UNSTRUCTURED DATA SECURITY AND PROLIFERATION IN STORAGE LOCATIONS

Virgílio Mendes Fijamo[1], Ms. Neha Chauhan[2]
[1]*Student*
[2]*Assistant Professor*

***Abstract*—**Data security is a huge concern in many areas using information and communication technologies that even governments adopt laws that protect privacy and data security, as security can be seen in both physical and logical terms, physical security data can be handled with data site protection such as use of CCTV, guards, door locks, drones even the protection of natural disasters such as flood fires etc. and in the logical protection implementation of access control mechanisms, passwords, digital signatures, hashing algorithms and encryption that will be our center of study according to the literature review. Security of unstructured data is more difficult than structured data and is very important to ensure integrity, confidentiality and availability in communication channel computer networks, software and hardware computer security, and internet security such as streaming. Data flows from one point to another without being intercepted by third parties. Security is a process or cycle that involves auditing, risk, policy, implementation and administration in a routine manner, if security is one of the points of the CIA not considered security, but in computing having total security is almost difficult and involves costs and material resources. This bibliographic review aims to create a combination of various algorithms such as RSA, SHA-256, AES, DES, ECC, cryptography, to identify robustness, to later suggest the most viable in unstructured data security.

Keywords*—Cryptography, Data security, Algoritm(SHA-256),Integrity, Confidentiality, Rivest-Shamir-Adleman(RSA), Advanced Encryption Standard.*

I. INTRODUCTION

With the high use of unstructured data in the big data universe [7] in almost every branch of information and communication technology, there is a threat to the security of this data due to malicious people or systems searching for this data for sales or personal sales, the CIA (Confidentiality, Integrity, and Availability) [8] rule has been breached [8] that the fundamental principle of digital data security thus creating tremendous discomfort for IT professionals to adopt mechanisms such as more sophisticated modern cryptography algorithms for data security [15], which is an asset of organizations as well as the rulers of some continents such as Europe and America are implementing new data protection and data security laws for violators. The need to protect data and ensure integrity, availability, and confidentiality requires compliance with certain terms such as communication channel network security, hardware and software computer security, and ultimately internet security - all this is a process as illustrated at figure1. The figure above illustrates what a security system should do that is a cyclical process.

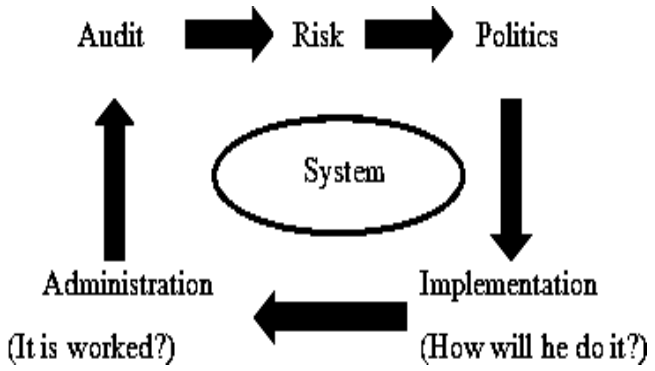


Figure 1: Security System Process

II. LITERATURE REVIEWE

Various studies and research conducted to improve data security at storage locations and communication channels using encryption algorithms and other logical security techniques. The researcher found the following study of literature as relevant in the proposed data security application.

A. R.Velummadhava Rao and K. Selvamani [1]

Focused on security issues and their solutions through encryption techniques. They suggest data security; the model comprises authentication, data encryption, data integrity, data retrieval, and user protection designed to improve data security in RSA-based encryption.

B. Noha MM. AbdElnapi, Fatma A. Omara and Nahla F.Omran [2]

Presented that the main purpose of using encryption is to fulfill primarily information security services namely integrity, confidentiality, availability, authorization and non-repudiation by comparing two types of encryption algorithms. Symmetric encryption that uses a single key to encrypt and dis crypt the most common are Data Encryption Standard (DES), Advanced Encryption Standard (AES), Triple DES, International Data Encryption Algorithm (IDEA), Rivest Cipher 4 (RC4) Asymmetric algorithms uses two public and private keys the different algorithms are RSA, Diffie-Hellman, Elgamal and others and Hybrid

algorithms and the combination of two or more encryption algorithms.

C. Rakhi Emelaya and DS Agrawal [3]

Mention that different types of encryption techniques such as AES, Blow fish, RC4, RSA, KB-ABE and hashing algorithms such as MD5, SHA-1, 2, 3 are provided to improve performance, time, encryption and decryption time. An optimized blow fish designed for longer key length is more secure, but encryption time and decryption time is slow. In order to overcome this problem in the blow fish algorithm Reducing two boxes (S) will increase speed and provide better data security.

D. Md. Ezazul ISLAM, MR and Shawkat Ali [4]

According to these authors considering two-dimensional security for volume and data variety, Big Data volume includes different varieties as structured, semi- structured and unstructured data and according to The study focused on unstructured data that has no fixed scheme such as XML, images, videos, audios and texts. The study concludes the application of conventional encryption algorithms such as 3DES, Snefru-256, CCM, Tiger, HMAC-SHA-1 for privacy, integrity and authenticity services that must be implemented according to data sensitivity to avoid creating overhead of system processing.

E. Oluyinka I. Omotosho [2], [5]

Both articles dealt with the combination of symmetric and asymmetric encryption algorithms [2] the researcher tried to classify the three types of cloud computing as public, private and hybrid and their services involved, software as a service, Platform as a service, and infrastructure as a service ceasing data on different devices and for cloud security used the Data Encryption Standard (DES) with 64-bit block size, and the comparison of AES that performs their calculations on non-bit bytes like DES and will mention the advantages and disadvantages of Blow fish algorithm and RSA.

F. Dorothy E. Denning and Peter J. Denning [6]

The study was based on the most common types of internal and external security breaches found by the investigated party and other breaches are not reported by banks in order not to create a bad reputation for customer service. The author concludes the intensification of access control, control flow, policy flow, inferences controls, cryptography control, DES (Data Encryption Standard) and asymmetric encryption algorithms to improve data security logically.

G. Amir Gandomi and Murtaza Haider [7]

According to this researcher who focused on unstructured data analysis, which constitutes about 95% of Big Data and they will highlight the need to develop appropriate and efficient analytical methods to evaluate large volumes of data heterogeneous big data. And the researcher makes a recommendation to create new tools such as efficient computational algorithms for predictive noise analysis in structured data, and this paper concludes the implementation of new statistical methods for classifying data to be considered big data.

H. Uma Somani, Kanika Lakhani and Manish Mundra [8]

The researcher attempts to evaluate the proper methodology for storing cloud data and its security by implementing digital signature with RSA Algorithm, addressing in both computer technologies grid computing and computing that underlies the development of cloud computing. Research [8], [1] discusses security challenges in the Confidentiality, Integrity, and Availability (CIA) model cloud, and in short, [8] gives a promising vision for reducing the infrastructure costs of large companies to choose to use cloud computing and future jobs of engineers in this area.

I. Domenico Cotroneo, Andrea Paudice and Antonio Pecchia [9]

Based on the sheer volume of security alerts displayed on monitors that allow system administrators to take data protection measures, [9] addresses consolidated filtering techniques to reduce the amount of security alerts generated when handling cloud data, and in turn concludes that the number of different alert types in a data set affects filtering performance.

J. Alireza Alireza Tamjidyamcholo and Rawaa Dawoud Al-Dabbagh [10]

Second points out that there are few publications dealing with data security risk management such as (NIST 800 - 30 and ISO / IEC 27005), for this reason. led to address risk reduction and risk assessment in uncertain environments by applying a GA (Genetic Algorithm), which considered treating risk reduction in this environment is laborious.

K. Jomin George and Takura Bhila [11]

There are sensitive areas such as health that data security such as the integrity and confidentiality of patient medical information must be taken into account to obtain satisfactory results of trust between Physician and patients. [11] Recommends that, *"to improve quality of health care, the health sector should undertake strong data and information protection measures outlined in the research"*.

L. Agnieszka Dardzinska [12]

Shows the possibility and importance of hiding certain attributes in a given tuple χ in the distributed information and communication system to ensure data security by making weighted value calculations within a given significance level limit λ .

M. Ravi Shankar Dhakar, Prashant Sharma and Amit Kumar Gupta [13]

Makes a comparison of Mod- ified RSA Encryption Algorithm (MREA) and RSA [1] which based on two mathematical problems factoring large numbers and attempting all possible private keys known as attacking brute force in terms of safety and performance. And [13] collusion *"MREA is safe compared to RSA because it is based on the factoring problem"*, just in terms of safety.

N. Mukul Gupta et al. [14]

Aims to present and evaluate a GA (Genetic Algorithms), code-based approach that enables organizations to choose the least cost investment and security that provides coverage of their needs. In conclusion *"Combining security technologies and vulnerabilities is a dual purpose problem"*, in this respect the researcher aims to maximize the degree of vulnerability and minimize costs in organizations in the security technologies to protect the assets that make up the organization.

O. Gurpreet Singh, Supriya [15]

Made evaluation of encryption algorithms [2] such as AES, DES and 3DES, a comparison was made for these encryption algorithms and AES is considered the best modern encryption algorithm in terms of speed and security [5], according to Kirchhoff, the security of the encryption system should depend on the secrecy of the key, not the encryption algorithm, regardless of symmetric or asymmetric encryption, however, the application of multiple algorithms increases the complexity of the time and space of the encryption system but system security becomes a reality. [15] Recommended to include more encryption algorithms and compression techniques to reduce memory requirements in execution.

III. METHODS OF SECURITY

In this section, we will show you the main security methods known as CIA, Confidentiality, Integrity and Availability, in confidentiality access to an organization's data should be restricted except authorized persons with a privilege level, Integrity ensure that the data is healthy without any changes with unauthorized people and Availability the data when requested must be available at all times to be used as a source of information that helps the decision making of any organization.

These principles are the basis for any data security but the author in this review suggests according to Figure 1 that it is the Process of cyclic security system Audit, Risks, Policies, implementation and administration, to point out that the correct combination of these structurally obeyed methods, significantly guarantees the corporate security of any organization's data, because they are routine procedures regardless of the type of data the organization deals with in our case unstructured data, most likely according to research from a company like IBM and the Ponemon Institute the most important cause probable data leakage is due to human error on the part of the users, which reaches up to 73.8% of any vulnerability that may happen in the organization.

A. Rivest-Shamir-Adleman (RSA)

In 1978, Adi Shamir, Leonard Adleman and Ronald Rivest developed an asymmetric cryptographic opera with two public and private keys that are used to encrypt, decrypt and generate file keys and now or RSA is a standard for all encryption algorithms and even including electronic transaction protocols such as Blockchain. The larger key size created by RSA and its decryption creates a computational overhead [9] and this remains a disadvantage of it. The following steps describe RSA encryption and decription [10]:

Choose random prime integers p and q approximately are the same size but not too close. Choose a smaller, random encryption exponent that has no factors in common with $p1$ or $q1$.

Calculate the product $n = pq$.

Calculate $\Theta(n) = (p - 1) \quad (q - 1)$.

Calculate $d = e - 1(mod\Theta(n))$.

A private key is kept secret or triple from integers (p, q, d). The encryption and decryption RSA algorithm scheme is illustrated as follows in Figure 2.

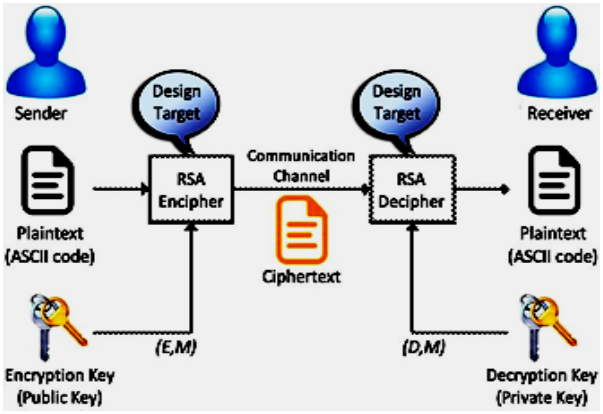


Figure 2: Encryption and decryption RSA algorithm scheme

An encryption function is $E(m) = m$ and $mod\ n$ for any message m. The description function is $D(c) = cd\ mod\ n$, where c is the cipher text. One public key all accesses are made by partners (n, e).

A private key is kept secret or triple from integers (p, q, d). The encryption and decryption RSA algorithm scheme is illustrated as follows in Figure 2.

B. Secure Hash Algorithm (SHA-256)

In simple terms, hash means the input of a sequence of any length and provides a fixed length output [9]. A cryptography hash function is a special class of functions

that has several characteristics such as Avalanche Effect, which means it even causes a slight change in your data entry, as changes that will be reflected in the hash will be affected an ideal for encryption in Algorithms. Hash functions (SHA-256) as hash functions are used in various distributed applications for password protection, cryptography password derivation, message authentication, and data integrity, digital signature such as these algo- rithms or their data outputs fixed is called message digest. The encryption of the hash function of the word "Virgilio786" is illustrated in figure 3 and the word "virgilio786" hash is shown in figure 4 to indicate that the *Avalanche Effect* [9] by changing only the uppercase letter "V" to the lowercase letter "v", the output is completely different which guarantees the security for decryption the message.

Figure3: Encryption message "Virgilio786" for hash

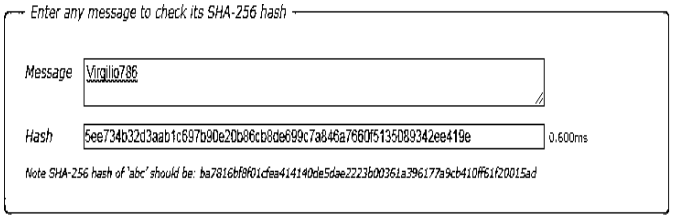
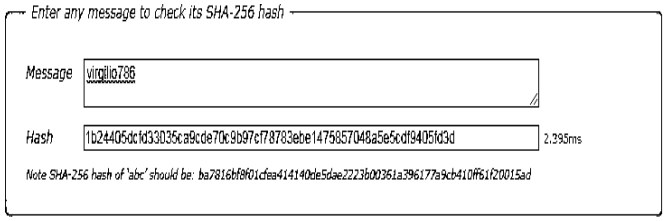


Figure 4: Encryption message "virgilio786" for hash



C. Advanced Encryption Standard (AES)

Advanced Encryption Standard (AES) is a three-block code intended to replace DES in 2001 for commercial applications and was used by the US government to protect classified information implemented in software and hardware worldwide for encryption of confidential information. Encrypting encrypts and decrypt the data and uses a block size of 128 bits and a key size of 128, 192, or 256 bits, respectively AES does not use a Feistel network structure, and instead, each complete set consists of four selected functions [16]: Byte substitution; Permutation; Arithmetic operations over a finite field, and XOR with a key.

The AES algorithm not only guarantees integrity, but less computational resource, which is the speed of operations used by the symmetric algorithm family is safe. Hardware and software implementation are faster AES.

D. The Data Encryption Standard (DES)

Data Encryption Standard (DES) has been the most widely used encryption algorithm until recently but is outdated [17] Displays Feistel's classic structure [16] DES uses a 64-bit block and a 56-bit key to Encryption and Decryption This completes the 16 encryption rounds in each 64-bit data block, in this block of which 8 bits are used for parity checking, two important methods of crypt analysis are: Differential crypt analysis and Linear crypt analysis.

DES has been shown to be highly resistant to both types of attack but does not show increased security.

E. Elliptic Curve Cryptography(ECC)

The addition operation in Elliptic Curve cryptography (ECC) is the counter part of modular multiplication in RSA, and multiple addition is the counterpart of modular exponentiation. To form a cryptography system using elliptic curves, we need to find a "hard problem" corresponding to factoring the product of two primes or taking the discrete logarithm.

Consider the equation $Q = kP$ where $Q, P \in E_p(a, b)$ and $k < p$. It is relatively easy to calculate Q given k and P, but it is relatively hard to determine k given Q and P. This is called the discrete logarithm problem for elliptic curves. Consider the group $E_{23}(9,17)$. This is the group defined by the equation $y^3 mod 23 = (x^3 + 9x + 17) mod 23$. What is the discrete logarithm k of $Q = (4, 5)$ to the base $P = (16,5)$?. The brute-force method is to compute multiples of P until Q is found. Thus, $P = (16,5)$; $2P = (20,20)$; $3P = (14,14)$; $4P = (19,20)$; $5P = (13,10)$; $6P = (7,3)$; $7P = (8,7)$; $8P = (12,17)$;

$9P = (4,5)$ Because $9P = (4,5) = Q$, the discrete logarithm $Q = (4,5)$ to the base $P = (16,5)$ is $k = 9$

In a real application, k would be so large as to make the brute-force approach in feasible.

COMPARISON OF DIFFERENT SECURITY ALGORITHMS

Table 1 illustrates types of some algorithm, symmetric, asymmetric, and hash function, computer system resource consumption, which are classified into low, medium, and high, privacy, integrity, and the maximum key size in (bits). According to this comparison for our study we will combine RSA and SHA-256 to have hybrid algorithm with maximum efficiency and robustness.

TABLE I. Comparison Table of Security Algorithms

Algorithms	Key Size (Bits)	Privacy	Integrity	Resource Consump - tion	Type of algorithms
RSA	15.360	14	yes	High	Asymmetric
SHA-256	1024	15	yes	Low	Hash Funtion
AES	256	—	yes	Low	Symmetric
DES	56	—	yes	Medium	Symmetric
ECC	512	yes	—	High	Asymmetric
OUR WORK				Low	Hybrid

CONCLUSION

From ancient times mankind has always been concerned with security of survival in natural disasters, ferocious animals and even conflict with other tribes, and in this age of information and communication technologies, with huge demand for circulation and storage of the famous big data. Organizations are vulnerable to threats of data theft or information from malicious people thus forcing them to invest their money in their security, and because of these concern researchers have unveiled their efforts in this area to create logical solutions such as creation of cryptography algorithms to stop this evil. Encryption as an approach to computer security comes at a cost in terms of resource usage, such as time, memory and CPU time, which in some cases may not be feasible to achieve the stated data protection objectives, and in our comparison of various researches we came up with ideas to make the combination of RSA algorithm which is the basis of encryption and most popular hash function for better data security of organizations.

ACKNOWLEDGMENT

Firstly to God who illuminated our way during this walk giving strength and insight so that the difficult moments were not moments of defeats; To teachers who shared knowledge throughout the academic journey. In a special way the course mentor Ms. Ayushi Nainwal great responsible for our learning to our Supervisor, Neha Chauhan, for assisting in the construction of this work, your 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; Lastly, our most sincere thanks to family members for their strength and support.

"God has made us perfect and does not choose the able, enables the chosen whether or not to do anything just depends on our will and perseverance." (Albert Einstein)

REFERENCES

[1] R. V. Rao and K. Selvamani, "Data Security Challenges and Its Solutions in Cloud Computing", Procedia Comput. Sci., vol. 48, no. Iccc, pp. 204–209, 2015.

[2] Noha MM. AbdElnapi, Fatma A. Omara and Nahla F.Omran, "A Hybrid Hashing Security Algorithm for Data Storage on Cloud Computing", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 4, April 2016.

[3] R. Emelaya and D. S. Agrawal,"A Survey Secure Data Storage Techniques in Cloud Computing", Int. J. Recent Innov. Trends Comput. Commun., vol. 3, no. 9, pp. 5376-5379, 2015.

[4] Islam, M.E., Islam, M.R. and Shawkat Ali, "An Approach to Security for Unstructured Big Data",A.B.M. Rev Socionetwork Strat, Springer, 10: 105, 2016.

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

[6] Dorothy E. Denning and Peter J. Denning, "Data Security", Computing Surveys, Vpl. II, No. 3, September 1979.

[7] Amir Gandomi and Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management 35, pp.137–144, 2015.

[8] Uma Somani, Kanika Lakhani and Manish Mundra, "Implementing Digital Signature with RSA Encryption Algorithm to Enhance the Data Security of Cloud in Cloud Computing", IEEE, 1st International Conference on Parallel, Distributed and Grid Computing, 978-1-4244-7674-9/10/, PDGC – 2010.

[9] Domenico Cotroneo, Andrea Paudice and Antonio Pecchia, "Empirical Analysis and Validation of Security Alerts Filtering Techniques", IEEE, Vol.16 Issue.5, Page(s): 856 - 870, Sept.-Oct. 1 2019.

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

[11] Jomin George and Takura Bhila, "Security, Confidentiality and Privacy in Health of Healthcare Data", International Journal of Trend in Scientific Research and Development (IJTSRD), pp.373-377, Volume-3, Issue-4, June 2019, www.ijtsrd.com .

[12] Agnieszka Dardzinska, "Optimization Algorithm and Data Security Problem in Distributed Information Systems", International IEEE Conference on Signal-Image Technologies and Internet- Based System, 978-0-7695-3122-9, 2007.

[13] Ravi Shankar Dhakar, Prashant Sharma and Amit Kumar Gupta, "Modified RSA Encryption Algorithm (MREA)", IEEE, International Conference on Advanced Computing & Communication Technologies pp.426-429, 978-0-7695-4640-7/12, 2012.

[14] Mukul Gupta, Jackie Rees, Alok Chaturvedi and Jie Chi ,"Matching information security vulnerabilities to organizational security profiles: a genetic algorithm approach", ELSEVIER, National Science Foundation grants, numbered EIA-0075506 and DMI-01222, pp.592–603, 2006.

[15] Gurpreet Singh, Supriya Kinger"Integrating AES, DES, and 3-DES Encryption Algorithms for Enhanced Data Security", International Journal of Scientific & Engineering Research, pp.2058–2062. Volume 4, Issue 7, July-2013, http://www.ijser.org.

[16] William Stallings,"Cryptography and Network security Principles and practice",Pearson Educa- tion, Inc., publishing as Prentice Hall,Fifth edition, pp.148-160, 2006.

[17] R.Sivakumar,B. Balakumar,V. Arivu Pandeeswaran,"A Study of Encryption Algorithms (DES, 3DES and AES) for Information Security",International Research Journal of Engineering and Technology (IRJET), pp. 4133–4137, Volume: 05 Issue: 04, Apr-2018, www.irjet.net .

Author: Engº Virgilio Mendes Fijamo
Born March 14, 1986, in the city of Quelimane Province of Zámbezia
He has a degree in Computer Engineering at Lúrio University, in Pemba, Mozambique
Math and Physics Teacher since 2011-2015, Computer Teacher since 2016-2017 at ISPS, Tete
And currently studying Post Graduation in Software Engineering at APG University, Shimla.