

Data protection on hadoop distributed file system by using encryption algorithms: a systematic literature review

Meisuchi Naisuty*, Achmad Nizar Hidayanto, Nabila Clydea Harahap, Ahmad Rosyiq, Agus Suhanto, and George Michael Samuel Hartono

Faculty of Computer Science Universitas Indonesia, Jakarta, Indonesia

*corresponding author's e-mail: meisuchi.naisuty@gmail.com

Abstract. Big data has capability to process huge amount of unstructured and structured data. Nowadays, technology is able to support business need by extracting massive amount of data and recognizing its pattern to predict future trends. It brings right insight in business strategy to gain tremendous benefit. Hadoop is a reliable technology which developed to distribute process and storage on big data efficiently. However, Hadoop doesn't have any built-in provision to encrypt data by default. Hadoop additional feature in encryption zone has security issue which key management does outside of HDFS. Sensitive and confidential data in HDFS can be exposed against security attack. Information security is fundamental concern and new set challenge for the world of big data. The main purpose of this paper is to protect Hadoop Distributed File System data by using encryption algorithm. This is to ensure data is secured at storage level of HDFS. Dealing with big data, it is important to choose fast enough encryption algorithm that has great performance. Research methodology is SLR (System Literature Review) by using methodology of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).

1. Introduction

Big data contains massive structured and unstructured data that being compiled, processed, analysed, and extracted from many sources. Hadoop was developed to process big data operation quickly and cost efficiently. Hadoop is open source framework that is able to distribute process and storage on big data sets for distributed, scalable, and reliable computing. Hadoop is commonly being used in large cluster scalability or public cloud services like Amazon, Facebook, Twitter, and Yahoo [1]. It is widely known being used because of its low cost, fault tolerance, computing power, flexibility, and scalability [2]. Hadoop has two main components: Hadoop Distributed File System (HDFS) and MapReduce. HDFS is storage layer to distribute large data in logical file system of multiple hard drives across many servers. MapReduce is used for processing and analysing large structured or unstructured data in parallel.

In Hadoop implementation, security was not the development focus [3]. HDFS does not provide any scheme to encrypt data while data in transit and stored in local storage. Meanwhile in reality, Hadoop users might not be aware of HDFS security issue which exposes sensitive and confidential data. It is important to secure HDFS storage level by encrypting its data. If data in motion is being captured or data stored being taken, it will be difficult to interpret the encrypted data by unauthorized party. This action needs to be taken to fulfil baseline of security and privacy standard, such as ISO [4]. ISO stands for International Organization for Standardization that implement data security compliance to protect enterprise data assets. Organizations need to make sure their data only can be accessed by authorized person in internal enterprise.



There are no comprehensive literatures that collecting various kind of encryption algorithms to be applied in HDFS. This paper research encryption algorithms for HDFS. Dealing with huge volumes of data, it is crucial to choose fast enough encryption algorithm that won't impacted performance. This paper structure starts from introduction to explain paper background. Second section is literature review to discuss theory behind this research. Next, third section is research methodology following by results section. Last one is conclusion and further research.

2. Literature Review

2.1. Hadoop as Big Data Platform

According to [5], big data has three characteristics: extremely large volumes of data (how much data), extremely high velocity of data (how fast data is processed), and extremely wide variety of data (various data types). Big data has capability to help organizations to collect, manage, store, and manipulate huge amount of data at the right speed and time which contribute to gain right insights in real time analysis and reaction. Big data merges structured and unstructured data from any kind of data sources. Hadoop was developed to capture and process big data by breaking down big data problem into small elements simultaneously. Hadoop is open source project built by Doug Cutting, a Yahoo engineer. This project managed under Apache Software foundation. Hadoop is designed to process data in parallel across computing nodes to speed up computation and reduce latency.

Hadoop philosophy is to provide a framework which are simple to use, easily scalable, provide fault tolerance, and high availability for usage in production [3]. It is powerful system that processing petabytes of data using low cost hardware very efficiently and quickly. Hadoop stores data locally on its DataNode and processes it locally. It is efficiently managed by the brain of Hadoop system called NameNode. Data read and write operation have to pass NameNode before stored in DataNode.

While in project initialization, data stored is insensitive big data like web usage data. However, situation has changed, Hadoop is used to store sensitive corporate data. It is said in [3] that security was not exactly the priority when Doug Cutting and team building Hadoop. Security was not the focus of development, it made Hadoop has lacked security model. It doesn't have any built-in provision to encrypt data by default. From reference of [6], HDFS security layer is just on HDFS file and directory level ownership and permission. This vulnerability can be covered by data encryption for data in transit and data at rest which stored in local storage. Encryption is common method to secure data [7].

In accordance with reference of [2] based on [8], Hadoop offers end to end transparent encryption called encryption zone. HDFS creates special directory which each file has unique DEK (Data Encryption Key) that will be encrypted using Encryption Zone Key. Then, encrypted key stored in NameNode as part of file metadata. To manage encryption keys, Hadoop uses Key Management Server (KMS). KMS does key management outside of HDFS which make new issue in KMS security layer. It needs trusted intermediary to handle Hadoop request between HDFS and KMS.

2.2. Security Attack in Big Data

A new set of challenges and obstacles in the world of big data is security and governance. To reduce the data exposure, encrypting everything in comprehensive way is a solution. However, it may affect performance [5]. Security mechanism is process to identify, prevent, or restrict security attack using authentication protocol, encryption algorithm, and digital signature [9]. Security attacks are classified into active and passive attack. Active attack attempts to change system resources such as files modification. This attack involves data stream changes or false stream creation. Passive attack attempts to obtain information by monitoring and eavesdropping. It is difficult to detect because no data modification involved. Messages are being sent and received normally across network. Both sender and receiver don't aware that a third party has monitored the traffic to gather information.

2.3. Encryption Algorithms

In references of [10]-[11], cryptography is science of information protection by sharing secret codes using encryption and decryption through unprotected channel. Encryption is basic cryptography that transforms plaintext into ciphertext meanwhile decryption transforms cipher text to plaintext.

Cryptology has 2 key systems: symmetric and asymmetric. Symmetric algorithm uses the same single key for both encrypting and decrypting information. Asymmetric algorithms use different key to encrypt and decrypt information by using public and private key. Sender will transmit message by encrypting it using recipient's public key. Recipient will decrypt the message using recipient's private key.

According to studies included in System Literature Review (SLR), there are 3 most used encryption algorithms: AES (Advanced Encryption Standard), DES (Data Encryption Standard), and RSA (Rivest, Shamir, Adleman). Another encryption algorithm founded in SLR researches is hybrid algorithm. Figure 1 represents encryption algorithms in cryptography classification of key system. Symmetric key algorithms to be analyzed are DES and AES. Asymmetric key algorithm that will be used in this paper is RSA.

2.3.1 Data Encryption Standard (DES)

This encryption algorithm was made by IBM and published in Federal Information Processing Standard in 1977. DES will divide plaintext into 64 bits blocks in encryption process. It is symmetric key that has 56 key bits and 8 parity bits. Every byte contains 7 key bits and 1 parity bit. Parity is used for detect error in a bit pattern. Each 64 bits block is split into 32 bits blocks. To produce ciphertext, symmetric key performs 16 rounds of transpositions and substitutions for each single character.

2.3.2 Advanced Encryption Standard (AES)

AES is symmetric key that consists of 128 bits message block. It is written into 4 x 4 square matrix of bytes. It performs 10, 12, 14 rounds of encryption whether using key size of 128, 196, or 256 bits. The base of this encryption is Rijndael Block Cipher which was developed to be simple, quick, and resistant to known attacks. It was created by Dr. Joan Daemen and Dr. Vincent Rijmen in 2000. Each process round of the encryption process requires four types operation to alter the state of array: sub bytes, shift rows, mix columns, and XOR round key.

2.3.3 Rivest, Shamir, Adleman (RSA)

RSA engages two large factoring prime numbers in asymmetric key algorithm to generate two keys: public key for encryption and private key for decryption. This algorithm is invented by Ronald Rivest, Adi Shamir, and Leonard Adleman which their initial name lead to the algorithm name. It was launched in 1978 in the journal Communication of the ACM. It has more than 1024 bits of key length and variable block size.

2.3.4 Hybrid Algorithm

Hybrid encryption is a method to join two or more encryption algorithm. This method usually combines symmetric key with message digesting technique to gain advantages over each encryption algorithm strength [12].

3. Research Methodology

3.1. Systematic Review

A systematic review is used for collecting and analysing data from studies that are included in the review. It applies systematic and explicit method based on formulated question to identify and select relevant research. Meta-analysis demands statistical techniques in systematic review to obtain integration result which is included in analysis [13].

3.2. PRISMA Protocol

Target of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) is to summarize existing literature results in specific research area. There are four phases in PRISMA flowchart diagram: identification, screening, eligibility, and included. These phases make review studies free from biases. The research starts from collecting literatures from eligible sources by defining criteria to exclude irrelevant literatures. Next step is removing duplicates and unrelated literatures then do full

text screening. Remaining literatures which passed 4 PRISMA phases will become further analysis of the research [14].

3.3. Database and criteria

Literatures in this paper are taken from databases of IEEE and Scopus. To obtain related literatures, searching criteria based on query “(hdfs AND encryption)”. This query will search literatures that contain metadata of HDFS and encryption. Objective of this research is to perform SLR (Systematic Literature Review) towards encryption algorithms to protect HDFS in order to answer the following research question:

RQ1: What encryption algorithms that can be used on Hadoop Distributed File System data?

The following is 2 inclusion criteria to filter literatures:

IC1: Literature is written in English.

IC2: Literature about HDFS encryption algorithm.

By using PRISMA flowchart diagram in figure 1, records identified from IEEE database are 31 literatures and Scopus database are 42 literatures. All records are being sorted to find duplication which is 15 literatures. Remaining literatures are 58 that contain 1 literature which is not meet IC1. Next, due to 8 inaccessible literatures, remaining literatures become 49 literatures. By doing references screening, there are 10 additional literatures. Final step is to exclude out of scope IC2 literature criteria. In the end, 15 eligibility literatures are used to this paper systematic review and analysis study.

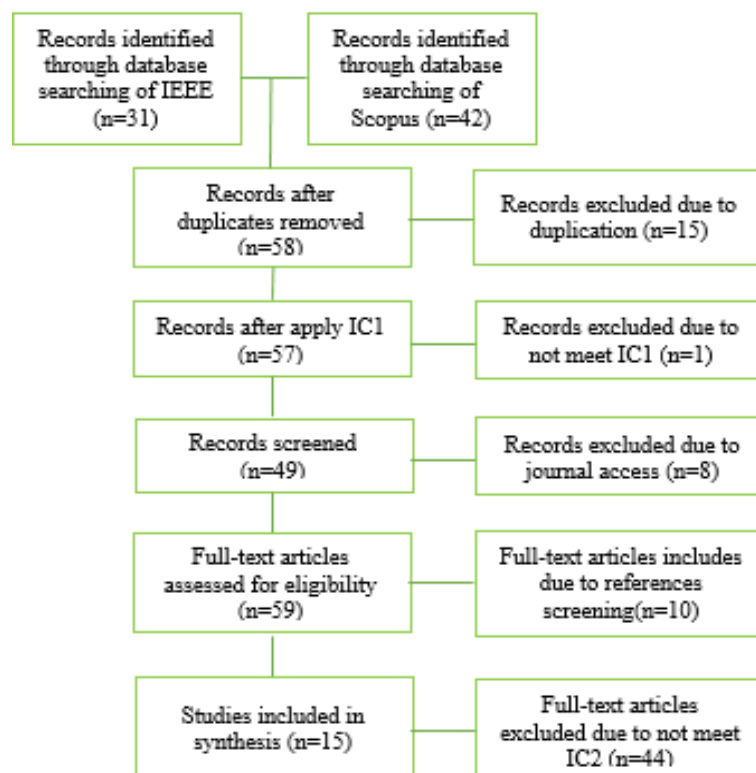


Figure 1. PRISMA flowchart diagram.

4. Result and Discussion

This research based on System Literature Review (SLR) using PRISMA protocol to assess 15 literatures related to encryption algorithm. Most common used encryption algorithm in SLR studies are DES, AES, and RSA. Five other researches are hybrid algorithms. Table 1 shows summary of research assessments. Shade grey in row section indicates preferable algorithm in its research.

Table 1. Summary of systematic literature review assessments

No	Title	DES	AES	RSA	Hybrid
1	Comparative Study of Encryption Algorithm over Big Data in Cloud Systems [15]	v	v	v	
2	Comparative Analysis of Performance Efficiency and Security Measures of Some Encryption Algorithms [19]	v	v		
3	A Survey on Performance Analysis of DES, AES and RSA Algorithm along with LSB Substitution Technique [16]	v	v	v	
4	Comparative Analysis between DES and RSA Algorithm [20]	v		v	
5	Efficiency and Security of Data with Symmetric Encryption Algorithms [21]	v	v		
6	Performance Evaluation of Symmetric Encryption Algorithms [22]	v	v		
7	Performance Evaluation of Symmetric Algorithms [23]	v	v		
8	Implementation Issues and Analysis of Cryptographic Algorithms based on different Security Parameters [24]	v	v	v	
9	Secure User Data in Cloud Computing Using Encryption Algorithms [17]	v	v	v	
10	Design and Implementation of Data-at-Rest Encryption for Hadoop [18]		v		
11	A Novel Triple Encryption Scheme for Hadoop-Based Cloud Data Security [25]				v
12	Design and Implementation of HDFS Data Encryption Scheme Using ARIA Algorithm On Hadoop [26]				v
13	An Approach for Big Data Security Based on Hadoop Distributed File System [1]				v
14	Towards a Trusted Hadoop Storage Platform: Design Considerations of an AES Based Encryption Scheme with TPM Rooted Key Protections [27]				v
15	Toward Data Confidentiality via Integrating Hybrid Encryption Schemes and HDFS [28]				v

4.1. DES vs AES vs RSA Performance

There are 3 parameters to compare encryption algorithm performance: encryption time, decryption time, and buffer size. Buffer size is the size of memory space for storing packets while waiting transmission over network or receiving packets from network. K. Sekar and M. Padmavathamma [15] experiment utilized packet size of 153 KB, 312 KB, and 868 KB while Padmavathi et al. [16] applied packet size of 118 KB, 153 KB, 196 KB, 312 KB, 868 KB. Encryption experiment data shown in table 2. Decryption time is in table 3. Data experiment for buffer size refer to table 4.

Table 2. Encryption time in second

Studies	Packet Size	DES	AES	RSA
Reference [16]	118 KB	3.2	1.7	10
Reference [15],[16]	153 KB	3	1.6	7.3
Reference [16]	196 KB	2	1.7	8.5
Reference [15],[16]	312 KB	3	1.8	7.8
Reference [15],[16]	868 KB	4	2	8.2

Table 3. Decryption time in second

Studies	Packet Size	DES	AES	RSA
Reference [16]	118 KB	1.2	1.2	5
Reference [15],[16]	153 KB	1	1.1	4.9
Reference [16]	196 KB	1.4	1.24	5.9
Reference [15],[16]	312 KB	1.6	1.3	5.1
Reference [15],[16]	868 KB	1.8	1.2	5.1

Table 4. Buffer size DES vs AES vs RSA

Studies	Packet Size	DES	AES	RSA
Reference [16]	118 KB	121	110	188
Reference [15],[16]	153 KB	157	152	222
Reference [16]	196 KB	201	200	257
Reference [15],[16]	312 KB	319	300	416
Reference [15],[16]	868 KB	888	889	934

RSA algorithm has highest score for number of encryption time, decryption time, and buffer size. Second place is DES then following by AES. Reference of [17] agree with this result that RSA consumes more encryption time and memory size than DES and AES. Research paper of [18], shown AES can be implemented in HDFS DataNode in proper functioning.

4.2. Symmetric vs Asymmetric Performance

According to performance results, RSA which is categorized as asymmetric algorithm has the highest number. Otherwise, symmetric algorithm of DES and AES are below RSA in terms of performance. A. Jeeva, V. Palanisamy, and K. Kanagaram [19] claims that RSA is more secured because it utilizes factoring prime number that become so complex to break the code. A. Kumar, S. Jakhar, and S. Makkar [20] made literature about DES compared to RSA algorithm. In their research, RSA consumes enormous amount of time to execute encryption and decryption operations. DES has better performance than RSA in throughput parameter. So, symmetric encryption performs better than asymmetric algorithm.

4.3. Symmetric vs Symmetric Performance

In this paper, two encryption algorithms categorized as symmetric key are DES and AES. Refer to table [2]-[4], AES performs better than DES in parameter of encryption, decryption time, and buffer size. Research of [19] prefers AES as better solution than DES. R. Chehal and K. Singh [21] made research about efficiency and security data between DES and AES. Their research shows that AES has better performance in throughput parameter. The same result is performed by D. Elminaam, H. Kader, and M. Hadhoud [22] that agreed AES has bigger throughput than DES. At last, S. Pavithra and E. Ramadevi [23] proved that AES algorithm performs more throughput and less processing time compared to DES. Higher throughput makes faster speed. Therefore, AES encryption algorithm is better than DES algorithm.

4.4. Comparative study

Based on references of [9]-[11] and [16]-[17], table 5 shows comparative study comparison towards encryption algorithms of DES vs. AES vs. RSA:

Table 5. Encryption algorithm comparison of DES, AES, and RSA

Factors	DES	AES	RSA
Created by	IBM	Dr. Joan Daemen and Dr. Vincent Rijmen	Ron Rivest, Adi Shamir, and Leonard Adleman
Published year	1977	2001	1978
Structure/Scheme	Fiestel	Substitution-Permutation	Factoring prime numbers
Key length	56 bits	128, 192, or 256 bits	>1024 bits
Rounds	16	10, 12, or 14	1
Block size	64 bits	128 bits	Variable
Cipher Type	Symmetric	Symmetric	Asymmetric
Key used	Same key	Same key	Different key

4.5. Hybrid Algorithm

Based on searching query in study literatures, there are 4 papers which assess solution by combining encryption algorithms to provide high security. Reference of [24] suggests the same solution to apply more than one encryption algorithms knows as hybrid cryptosystem to secure data.

4.5.1 Novel Triple Algorithm

Novel triple algorithm scheme is hybrid encryption of DES, RSA, and IDEA. File encryption uses DES and data key encryption uses RSA. IDEA (International Data Encryption Algorithm) performs encryption of user's RSA private key. This research has successfully performed implementation of triple encryption scheme. It is feasible as its performance meet HDFS reading and writing characteristics [25].

4.5.2 ARIA Algorithm

ARIA algorithm is standard data encryption scheme for local usages in South Korea. This research lets user to choose encryption algorithm between AES and ARIA in HDFS data encryption. ARIA can support variable length data, meanwhile AES is limited to 128 bits data block. ARIA algorithm has 2-3% performance reduction in query processing compared to AES [26].

4.5.3 AES and OTP Algorithm

This research integrates AES and OTP algorithm for encryption and decryption file. OTP (One Time Pad) generates unique key that only being used for one-time. Research experiment claims this method have better performance in reading and writing time compared to just AES algorithm [1].

4.5.4 AES and TPM Rooted Key

Encryption scheme is a combination of based encryption scheme of AES and TPM (Trusted Platform Module) rooted key. This integration method has performance issue for 16% overhead in encryption process and 11% for decryption process for 128MB block data. It needs more research to limit the overhead and identify TPM operations speed [27].

4.5.5 HDFS-RSA and HDFS-Pairing

Experiment of HDFS performance overhead by using HDFS-RSA and HDFS-Pairing. This integration is suitable for read-many and write-once applications. It has considerable writing overhead and acceptable reading overhead operations. HDFS-RSA and HDFS-Pairing performance is limited because the header files are small files while HDFS has great performance for massive files [28].

5. Conclusion

This paper is research of Systematic Literature Review towards encryption algorithm in implementing data protection in Hadoop Distributed File System. There are 3 most used encryption algorithms: DES, AES, and RSA. This research verified that AES has higher performance than DES and RSA. Performance parameter are decryption time, encryption time, and buffer size. In performance comparison of symmetric key (DES and AES) vs asymmetric key (RSA), symmetric key performs better than asymmetric key. Between symmetric key of DES vs AES, AES has better performance than DES. It is supported by high throughput that bring faster speed.

Another consideration is about strong security protection. RSA has more than 1024 bits key length, asymmetric key, and complicated processes for encrypting and decrypting. It is more difficult to decode RSA than DES and AES which impacted to slower performance. Hybrid algorithm also offers high security level by combining more than one encryption algorithm. Overall, user may choose asymmetric algorithm or hybrid algorithm for high level protection but lack in terms of performance.

6. References

- [1] H. Mahmoud, A. Hegazy and M. H. Khafagy 2018 An approach for big data security based on Hadoop distributed file system *International Conference on Innovative Trends in Computer Engineering (ITCE)* pp 109-114
- [2] S. Suganya and S. Selvamuthukumaran 2018 Hadoop Distributed File System Security - A Review *International Conference on Current Trends towards Converging Technologies (ICCTCT)* pp. 1-5
- [3] B. Lakhe 2014 *Practical Hadoop Security* (New York: Apress)
- [4] R. R. Parmar, S. Roy, D. Bhattacharyya, S. K. Bandyopadhyay, and T.-H. Kim 2017 Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions *IEEE Access* vol. 5 pp. 7156-7163
- [5] J. Hurwitz and A. Nugent 2013 *Big Data for Dummies: a Wiley Brand*. Hoboken, NJ: John Wiley and Sons.
- [6] A. Holmes 2013 *Hadoop in Practice* (Shelter Island: Manning Publications)
- [7] B. Spivey and J. Echeverria 2015 *Hadoop Security* (O'Reilly Media)
- [8] M. M. Shetty and D. H. Manjaiah 2006 Data security in Hadoop distributed file system 2016 *International Conference on Emerging Technological Trends (ICETT)* pp. 1-5
- [9] W. Stallings 2006 *Cryptography and Network Security: Principles and Practice*
- [10] L. Miller and P. H. Gregory 2016 *CISSP for Dummies*

- [11] S. Vaudenay 2010 *A Classical Introduction to Cryptography: Applications for Communications Security* (New York: Springer)
- [12] S. Kumarsinha, M. Shrivastava, and K. K. Pandey 2013 A New Way of Design and Implementation of Hybrid Encryption to Protect Confidential Information from Malicious Attack in Network *International Journal of Computer Applications* vol. 80 no. 3 pp. 48–55
- [13] D. Moher, et al. 2010 Preferred reporting items for systematic reviews and metaanalyses: The PRISMA statement *International Journal of Surgery* pp.336-341
- [14] M. Sokouti and B. Sokouti 2018 A PRISMA-compliant systematic review and analysis on color image encryption using DNA properties *Computer Science Review* vol. 29 pp. 14–20
- [15] K. Sekar and M. Padmavathamma 2016 Comparative Study of Encryption Algorithm over Big Data in Cloud Systems *International Conference on Computing for Sustainable Global Development (INDIACom)* pp. 1571-1574
- [16] B. Padmavathi and S. Ranjitha Kumari 2013 A Survey on Performance Analysis of DES, AES, and RSA Algorithm along with LSB Substitution Technique *International Journal of Science and Research (IJSR)* vol. 2 pp. 170–174
- [17] R. Arora and A. Parashar 2013 Secure User Data in Cloud Computing Using Encryption Algorithms *International Journal of Engineering Research and Applications (IJERA)* vol. 3 pp. 1922–1926
- [18] S. H. Kamaruzaman, W. N. S. W. Nik, M. A. Mohamed, and Z. Mohamad 2018 Design and Implementation of Data-at-Rest Encryption for Hadoop *International Journal of Engineering & Technology* vol. 7 p. 54
- [19] A. Jeeva, V. Palanisamy, and K. Kanagaram 2012 Comparative Analysis of Performance Efficiency And Security Measures Of Some Encryption Algorithms *International Journal of Engineering Research and Applications (IJERA)* vol. 2 pp. 3033–3038
- [20] A. Kumar, S. Jakhar, and S. Makkar 2012 Comparative Analysis between DES and RSA Algorithm *International Journal of Advanced Research in Computer Science and Software Engineering* vol. 2 pp. 386–390
- [21] R. Chehal and K. Singh 2012 Efficiency and Security of Data with Symmetric Encryption Algorithms *International Journal of Advanced Research in Computer Science and Software Engineering* vol. 2 pp. 472–275
- [22] D. Elminaam, H. Kader, and M. Hadhoud 2008 Performance Evaluation of Symmetric Encryption Algorithms *IJCSNS International Journal of Computer Science and Network Security* vol. 8 pp. 280–286
- [23] S. Pavithra and E. Ramadevi 2012 Performance Evaluation of Symmetric Algorithms *Journal of Global Research in Computer Science* vol. 3 pp. 43–45
- [24] K. Kalaiselvi and A. Kumar 2015 Implementation Issues and Analysis of Cryptographic Algorithms based on different Security Parameters *International Conference on Current Trends in Advanced Computing* pp. 23–28
- [25] C. Yang, W. Lin, and M. Liu 2013 A Novel Triple Encryption Scheme for Hadoop-Based Cloud Data Security *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*
- [26] Youngho Song, Young-Sung Shin, Miyoung Jang and Jae-Woo Chang 2017 Design and implementation of HDFS data encryption scheme using ARIA algorithm on Hadoop *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* pp. 84-90
- [27] J. Cohen and S. Acharya 2013 Towards a Trusted Hadoop Storage Platform: Design Considerations of an AES Based Encryption Scheme with TPM Rooted Key Protections *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing* pp. 444-451
- [28] H.-Y. Lin, S.-T. Shen, W.-G. Tzeng, and B.-S. P. Lin 2012 Toward Data Confidentiality via Integrating Hybrid Encryption Schemes and Hadoop Distributed File System *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*

Acknowledgement

This study was supported by Pendidikan Magister Menuju Doktor untuk Sarjana Unggul (PMDSU) Research Grant No NKB-3031/UN2.R3.1.HKP.05.00/2019 Ministry of Research, Technology and Higher Education Republic of Indonesia.