**Bohdan Ivanishchev, Pavlo Rehida,**
**Artem Kaplunov, Oleksandr Honcharenko.**

# ANALYSIS OF DATA BALANCING PROBLEMS IN DISTRIBUTED DATA STORAGES

The article analyzes existing problems that confront data storages, data balancing problems in these storages, as well as ways to solve these problems.

**Key words:** distributed data storages, data balancing, facility location problem.

Fig.: 0. Tabl.: 0. Bibl.: 6.

**Analysis of existing types of distributed data storages.** The main problems that confront data storages are the permanent increase of data amount, data decentralization, need for reliable data storing, etc. For a solving problem of the permanent increase of data, storages must have a very large amount and will be easily scalable. The problem of data decentralization can be solved by the placement of data storages as close as possible to both data sources and data users. The need for reliable data storing will be solved by storing data in multiple copies i.e. by data replication.

The general solution to these problems is distributed data storages. A distributed data storage is a computer network where information is stored on more than one node [1]. Currently, distributed data storages are mainly represented by distributed databases, distributed in-memory grids, distributed file systems.

But there are some problems that confront distributed data storages:

1. structure of distributed data storage is constantly changing, attached storage devices can be disconnected, and new storage devices can be attached;

2. replicas of data that is deleted must also be deleted;

3. new data that is added to the storage must be replicated;

4. accordingly, an imbalance in the amount of data on different storage devices may occur periodically.

The solution to these problems is the data balancing procedure.

**Data balancing problem.** The data balancing procedure contains data location and data migration tasks. Data location task can be solved by reduction to capacitated facility location problem. Data migration task can be solved by reduction to graph coloring problem.

**Capacitated facility location problem.** Capacitated facility location problem can be solved by using integer programming. In this context capacitated facility location problem is often posed as follows: there are $n$ facilities and $m$ customers; $f_i$ denote the cost of opening facility $i$ $(i = 1, ..., n)$; $c_{ij}$ denote the cost to ship a product from facility $i$ to customer $j$ $(j = 1, ..., m)$; $d_j$ denote the demand of customer; $u_i$ denote

the capacity of facility *i* [2]. We wish to choose which of the n facilities to open, and which facilities to use to supply *m* customers, in order to satisfy some fixed demand at minimum cost.

Mathematical formulation of the capacitated facility location problem:

$$\begin{gathered} min\left[\sum_{i=1}^{n}\sum_{j=1}^{m} c_j \cdot d_j \cdot y_{ij} + \sum_{i=1}^{n} f_i \cdot x_i\right], \\ s.t\ \sum_{i=1}^{n} y_{ij} = 1\,, j = 1,\ldots,m, \\ \sum_{j=1}^{m} d_j \cdot y_{ij} \le u_i \cdot x_i, i = 1,\ldots,n \\ y_{ij} \ge 0, i = 1,\ldots,n, j = 1,\ldots,m, \\ x_i \in \{0,1\}, i = 1,\ldots,n, \end{gathered} \quad (1)$$

where $x_i$ = 1 if facility i is open and $x_i$ = 0 otherwise, $y_{ij}$ represents the fraction of the demand $d_j$ filled by facility.

**Existing solves of facility location problem.** Capacitated facility location problem is NP-hard task. Generally solutions of this problem base on heuristic algorithms and integer programming.

Eiman J. Alenezy proposed one of the last solutions of capacitated facility location problem which is based on using the Lagrangian Decomposition and Volume Algorithm [3].

One more solution was proposed by M. Luis, M. Fadzli Ramli, and A. Lin. This solution solves the capacitated facility location problem by using an efficient greedy randomized adaptive search procedure (GRASP). Greedy randomized adaptive search procedure (GRASP) is a multi-start heuristic to solve hard combinatorial optimization problems where each iteration consists of a constructive phase and a local search phase [4].

**Graph coloring problem.** Graph coloring problem (or graph's vertices coloring problem) is a problem of assigning colors (or numbers) to graph's vertices such that no two vertices sharing the same edge have the same color.

**Existing solves of graph coloring problem.** The last papers that studied the graph coloring problem proposed solutions to this problem that are based on the Douglas–Rachford algorithm [5] and on improved hybrid ant-local search algorithm [6].

**Possible ways to improve existing solutions.** There are two problems with data balancing procedure that need attention:

1. definition the demand of customer $d_j$ in capacitated facility location problem doesn't accurate enough for data location task because the demands of customer for the different facilities (elements) of data are different (so must be $d_{ij}$ but not $d_j$);

2. when we are planning the migration of various data elements we can add priorities that will allow us to restore storage integrity more quickly after disconnecting some storage devices.

**Conclusions.** This article was analyzing problems that confront data storages and solutions to these problems. The general solution to these problems is distributed data storages. But distributed storages are confronted with other problems and the final solution to part of them is data balancing.

Data balancing problem base on known problems: capacitated facility location problem and data migration problem. These problems are NP-hard but they have many different known solutions part of which are given in this article.

The article also highlighted some problems of existing solutions to data balancing problem. Solving these problems will allow increase efficiency of data balancing procedure in distributed data storages.

## References

1. Pessach Y. (2013). Distributed Storage: Concepts, Algorithms, and Implementations.

2. Conforti, Michele; Cornuéjols, Gérard; Zambelli, Giacomo (2014). Integer Programming | SpringerLink. Graduate Texts in Mathematics. 271.

3. Alenezy E. J. (2020) 'Solving Capacitated Facility Location Problem Using Lagrangian Decomposition and Volume Algorithm', Advances in Operations Research.

4. Luis, Martino & Ramli, Mohammad Fadzli & Lin, Abdullah. (2016). A greedy heuristic algorithm for solving the capacitated planar multi-facility location-allocation problem.

5. Aragón Artacho, F. J., Campoy, R., & Elser, V. (2020). An enhanced formulation for solving graph coloring problems with the Douglas–Rachford algorithm.

6. Fidanova, S., & Pop, P. (2016). An improved hybrid ant-local search algorithm for the partition graph coloring problem. *Journal of Computational and Applied Mathematics*, *293*, 55-61.

## AUTHORS

**Bohdan Ivanishchev** – PhD student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: art.kaplunov@gmail.com

**Rehida Pavlo** – assistant professor, Department of Computer Engineering, NationalTechnical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: pavel.regida@gmail.com

**Artem Kaplunov** – PhD student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: art.kaplunov@gmail.com

**Oleksandr Honcharenko** – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

E-mail: alexandr.ik97@ukr.net