Measurement: Sensors 27 (2023) 100738

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Information security for big data using the NTRUEncrypt method

Mohammed Khalid Yousif [a], Zena Ez Dallalbashi [a], Shahab Wahhab Kareem [b,c,*]

[a] Department of Electronic Techniques, Mosul Technical Institute, Northern Technical University, Iraq
[b] Department of Technical Information Systems Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq
[c] Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, Erbil, Iraq

ARTICLE INFO

Keywords:
Big data
NTRU Encryption
Public key
Hadoop
Cloud computing

ABSTRACT

Cloud computing processes vast quantities of data and offers a variety of flexible, secure, on-demand, and cost-effective collaboration options for consumers. Due to the increasing prevalence of hosted services, data security has become an increasingly critical concern. Hadoop, the engine at the heart of cloud computing, causes serious problems for the cloud. Any public, private, or hybrid cloud environment can use this security solution without any hassle (IaaS). Furthermore, it is compatible with the vast majority of Cloud computing's capabilities. Increase cloud security using NTRU encryption. This study made advantage of the (NTRUEncrypt) algorithms residing in Hadoop to speed up the file encryption and decryption processes. If HDFS is engaged in the Map Task, then HDFS will take care of both the encryption and decryption processes. Data on the cloud can be kept private and secure thanks to the proposed protection technique, which makes use of cryptography. Combining the proposed technique with preexisting infrastructure and web-based.

1. Introduction

"Big data" in the context of digital information refers to a large quantity of data that is reported often but not usually from the same source. It is predicted that the quantity of potential digital data would expand by more than 40% in the space of a single season, contributing to the widespread nature of the current data flood. It is still the case that the vast majority of these datasets are created at no cost and in real time purely as a byproduct of the user's everyday activity [1]. Regardless of the availability of computational resources or storage infrastructure, "big data" [2] refers to datasets that are challenging to save and prepare using typical software tools. Statisticians consider Big Data as an example of data produced by computer systems whose primary objective is not mathematical inference but rather an end result. Mobile call detail records (MCDRs), Internet/text (website), GPS mapping data, social media (Twitter, Facebook, and Instagram), crowdsourcing data, Internet of Things (IoT), satellite and remote sensing data, and so on are all instances of Big data that are available to the public [3]. In view of the ubiquitous use of computers, the internet, and cloud computing, there is a growing consensus that security must be emphasized in order to give acceptable assurances of data availability, data integrity, and data confidentiality. The development of cloud computing ideas is intrinsically tied to every major improvement in machine learning techniques. Cloud computing has numerous benefits, notably for saving money and streamlining processes, but users are understandably concerned about the safety and privacy of their data while it is being stored and processed on the cloud. These concerns arise because of vulnerabilities in the form of things like hackers, insider threats, and protection holes [4].

The NTRU technique, together with the homomorphic features of the counting methods introduced in Ref. [5], is utilized to guarantee the secrecy and security of online votes. The goal of this research was to improve the efficiency of a classic universal re-encryption methodology by decreasing the size of the ciphertext and to create a way that works well with supply-constrained devices. The fundamental motivation behind [6] is to reduce the expense of Searchable encryption on Hadoop by cutting down the time required to encrypt vast amounts of data by preparing them in parallel. In Ref. [7], multiple methods were employed to provide better cloud security. These methods included RSA, AES, and DES.

2. The public-key cryptosystem of the NTRUEncrypt system

An alternative to RSA and elliptic curve cryptography (ECC) is the NTRU lattice-based public key cryptosystem NTRUEncrypt (also known as the NTRU encryption method) (which is not known to be breakable using quantum computers). The idea is predicated on the assumption

* Corresponding author. Department of Technical Information Systems Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq.
E-mail addresses: mohamed.khalid@ntu.edu.iq (M. Khalid Yousif), zeina.ez@ntu.edu.iq (Z.E. Dallalbashi), shahab.kareem@epu.edu.iq (S.W. Kareem).

https://doi.org/10.1016/j.measen.2023.100738
Received 23 December 2022; Received in revised form 20 January 2023; Accepted 2 March 2023
Available online 8 March 2023
2665-9174/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

that some polynomials in a truncated polynomial ring are difficult to factor into a quotient of two polynomials with very small coefficients. Lattice reduction presents a significant algorithmic task related to but distinct from cracking a cryptosystem. Selecting appropriate values for parameters helps prevent some of the documented attacks [5].

Contrasted with other asymmetric encryption algorithms like RSA, ElGamal, and elliptic curve cryptography, elliptic curve encryption and decryption are much faster because they simply require simple polynomial multiplication. However, in its currently deployed form, NTRUEncrypt has not yet been subjected to the same level of cryptographic scrutiny [6].

NTRUSign, a digital signature algorithm, is a related algorithm.

All polynomials R = Z [X]/(X^{N-1}) is a truncated polynomial ring where the convolution multiplication is performed. have integer coefficients, and the maximum degree of the polynomials in the ring is N-1. This is the basis for NTRU operations.

$$a = a0+a_1X+a_2X^2+a_3X^3+ \ldots +a_{n-2}X^{N-2}+a_{n-1}X^{N-1} \quad (1)$$

Instead, NTRU is described by four groups of polynomials and three integer parameters (N, p, q) in which N is prime, q is always larger than p, and p and q are coprime (a polynomial part of the private key, a polynomial for creating the public key, and a polynomial for generating the symmetric key).

A modern public key cryptosystem, NTRUEncrypt has only been around for a short while. The original NTRU system, created by three mathematicians in the mid-1990s, was dubbed by its acronym (Jeffrey Hoffstein, Jill Pipher, and Joseph H. Silverman). These mathematicians and Daniel Lieman created the NTRU Cryptosystems, Inc. in 1996, and they were granted a patent [1] on the cryptosystem that has since expired.

The recent decade has seen a lot of effort put into strengthening cryptography. Several tweaks have been made to the cryptosystem since its initial Four sets of polynomials (a polynomial portion of the private key, a polynomial for constructing the public key, and two polynomials that are coprime with N) and the properties that N is prime, q is always greater than p, and p and q are coprime, introduction to strengthen its security and boost its performance. The majority of the performance enhancements concentrated on increasing the rate of the process. There is published literature describing NTRUEncrypt decryption problems as recently as 2005. Concerning safety, with NTRUEncrypt 1.0, new parameters have been implemented that appear secure against all currently known attacks and a moderate increase in computational capacity.

IEEE P1363 standards for lattice-based public-key cryptography have been met, and the system is now approved for use (IEEE P1363.1). Mobile devices and Smart-cards can take advantage of the NTRUEncrypt Public Key Cryptosystem because of its speed (for benchmarking results, visit http://bench.cr.yp.to) and minimal memory utilization (for more on this, see below). For use in banking and other financial institutions, NTRUEncrypt was officially recognized as an X9.98 Standard back in April 2011. If Alice wants to send Bob a secret message, she'll need to create a public and private key. While Alice and Bob have access to the public key, only Bob has access to the secret one. Two polynomials, f, and g, with degrees no greater than N-1 and coefficients in [-1, 0, 1], are needed to produce the key pair. They are equivalent to representations in R of the residue classes of polynomials modulo X^{N-1}. In addition to satisfying the requisites for a polynomial f in Lf, the polynomial f must have inverses modulo q and modulo p (computed with the Euclidean technique), which means that f. fp = 1 mod p and f. fq = 1mod q. If Bob chooses an f that isn't invertible, he has to try again with a different f.

Bob's private keys are f and fp (and g). By solving equation (2) may derive the public key h.

$$h = pfq.g \bmod q \quad (2)$$

Alice sends Bob a secret message by encoding it as a polynomial m with coefficients in the range [-p/2,p/2]displaystyle [-p/2,p/2]. As it is, binary and ternary representations of message polynomials are both viable in modern encryption applications. After creating the message polynomial, Alice chooses a polynomial r at random with small coefficients (not limited to the set −1,0,1) in order to conceal the message. Using Bob's public key h, here is the encrypted message e:

$$e = r2+h+m/q \quad (3)$$

For privacy, Alice's messages are encoded in this cipher text before being forwarded to Bob.

Therefore, Alice cannot divulge r, as doing so would allow anyone with knowledge of r to calculate the message m by assessing e minus rh. Bob has knowledge of the private key in addition to the information that is already out there. How he can get m is as follows: To begin, he performs a multiplication using the encrypted message (represented by e) and a fraction of his private key (represented by f

$$"a = f.e \bmod q" \quad (4)$$

The following calculation is represented by this equation once the polynomials have been rewritten:

$$a = f(r(h+m)) / q \quad (5)$$

If a = f (rpfqg+m)mod q, then a = pr2+f2+m2 mod q.

Alice selects the coordinates of her message m in the interval [-p/2, p/2], therefore if she selects the coefficients of a between 0 and q - 1, the original message may not be successfully recovered. Because the coefficients of the polynomials r, g, f, and m and the prime p are all tiny in comparison to q, this means that all coefficients of (pr.g+f.m) already lie inside the region [-q/2, q/2]. When a message is reduced to a smaller size using the modulus q, all coefficients remain the same, allowing for a successful decryption of the original.

Since pr. g (mod p) = 0, we may proceed to the following step and determine a modulo p as follows:

$$b = a \pmod p = f.m \pmod p \quad (6)$$

By multiplying b by his second private key, fp, Bob may decrypt Alice's message.

$$c = fp.b = fp.f.m \pmod p \pmod p \quad (7)$$

Since the property f. fp = 1 holds, c = m (mod p) (mod p).

3. Big data

Big data is a different kind of world. Simply put, "big data" refers to a collection of information that is either too large or too complex to be stored in a traditional database table format, such as structured, unstructured, or semi-structured information. Among the various explanations prepared for large data, the author of this work investigates the composition's property, related, and structural durability. The notions of "Big Data" [7] will determine the future of the IT sector. New data from IDC demonstrates that social media sites produce enormous amounts of data daily. This includes 6.9 billion Google searches, 3.6 billion Instagram likes, 4 million hours of content posted to YouTube, 500 million tweets from Twitter, 5.7 billion Facebook likes, and 4.3 billion messages written on Facebook. Because of the trend that emerged after the "Data Explosion," we now have Big Data. The only thing that is constantly expanding and altering our culture is data [20]. Big data has been defined in various ways by some of the most well-known names in IT, such as EMC, IBM, and others. The characteristics of size, velocity, diversity, and value serve as defining characteristics of big data. In 2011, EMC and IDC collaborated on a study titled "Big Data Technologies," which explains how to "high-speed data capture, discovery, and analysis" to efficiently extract value from huge amounts of different data [8]. Concerns about the privacy and security of Big data have been the

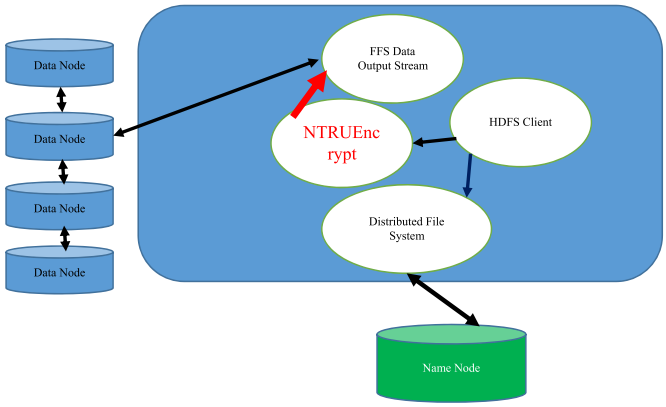

Fig. 1. Illustration of the HDFS encryption process.

subject of numerous news articles and scholarly studies in recent years (see, for example, Chen, Mao, Zhang, and Leung [9]; Goldfield [3]; Kambatla, Kollias, Kumar, and Grama [10]). Academics occasionally discover techniques for making sense of massive amounts of data from well-known American companies like Orbitz, Netflix, and Target.

Despite the prevalence of these and other big data security incidents, a recent assessment of the latest big data by Chen, Mao, Zhang, and Leung revealed major gaps in the literature. The "3Vs" were developed by Gartner in 2011 to describe the three primary aspects of Big Data.

Bigdata is a freely accessible data set with a size that can be measured in terabytes or even zettabytes, which is described by the Volume. The sheer volume of this data makes it difficult to store and analyze using standard visual approaches [11,12]. Data sets are heterogeneous in the sense that they include many various kinds of information, such as text, photos, audio, video, clickstreams, and log files [2,12–19,26–30].

Pingdom estimates that in 2012 there were 1.2 trillion Google searches, 1.3 Exabytes of mobile bandwidth consumed by mobile devices in a single month, 2.2 billion email users sent and received 144 billion emails per day, and 7 Petabytes of photo content was uploaded to Facebook. Credibility and usefulness have been added to the original 5Vs of big data [8,21–25].

4. Proposed algorithms

Estimates suggest that encryption must precede the storage of every file in HDFS [6]. The HDFS Client is in charge of the encryption process and depends on NTRU cryptography systems.

Refer to Fig. 1 for a suggested encoding and decoding setup. The NTRU is a coding scheme in which a random number is used to encrypt the plaintext. It is the public key that is used to implement encryption. Name Nodes in HDFS are responsible for storing the Metadata that governs the file system namespace and restricts client access to the encrypted file.

Data is collected in the HDFS after the suggested coding scheme has been applied, and then HADOOP files for HDFS are stored in batches. When a user requests data, the server will access encrypted data to perform the decryption, and the user can then recover the decrypted data using the private key [21–30].

5. Experimental results and analysis

They have been utilizing HDFS and Map Reduce to gauge the efficacy of encrypted HDFS on nodes equipped with an i7 core processor, 8 TB of hard disk, and 16 GB of memory. How long it takes the using the NTRU technique to encrypt the Hadoop-separated data files may be thought of as the encryption time, while decryption time can be thought of as how long it takes to convert the cipher text back into plaintext.

Data from a study that compared the NTRU, ElGamal, and Piallier cryptosystem across different file sizes is displayed in Fig. 2. The proposed method clearly showed efficient time usage in comparison to the RSA throughout the board, from 10 MB up to 5.12 GB at a step size increase.

Results of a comparison between NTRU and the Pillair and ElGamal cryptosystem for various file sizes are displayed in Fig. 2. For data sizes between 10 MB and 5.12 GB, with the step size increasing with each repetition, the suggested technique consumed less time than the RSA.

Therefore, the proposed approach (the Proposed algorithm) is faster than the traditional ElGamal and Paillier in completing the encoding stage. Fig. 3 displays the ElGamal and Paillier decoding procedure in action. We've been making do with encoded files of varied sizes up until now. The proposed method boasts quicker decoding speeds than the aforementioned algorithm.

6. Conclusion

Hadoop allows us to address large data challenges in businesses and government agencies, but the platform's security infrastructure is inadequate. An eavesdropper or other unwelcome third party may attempt to access data stored in Hadoop. Since Hadoop doesn't employ any sort of security controls, you never know if the data you're looking at is legitimate. Before being gathered in HDFS, the content of files is

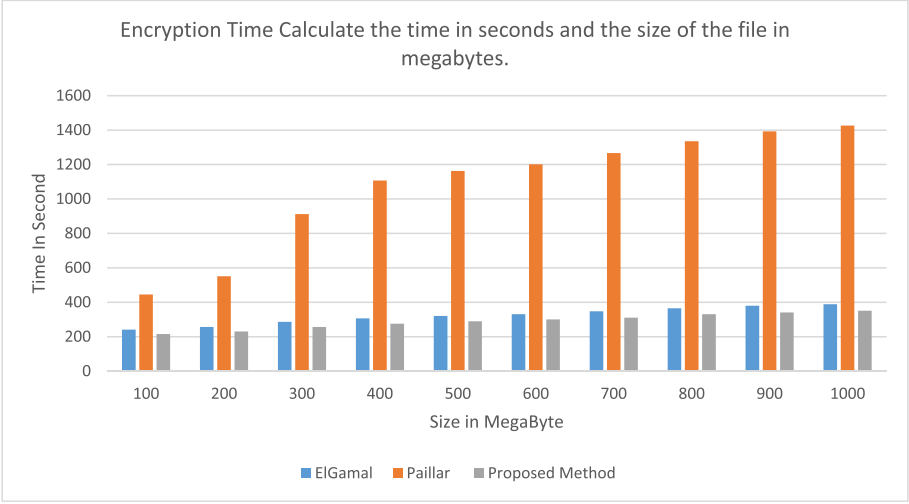

Fig. 2. Encryption time of ElGamal. Paillier and proposed method.

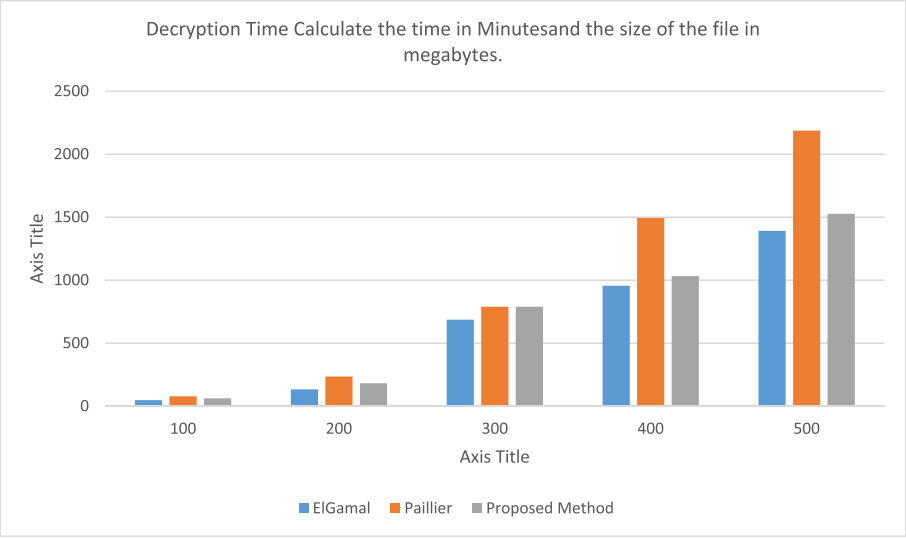

Fig. 3. Encryption time of ElGamal. Paillier and proposed method.

encrypted using the suggested asymmetric approach, making them protected from a wide range of network intrusions. Following the application of the encryption technique, data or files can be collected rapidly in Hadoop without concern about security issues. To name just a few examples, "platform as a service," "infrastructure as a service," and "software as a service" are three of the most prominent service models in a cloud computing system, and the proposed encryption method enables all of them (PaaS). The protection and management of data is given, as is the protection and management of transfer keys for sensitive data (authentication, integrity, availability, and confidentiality). When applied to files of varying sizes, the proposed method's encoding and decoding times skyrocketed, as did the amount of computational complexity required (twice as high in the decoding stages as in the encoding).

CRediT authorship contribution statement

Mohammed Khalid Yousif: Paper written, simulationsimultion, out put figures, related work. **Zena Ez Dallalbashi:** english editing. **Shahab Wahhab Kareem:** referencess arrangementarangement .

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] S.W. Kareem, Hybrid Public Key Encryption Algorithms for E-Commerce, University of Salahaddin–Hawler, Erbil, 2009.
[2] Danish Ahamad, M.D. Mobin Akhtar, Shabi Alam Hameed, A review and analysis of big data and MapReduce, Int. J. Adv. Trends Comput. Sci. Eng. 8 (1) (2019) 1–3.
[3] N. Goldfield, Big data—hype and promise, J. Ambul. Care Manag. 37 (3) (2014) 195–196.
[4] Huixiang Zhou, Qiaoyan Wen, A new solution of data security accessing for Hadoop based on CP-ABE, in: 5th International Conference on Software Engineering and Service Science, 2014.
[5] E. Jaulmes, A. Joux, A chosen-ciphertext attack against NTRU, in: 20th Annual International Cryptology Conference on Advances in Cryptography, 2000.
[6] Jeffrey Hoffstein, Jill Pipher, Joseph H. Silverman, "A Ring Based Public Key Cryptosystem.," Algorithmic Number Theory, ANTS III), 1998.
[7] Shadan Mohammed Jihad Abdalwahid, Raghad Zuhair Yousif, Shahab Wahhab Kareem, Enhancing approach using hybrid paillier and RSA for information security in bigdata, Applied Computer Science 15 (4) (2019) 63–74.
[8] Aditya Bhardwaj, Vineet Kumar Singh, Vanraj, Yogendra Narayan, Analyzing BigData with hadoop cluster in HDInsight azure cloud, in: Annual IEEE India Conference, INDICON), India, 2015.
[9] M. Chen, S. Mao, Y. Zhang, V.C. Leung, Open issues and outlook in big data, in: Chen (Ed.), Big Data: Related Technologies,Challenges and Future Prospects, vol. 1, Springer, 2014, pp. 81–89.
[10] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics, J. Parallel Distr. Comput. 74 (7) (2014) 2561–2573.
[11] S.W. Kareem, Secure cloud approach based on okamoto-uchiyama cryptosystem, J. Appl. Comput. Sci. Methods 14 (29) (2020) 9–13.
[12] M.M. Shetty, D.H. Manjaiah, Data security in Hadoop distributed file system, in: IEEE Int. Conf. Emerg. Technol. Trends Comput. Commun. Electr. Eng, ICETT, 2016, 2016.
[13] Goodubaigari Amrulla, Murlidher Mourya, Rajasekhar Reddy Sanikommu, Abdul Ahad Afroz, A survey of : securing cloud data under key exposure, Int. J. Adv. Trends Comput. Sci. Eng. 7 (3) (2018) 30–33.
[14] Rana M. Pir, Rumel M.S. Pir, Imtiaz U. Ahmed, A survey on homomorphic encryption in cloud computing, IJEDR 2 (2) (2014) 2173–2177.
[15] Raghad Z. Yousif, ShahabWahhab Kareem, Ammar O. Hasan, Design Security System Based on AES and MD5 for Smart Card, charmo university, Sulaimanyia, 2016.
[16] M. Bhandarkar, MapReduce programming with Apache Hadoop, in: International Symposium on Parallel & Distributed Processing, IPDPS), Atlanta, 2010.
[17] J. Gokulakrishnan, V.T. Bai, A survey report ON VPN security & its technologies, Indian Journal of Computer Science and Engineering (IJCSE) (2014) 3–5.
[18] Roojwan S. Ismael, Rami S. Youail, Shahab Wahhab Kareem, Image Encryption by Using RC4 Algorithm, EUROPEAN ACADEMIC RESEARCH, 2014, pp. 5833–5839, vol. Vol. II, no. Issue 4.
[19] Sourabh Chandra, Sk Safikul Alam, Smita Paira, Goutam Sanyal, A comparative survey of symmetric and asymmetric key cryptography, in: International Conference on Electronics, Communication and Computational Engineering, ICECCE), 2014.
[20] Majedah Alkharji, Hang Liu, Mayyada Al Hammoshi, A comprehensive study of fully homomorphic encryption schemes, Int. J. Adv.Comput. Technol. 10 (1) (2018) 1–24.
[21] Ms SnehaPatil, Vidyullata Devmane Pg, An implementation of online voting system using okamoto-uchiyama encryption scheme, Int. J. Comput. Technol. 17 (2) (2018) 7326–7334.
[22] Shweta Malhotra, M.N. Doja, Bashir Alam, Mansaf Alam, Bigdata analysis and comparison of bigdata analytic approches, in: International Conference on Computing, Communication and Automation (ICCCA2017), 2017.
[23] Kenneth David Strang, Zhaohao Sun, Meta-analysis of big data security and privacy scholarly literature gaps, in: IEEE International Conference on Big Data, Big Data), 2016.
[24] Rifki Suwandi, Surya Michrandi Nasution, Fairuz Azmi, Okamoto-uchiyama homomorphic encryption algorithm implementation in E-voting system, in: International Conference on Informatics and Computing, (ICIC), 2016.
[25] Venkata Narasimha Inukollu, Sailaja Arsi, Srinivasa Rao Ravuri, Security issues associted with big data in cloud computing, Int. J. Netw. Secur. Appl. 6 (3) (May 2014) 45–56, 2014.
[26] Shahab Wahhab Kareem, Yahya Tareq Hussein, Survey and new security methodology of routing protocol in AD-hoc network, in: QALAAI ZANIST JOURNAL, Erbil, 2017.

[27] Majedah Alkharji, Hang Liu, Homomorphic encryption algorithms and schemes for secure computations in the cloud, in: International Conference on Secure Computation and Technology, 2018. Virginia.
[28] V.G. Savant, Approaches to solve big data security issues and comparative study of cryptographic algorithms for data encryption, Int. J. Eng. Res. Gen. Sci. 3 (3) (2015) 425–428.
[29] Omar A. AlKawak, Bilal A. Ozturk, Zinah S. Jabbar, Husam Jasim Mohammed, Quantum optics in visual sensors and adaptive optics by quantum vacillations of laser beams wave propagation apply in data mining, Optik 273 (2023), https://doi.org/10.1016/j.ijleo.2022.170396.
[30] Kareem, Shahab Wahhab, Raghad Zuhair Yousif, Shadan Mohammed Jihad Abdalwahid, An approach for enhancing data confidentiality in hadoop, Indonesian J. Electrical Eng. Comput. Sci. 20 (3) (2020) 1547–1555.