Polytechnic Journal. 2020. 10(1): 81-87
ISSN: 2313-5727
http://journals.epu.edu.iq/index.php/polytechnic

POLYTECHNIC JOURNAL

RESEARCH ARTICLE

Enhancing Approach for Information Security in Hadoop

Raghad Z. Yousif[1,2]*, Shahab W. Kareem[3], Shadan M. Abdalwahid[3]

[1]Department of Applied Physics Communication, College of Science, Salahaddin University, Erbil, Kurdistan Region, Iraq, [2]Department of Information Technology, Catholic University in Erbil, Erbil, Kurdistan Region, Iraq, [3]Department of Information System Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil, Kurdistan Region, Iraq

***Corresponding author:**
Raghad Z. Yousif,
Department of Applied Physics, College of Science, Salahaddin University, Erbil, Kurdistan Region, Iraq.
E-mail: raghad.yousif@su.edu.krd

Received: 10 September 2019
Accepted: 17 June 2020
Published: 30 June 2020

DOI
10.25156/ptj.v10n1y2020.pp81-87

ABSTRACT

Developing a confident Hadoop essentially a cloud computing is an essential challenge as the cloud. The protection policy can be utilized during various cloud services such as Platform as a Service (PaaS), Infrastructure as a Service (IaaS), and Software as a Service (SaaS) and also can support most requirements in cloud computing. This event motivates the need of a policy which will control these challenges. Hadoop may be a used policy recommended to beat this big data problem which usually utilizes MapReduce design to arrange huge amounts of information of the cloud system. Hadoop has no policy to ensure the privacy and protection of the files saved within the Hadoop Distributed File System (HDFS). Within the cloud, the safety of sensitive data may be a significant problem within which encryption schemes play an avital rule. This paper proposes a hybrid method between pair well-known asymmetric key cryptosystems (RSA and Rabin) to cipher the files saved in HDFS. Therefore, before storing data in HDFS, the proposed cryptosystem is employed to cipher the information. In the proposed system, the user of the cloud might upload files in two ways, secure or non-secure. The hybrid method presents more powerful computational complexity and smaller latency as compared to the RSA cryptosystem alone.

Keywords: Big data; Cryptography; Hadoop; Rabin; RSA

INTRODUCTION

Cloud computing has brought growing awareness for the previous few years. Cloud computing presents users including a large area of sources such as computing principles, computing power, storage, and internet applications. Cloud computing is instantly being utilized in a very large volume in several areas. In way of life, huge amounts of knowledge are produced. Consumers use cloud computing services to gather the aforementioned extremely huge amount of knowledge. A number of the important difficulties cloud computing faces are to secure, guard, and prepare the information that is the user's estate (Merla and Liang, 2017). Big data requirement is also involved the mix of very sensitive and valuable data from social sites and therefore the effects of presidency and safety. This accumulated data should be ciphered by utilizing an acceptable method to secure them. The characteristics of massive data will be classified in terms of 4 V's (Hilbert, 2016): Volume, velocity, variety, and veracity. Every subject has its job of resting in big data. Thus, volume: The quantity of produced data and the data saved in it which is in the level of several terabytes or petabytes in size. Variety: This is often the outline of knowledge and its applications, unstructured, structure, and semi-structure. Velocity: This suggests the input and therefore the output speeds of knowledge currents produced and saved within the system. Within this context, a plan is presented in an exceedingly form that big data operations can ultimately, store data separately of the incoming or outgoing rate. Veracity: It is the term of information quality, this context is additionally applies to data privacy, integrity, data confidentiality, and availability. Enterprises requirement grantee that the information and also the analyses transferred on the information is precise. Criminal societies produce cover businesses to the property and buy of hijacked secret information (Marti et al., 2011). Government intelligence services trust individual, corporate, and adverse government eavesdropping and aggressive power systems. Moreover, protection across cloud services is in its growing stage, a large amount of protection vulnerabilities would risk data within the cloud. Furthermore, analyzing/processing large data at the cloud data center could be a significant problem. Several issued frameworks like HADOOP have newly been available (Li et al., 2015), like Google filing the system (Yang et al., September 9-11, 2013, pp. 437-442) which is created to gather and treat the large data. Nevertheless, the distributed HADOOP framework is common with business and analysis centers. HADOOP involves couple assortments of functionalities, (i) for warehouse of huge and unstructured data sets (HDFS), has been applied, and (ii) MapReduce framework for hug data administration.

HADOOP regularly operates including statements that own thousands of information nodes and petabytes. As a quest survey, Yang et al. (Yang et al., September 9-11, 2013, pp. 437-442) recommended a triple encryption scheme for improving the protection of Hadoop. Thus, the encryption of HDFS files is performed through utilizing data encryption algorithm (DEA), whereas RSA has been employed in the encryption of information key. Ultimately, the RSA private key's secured using the concept (International Encryption Algorithm). Huixiang et al. (Huixiang et al., 2014), they perform CP-ABE (Ciphertext policy attribute-based encryption) scheme for access control rather than standard schemes like PKI, which wants all related consumer's data to be transferred to the source provider, thus damaging the secrecy of the user, and uses more bandwidth and overhead processing. Masoumeh et al. (Masoumeh et al., 2014) showed the reason that currently, the middle technology of cloud computing is services protection and data secrecy. A security tool trusted Kerberos protocol for authentication firewalls of perimeter level security was performed. Security leak was controlled by performing the Apache Sentry for access control, triple encryption of knowledge using RSA, DES, and IDEA algorithms, was proposed in protecting classification system trusted fully homomorphic encryption. Parmar et al. (Parmar et al., 2017) proposed a brand new method which will be utilized to secure Hadoop, an economical technique that works within the Hadoop cluster to offer it 3-D security Usama and Zakaria (Usama and Zakaria, 2017) proposed encryption and compression for Hadoop. HADOOP does not combine security tools. The appliance of the ciphering method in HADOOP encryption later saving them at HDFS has been published in several works. Ciphering schemes perform several replacements and do some guidance on the clear message to converts it into ciphertext which must be random and meaningless. Various ciphered systems were formed and are used for the sake of knowledge security. Therefore, the couple's main categories are as follows: (i) Symmetric key cryptosystems (Sourabh et al, 2014) such as encryption standard (DES), triple DES, and advanced encryption standard (AES), (ii) asymmetric key algorithms (Sourabh et al, 2014) such as RSA and elliptic curve Diffie–Hellman. The proposed method is often deemed essentially a trial to develop what was performed by the paper (Kareem 2009) at both of encipherment/decipherment methods for securing files of massive data-based Hadoop-integrated AES and OTP algorithms (Hadeer et al., 2018). An architecture to secure Hadoop was examined in paper (Park and Lee,, 2013). Thus, for encryption and decryption, AES encryption/decryption classes are added. Complete two HDFS-RSA and HDFS matching integrations (Shetty and Manjaiah, 2016) applied as some differing kinds of extensions of HDFS. Analyses confirmed a suitable overhead for reading operations and significant overhead for writing operations (Yang et al., September 9-11, 2013, pp. 437-442). The subsequent is the organization of this paper: Section II outlines the protection framework. Section III, supported HDFS and MapReduce, presents the massive data at HADOOP. Section IV discusses the proposed optimized hybrid encipherment algorithm and compare it with the classical public key cryptosystems before apply it to secure big data at HADOOP. Section V presents the discussion of the simulation results, Finally, section VI lists the conclusions.

SECURITY ISSUES

Technology does not own the boundary even after continued to significant – high inside the latest six to seven decades. While technology upgrades, the crime also grows within each area. To bypass the aforementioned type of crimes, users must to meticulously preserve working the common techniques such as holding further cooperation, modernizing the antivirus during various periods, implementing the security by investing firewalls, and focusing on parental controls. Nevertheless, data confidentiality is not warranted also sometimes data could be seized also in regular computer protection. There is an alternative method of solving this set of cybercrime difficulties which means identified as cryptography technology. This is the highest technology that can take command of both nodes and mediums and does not release the hacker or theft to steal the information without the security key. This cryptography technology performs a vital role both on the sender and receiver sides while encrypt and decrypts the data as well. There are a lot of cryptography algorithms that exist for encrypting and decrypting the data during data transmissions in the past two decades. Even though cryptography is developed for producing strong protection within data transmission, there are many cybercrime problems that have been issued including the enhancing of advanced technology and the identical is adjusting to using for malfunctions. Apparently, the technology which held obtained for implementing protection in communication began to meet slightly collapses. Without improving technology, it is not reasonable to evade those types of issues (Bhandarkar, 2010) (Raghad et al., 2016) (Roojwan et al., 2014).

Given the intensified usage of the internet, computer, and cloud computing technology, protection estimates essentially an excellent requirement to guarantee integrity, confidentiality, and convenience of the data systems supplies. The increase in computational methods for machine learning has combined with the development of cloud-based computing principles. Notwithstanding

the effective computing solution and commercial benefits compared with cloud computing, users are very concerned about the protection and confidentiality of data collected and prepared in the cloud. Those safety concerns are affected by amazing protection opportunities such as insider threats, protection gap, and possible hackers (Shahab and Hussein, 2017).

Cryptography permits a user to interact on the internet, carrying important and private knowledge securely. Accordingly, cryptography authorizes people to utilize unrestricted or restricted tools such as the internet to arrange online purchasing and avoids living victims of offenders and password sniffers. This is performed by utilizing the most advanced technological progress in computer science. Cryptography, additionally, recognized as cryptology, so it supports users and organizations to cipher and deciphers protected information within codes, numbers, and ciphers so data can be transferred securely. Cryptography is admitted by encryption and decryption keys. The method of coding and conversion of traditional text into the unreadable format is named encryption; while the method of decoding and switching the unreadable text to readable information using a specific digital key is called decryption (Gençoğlu, 2019). A couple of important methods in cryptography that are utilized to transform information into encryption are symmetrical and asymmetrical cryptography. During asymmetrical cryptography, a couple of digital keys are utilized by the end user. One digital key is applied to encryption while another is selected for decryption. These digital keys are named public and private keys. Both keys are different from each other. Hence, the common sense is that asymmetrical cryptography is honestly secure and protected. One of the methods utilized in asymmetrical cryptography is the assignment of a key to a special type of data. Another interesting idea during asymmetrical cryptography is the method of a random digital key specified by the public key keeper or the sender. It is also called pair digital keys that must be used to encrypt and decrypt the information (Gençoğlu, 2019).

BIG DATA AND HADOOP

Parallel computation of cryptographic systems on multi-core computing policies is often a hopeful approach to decrease performance time and from the facility consumption of certain algorithms. Hadoop structure consists mainly of two primary elements which are HDFS to gather big data and MapReduce to research big data. HDFS may be a file administration system utilized for the shared warehouse of huge datasets on the Hadoop group in with a default chunk size of 64 MB (Aditya et al., 2015). After collecting the input files in HDFS, then it managed with MapReduce software, eventually, the results are transferred to the output folder of HDFS (Dubey et al., 2015).

MapReduce is that the kernel scheme utilized by the Hadoop engine to share a cluster of labor. The computer file, which occupies during the cluster on a shared filing system, is split into collections of equal size to market and clarify in an exceedingly fitting, and comparatively error-free use the big volumes of knowledge processing in correspondence on large clusters of hardware. As defined by the name, MapReduce involves two stages of knowledge computation in Hadoop, the primary stage is that the map and also the second stage is reducing, that is, an outsized volume of knowledge collections is transformed into structured key value sets and provided as inputs (Dubey et al., 2015). Figure 1 presents the MapReduce estimate data flow. The mapper does not communicate immediately to the disk just catches the advantage of buffering the writings. Each mapper has annular memory buffer including a default 100 MB size which will be adjusted through turning the characteristic of (io. sort. mb). It offers a resourceful flush. While the buffer is loaded up to a particular threshold, it starts dropping the content of buffer to the disk. Before the spill occurs to the disk, the thread separations the info in keeping with the reducers it requires traveling background thread executes a range of in-memory inside the key-based partition before the spill takes place to the disk. If a combiner is started, it applies the output of the in-memory sort (Dubey et al., 2015).

PROPOSED ALGORITHMS

Hadoop is that the principal provider of large-scale cloud processing and warehouse, it, since, applies amazing methods of encryption to ensure protection. The paper

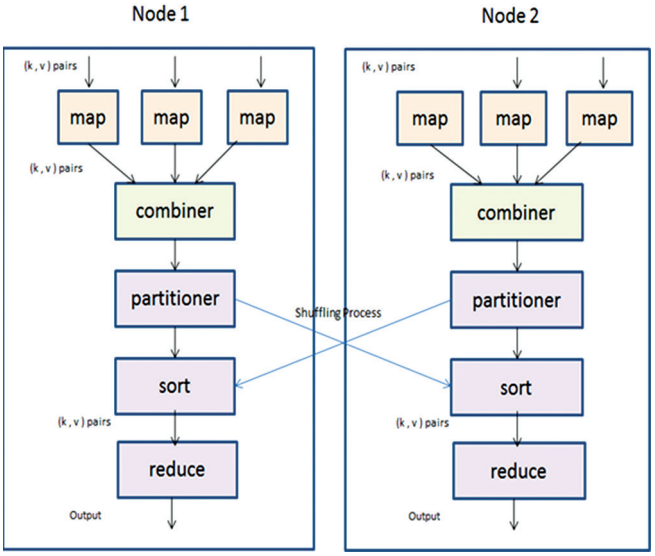

Figure 1: Data flow of MapReduce computation (Jeffrey and Sanjay, 2008)

presents a brand new method. The strategy relies on cascading two public key cryptosystems (RSA and Rabin). Hybridization's thought to beat the conditions of practicing specific cryptosystem alone and to extend protection. It is contemplated that each one the records corresponded to HDFS shall be earlier ciphered. The HDFS customer is qualified to key generations (public and personal keys). Then, the recommended hybrid method is applied to cipher the file while buffering it to HDFS practicing an unstructured file. The HDFS begins transmitting the cipher files to the info nodes. These steps are presented in Figure 2. HDFS consists of a reputation node that saves metadata which controls the name space of the filing system and controls clients' access to the encrypted file. The encrypted file is produced from one or more blocks collected in a very collection of knowledge nodes. The recommended hybrid method is printed in Figure 3. From Figure 3, the keys (public and private) creation model relies on the mechanism utilized by the RSA. The powers of factorizing huge numbers could change the protection afforded by the RSA algorithm. Cryptosystem and it is represented by Algorithm 1.

Algorithm 1: Key generation of proposed algorithm

INPUT: Select large random prime numbers p and q

OUTPUT: A public key, (n; e), and a private key (p,q,d)

User A sends the message to user B.

1. Generate two large random (and distinct) primes p and q, each roughly the same size.

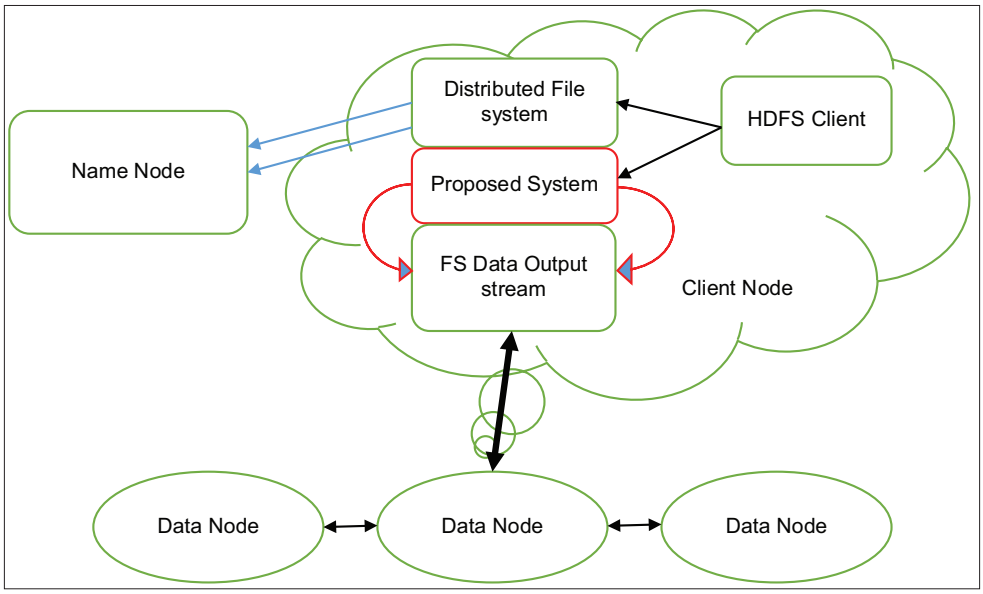

Figure 2: Encryption procedure in HDFS (Shadan et al., 2019)

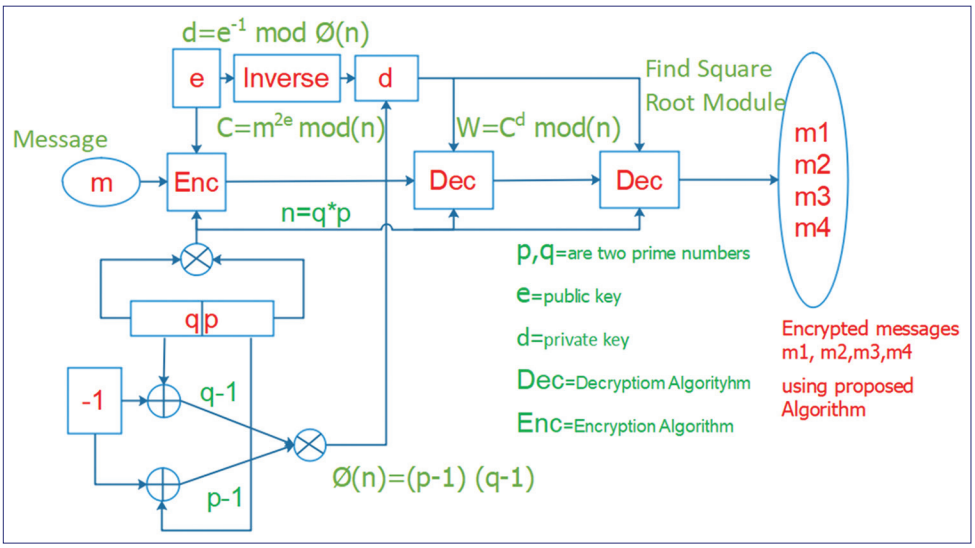

Figure 3: Process of hybrid public key algorithm (Kareem, 2009)

2. Compute n = p*q and Ø = (p−1) *(q−1).
3. Select a random integer e, 1 < e < Ø, such that gcd (e; Ø) = 1.
4. Use the extended Euclidean algorithm to compute the unique integer d, 1<d< Ø, such that ed≡1 (mod Ø).
5. User's public key are (n; e); user's private key is(d, p, q).

The encryption process is also based mainly on RSA encryption procedure but here the clear text or the message (m) is raised to power 2e rather than e (where e is the public key) as in RSA cryptosystem, hence, the ciphertext c is calculated according to Equation (1):

$$c = m^{2e} \bmod(n) \quad (1)$$

Algorithm 2 shows in detail the encryption procedures:

Algorithm 2: Encryption process of the proposed algorithm

INPUT: Plaintext to encrypt and receiving user's public key (n; e).

OUTPUT: Encrypted ciphertext.

User A sends the message to user B.

To encrypt B should do the following:
a. Obtain A's authentic public key (n; e).
b. Represent the message as an integer m in the interval [0; n−1].
c. Compute C = (m^{2e})mod n.
d. Send the cipher-text c to A.

The decryption process implements the same scenario of RSA decryption in which the ciphertext is raised to the power of the private key (d) but the output is not the direct message (m) but the message raised to power 2 according to Equation (2):

$$W = m^2 = c^d \bmod(n) \quad (2)$$

To recover the message m, four messages are generated m^1, m^2, m^3, or m^4. Hence, the correct plain text is one of them. This procedure is explained in Algorithm 3:

Algorithm 3: Decryption process of the proposed algorithm

INPUT: Received encrypted cipher-text and the receiver's private key a.

OUTPUT: Original plaintext.

To recover plaintext m from c, B should do the following:
a. Use the private key d to compute W=cd mod n.
b. To find the four square roots m^1, m^2, m^3, and m^4 of W modulo n.
c. The message sent was either m^1, m^2, m^3, or m^4.

After utilizing the recommended method, data are collected in the cloud. Therefore, through HDFS, data will be collected in a cluster. Whenever the user demands data, the server will begin the encrypted data to the decryption procedure. The user then uses the private key to recover the decrypted data utilizing a hybrid method which is the proposal of this paper.

EXPERIMENTAL RESULTS AND ANALYSIS

The MapReduce and HDFS are utilized for the performance evaluation of encrypted HDFS. Each node has i3 core, 4 processors, 4 GB of memory, and 750 G of the disk. Encryption Time: The time is practiced by the RAS alone or the hybrid algorithm to cipher the Hadoop divided dataset files into ciphertext employing a key. It is determined in seconds. Decryption time: The time is practiced by the RAS alone or the hybrid method to decrypt the Hadoop separation dataset files following into the plaintext using the private key. It is measured in seconds. Thus, the encryption time is analogous to the system current time before encryption deducted from it the system modern time after encryption. Whereas, the decryption time is similar to the system current time before decryption subtracted from it the system current time after decryption. Table 1 describes the implications of the association between encryption schemes, the RSA alone, and also the Hybrid system with various file sizes. It is clear that the proposed method showed effective time consumption compared to the RSA for file size stars from 100 MB and ends with 1 GB with a step size of 100 MB. Furthermore, the proposed method (hybrid system) within the encryption

Table 1: Running time in second for encryption

File size in MB	RSA	RSA modification proposed methods in second
100	220.5882	203.505
200	215.9926	199.2653
300	425.7813	392.807
400	394.0717	363.5532
500	442.3254	408.0699
600	473.3456	436.6878
700	488.2813	450.4668
800	558.0357	514.8192
900	627.7902	579.1716
1000	697.5446	643.524

Table 2: Running time in minutes for decryption

File size in MB	RSA	RSA modification proposed methods
100	47.64705	42.69321746
200	133.6091	119.7178495
300	686.201	614.857132
400	955.9391	856.5507387
500	1391.367	1246.707485

Table 3: Computational complexity of the proposed method, RSA and Rabin

Method	Encryption	Decryption
RSA	$T(c)=O (\log n)^3$	$T(M)=O (\log n)^3$
Rabin	$T (c)=O (\log n)^2$	$T(M)=O (\log n)^3$
Hybrid system	$T(c)=O (\log n)^3$	$T(M)=2 O (\log n)^3$

stage is quicker than the default RSA. Table 2 shows the period for RSA and also the proposed method within the decryption stage. The encrypted files applied to the present stage are of various sizes. By applying both RSA and also the hybrid system (the proposed method), it is obvious that the decryption time needed by the hybrid ciphered method is shorter than that is needed by RSA. Table 3 shows the computational complexity of the hybrid cipher (proposed method) with RSA and Rabin cryptosystems from which it is clear that the proposed method has doubled the computational complexity as compared to the individual systems (RSA or Rabin).

CONCLUSION

While Hadoop enables us to beat the difficulties encountered by big data in businesses and organizations, it is no protection tool. An intruder or snoop may settle the info collected in Hadoop. The authenticity of information is continually at stake. Before saving the info in HDFS, the proposed hybrid cipher asymmetric key algorithm encrypts the content of the file by achieving it of various network intrusions. The files or data can, therefore, presently be saved in Hadoop without bothering of protection problems through implementing the encryption method to the files before it is collected in Hadoop. The proposed hybrid system carries the most important cloud computer system service models like Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS). It additionally helps data administration and protection issues (authentication, integrity, availability, and confidentiality) in security and key management for data transfer. The proposed method presented a beautiful time applying with various file sizes within the encryption and decryption stages with greater complexity (double the computational complexity in decryption stages). The longer term work would be integrating both of ElGamal and RSA asymmetric key cryptosystem. As a comparison to the leads to Usama and Zakaria, 2017, the proposed system demands an extended time because it is the next complexity than the system proposed by Usama and Zakaria, 2017. The limitation of the proposed hybrid system is the time taken by the decryption method to spot the proper plaintext form the four options messages produced through Rabin method decryption.

REFERENCES

Aditya, B., V. K. Singh and V. Y. Narayan. 2015. Analyzing BigData with Hadoop Cluster in HDInsight Azure Cloud. IEEE, India.

Bhandarkar, M. 2010. MapReduce programming with apache Hadoop. IEEE, Atlanta. p1-2.

Dubey, A. K., V. Jain and A. P. Mittal. 2015. Stock Market Prediction using Hadoop Map-Reduce Ecosystem. 2015 2nd International Conference on Computing for Sustainable Global Development. p616-621.

Gençoğlu, M. T. 2019. Importance of Cryptography in Information Security. IOSR J. Comput. Eng. 21(1): 65-68.

Hadeer, M., A. Hegazy and M. H. Khafagy. 2018. An Approach for Big Data Security Based on Hadoop Distributed File System. Aswan University, Aswan.

Hilbert, M. 2016. Big data for development: A review of promises and challenges. Dev. Pol. Rev. 34(1): 135-174.

Huixiang, Z. and Q. Wen. 2014. A new solution of data security accessing for Hadoop based on CP-ABE. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey. p525-528.

Jeffrey, D. and S. Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. Association for Computing Machinery, New York. p107-113.

Kareem, S. W. 2009. Hybrid Public Key Encryption Algorithms For E-Commerce. University of Salahaddin-Hawler, Erbil.

Li, B., M. Wang, Y. Zhao, G. Pu, H. Zhu and F. Song. 2015. Modeling and Verifying Google File System. 2015 IEEE 16th International Symposium on High Assurance Systems Engineering. p207-214.

Marti, M., D. McCoy, K. Levchenko, S. Savage and G. M. Voelker. 2011. An Analysis of Underground Forums. Association for Computing Machinery, New York. p71-80.

Masoumeh, R. J., L. M. Khanli, M. K. Akbari and M. S. Javan. 2014. A Survey on Security of Hadoop. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey. p716-721.

Merla, P. and Y. Liang. 2017. Data analysis Using Hadoop MapReduce Environment. Institute of Electrical and Electronics Engineers, Boston. p4783-4785.

Park, S. and Y. Lee. 2013. Secure Hadoop with Encrypted HDFS. Springer, Berlin, Germany. p134-141.

Parmar, R. R., S. Roy, D. Bhattacharyya, S. K. Bandyopadhyay and T. H. Kim. 2017. Large-scale Encryption in the Hadoop Environment: Challenges and Solutions. IEEE Access. p7156-7163.

Raghad, Z. Y., S. W. Kareem, A. O. Hasan. 2016. Design Security System Based on AES and MD5 for Smart Card. Charmo University, Sulaimanyia.

Roojwan, S. I., R. S. Youail, S. W. Kareem. 2014. Image encryption by using RC4 algorithm. Eur. Acad. Res. 2(4): 5833-5839.

Shadan, M. J. A., R. Z. Yousif and S. W. Kareem. 2019. Enhancing approach using hybrid pailler and rsa for information security in bigdata. Appl. Comput. Sci. 15(4): 63-74.

Shahab, W. K. and Y. T. Hussein. 2017. Survey and New Security methodology of Routing Protocol in AD-Hoc Network. 1st International Conference on Information Technology, Erbil.

Shetty, M. M. and D. H. Manjaiah. 2016. Data Security in Hadoop Distributed File System. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey. p939-944.

Sourabh, C., B. Siddhartha and S. Paira. 2014. A Study and Analysis on Symmetric Cryptography. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey. p1-8.

Sourabh, C., S. S. Alam, S. Paira and G. Sanyal. 2014. A Comparative Survey of Symmetric and Asymmetric Key Cryptography. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey. p83-93.

Usama, M. and N. Zakaria. 2017. Chaos-Based Simultaneous Compression and Encryption for Hadoop. PLoS One. 13(3): e0195420.

Yang, C., W. Lin and M. Liu. 2013. A Novel Triple Encryption Scheme for Hadoop Based Cloud Data Security. Emerging Intelligent Data and Web Technologies (EIDWT), 2013 4th International Conference on Xi'an, China. p437-442.