TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

ACTA TECHNICA NAPOCENSIS

Series: Applied Mathematics, Mechanics, and Engineering
Vol. 66, Issue Special I, Sseptember, 2023

CONTRIBUTIONS TO THE MODELING AND OPTIMIZATION OF DATA SAVING SOLUTIONS WITHIN A MODERN ORGANIZATION

Radu-Costin MOISESCU, Constantin OLTEANU, Petrică TERTEREANU, Aurel-Mihail ȚÎȚU

Abstract: *In the knowledge-based economy that is digital in nature, data is a representative factor for organizations' capital. It can be said that data could represent an important production factor in the field of goods, but especially in the field of digital services. The increase in the volume of data processed within organizations can lead to new challenges in terms of organizational management, but especially in the methods of optimizing the processes by which data is backed up in more secure environments to ensure successful disaster recovery. The scientific paper actually presents the way to an evaluation of the possibilities of modeling and optimizing the backup processes of the volume of data from the perspective of compliance with the international information security standards (ISO27001:2018). The scientific work has behind it applied research and can represent a point of view of the authors argued for obtaining in the field of computer data backup technologies.*
Key words: *data volumes, backup, ISO27001:2018, optimize, processes, deduplication.*

1. INTRODUCTION

Some of the most important obstacles for a better and wider use of information technology in the fields of intellectual property include, among others, the lack (unavailability) of services based on information technology, the lack of interoperability of solutions across states members, as well as the fragmentation of this market of information space and technology-based solutions [1].
Information security is focused on ensuring the integrity availability and confidentiality of computer data, regardless of their form. The security of information systems aims at guaranteeing both their availability and correct functioning, but also the possibility of recovering this data in the event of disasters and ensuring the continuity of the organization's operations. Network security focuses on the protection of data during transmission [2].
According to the ISO 27001:2018 standard, the mechanisms and tools by which the protection of data whether stored or transmitted within an information system is ensured is a way to ensure the confidentiality, integrity and availability of data.
Certain procedures are also indicated to mitigate different types of security threats.

2. EXPERIMENTAL RESEARCH

At the level of national intellectual property organizations, a variety of IT solutions can be implemented to ensure data security. We present in the current work an experimental research based on a centralized storage and data saving solution of such an organization. The solution we tested integrates a SAN-type centralized storage system and a data saving and recovery system on and from two storage media: a saving system on disk volumes and one on LTO6 magnetic tapes [3]. To carry out the experiment I used an HP OverLand Neo magnetic tape drive. This model has the advantage of ensuring the redundancy of the writing and reading system on LTO6 media through the presence of two

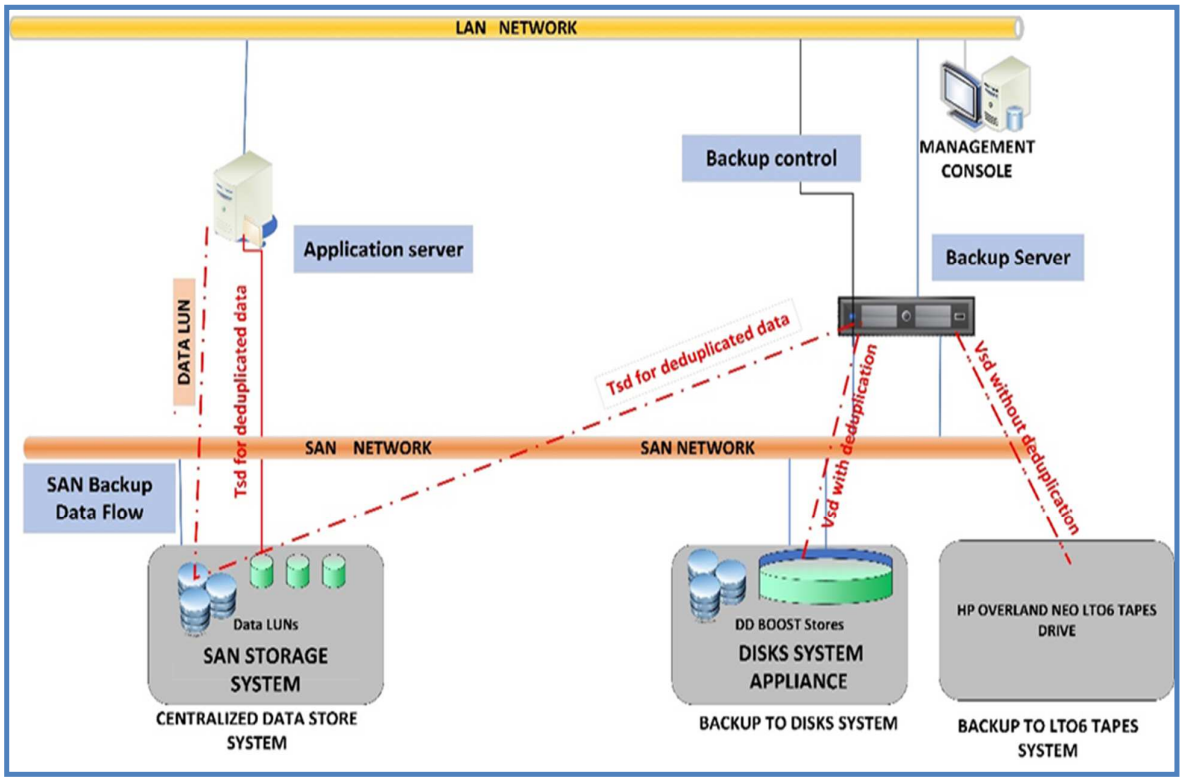

Fig.1. Diagram of the system Backup system with deduplication technology for experiment

independent reading and writing units and two magnetic tape magazines that can store 15 tapes each. the drive has 2 tape read/write drives and an internal robot.

In figure 1, I graphically represented the scheme of the equipment used in the experimental study and the data flows that are transmitted within the processes of writing and reading data on the two mentioned media. The control of data flows is carried out through the specific backup application (Microfocus Data Protector) which manages the operation of the agents installed on the servers from which it is desired to extract or rewrite the saved data. Also, through the application, the backup sessions are programmed and monitored in order to ensure the saving (writing) of data on the two storage media. In the event of disasters, the recovery of data previously written to the two storage media described, saving the media and rewriting it to the original location or to

Table 1

Experimental recordings of the objective function Vsd and Vsb vs X_1 and X_3

Month	X_1 (TB)	X_2 (TB)	X_3 (Applications no.)	Vsd (TB)	FITS1	RESI1
1	1.2034	0.0008	876	0.892	0.893094	-0.0010943
2	1.2042	0.0010	1040	0.894	0.902565	-0.0085654
3	1.2052	0.0007	763	0.897	0.899264	-0.0022637
4	1.2059	0.0006	626	0.899	0.898740	0.0002602
5	1.2065	0.0009	968	0.921	0.912722	0.0082775
6	1.2074	0.0009	988	0.924	0.918210	0.0057898
7	1.2083	0.0010	1030	0.926	0.924389	0.0016110
8	1.2093	0.0009	863	0.928	0.924543	0.0034574
9	1.2102	0.0009	930	0.932	0.931507	0.0004933
10	1.2111	0.0010	998	0.936	0.938502	-0.0025022
11	1.2121	0.0008	881	0.939	0.940226	-0.0012265
12	1.2129	0.0009	903	0.941	0.945237	-0.0042371

another medium is provided by the centralized backup application.
The values we recorded during the experiment are described in table 1.

3. MATHEMATICAL MODELING

3.1 Mathematical modeling Vsd versus X_1 and X_3

To carry out this experiment, the input parameters of interest that will be processed were established from the beginning. Thus, it was established that they are: the size of the data volume, its estimated monthly increase (increase due to the monthly submissions of protection applications in the field of intellectual property), the number of protection submissions in the studied field.

In table 1 we presented the experimental values recorded during the experiment.

In order to have a high accuracy of the results, we chose to perform a number of 12 tests that represent a simulation of the growth of the data volume for one year. In table 1 we presented the input and output parameters:

- Input parameters are:

- X_1 = the data volume measured in TB where $X_1 \in$ (0.9TB... ...12TB);

- X_2 = the estimated growth per month of the data volume measured in TB where $X_2 \in$ (0.009TB… ...0.02TB);

- X_3 = the estimated monthly number of protection requests submitted in case $X_3 \in$ (100... ...2010) in one year.

- The output parameter (named objective function) is noted Vsd and represents the size of the volume of data written on disks after deduplication process. This data volume is measured in TB [4]. At the same time, we noted with Vsb the volume measured in TB of data saved with compression but without deduplication technology on LTO6 model magnetic tapes.

The response surface plot for objective function Vsd is used to evaluate relationships between three variables simultaneously (Vsd, X1 and X3). Like a three-dimensional scatterplot, a 3D surface plot has three axes.

Additionally, interpolation is used in order to produce a continuous surface or a grid of Vsd values that fit the data obtained.

Figure 2 represents the response area graph and shows how the variables X_1 and X_3 relate to Vsd. In this experiment, the Minitab software was used to study and calculate the variation of the Vsd and to draw the diagram of the response area.

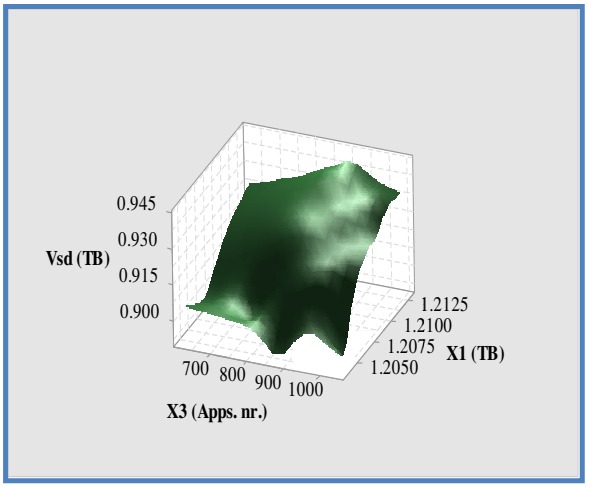

Fig.2. Response surface plot for objective function Vsd vs. X_1 and X_3

By means of this diagram we can identified the maximum value of the data written on disks with deduplication (at 0.941TB) can be located and the value of the variable X_3=904 the estimated number of protection requests corresponding to the twelfth month [5].

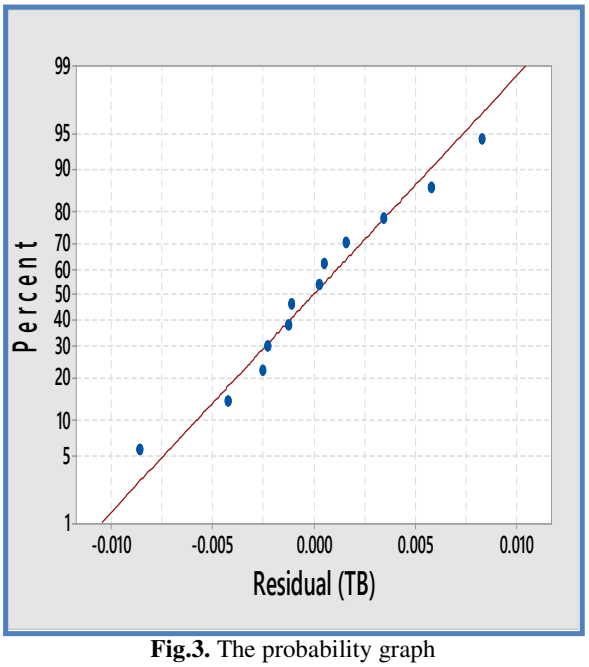

Fig.3. The probability graph

The adjusted distribution line and especially the matching of the points with this line are presented in figure 3. The graph represents the

probability that the specific theoretical distribution closely follows the straight red line and that the points are located on either side near this line.

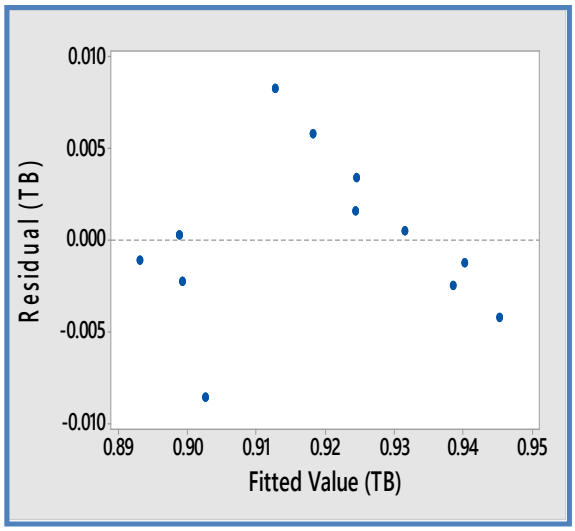

Fig.4. Diagram of the system

The variation of the residual values compared to the adjusted values are represented graphically in Figure 4. It can be seen that the adjusted values are represented on the x-axis and the residual values are represented on the y-axis. The hypothesis that the residual values are randomly distributed with constant variability is verified by the experimental data obtained and presented in figure 4. [6] In the graph in figure 5, the red line determines the way in which the input factors influence each other, representing the statistical model determined by the data collected during the experiment.

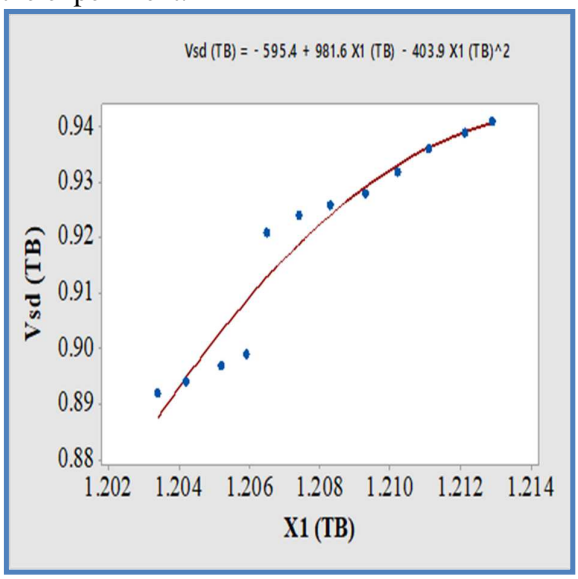

Fig.5. Response surface plot for objective function Vsd vs. X_1

In the graph in figure 5, the individual values that Vsd has according to the parameter X1 are represented, the red line representing the estimation of the values that the studied mathematical model should follow. Thus, the degree of matching of the determined model with the input parameters is represented by the red line [7].

During the experimental research, the following regression equation was identified:

$$Vsd\ (TB) = -5.632 + 5.399\ X_1\ (TB) + 0.000031\ X_3\ (Apps.\ No.) \quad (1)$$

The residual histogram shown in Figure 6 is used to visualize the correctness of the data and if there are errors in this data.

From the analysis we performed, following this experiment, we obtained a reduction in the volumes of data saved on disks by applying deduplication processes with approximately 30.31% volume resulting from the calculation:

$$RVsd\ (\%) = 100\% - \frac{\sum_{k=1}^{12} Vsd\ (TB)}{\sum_{k=1}^{12} X1\ (TB)} * 100 \quad (2)$$

$$RVsb\ (\%) = 100\% - \frac{10{,}105}{14{,}4994} * 100 \quad (3)$$

$$RVsd\ (\%) = 100\% - 69{,}69\% = 30{,}31\% \quad (4)$$

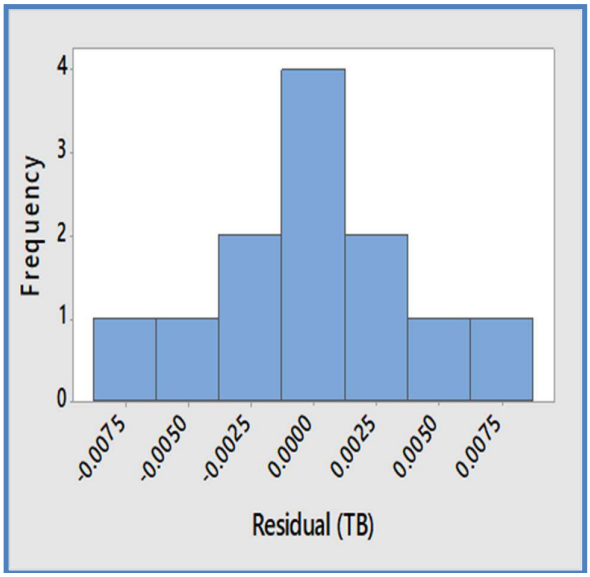

Fig.6. Response surface plot for objective function Vsd vs. X_1 and X_3

At the same time reduction in the volumes of data saved on magnetic tapes by applying the data

compression processes with approximately 16.46% volume resulting from the calculation:

$$RVsb\ (\%) = 100\% - \frac{\sum_{k=1}^{12} Vsb\ (TB)}{\sum_{k=1}^{12} X1\ (TB)} * 100 \quad (5)$$

$$RVsb\ (\%) = 100\% - \frac{12,1142}{14,4994} * 100 \quad (6)$$

$$RVsb\ (\%) = 100\% - 83,54\% = 16,46\% \quad (7)$$

3.2 Mathematical modeling Vsd versus X_2 and X_3

The experimental results that we recorded during the experiment are described in table 2.

Table 2

Experimental recordings of the objective function Vsd and Vsb vs X_2 and X_3

Month	X_1 (TB)	X_2 (TB)	X_3 (Applications no.)	Vsd (TB)	FITS1	RESI1
1	1.2034	0.0008	876	0.892	0.911290	-0.0192902
2	1.2042	0.0010	1040	0.894	0.926610	-0.0326097
3	1.2052	0.0007	763	0.897	0.906940	-0.0099398
4	1.2059	0.0006	626	0.899	0.905151	-0.0061514
5	1.2065	0.0009	968	0.921	0.917882	0.0031176
6	1.2074	0.0009	988	0.924	0.915747	0.0082527
7	1.2083	0.0010	1030	0.926	0.927677	-0.0016773
8	1.2093	0.0009	863	0.928	0.929092	-0.0010916
9	1.2102	0.0009	930	0.932	0.921939	0.0100609
10	1.2111	0.0010	998	0.936	0.931093	0.0049066
11	1.2121	0.0008	881	0.939	0.910756	0.0282436
12	1.2129	0.0009	903	0.941	0.924821	0.0161786

Figure 7 represents the response area graph and shows how the variables X_2 and X_3 relate to Vsd. In this experiment, the Minitab software was used to study and calculate the variation of the Vsd and to draw the diagram of the response area. By means of this diagram we can identified the maximum value of the data written on disks with deduplication (at 0.941TB) can be located and the value of the variable X_2=0,0009TB and X_3=903 the estimated number of protection requests corresponding to the twelfth month

The adjusted distribution line and especially the matching of the points with this line are presented in figure 8.

The graph represents the probability that the specific theoretical distribution closely follows the straight red line and that the points are located on either side near this line.

The variation of the residual values compared to the adjusted values are represented graphically in Figure 4. It can be seen that the adjusted values are represented on the x-axis and the residual values are represented on the y-axis.

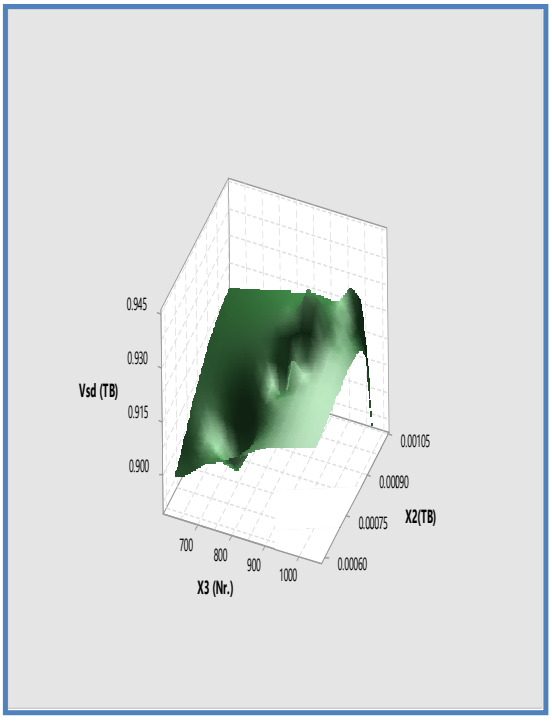

Fig.7. Response surface plot for objective function Vsd vs. X_2 and X_3

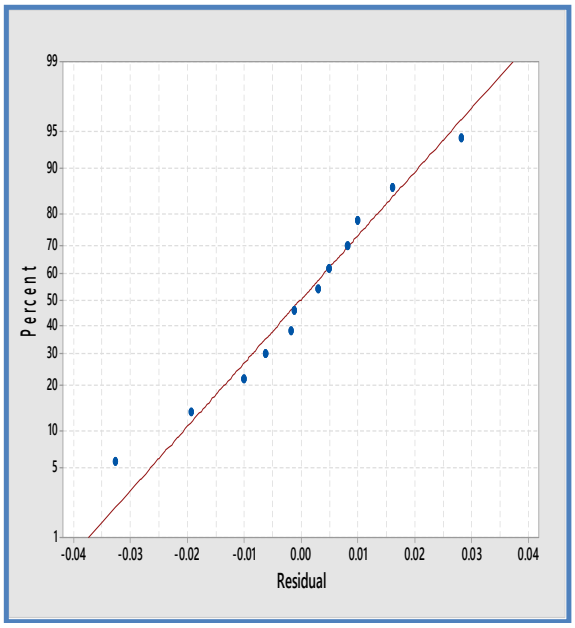

Fig.8. The probability graphs

The hypothesis that the residual values are randomly distributed with constant variability is verified by the experimental data obtained and presented in figure 9.

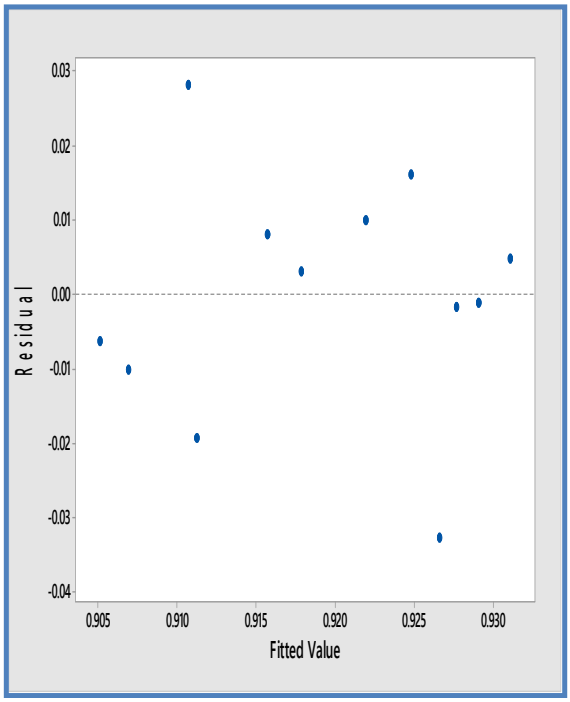

Fig.9. Diagram of the system

The residual histogram shown in Figure 10 is used to visualize the correctness of the data and if there are errors in this data.

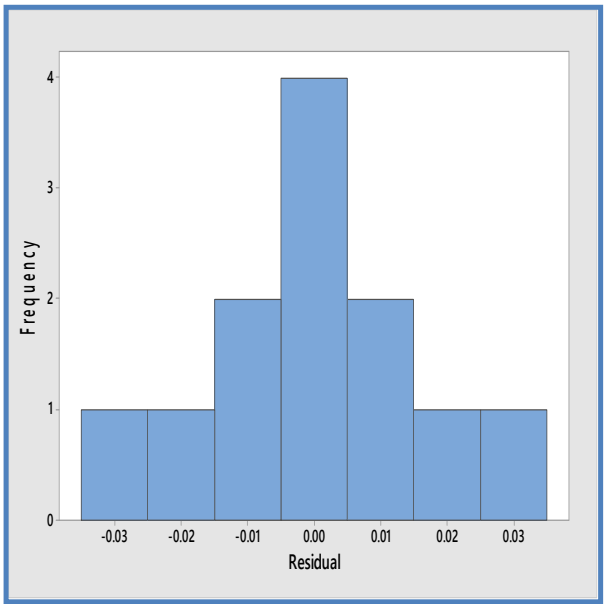

Fig.10. Response surface plot for objective function Vsd vs. X_1 and X_3

We can conclude that data backup and recovery systems on disk volumes have the following advantages compared to tape backup systems:

the solution offers substantially improved data transmission speeds with the possibility of quick restoration of data stored on disks;

- relatively low implementation and maintenance costs;
- possibility of implementing new technologies such as data deduplication, which reduces the volume of data to be saved as well as the time to save and restore them;
- low degree of backup and restore sessions performed with errors or even failed;
- improve the security level for the data saved on the disk volumes with their deduplication because in case of interception of these data by a possible attacker, they cannot be accessed directly.

3.3 Process efficiency analysis

The main purpose of the experimental research is to analyze the technical possibilities for reducing the volumes of data that must be saved and restored to avoid their loss in the event of disasters. In this sense, a decrease in data volumes thus represents one of the key factors in the efficiency of the processes of saving and restoring this data.

Table 3

Analysis table of variation of data volumes and restore times for deduplicated data

Month	Vsdd (TB)	X1 (TB)	Data volume after deduplication (%)	Reduction after deduplication (%)	Trd (minutes)	Trdd (minutes)	The time to restore deduplicated data (%)	Percentage reduction after deduplication (%)
1	0.892	1.2034	74.12	25.88	141.27	112.41	79.57	20.43
2	0.894	1.2042	74.24	25.76	143.21	113.12	78.98	21.02
3	0.897	1.2052	74.42	25.58	144.37	114.22	79.11	20.89
4	0.899	1.2059	74.55	25.45	145.56	115.35	79.24	20.76
5	0.921	1.2065	76.33	23.67	146.37	116.26	79.44	20.56
6	0.924	1.2074	76.52	23.48	147.45	117.13	79.43	20.57
7	0.926	1.2083	76.63	23.37	149.02	117.45	78.81	20.19
8	0.928	1.2093	76.73	23.27	150.17	119.04	79.27	20.73
9	0.932	1.2102	77.01	22.99	151.26	120.15	79.43	20.57
10	0.936	1.2111	77.28	22.72	152.19	121.31	79.71	20.29
11	0.939	1.2121	77.41	22.59	153.32	122.23	79.72	20.28
12	0.941	1.2129	77.58	22.42	155.42	123.48	79.44	20.56

In table 3 we presented the results obtained from the experiment of the variation of data volumes and restoration times for deduplicated data.

Thus, it can be seen that there is a variation in the size of data volumes saved on disks with the application of deduplication technology.

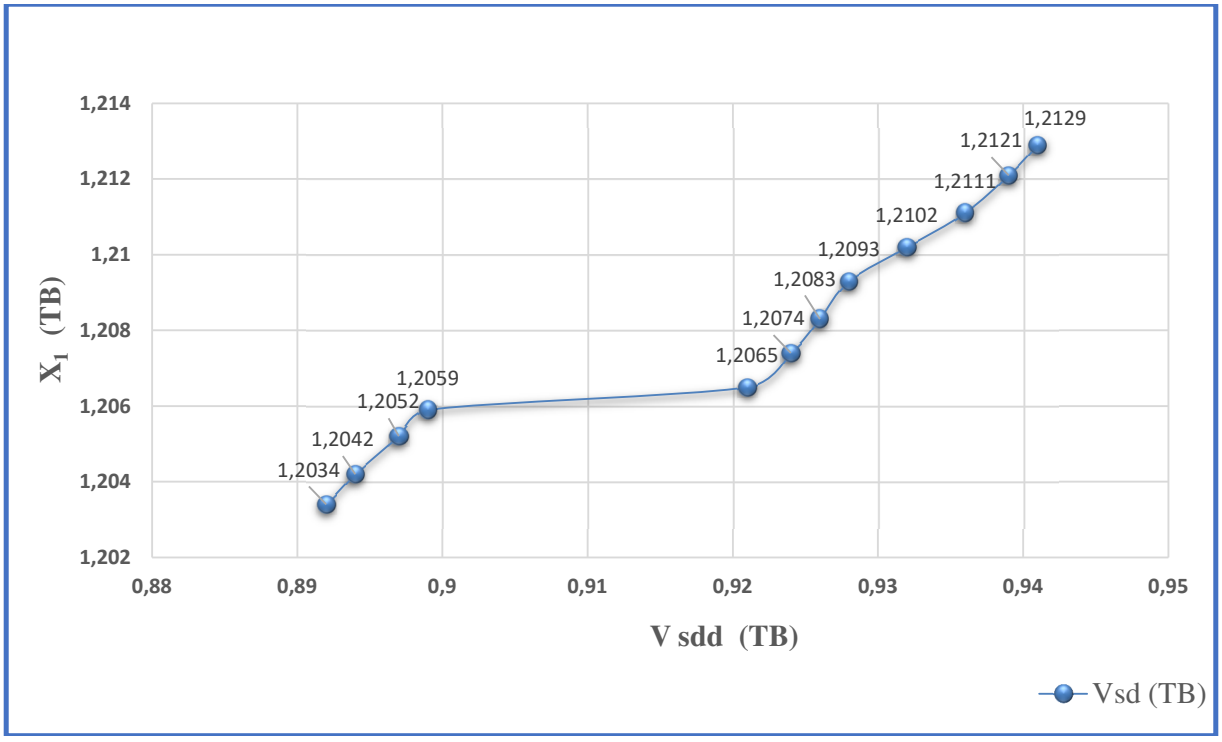

Fig.11. Plot of data volume size value X_1 versus data volume saved with Vsdd deduplication

In figure 11 we graphically represented the evolution of the sizes of data volumes saved on disks with the application of deduplication technologies Vsdd (TB) in relation to the initial size of data volumes to be saved X1 (TB).

To calculate the efficiency that, following the implementation of the data deduplication technology, we obtained for the size of the data volumes saved and restored, we used the calculation formula and we noted the efficiency with Efc., the effort made with Efm. and the Efo. the effort obtained:

$$Efc. = \frac{\text{Ef m.}}{\text{Efo.}} = \frac{\text{X1 (TB)}}{Vsdd\ (TB)} \quad (8)$$

$$Efc. = \sum_{k=1}^{12} \frac{\text{X1 (TB)}}{\text{Vsdd (TB)}} = \frac{14.4965}{11,029} =$$

$$1,3143 \Longrightarrow \text{Efc.} = +31,43\% \quad (9)$$

In figure 11.2 we graphically represented the evolution of the sizes of the restoration times of the deduplicated data volumes and the evolution of the sizes of the restoration times of the data volumes saved without deduplication.

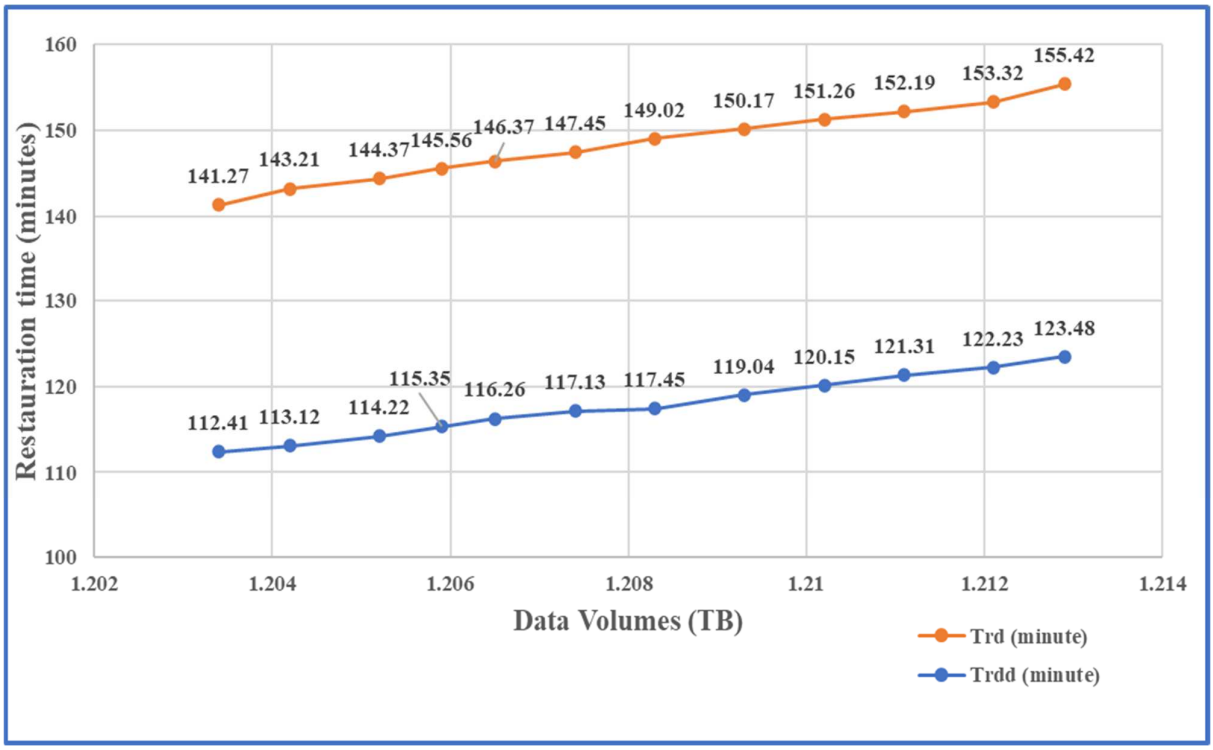

Fig.12. Graph with X_1 data volume size values versus restore time from deduplicated disks (Trdd) and restore time from non-deduplicated disks (Trd)

The classical experiment method that we used in this research to determine the output parameters or objective functions using the MiniTab application is a valuable tool for analyzing and validating the results of numerical models for intellectual organizations data storage and retrieval systems in cases of disasters.

In the present study, the data saving and recovery conditions are specific to the cases when these operations are carried out on two data writing and reading media, LTO6 magnetic tapes and SAS and SSD technology disk volumes, were addressed.

As observed during the experiment, specific to the processes of writing and reading data from disk volumes is the fact that the implementation of duplicating technologies brings a significant improvement in the time of writing and reading saved data that can be obtained in disk volumes (Trdd), but above all it should be emphasized that what can be obtained is a reduction in the volumes of data saved in this type of storage medium (Vsdd). Reducing the volumes of data that must be saved and restored to avoid their loss in the event of disasters is thus one of the key factors in making these processes more efficient.

4. CONCLUSION

Current developments and the increasing dependence of intellectual property organizations on the proper functioning of information systems has led to the increase of cyber security measures that they must implement, the issue of the security of these systems becoming more and more important. Only by investing in comprehensive security solutions will we be able to have more secure information technology systems.

Some of the most important obstacles for a better and wider use of information technology in the fields of intellectual property include, among others, the lack (unavailability) of services based on information technology, the lack of interoperability of solutions across states members, as well as the fragmentation of this market of information space and technology-based solutions;

Information security is focused on ensuring the confidentiality, integrity and availability of computer data, regardless of their form. The security of information systems aims at guaranteeing both their availability and correct functioning, but also the possibility of recovering this data in the event of disasters and ensuring the continuity of the organization's operations. Network security focuses on the protection of data during transmission. The mechanisms and tools by which the protection of data stored or transmitted within an information system is ensured as a way of ensuring the confidentiality, integrity and availability of data. Certain procedures are also indicated to mitigate various types of security threats.

Information security is focused on ensuring the confidentiality, integrity and availability of computer data, regardless of their form. The security of information systems aims at guaranteeing both their availability and correct functioning, but also the possibility of recovering this data in the event of disasters and ensuring the continuity of the organization's operations. Network security focuses on the protection of data during transmission. The mechanisms and tools by which the protection of data whether stored or transmitted within an information system is ensured is a way to ensure the confidentiality, integrity and availability of data. Certain procedures are also indicated to mitigate. Information security is focused on ensuring the confidentiality, integrity and availability of computer data, regardless of their form. The security of information systems aims at guaranteeing both their availability and correct functioning, but also the possibility of recovering this data in the event of disasters and ensuring the continuity of the organization's operations. Network security focuses on the protection of data during transmission. The mechanisms and tools by which the protection of data whether stored or transmitted within an information system is ensured is a way to ensure the confidentiality, integrity and availability of data. Certain procedures are also indicated to mitigate different types of security threats.

Security is a major concern for knowledge-based organizations that process and store data electronically. The decision to keep their digital files in internal information structures, so that they can personally oversee their security, turns out to be the best option for maintaining direct control over data security. However, this decision also has direct implications on the necessary investments to be made both in the organization's informational infrastructure, but especially in the qualified human resource necessary to operate this infrastructure.

Security is a major concern for knowledge-based organizations that process and store data electronically. The decision to keep their digital files in internal information structures, so that they can personally oversee their security, turns out to be the best option for maintaining direct control over data security. However, this decision also has direct implications on the necessary investments to be made both in the organization's informational infrastructure, but especially in the qualified human resource necessary to operate this infrastructure.

The problem of managing the quality of computer data saved and restored in and from the backup systems implemented in intellectual property organizations is a topical issue to be analyzed in the context of the need to increase measures to increase the level of cyber security. In this sense, the analysis of the possibilities to improve the quality of the data is essential both

for identifying the factors that can influence the processes of saving and recovering data in electronic format and for reducing the risk factors that can alter these data.

Implementing a solid data rescue and recovery plan in case of disaster and continuity of operations of intellectual property organizations is an effective way to secure the data structures of these organizations.

5. REFRENCES

[1] Țîțu M., Oprean C., Boroiu A., applied experimental research in improving the quality of products and services, AGIR (2011).

[2] Savage, A., The Data Deduplication Handbook - Everything You Need to Know About Data Deduplication, (2016).

[3] Thwel T. T., Data Deduplication Approaches: Concepts, Strategies, and Challenges, Academic Pr. Inc. (2020).

[4] Feng D., Data deduplication for High Performance Storage System, Springer Singapore, (2022).

[5] Daehee K., Sejun S., Baek-Young C., Data deduplication for data optimization for storage and networks systems, Springer (2022).

[6] O'Reilly J., Network Storage, Elsevier Science & Technology, (2016).

[7] Rehman M. K., Problem solving and data analysis using Minitab: A clear and easy guide to six sigma methodology, Wiley, (2013).

Contribuții la modelarea și optimizarea soluțiilor de salvare a datelor în cadrul unei organizații moderne

***Rezumat**: În economia bazată pe cunoaştere, care este de natură digitală, datele sunt un factor reprezentativ pentru capitalul organizațiilor. Se poate spune că datele ar putea reprezenta un factor important de producție în domeniul bunurilor, dar mai ales în domeniul serviciilor digitale. Creșterea volumului de date procesate în cadrul organizațiilor poate duce la noi provocări în ceea ce privește managementul organizațional, dar mai ales în ceea ce privește metodele de optimizare a proceselor prin care datele sunt susținute în medii mai sigure pentru a asigura recuperarea cu succes în caz de dezastru. Lucrarea științifică prezintă, de fapt, calea către o evaluare a posibilităților de modelare și optimizare a proceselor de backup al volumului de date din perspectiva respectării standardelor internaționale de securitate a informației (ISO27001:2018). Lucrarea științifică are în spate cercetări aplicative și poate reprezenta un punct de vedere al autorilor argumentat pentru obținerea în domeniul tehnologiilor de backup al datelor informatice.*

Radu - Costin MOISESCU, Sc.D. Student, National University of Science and Technology POLITEHNICA Bucharest, Faculty of Industrial Engineering and Robotics, 313 Splaiul Independenţei, 6th District, Bucharest, Romania, e-mail: radu_moisescu@yahoo.com

Constantin OLTEANU, Sc.D. Student, National University of Science and Technology POLITEHNICA Bucharest, Faculty of Industrial Engineering and Robotics, 313 Splaiul Independenţei, 6th District, Bucharest, Romania, e-mail: ocosti@gmail.com

Petrică TERTEREANU, Sc.D. Student, National University of Science and Technology POLITEHNICA Bucharest, Faculty of Industrial Engineering and Robotics, 313 Splaiul Independenţei, 6th District, Bucharest, Romania, e-mail: tertereanupetrica@yahoo.com

Aurel - Mihail ȚÎȚU, Professor, Corresponding Author, "Lucian Blaga" University of Sibiu, 10 Victoriei Street, Sibiu, Romania, e-mail: mihail.titu@ulbsibiu.ro; The Academy of Romanian Scientist, 3 Ilfov Street, Bucharest, Romania