage - a model of online storage in which data is stored on numerous distributed in the network, servers provided for use by clients, as well as cloud computing, which provide distributed data processing, in which the user receives computer resources and power as an Internet service [2].

Among a number of problems that arise during the implementation of CLOUD systems, the most relevant are those related to the distribution of system resources, which is caused by both the significant growth of users and the expansion of services provided by modern CLOUD systems [3, 4]. In particular, there is the task of optimal allocation of resources in CLOUD systems that support hybrid cloud data storage (HCSD). A special place among such tasks is the task of optimal distribution of various information flows with the aim of minimizing the cost of their passage.

The purpose of this article is to develop a method optimal distribution of several different transactions to the hybrid cloud data storage, connected by the general limitation of bandwidth.

The optimization criterion will be the minimum cost. Information flows have specific points of departure and destination, they are heterogeneous flows with common limitations on the bandwidth of the used communication channels (CC).

Let's consider the algorithm of its solution. Given a distributed computing network with a cloud component, which is characterized by an undirected, edge-weighted graph $G = \langle Z, Y, b, c \rangle$ bandwidths of CC and the costs of transmitting a unit of information on them and the graph $Г = <Z, W, r>$ requirements for the transfer of m information flows between nodes of a distributed computing network. Here are $Z = \{Z_1, \ldots, Z_l\}$ - renumbered set of graph vertices *G* and Г, being in isomorphism to nodes distributed computing network, $l = |Z|$ - number of graph vertices *G* and *Г*; $Y = \{y_1, \ldots, y_n\}$ - renumbered set of graph edges *G*, are in isomorphism to CC distributed computing network, $n = |Y|$ - number of graph edges *G*; $b : Y \rightarrow R_+$ - weight function that determines each edge $y_i \in Y$ throughput b_i, $1 \le i \le n$; $c : Y \rightarrow R_+$ - weight function that determines the cost of transferring a unit of flow C_i edgewise $y_i \in Y$; $W = \{W_1, \ldots, W_m\}$ - renumbered set of graph edges *Г*. Edge $W_j \in W$, $1 \le j \le m$, connects two vertices of a graph *Г*, if between corresponding nodes distributed computing network information is exchanged. The amount of information flow to be sent along the edge W_j, given by the weight function $r : W \rightarrow R_+$, defining each edge $W_j \in W$ graph *Г* required flow r_j. The number of information flows that need to be distributed in the network is equal to $m = |W|$.

Required for given graphs ***G*** and ***Г*** distribute the flows in the network in such a way that, if the requirements of the graph are satisfied, ***Г*** the cost of transmitting the flow in the network was minimal. To present the formal formulation and algorithm for solving this problem, we introduce a number of definitions.

Distribution γ flows of information transmitted to a distributed computing network defined by graphs *G* and *Г*, of information flows transmitted to a distributed computing network defined by graphs *H*(γ). Here $R'(\gamma) = \{r'_1, \ldots, r'_m\}$ - a set that establishes the volumes of transmission of information flows in the distribution γ between each pair of vertices for which in the graph *Г* there is an edge. Element $r'_j \in R'(\gamma)$ defines in distribution γ real flow between j - th pair of graph vertices *Г*. $Q(y) = \{q_1, \ldots, q_n\}$ - set, each element of which q_i is the total flow along the edge $y_i \in Y$ of graph *G* in distribution γ. $H(\gamma) = \{H^1, \ldots, H^m\}$ - set family $H^j = \{H_1^j, \ldots, H_{z(j)}^j\}$, made up of routes where it is transmitted *j*-th flow in distribution γ. Here *Z*(*j*) - number of transmission routes *j*-th flow in distribution γ. Each route $H_p^j, 1 \le p \le Z(j)$, corresponding threesome $\langle A_p^j, d_p^j, x_p^j \rangle$, where

$A_p^j = \left\{ a_{1p}^j, ..., a_{d_p^j p}^j \right\}$ - a set that defines the composition of the route (in the list of edges) of the graph *G*, each element of which a_{sp}^j, $1 \le s \le d_p^j$, is an edge of the set *Y* of the graph *G*; d_p^j - route length H_p^j; x_p^j - value j - th flow along the route H_p^j. Note that in the set A_p^j elements a_{sp}^j located according to their order of travel.

Using route parameters H_p^j define the value

$$q_i \in Q(\gamma) : \forall y_i \in Y \exists \; q_i = \sum_{j=1}^{m} \sum_{p=1}^{z(j)} x_p^j h_p^j ,$$

where
$$h_p^j = \begin{cases} 0, & y_i \notin A_p^j; \\ 1, & y_i \in A_p^j \end{cases}$$

and meaning $r_j' \in R'(\gamma) : \forall W_j \in W \exists r_j' = \sum_{p=1}^{z(j)} x_p^j$.

Let us give a formal statement of the problem. Graphs are given *G* = <*Z*, *Y*, *b*, *c*> and *Г* = <*Z*, *W*, *r*>. It is required to distribute flows in the network, that is, to form sets *R'*(γ), *Q*(γ) and family *H*(γ) so that the value $F(\gamma) = \sum_{j=1}^{m} \sum_{p=1}^{z(j)} C_p^j X_p^j$ took the minimum value under the following conditions:

1) $\sum_{p=1}^{z(j)} X_p^j - R_j = r_j$;

2) $\forall y_i \in Y \exists q_i \in Q(\gamma) \wedge b_i \in b(Y) \big| q_i + S_i = b_i$;

3) $X_p^j, \; R_{j,} \; S_i \ge 0$,

where C_p^j - unit transfer cost *j*-th flow by H_p^j route; $R_{j,} \; S_i$ - weak variables.

Value C_p^j is defined by the expression

$$C_p^j = \sum_{s=1}^{d_p^j} C_s ,$$

where C_s - cost of transmitting a unit of information flow along an edge a_{sp}^j, included in the route H_p^j.

To solve this problem using the modified simplex - Danzig method [2] it is necessary to specify the initial admissible distribution **γ₀**, more specifically, a family of routes *H*(γ₀). Distribution γ₀ can be found by solving the problem of the maximum heterogeneous information flow allowed for a given network. Formally, the problem of the maximum admissible information flow can be formulated as follows.

Graphs are given *G* = <*Z*, *Y*, *b*, *c*> и *Г* = <*Z*, *W*, *r*>. Required to maximize $\lambda = \sum_{j=1}^{m} \sum_{p=1}^{Z(j)} X_p^j$ under conditions:

1) $\forall y_i \in Y \exists q_i \in Q(\gamma) \wedge b_i \in b(Y) \big| q_i + S_i = b_i$;

2) $X_p^j, S_i \ge 0$.

Denote by the function **μ** the share of the total required flow, which is realized in the distribution **γ**. In the course of the further presentation of the algorithm, we will use the vector form of data representation and distribution results.

Associate the functions b and c the vectors $\tilde{b}$ and $\tilde{c}$, where $\tilde{b} = \left(\tilde{b}_1, ..., \tilde{b}_n, \tilde{b}_{n+1} \right)$ - graph's edge capacity vector *G*, $\tilde{b}_1 = b_i$, $\tilde{b}_{n+1} = \mu$ (at the start of execution $\tilde{b}_{n+1} = 0$); $\tilde{C} = \left(\tilde{C}_1, ..., \tilde{C}_n \right)$ - cost vector (lengths) of graph edges *G*, $\tilde{C}_i = C_i$. We also have: the identity matrix $\bar{B}_0$ size (*n*+1)×(*n*+1) (an additional dummy variable is introduced), wherein $\bar{b}_{\alpha,\beta} = 1$, if α = β, 1 ≤ α ≤ (*n*+1), 1 ≤ β ≤ (*n*+1), and $\bar{b}_{\alpha,\beta} = 0$ if $\alpha \ne \beta$, the matrix of the current basis $\bar{B}^{-1}$ (at the start of the algorithm $\bar{B}^{-1} = \bar{B}_0$). Distribution γ determined by the current simplex table (matrix $\bar{B}^{-1}$ and vector $\tilde{b}$).

The algorithm for constructing the maximum allowable flow is iterative. At each *k-th* iteration, the simplex table is modified.

Step 1. Taking the values of the vector as the length of the edges $\tilde{c}$, find in the graph G for each required graph flow *Г* (edges *W*ⱼ∈*W*) the shortest route, for example, using the Deijkstra method [3] and not taking into account the restrictions on the bandwidth of the edges, we implement the entire flow along this route $r_j \in r(W)$.

Route parameters are stored in a set family H(γₖ). As a result, we have the distribution of the total flow over the edges of the graph *G*, represented by the vector

$$\tilde{q}^k = (\tilde{q}_1^k, ..., \tilde{q}_n^k, \tilde{q}_{n+1}^k) ,$$

$$\text{where } \tilde{q}_i^k = \sum_{j=1}^{m} r_j h_j \left| h_j = \begin{cases} 0, & y_i \notin \mathrm{A}^j; \\ 1, & y_i \in \mathrm{A}^j , \end{cases} \right. , \; \tilde{q}_{n+1}^k = -1 ,$$

A^j - composition of the shortest route between the *j -th* pair of nodes of the graph *Г*, $A^j \in H(\gamma_k)$.

Vector $\tilde{q}^k$ entered into the simplex - table.

Step 2. Among the positive components of the vector $\tilde{q}^k$ determined $\tilde{q}_i^k$ with the minimum value of the ratio $\tilde{b}_i^{k-1} / \tilde{q}_i^k$ (here $\tilde{b}_i^{k-1}$ - componentof the vector $\tilde{b}^{k-1}$ the current capacities of the edges of the graph **G**). Number **i** of the selected component determines the leading row of the simplex table. Let us formulate a new value of the vector $\tilde{q}^{k*}$. To do this, we separate everything except $\tilde{q}_i^k$ vector components $\tilde{q}^k$ at $\left(-\tilde{q}_i^k\right)$ (negative leading value), and set the leading value to $1/\tilde{q}_i^k$.

Step 3. Let us replace in the original identity matrix $\bar{B}_0$ **i** - th column vector $\tilde{q}^{k*}$ and get the matrix $\bar{B}_0^k$. We can form the current simplex table, that is, we form: the matrix of the current basis $\bar{B}_k^{-1} = \bar{B}_0^k \cdot \bar{B}_{k-1}^{-1}$ и new bandwidth vector $\tilde{b}^k = \bar{B}_0^k \cdot \tilde{b}^{k-1}$, in which the vector $\tilde{q}^k$ changes the current base variable of the leading row. For the constructed flow distribution, we obtain the value of the objective function $\mu_k = \tilde{b}_{n+1}^k$, those share of the total required flow, which is realized in the distribution γ*k*, defined by the current simplex tableau and family *H*(γ*k*).

If μ*k* ≥ 1, then the work of the algorithm for constructing the maximum admissible flow ends here.

At μ*k* < 1 from edges set *Y* graf *G* is removed edge yi, coincider i - th the leading row of the simplex table, and the process of constructing the maximum allowable flow continues cyclically until then (steps 1 - 3), until either μ*k* ≥ 1, or it will not be possible to build any route that allows you to distribute the remaining required flows $r_j \in r\left(W\right)$ of the graf *Г*. In the latter case, we can say that the given distributed computing network, defined by graphs *G* and *Г*, allows you to distribute the required information floVVws of no more than **μ*k*** from the given value, that is

$$\forall W_j \in W \exists r_j \in r\left(W\right) \;\Bigg|\; \sum_{p=1}^{z(j)} X_p^j \le \mu_k \cdot r_j \,.$$

The resulting flow distribution **γ0 =γk**, route family defined **H(γ0)** together with the bandwidth vector $\tilde{b}$, used to build minimum cost flows.

Initial data for solving the problem of constructing minimum cost flows: the number of simplex-table variables equal to **(*n*+*m*+1)** additionally, a dummy variable is introduced); graph's edge capacity vector ***G*** $\tilde{f} = \left(\tilde{f}_1,..., \tilde{f}_n, \tilde{f}_{n+1},..., \tilde{f}_{n+m}, \tilde{f}_{n+m+1}\right)$, где $\tilde{f}_i = \tilde{b}_i$; $\tilde{f}_{n+j} = -r_j$; $\tilde{f}_{n+m+1} = F$ (at the start of the algorithm $\tilde{f}_{n+m+1} = 0$); cost vector (lengths) of graph edges **G** $\tilde{C} = \left(\tilde{C}_1,...\tilde{C}_n\right)$, where $\tilde{C}_i = C_i$; identity matrix $\bar{B}_0$ sizes (*n*+*m*+1)×(*n*+*m*+1); current basis matrix $\bar{B}^{-1}$ (at the start of execution $\bar{B}^{-1} = \bar{B}_0$); route family *H*(γ0).

Step 1. Route family's **H(γ0)** are sequentially entered into the simplex table. The route entered into the table, for the transmission of the j -th stream, is represented by the vector $\tilde{X}_p^j = \left(\tilde{X}_1,..., \tilde{X}_n, \tilde{X}_{n+1},..., \tilde{X}_{n+m}, \tilde{X}_{n+m+1}\right)$, where component $\tilde{X}_i$ is equal to one if the corresponding index **i** the edge is included in this route and is equal to zero in other cases; $\tilde{X}_{n+j} = -1$ - coefficient at X_p^j in the equation $R_j - \sum_{p=1}^{z(j)} X_p^j = -r_j$; $\tilde{X}_{n+m+1} = C_p^j$ - unit transfer cost **j** - th flow along the input route. Before entering into the simplex - table vector $\tilde{X}_p^j$ is adjusted according to the expression $\tilde{X}_p^{jk} = \bar{B}_0^{k-1} \cdot \tilde{X}_p^j$. Among the positive components *i* of the vector $\tilde{X}_p^{jk}$ such $\tilde{X}_i^k$, for which the relation $\tilde{f}_i^{k-1}/\tilde{x}_i^k$ has a minimum value. The number i of the selected component determines the leading row of the simplex table.

Step 2. Let's input the vector $\tilde{X}_p^{jk}$ into the basis and form the current simplex tableau in accordance with the following procedure.

Get the new value of the vector $\tilde{X}_p^{jk*}$. To do this, we separate everything except $\tilde{X}_i^k$, vector components $\tilde{X}_p^{jk}$ at $(-\tilde{X}_i^k)$ (negative leading value), and set the leading value to $1/\tilde{X}_i^k$. Let us replace in the original identity matrix $\bar{B}_0$ **i** - th column vector $\tilde{X}_p^{jk*}$ and get the matrix $\bar{B}_0^k$. Let's form the current simplex table (we form: the matrix of the current basis $\bar{B}_k^{-1} = \bar{B}_0^k \cdot \bar{B}_{k-1}^{-1}$ and a new bandwidth vector $\tilde{f}^k = \bar{B}_0^k \cdot \tilde{f}^{k-1}$), in which the vector $\tilde{X}_p^{jk}$ changes the current base variable of the leading row. For the constructed flow distribution, we obtain the value of the objective function $F\left(\gamma_k\right) = \tilde{f}_{n+m+1}^k$, which determines the cost of transferring information flows in the distribution γ*k*. After finishing entering the routes included in the family *H*(γ0), determining the initial basic distribution of information flows, the basic flows are redistributed in order to minimize the functional *F*.

Step 3. Components are analyzed (*n* + *m* + 1) - th row of the matrix of the current basis $\bar{B}_k^{-1}$. If among the analyzed components $\bar{b}_{n+m+1,t}^k$, 1 ≤ *t* ≤ (*n*+*m*+1), there are components with a value less than zero, then from $\bar{b}_{n+m+1,t}^k < 0$ choose a component $\bar{b}_{n+m+1,r}^k$, 1 ≤ *r* ≤ (*n*+*m*+1), satisfying the condition $\bar{b}_{n+m+1,r}^k = \min_t \bar{b}_{n+m+1,t}^k < 0$. Column *r* - leading.

To enter the leading column into the basis, in order to minimize *F*, among the components $\bar{b}_{t,r}^k > 0$ find such $\bar{b}_{a,r}^k$, 1 ≤ *a* ≤ (*n*+*m*+1), for which the relation $\tilde{f}_{t,r}^k / \bar{b}_{t,r}^k$ takes the minimum value, $\tilde{f}_{a,r}^k / \bar{b}_{a,r}^t = \min_t \left(\tilde{f}_{t,r}^k / \bar{b}_{t,r}^k\right)$. Line *a* - leading. Let's form new component values $\bar{b}_{t,r}^{k*}$ the leading column of the matrix $\bar{B}_k^{-1*}$. To do this, we separate everything except $\bar{b}_{a,r}^k$ components $\bar{b}_{t,r}^k$ at

$(-\bar{b}_{a,r}^{k})$ (negative value of the leader), and the value of the leader will be set equal to $1/\bar{b}_{a,r}^{k}$.

<u>Step 4.</u> Let us replace in the original identity matrix $\bar{B}_0$ a - th column to newly formed column **r** matrix $\bar{B}_k^{-1*}$ and get the matrix $\bar{B}_0^{k+1}$. Now we can form the current simplex table, that is, we form the matrix of the current basis $\bar{B}_{k+1}^{-1} = \bar{B}_0^{k+1} \cdot \bar{B}_k^{-1}$ and a new bandwidth vector $\tilde{f}^{k+1} = \bar{B}_0^{k+1} \cdot \tilde{f}^{k}$, in which the pivot row base variable a is replaced by the pivot column variable r. For the constructed flow distribution, we obtain a new value of the objective function

$$F\left(\gamma_{k+1}\right) = \tilde{f}_{n+m+1}^{k+1} .$$

If in the resulting distribution γ_{k+1} в $(n+m+1)$ - th line has at least one component $\bar{b}_{n+m+1,t}^{k+1}$ with a value less than zero, then the operation of the algorithm is repeated, starting from step 3. If the components $\bar{b}_{n+m+1,t}^{k+1} \geq 0$, then the vector is formed $\tilde{C}^{*} = \left|\tilde{C}_1^{*}, ..., \tilde{C}_n^{*}\right|$ the current cost of transferring a unit of information flow for each edge of the set Y graf G, $\tilde{C}_i^{*} = \tilde{C}_i + \bar{b}_{n+m+1,i}^{k+1}$. Taking as the length of the edges the value of the vector $\bar{C}^{*}$, find in the graph G for each required stream j graph Γ the shortest route, for example, using Deijkstra's method.

Found route H_{p+1}^{j} represented by a vector

$$\tilde{X}_{p+1}^{j} = \left(\tilde{x}_1, ..., \tilde{x}_n, \tilde{x}_{n+1}, ..., \tilde{x}_{n+m}, \tilde{x}_{n+m+1}\right) .$$

Then if $H_{p+1}^{j} \in H\left(\gamma_0\right)$, then the optimal solution to the problem of constructing minimum cost flows is obtained and the algorithm ends, otherwise the route found H_{p+1}^{j} is introduced into the route family H(γ₀) and the process of constructing the minimum cost flows continues from the step 1. Resulting optimal distribution $\gamma = \gamma_{k+1}$ defined by family routes $H(\gamma)$, sets $R'(\gamma)$, $Q(\gamma)$ and defined by family routes $F(\gamma)$.

Conclusions

The article proposes a method for splitting transactions to a hybrid cloud data warehouse.

The method is based on the construction of network graphs.

The optimization problem is reduced to the distribution of flows in the network in such a way that, if the requirements of the cloud hybrid data storage are met, the cost of transmitting the flow in the network is minimal. To solve it, an iterative algorithm for constructing the maximum admissible flow is proposed. At each iteration, the simplex table of network graphs is modified.

The direction of further research is the development of a method for optimizing the structure of cloud data storage.

References

1. Широкова Е.А. Облачные технологии / Е.А. Широкова // Современные тенденции техн. наук: мат. межд. науч. конф.; Уфа, 2011 г. – Уфа: Лето, 2011. – С. 30 – 33.
2. Риз Д. Облачные вычисления [Текст] / Джордж Риз. – СПб.: 2011. – 288 с.
3. Google Cloud Platform [Электронный ресурс]. – Режим доступа: http://cloud.google.com. – 12.04.2013.
4. Зиков І. С., Кучук Н. Г., Шматков С. І. Синтез архітектури комп'ютерної системи управління транзакціями e-learning. *Сучасні інформаційні системи.* 2018. Т. 2, № 3. С. 60–66. DOI: https://doi.org/10.20998/2522-9052.2018.3.10.
5. Кучук Г.А. Управление ресурсами инфотелекоммуникаций : монография / Г.А. Кучук, Р.П. Гахов, А.А. Пашнев. – М.: Физматлит, 2006. – 220 с.
6. Кучук Г.А. Інформаційні технології управління інтегральними потоками даних в інформаційно-телекомунікаційних мережах систем критичного призначення : монографія / Г.А. Кучук. – Х.: ТОВ «Щедра садиба плюс», 2013. – 264 с.. – ISBN 978-617-7188-18-5.
7. Ткачов В. М., Коваленко А. А., Кучук Г. А., Ні Я. С. Метод забезпечення живучості високомобільної комп'ютерної мережі. *Сучасні інформаційні системи.* 2021. Том 5, № 2. С. 159-165. DOI: https://doi.org/10.20998/2522-9052.2021.2.22
8. Саймак А. Обработка транзакций / А. Саймак // СУБД. – 1997. – № 2. – С. 70 - 82.
9. Сергиенко И.В. Модели и методы решения на ЭВМ комбинаторных задач оптимизации / И.В. Сергиенко, М.Ф. Каспшицкая. – К.: Наук. думка, 1981. – 287 с.

Received (Надійшла) 04.09.2022
Accepted for publication (Прийнята до друку) 23.11.2022

Метод розподілу транзакцій до гібридного хмарного сховища даних

Н. Г. Кучук, А. П. Шиман

Анотація. У статті розглянуто **актуальне завдання** оптимального розподілу ресурсів у CLOUD-системах, котрі підтримують гібридні хмарні сховища даних. **Метою даної статті** є розробка методу оптимального розподілу кількох різнорідних транзакцій до гібридного хмарного сховища даних, пов'язаних загальним обмеженням пропускних здібностей. Критерієм оптимізації буде мінімальна вартість. Інформаційні потоки мають певні пункти відправлення та призначення, є різнорідними потоками із загальними обмеженнями на пропускні здібності каналів зв'язку, що використовуються. Отримані результати. Метод ґрунтується на побудові мережевих графів. Завдання оптимізації зводиться до розподілу потоків у мережі таким чином, щоб при задоволенні вимог хмарного гібридного сховища даних вартість передачі потоку в мережі була мінімальною. Для її вирішення запропоновано ітераційний алгоритм побудови максимального допустимого потоку. На кожній ітерації проводиться модифікація симплексної таблиці зв'язку мережевих графів. Напрямок подальших дослідів – розробка методу оптимізації структури гібридного хмарного сховища даних

Ключові слова: хмарні технології, сховище даних, гібридність, CLOUD-система, мережний граф.