# IMPLEMENTING UNSTRUCTURED DATA SECURITY IN SMALL AND MEDIUM ORGANIZATIONS

*Management and Protection of Sensitive And Confidential Data*

[1]Virgílio Mendes Fijamo, [2]Ms Neha Chauhan
[1]Student, [2]Assistant Professor
[1]Computer Science Department,
[1]Alakh Prakash Goyal Shimla University, Shimla, India.

***Abstract :*** Protecting data from any organization is not an easy task. It requires immense knowledge and strategies on the part of data security engineers in information and communication technologies, because many times they require very high costs and resources. protect this very valuable asset for any line of business dealing with sensitive and confidential data and information such as credit card numbers, medical information, regulated contents of state or country secrets, many of which are confused between privacy policies and data protection policies, the two concepts relate to each other because they are based on confidentiality, availability and data integrity. In recent years, due to technological advances and the globalization of electronic commerce, government policies have also changed their data protection laws in favor of technologies, in a set of millions of files that can be found in a storage place like hard drives, pen drives and more makes it imperative to define what sensitive data is? Who has access to them? Who is using it and for what purposes? With these range of questions, you will be able to define appropriate policies for tracking and monitoring to provide reliable security suggestions. The framework used in this work analyzes the behavior of users and integrated systems that access data and can alert competent entities to bad breach behavior, and to protect files and email servers from cyber-attacks and internal threats by organizations' attempts to rule access control not allowed to sensitive data.

***IndexTerms*** **- Organization, Data, Security, Framework, Confidentiality, Sensitive**

## I. INTRODUCTION

The data in a computer system can be protected in two physical and logical forms, making it more difficult to execute the security logic because it fights an invisible attacker and because we do not know how to qualify the intention to execute with our data, and for our study to base in unstructured data that makes it even more complex to maintain security due to its veracity, volume, value, variety and speed by treating a part of big data, treating data that does not have a structural shape, our research focuses on classifying the data specified in a set of thousands of masses, due to this classification technique freed us from the base and the limits for those who have total or partial capacity to use data in an impaired organization. Classifying data reduces a significant number of risks of making confidential or confidential data available to any employee who may increase the risk of attack vulnerability, moving a very large number of data without categorizing it can incur some incalculable risk costs for our services, but also if the country's legislation does not allow the identification of personal information or the execution of arrests of violators in case of using data that are not difficult to use or measures of organization level that do not have application of protection laws of data and cybercrimes.

Nowadays, a very large number of data circulates in the areas of electronic commerce, military areas, health, education, finance, research and more, each area with its level of sensitivity of the data in the storage locations and its governance, allowing visibility permission, usable audit trail is the one who edits, deletes, copies and reads, test permission and analytical identification of the data owner.

In ways of synthesizing this research work, it presents a deep clarity in terms of concepts of unstructured data security in the computational areas or of information and communication tracked in the data in a reproduced, personalized and flexible way.

## 1. PROBOLEM

**Problem:** How important is it to monitor and track creating unstructured data security in Organizations?

## 2. HYPOTHESES

**Hypothesis 1**: Security of unstructured data minimizes the problems of the risk of data invasion in the hands of unauthorized persons.

**Hypothesis 2:** With a well-defined data security structure, it ensures greater reliability and efficiency in the use of the Organization's computational resources, increasing the productivity of its employees.

**Hypothesis 3:** The security of confidential and sensitive data minimizes the cost and time spent in solving problems that may affect the availability, integrity and confidentiality of the services offered by the organization to the minimum possible.

## 3. OBJECTIVES

### 3.1. GENERAL OBJECTIVE

Analyze the importance of creating security of sensitive and confidential data in a set of unstructured data in state institutions in Mozambique.

### 3.2. SPECIFIC OBJECTIVES

Find out how much the security of unstructured data minimizes problems in an organization, making it more stable and reliable, causing the minimization of costs and optimization of the productivity of the services provided.

Analyze the misuse of sensitive and confidential data in organizations to implement security mechanisms.

Study the use of the tool Varonis IDU Classification Framework for analyzing sensitive data in files or directories in any storage location, assessing the reliability and efficiency in detecting or verifying sensitive data sought by malicious attackers.

## 4. CLASSIFICATION OF DATA

Data classification is the process of categorizing and classifying data according to its variety, veracity and type or any other different class. The classification of data allows analysis and separation of sensitive and confidential data according to the specific type of business of the organization, since organizations are created to provide products and services [7] and whose organization can be of an economic or social.

Organizations of an economic nature are those that have a specific company character and seek a profitable purpose, where in their execution they take risks, pay legislative taxes according to the country and are directed by a CEO who is the financial director who directs by a philosophy of business and another of a social nature aimed at common actions of non-profit public utility more distant in humanitarian aids that can operate in any territory of a country that have a diplomatic effect in the social environment and supports. For our study, we will focus on organizations of an economic nature due to the risks that they may have in the invasion of confidential data.

To classify data, one must have a purpose and purpose for our case study is to maintain the security of sensitive data by minimizing the operating costs in its implementation, mainly in the process of managing organizational data.

### What is unstructured Data

According to the classification of the data, it can be structured, semi-structured and unstructured, for normally unstructured data, data that does not come from a pre-defined model, that is not easily tabulated, such as information found in a file text - but that can bring data about dates, accounts, different numbers, etc. and even audio and video files. In short it is all the data contained in the files used in the user's day-to-day activities. The other authors define unstructured data, which is data that does not necessarily have a format or sequence, does not follow rules and is not predictable. This type of data is currently receiving a lot of attention due mainly to the proliferation of mobile devices responsible for creating a wide variety of data. However, there are other data sources such as machine sensors, smart devices, collaboration technologies and social networks. These data are not related but diversified data. Some examples of this type of data are: Texts, videos, images, etc. [8]. The challenge in managing unstructured data is much more complex. The flow of information can be considered free and it is enough for a malicious user or a less attentive employee to copy the file for all security barriers to be ignored. Thus, working with a platform that indicates where sensitive data is hosted and who is accessing it is mandatory. You need to detect unauthorized behavior, the misuse of privileges or file escalation, corporate email and shared services, such as Active Directory. This data visibility must be presented to IT and security teams via detailed reports and records that contribute to meeting security and privacy compliance standards.

### Structured Data

These types of data contain an organization that is very easy to be recopied because it represents an organizational shape and structure such as rows and columns that identify different points of information, this type of data is easier to query using commands like SQL, more like this data found in structured databases.

**Table 1** - Differences between structured, semi-structured and unstructured data

| Structured Data | Semi-Structured Data | Unstructured Data |
|---|---|---|
| Predefined structure | There is not always a scheme | There is no scheme |
| Regular structure | Irregular structure | Irregular structure |
| Independent data structure | Structure embedded in the data | The structure is dependent on the Source of data |
| Reduced structure | Extensive structure (particular in each data since each one can have its own organization) | Extensive structure depends a lot on the data type |
| Little evolutionary and very rigid | Very evolutionary, the structure can change very often | Very evolutionary, the structure changes quite frequently |
| Has closed schema and integrity constraints | There is no associated data schema | There is no associated data schema |
| Clear distinction of the data structure | The distinction between data structure is not clear | It is not possible to distinguish between data structures |

### Unstructured Data Management

Data is everywhere in computer systems today and is an essential ingredient for any organization's business strategy. As the data is not created in the same way others may have come from the Internet of Things (IoT),[18] [23] its management is varied.

However, access to this data must always be monitored, tracked and protected. As organizations change their way of working to a more collaborative and modern modality, the number of files or folders grows exponentially, along with access control policies, and this content is usually stored in the cloud or on the local computer system due Internet costs, without proper protection and care to take on your vulnerability. Also, migration to the cloud presents an additional risk, after all, any consultation, alteration or exclusion needs to be done over the internet where most of the malicious people are to satiate their appetites by invading third party data. Despite the danger, this access is extremely important for the modernization of organizations' businesses managing these files used to be reactively easy in an environment with structured systems and applications. Today, with the proliferation of applications and mobile devices such as smart car, smart house, smart phone and others, user generated data includes a range of files that can include documents, reports, presentations, spreadsheets, PDFs, photos, basically archived from individually.

**Security Responsibility**

In order to maintain data security, all stakeholders in the organization are involved, employees in internal and external services from all areas involved in the organization, especially key sectors such as human resources, marketing, the IT department are more vulnerable due to their data sensitivity, only with these responsibilities will the security mechanism be able to function properly, taking into account that the biggest access door to internal information or data of a certain organization, there has been a facilitation of someone who knows all corners of the organization. However, in our analysis of what has happened in recent years, we can see that the lines between privacy and security are becoming more and more parallel; this does not mean that these concepts are the same, but that they are intricate.

Security consists of implementing the appropriate technical controls, such as multifactorial authentication, strong encryption and registration to protect data privacy, on the other hand, comes down to how this data is stored, accessed, and transmitted over the network or locally, its confidentiality applies.

**Data Security Impacts**

The big impact is on data leakage, which is a real and growing problem nowadays due to the huge increase and flow of digital information every month, news about confidential or confidential data leaks becomes public. These are the known cases, that is, that have a visible impact but many more similar incidents occur daily without being disclosed to the public, thus keeping the organizations secret so as not to create discrediting of the services of the organizations, and the vast majority of leakages. data is accidental where employees with little knowledge in the areas of information and communication technologies end up exposing the security of the organization, it is not only the result of intentional and harmful actions that prove many of the adversaries' business in the business due to economic or malicious interactions, they entice employees in money or promises of high leadership positions in your organization, making it easier for the adversary's data, employees less committed to the business enter the scheme. Unintentional data loss is perhaps the most dangerous, because those affected are not necessarily aware of or able to act on the problem due to the lack of financial conditions for the investigation of criminals. The loss of data can represent a very high cost for organizations; this failure generates direct and indirect costs to intellectual property or industrial information itself, in addition to the cost to deal with the consequences of the loss. Indirect losses include loss of credibility, erosion of competitive advantage and regulatory breaches in the market.

**Varonis IDU Classification Framework**

It is a powerful tracking and confirmation platform that can simulate access control changes for specific users of files, computational resources such as data, email servers etc. One of the mannequin features is to be able to alert you in the event of attempts to violate rules to allow access to data in storage locations as external and internal hard drives.

And manly can perform dependency checks and allow multithreaded work, scheduling tasks and reverting your changes in a few words. Manly IDU is a framework [9] for classifying data according to your level of sensitivity is sensitive and confidential data of any organization in the case of our study we deal with small and medium organizations. We can illustrate in the figure below the access control mechanisms in a SandBox.

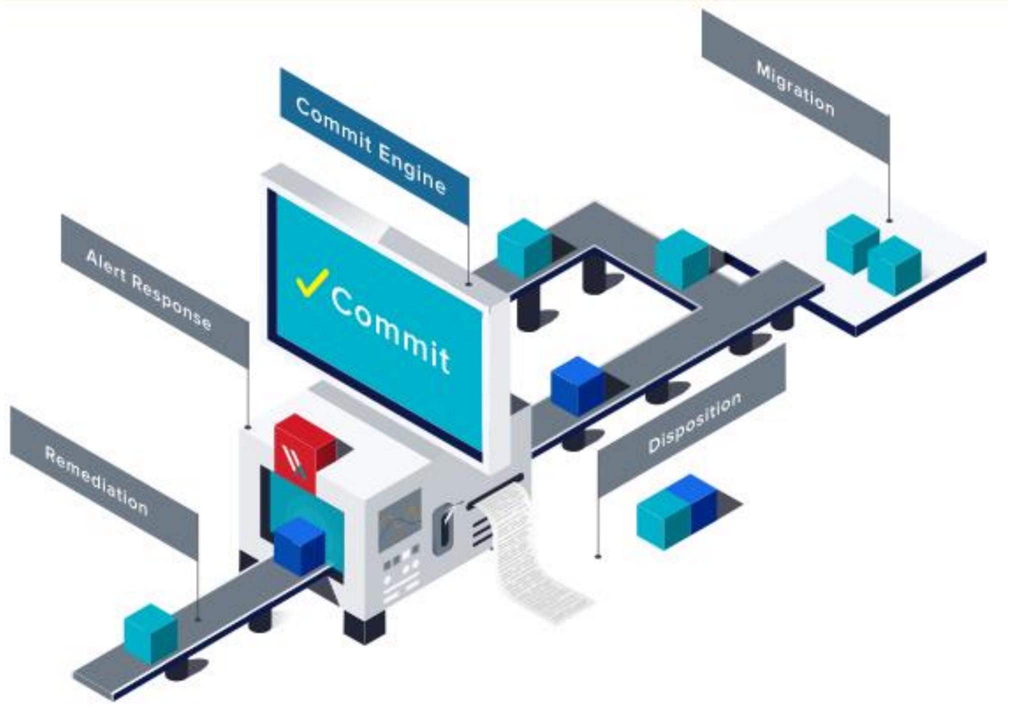


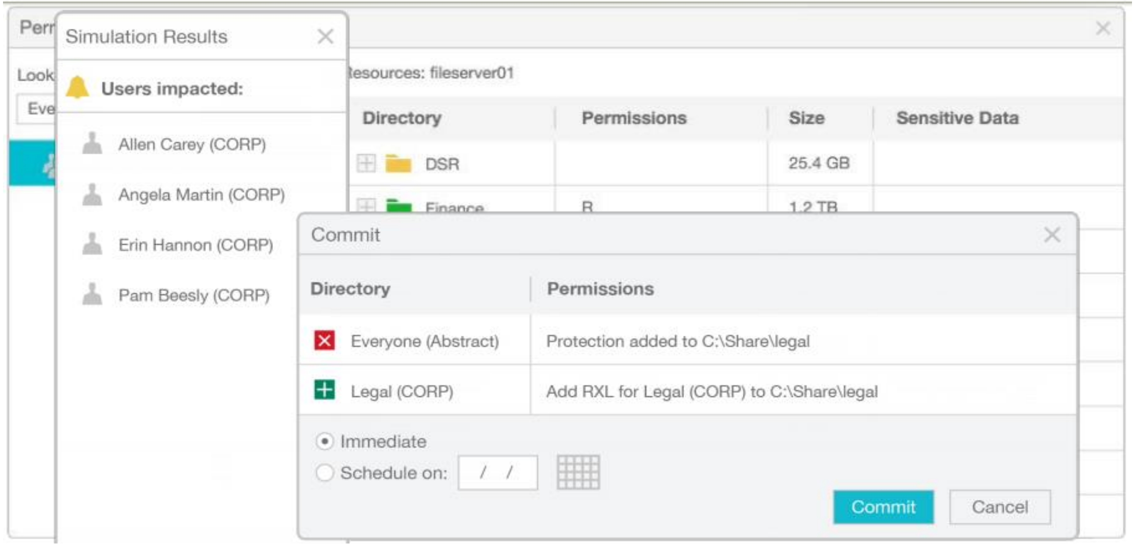
Figure.1 Confirmation Mechanism of acces

***Abbreviations and Acronyms***

SQL – Structured Query Language
XML – Extensible Markup Language
CEO – Chief Executive Officer
IT – Information Technology
IoT – Internet of Things
IS – Information Security
CIA – Confidentiality, Integrity, Availability
AI – Artificial Intelligence
PDFs – Portable Document Format
BI – Business Intelligent
NoSQL– Not Only Structured Query Language

**Screen to privileges**

A screen to eliminate global access for Varonis users, whose elimination can be scheduled or can have an immediate effect, discovering and revising permissions for critical folders in a Varonis IDU Classification Framework in a graphical environment is very easy where any users with administrative privileges can do it without having a strong knowledge of information and communication technologies, as we can illustrate the figure below in removing user privileges Erin Hannon.



Simulation Results

Users impacted:

Allen Carey (CORP)
Angela Martin (CORP)
Erin Hannon (CORP)
Pam Beesly (CORP)

Resources: fileserver01

| Directory | Permissions | Size | Sensitive Data |
|---|---|---|---|
| DSR | | 25.4 GB | |
| Finance | R | 1.2 TB | |

Commit

| Directory | Permissions |
|---|---|
| ☒ Everyone (Abstract) | Protection added to C:\Share\legal |
| + Legal (CORP) | Add RXL for Legal (CORP) to C:\Share\legal |

● Immediate
○ Schedule on: / /

Figure 2 Screen to eliminate global access

## II. RESEARCH METHODOLOGY

This research aims to create a knowledge in the small and medium Mozambican organizations to invest in the security of their data using Varonis IDU Classification Framework which is very powerful in maintaining the classification and tracking of unstructured data of the sensitive and confidential type in the storage hard drives or on servers and files, the main seven activities of which will be illustrated in the figure below of how this Framework works, but the study will focus more on content classification and support giving an alert in case of violation of access to privacy.

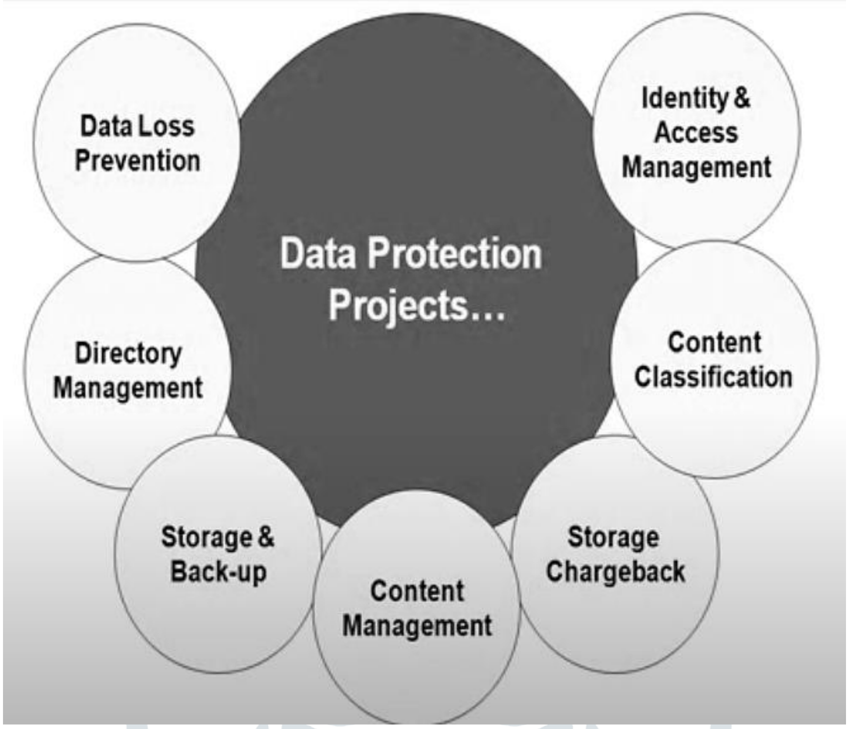


Figure 2 Data Protection Project Steps

According to [4] proposed XML data filtering mechanisms of a model not centralized on the internet and another in its similarity to the type of meta data, since this data model that the author made the filtering process is non-data structured, filtering a set of XML model documents involves cost and effect compared to the ideal number of data that will be useful for our purpose. Filtering a large volume of data requires a lot of knowledge in its handling until the final stage of processing in addition to very high computational resources on hardware.

This process of filtering content is very time consuming compared to the IDU framework of our study which is very fast and secure and with a variety of data in a graphical way.

### 3.1Population and Sample

A set of three (3) organizations was selected for direct observation for the use of monitoring and control of sensitive data in a universe of six (6) organizations. The criterion for choosing these organizations has to do with the diversity of the type of business, since it is about commerce, education and research that on average 75% of organizations are located in that country. What leads us to take a very small sample is due to the physical location of the province with 102 Km$^2$, and with little study on data security and its protection. The study composed by almost all government organizations, which still limited us a lot in challenging the implementation of security which involves additional costs for its implementation considering a very critical year 2020 where these organizations depend very fundamentally on the budget of the central government of the country.

### 3.2 Data and Sources of Data

For this study, primary data are collected in northern Mozambique in three categories, including the National for Economic Activities, the Ministry of Science and Technology and Higher and Technical Professional Education and the Cabo Delgado Provincial Directorate of Education and Human Development, then done between February and March 2020. And it detected an almost inadequate model of data storage in the stored storage locations subject to vulnerability.

.

### 3.3 Theoretical framework

Implementation of the structural model of deep information security structure [10], for security meters that meet the operational operations characteristic of individuals, and the basic principle for maintaining security, which is integrity, confidentiality and availability (CIA), delivered its limitation in comparison with the deep structural model that analyzes locks, padlocks, keys, passwords and firewalls for this purpose, it is necessary to implement a deep security structure due to its conceptual clarity, in keeping the personalization and flexibility in solving security problems.

In the analyzes made for this type of model, its effectiveness in tracking faults that may arise in a system that does not help to create information and communication security is shown.

Information security is provided by many researchers who meet the needs of risk analysts in early childhood education activities, followed by [11] characterized as information security approaches in three checklist categories, risk analysis and formed models and more later other categories added to include behavioral issues of the users of the site due to the advancement of information and communication technologies in the implementation of machines related to Artificial Intelligence (AI).

No scheme by Jiafu Jiang, Linyu Tang, no. all [14] the fog server is responsible for processing requests and resource allocations and the cloud audit center is responsible for auditing the performance of the new servers and fog nodes. The big challenge Based on the proposed security framework, our proposed scheme can resist or attack a single malicious node and attack fog servers and computing devices. In addition, experiments show that the scheme is efficient.

NoSQL and data repositories NewSQ[15] presents alternatives that can handle a large volume of data. Because of the large number and diversity of existing NoSQL and NewSQL solutions, it is difficult to understand the domain and even more challenging to choose an applicable solution for a specific task. Therefore, this article analyzes NoSQL and NewSQL solutions in order to [16]:

- ✓ Provide perspective in the field;
- ✓ Provide guidance for professionals and researchers to choose the appropriate data storage and
- ✓ Identify challenges and opportunities in the field.

Specifically, how more important solutions are compared with a focus on data models, queries, scaling and security-related features. Features that direct the ability to request read and write requests or dimensioning or storage of investigated data, in particular partitioning, replication, consistency and concurrency control.

In addition [17], use cases and scenarios in which NoSQL and NewSQL data repositories were used to suit various solutions for different sets of applications examined are discussed. Consequently, this study has specific challenges in the field, including the difference and inconsistency of terminologies, the restrictions, the usage restrictions, the spatial differences and the comparison and the lack of standardized query languages.

## III. RESULTS AND DISCUSSION

Varonis released the results of data collection carried out over a year in risk assessments conducted for potential customers in a limited number of its systems. The 2018 results show an impressive level of exposure for corporate files, including an average of 9.9 million files per assessment that were accessible to any employee in the organization.

Varonis Dat Advantage offers full visibility of who is authorized and in fact accesses unstructured files and data[22]. The Varonis Data Classification Framework identifies regulated and sensitive information, such as credit card numbers and medical patient data, and maps exposures in the file system. Even while remediation projects are being carried out, Varonis DatAlert can detect and stop internal threats, unauthorized privilege escalations and ransomware attacks, such as Cryptolocker.

From the insights generated from the dozens of risk assessments conducted during the year in small and medium-sized organizations, Varonis found that, on average, each organization had:

- ✓ 35.3 million files stored in 4 million folders - that is, about 8.8 files per folder;
- ✓ 1.1 million folders, or an average of 28% of all folders, open to all network users;
- ✓ 9.9 million files accessible to all employees of the organization, regardless of their position;
- ✓ 2.8 million folders, or 70% of all folders, containing obsolete data - untouched in the last six months;
- ✓ 25 thousand user accounts, of which 7,700 of them (or 31%) have not logged in in the last 60 days, suggesting that they are former employees, people who have changed positions, or consultants and employees who no longer provide services to the organization.

Leaving a file open for all users on the network is a common convenience when setting permissions. This massive access also makes it much easier for hackers to steal corporate data.

During the assessments, Varonis found some negative points that stood out:

- ✓ In one of the organizations, each employee had access to 82% of the 6.1 million folders;
- ✓ Another organization had more than 2 million files containing sensitive data (credit cards, social security and account numbers) with free access for everyone in the company;
- ✓ 50% of the folders of another company were allowed access for everyone, and more than 14 thousand files in these folders were sensitive;
- ✓ A single organization had more than 146,000 inactive users - accounts that had not been accessed in the past 60 days. The number is three times the average number of total employees in the Fortune 500 organization.

## IV. ACKNOWLEDGMENT

To my teachers who shared knowledge throughout the academic day. In a special way to Mr. Keshav Kishore, who is largely responsible for our learning in the programming language; Our Advisor Ms. Neha Chauhan, for her assistance in the construction of this work, her attention and disposition, regardless of the day and time, even from a distance, was always present, especially in corrections; To our new friendships conceived at the University. May they last as long as they were intense; finally, our most sincere thanks to my children, Aymane, Elizabete, Ganissa and family, for their strength and moral support.

"*The world is not threatened by bad people, but by those who allow the evil*". Albert Einstein

## REFERENCES

**[1]** Study Regarding Data Security and Safety in Small and Medium-Sized Companies

**[2]** Rodrigo Werlinger, Kirstie Hawkey, Konstantin Beznosov.Human, Organizational and Technological Challenges of Implementing IT Security in Organizations

**[3]** Yang Chen, Wenmin Li, Fei Gao, Wei Yin. December 2019. Kaitai Liang, Efficient Attribute-Based Data Sharing Scheme with Hidden Access Structures**.** The Computer Journal, Volume 62, Issue 12, Pages 1748–1760

**[4]** Georgia Koloniari, Evaggelia Pitoura.2004. Filters for XML-based Service Discovery in Pervasive Computing. The Computer Journal, Volume 47, Issue 4, Pages 461–474

**[5]** Manoj Thomas, Gurpreet Dhillon. October 2012.Interpreting Deep Structures of Information Systems Security. The Computer Journal, Volume 55, Issue 10, Pages 1148–1156

**[6]** Oluyinka. I. Omotosho. September- 2019.A Review on Cloud Computing Security**.** International Journal of Computer Science and Mobile Computing, Vol.8 Issue.9,pg. 245-257

**[7]** MARCONDES, José**.** April 25, 2019.Organization - Concepts, characteristics and types of organizations.

**[8]** Guolinag, L., Beng, C. O., Jianhua, F., Jianyoung, W., & Lizhu, Z. 2008. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. SIGMOD '08 Proceedings of the 2008 ACM SIGMOD, pp. 903-914

**[9]** https://www.varonis.com/customers/

**[10]** Manoj Thomas∗ and Gurpreet Dhillon.2011.Interpreting Deep Structures of Information Systems Security, Published by Oxford University Press on behalf of The British Computer Society

**[11]** Dhillon, G. and Backhouse, J. 2001 Current directions in IS security research: towards socio-organizational perspectives. Inf. Syst. J., 11,pages 127–153.

**[12]** Bishop, M. 2003Computer Security. Art and science. Addison-Wesley, Boston, MA.

**[13]** Bonomi, F., Milito, R., Zhu, J. and Addepalli, S. 2012. Fog Computing and Its Role in the Internet of Things. In Proc. of the First Edition of the MCC Workshop on Mobile Cloud Computing, Helsinki, Finland, August 17, pp. 13–16

**[14]** Jiafu Jiang, Linyu Tang, KeGu and WeiJia Jia.2019.Secure Computing Resource Allocation Framework For Open Fog Computing.

**[15]** Tudorica BG, Bucur C .2011. A comparison between several NoSQL databases with comments and notes. 2011 10th International Conference RoEduNet. IEEE:1–5

**[16]** Sadalage PJ, Fowler M .2013. NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Addison-Wesley, Upper Saddle River, NJ

**[17]** Aslett M.2011. How will the database incumbents respond to NoSQL and NewSQL.

**[18]** Shah-Mansouri, H. and Wong, V. W. 2018.Hierarchical fogcloud computing for IoT systems: a computation offloading game. IEEE Internet Things J., 5, 3246–3257.

**[19]** Ohlhorst FJ.2013.Big Data Analytics:Turning Big Data into Big Money. John Wiley & Sons, Inc, Hoboken, New JerseyUSA

**[20]** Dorothy E. Denning and Peter J. Denning,1979.Data Security, Computing Surveys, Vpl.II, No. 3

**[21]** Alireza Tamjidyamcholo and Rawaa Dawoud Al-Dabbagh.2012.Genetic Algorithm Approach for Risk Reduction of Information Security, International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(1): 59-66

**[22]** Stamp, Mark . 2006."Information security: principles and practice," Published by JohnWiley and Sons, Inc., Hoboken, New Jersey, Published simultaneously in Canada

**[23]** Yousefpour, A., Ishigaki, G., Gour, R. and Jue, J. P. 2018. On reducing IoT service delay via fog offloading. IEEE Internet Things J., 5, 998

**[24]** MYERS, Molinari, L.2008. "Testes Funcionais de Software". Ed. Visual Books. Florianópolis

**[25]** Alotaibi, A., Barnawi, A. and Buhari, M.2017. Attribute-based secure data sharing with efficient revocation in fog computing. J. Inform. Secur., 8, 203–222.