Comparative Study of Stream, Block and Hybrid Cipher Techniques in Hadoop Distributed File System

Omar Helmy Khafagy [1*], Fatma A. Omara [2], Mohamed Hasan Ibrahim [1]

[1]Faculty of Computers and Information, Fayoum University, Fayoum, Egypt.
[2]Faculty of Computers and Information, Cairo University, Cairo, Egypt.

O_h_khafagy@hotmail.com, fatma_omara@hotmail.com,mhi11@fayoum.edu.eg.

Abstract— big data streaming is the most trending term nowadays, collecting a stream of data from different places and devices need to be processed in real time. Hadoop is the suitable framework for this large stream of data because it supports handling of big data as storage in HDFS and real times processing with map-reduce functions. Big data streaming needs a security model to achieve data confidentiality. According to the work in this paper, a comparative study has been done among different security approaches (i.e., block, stream, and hybrid) which have been applied for Hadoop. The implementation of these approaches is based on the performance and the output data for each approach. The results of the comparative study show that the streaming security approach outperforms other approaches; block and hybrid.)

***Keywords**; Security; Big Data streaming; Hadoop; HDFS; MapReduce;*

1. Introduction

Cloud computing became a critical technology for the E-business and used for massive processing in large-scale data .it helps for cost saving by sharing the resources of hardware and software. Experience in managing cloud computing is not a challenge for the users because it's easy to use. Pay per use is the best solution for saving the cost. The users as they use the hardware resources and the software applications and services they pay for that usage only [1].

The vendors of the cloud commuting offered all support and the management of the back-end to help the users to use the services that scalable as the user demand. Cloud computing services categorized as Platform as a Service (PaaS), Infrastructure as a Service (IaaS) or Software as a Service (SaaS) [2]. Companies and users of cloud computing face some challenges instead of its benefits. Cloud computing not just the best way to reduce IT costs. Nowadays, it is the businesses solution of interacting in real time directly with the customers. The cloud computing challenges are shown in figure.1.

Authentication: The process of proving one's identity. This means that before the system sends and receives data, the receiver and sender identity should be verified.

Confidentiality: Ensures that only the intended receiver reads the message and no one else does, this is usually how most people identify a secure system.

Integrity: Ensures that the message received by the user was not altered or manipulated with. The basic form of integrity is packet checksum in IPv4 packets.

Performance: Different models can decrease performance time if there is no probable plan that negatively impacts performance. Service Reliability/Availability: Since the main problem with most systems is getting hacked by an intruder, who can cause a downtime in availability, such systems provide a way to provide their users with the quality of service they expect.

This paper will focus on cloud computing security special in Confidentiality

Figure1. Cloud Computing Challenges

1.1 Hadoop Architecture

Hadoop [3] is an open source platform for managing large data sets and distributed storage. It supports distributed data processing. Hadoop consists of two components. Hadoop Distributed File System HDFS and Map Reduce.

1.1.1 Hadoop Distributed File System

HDFS is responsible for storing large amount datasets through multi-clusters that consist of a single Name Node as a master server and multi Data Node for storing data blocks. Figure.3 describes the HDFS procedures for uploading a file. The user can upload a file in HDFS through distributed file system module that uses load balance technique[4] and reserves the data blocks on metadata that located at name node server, the stream of data uploaded to HDFS on Data Nodes servers [5,6].

1.2 Big data Streaming

Big data has appeared from massive data sets that become a big challenge for data management tools and data processing in real time. Big data streaming is quickly processing for continuously generated data. Data analysis in real-time streaming data is a single pass analysis [7].

Big Data consists of 4 vectors [8] are velocity, variety, volume and value are described:

- Velocity is the speed of generating data.
- Variety is the different types of data.
- Volume is the size of generated data.
- Value is the outputs for gains from massive data.

1.3 Big Data Challenges

Big data is gated on some issues [9] are showed in Figure.2 that describes:

- Availability the data is limited availability for skill clients to manage big data.
- Data development timing.
- Security to protect big data from unauthorized users.
- Performance of processing big data[10,11,12].
- Scalability of resources for processing data.

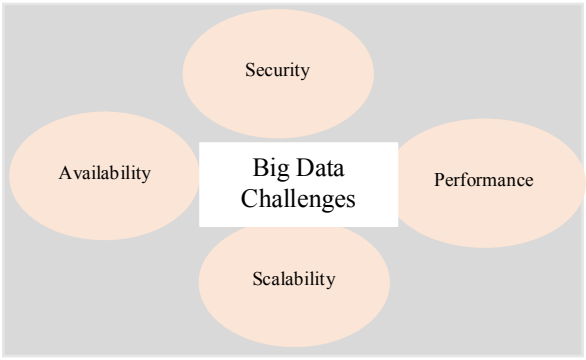

Figure2. Big Data Challenges

1.4 Big Data Security

Big data security [13] that is becoming a critical issue. The most companies and banks need to protect data, so big data security analytics solution to secure data that consists of three characteristics of scale, performance, and flexibility.

١.4 ١.Security Property

Security property can be achieved the cryptography goals that consists of four core purposes of cryptography: Authentication, integrity, availability, and confidentiality that is shown in Figure.3 Any security system must use objective of this goal to achieved security in data [14].

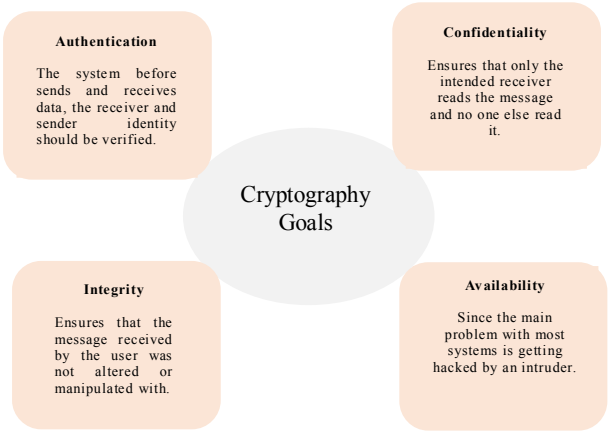

Figure 3. Cryptography Goals

١.4 ٢.Hadoop Security

Original Hadoop designed without any security. The data in Hadoop not secure and not protect from

hackers. Some previous work can be implemented and tested to techniques to secure data in the Hadoop distributed file system [15].

The rest of this paper is organized as follow:

Section 2 describes related work Stream, Block and Hybrid cipher , Section 3 describes a comparative analysis of HDFS Security Techniques and experimental results section 4 conclusion and future work.

2 Related Work

Some previous work used the block or stream data by new mechanisms to secure data in Hadoop, it includes a lot of techniques to achieve the security method like confidentiality and authentication. We show the performance and complexity that affected each mechanism.

2.1 Stream Cipher

We show the related work when use the data stream in HDFS using new techniques to secure data.

Pardeep et al. [16] used different stream ciphers to secure the data over transmissions, in different ways like: to ensure that there is no unauthorized access, to achieve the integrity of data. Enhancement implementations on RC4 are also shown in the paper. This paper also discusses the difference in the performance between the RC4 stream cipher algorithm and the stream cipher algorithms.

Anandu Jayan et al. [17] implemented a secure Hadoop using RC4 as a stream cipher to enhance the performance of encryption/decryption by using the MapReduce functions to ensure the confidentiality of data and allow the users to protect their data. In this paper, RC4 encryption algorithm modified to use in parallel to enhance the security of data on HADOOP and reduced the cost algorithm utilizing the map reduce. RC4 depends on the resources of the Nodes.

2.2 Block Cipher

We show the related work when use the data block in HDFS using new techniques to secure data.

Seonyoung Park and Youngseok Lee [18] implemented the AES encrypt/decrypt class and adds to the Compression Codec in Hadoop to secure the HDFS architecture. The testing and Resulting in Hadoop showed MapReduce job on encrypted HDFS generates affordable computation overhead less than 7%. The encrypted and decrypted the file before it is written or read in the HDFS. Client request to encrypted or decrypted file I HDFS blocks at each Data Nodes used 128-bit AES with ECB mode with HDFS blocks.

Chao YANG, Weiwei LIN et al. [19] implanting triple encryption scheme using combines HDFS file encrypting used three algorithms are DEA, RSA, and IDEA to users encrypt the file using RSA private key, then it integrated into Hadoop based cloud data storage. The hybrid Encryption method used a symmetric cipher to encrypt data with a unique key and the key is used asymmetric cipher encrypted by user's public key. Hybrid Encryption used the symmetric and asymmetric cipher to encrypt file uses DES algorithm to get the data key, Data key encrypted using the RSA algorithm. The user keeps the private key in order to decrypt the Data key.

Pradeep Adluru et al. [20] provide the random encryption techniques that achieved an authentication of distributed data access. This paper discussed the issue in the security and privacy of Big Data. Hadoop Eco System that provides is security and privacy as the Name Node and the Data Nodes in Hadoop, this system is a need for trust between the client and the Name Node used Hashing. Encryption techniques on the data that are implemented using random algorithms like RSA, AES, and RC6.

2.3 Hybrid Cipher

Hybrid data used in some of the related work when using the data stream combined with the data block to enhance data security in HDFS.

Hadeer Mahmoud et al. [21] presented a new approach that consists of two algorithms; one of the

blocks cipher the AES algorithm with ECB mode and second the stream cipher OTP algorithm to achieve the data confidentiality in HDFS. They approach enhanced the file encrypted size that becomes increased by 20% of the original file. The performance in encryption/decryption data can be improved in Hadoop. Figure.4 is showed they proposed approach mechanism used to encrypt the data.

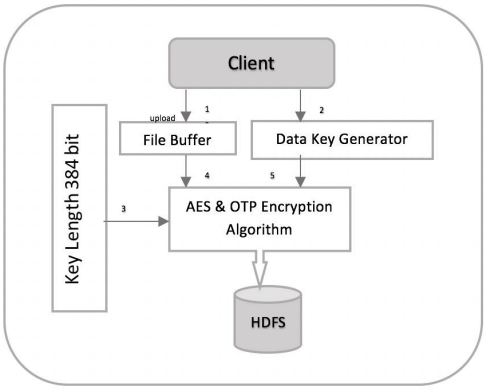

Figure 4. A Proposed Approach Mechanism

Hsiao-Ying Lin et al. [22] designed and implemented the integration of HDFS-RSA and HDFS-Parining to achieve the data confidentiality in Hadoop. The result of the integration showed an overhead on reading and writing operations. This paper designed the fuse-dfs module to modify the reading and writing operations, they described the two implementations to achieve this module. First, HDFS-Parining used the PBC library. Second, HDFS-RSA uses the OpenSSL took it. The Hybrid used a block cipher and stream cipher to encrypted data as shown in figure.5.

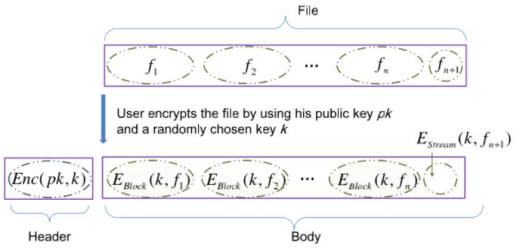

Figure 5. Hybrid Encryption Scheme

3 Experimental analysis and result

3.1 Experimental Environment

For the performance evaluation of encrypted HDFS, we configured Hadoop 1.2.1 as a Single-Node Cluster to use the HDFS and MapReduce functions. Each node has core i7, 4 processors, 8 GB memory, and 750G hard disk.

3.2 Dataset

We use the TPC-H benchmark [23] dataset to evaluate our implementation with original Hadoop.

3.3 Comparative Analysis of HDFS Security Techniques

the summary of the classification study on security techniques used to secure the data in HDFS of the method; strength, weaknesses, and security property are described in table.1.

The performance comparison used in the previous work when used AES [18], RC4 [17], and the hybrid approach [21] algorithms compare with generic file in HDFS. Read performance that compared in table.2 and writes performance that compared in table.3.

Figure.6 described the comparison of the read performance between related work [17][18][21] with generic in the chart. Figure.7 showed the written comparison between the same related work and generic file in Hadoop.

These results show that the using of stream cipher algorithms to secure the Hadoop Distributed file system has the best performance in reading and writing data in HDFS.

4 Conclusion

This paper introduced the phases to secure the data in HDFS, it provided the difference between Stream and block cipher that used by the latest techniques in the previous work. Big Data Security is the critical issue in Hadoop, so they need to implement a new mechanism using different ways of stream or block cipher to secure data.

We wished that this paper would support the latest security techniques of big data and its system better in order to enhance it in the future.

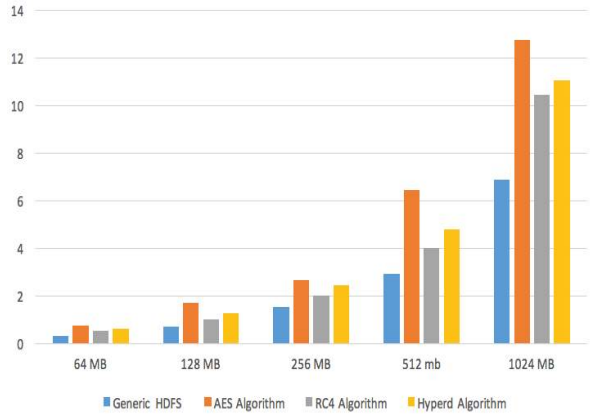

Figure 6. Comparison File Read Performance

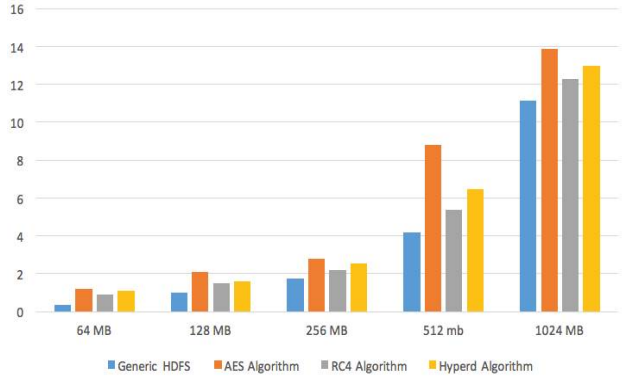

Figure 7. Comparison File Write Performance

TABLE 2. File write performance comparison between generic HDFS, AES algorithm,RC4 algorithm and Hybrid technique

File Size (MB)	*Generic HDFS (minutes)*	*AES algorithm (minutes)*	*RC4 algorithm (minutes)*	*Hybrid (minutes)*
64	0.3292	1.2077	0.8920	1.0850
128	1.0239	2.0843	1.4982	1.5980
256	1.7502	2.7878	2.2099	2.5332
512	4.1879	8.8221	5.3652	6.4801
1024	11.1232	13.8540	12.2805	12.9877

TABLE 3. File read performance comparison between generic HDFS, AES, RC4 algorithm and Hybrid technique

File Size (MB)	*Generic HDFS (minutes)*	*AES algorithm (minutes)*	*RC4 algorithm (minutes)*	*Hybrid (minutes)*
64	0.3146	0.7604	0.5347	0.6324
128	0.7007	1.7214	0.9914	1.2790
256	1.5385	2.6378	2.0234	2.4335
512	2.9045	6.4622	3.9850	4.7789
1024	6.8923	12.7654	10.4457	11.0539

Table 1. Summary of the Classification Study on Security Techniques

Work/Year	Cryptography Cipher	Method	Strengths	Weaknesses	Security Property
Hadeer: An approach for Big Data Security **[19] 2018**	Block Cipher With Stream Cipher	Implement AES algorithm with OTP algorithm	Used hybrid encryption algorithm. Enhance the performance of encryption/decryption data.	File encrypted size increased by 20% of the original data.	onfidentiality, Integrity.
Anandu Jayan: RC4 in Hadoop Security using MapReduce **[15] 2017**	Stream Cipher	Modified in the RC4 encryption algorithm.	File encrypted size is the same as the original file.	The key is not secure.	onfidentiality, Integrity, and Authentication.
Xiaowen Zhang Big Data Security and Privacy **[18] 2015**	Block Cipher	Implement random techniques like RSA and AES to encrypt data. Trust mechanism between the user and name node.	Achieve privacy and security in HDFS.	The connection between the user and Name Node takes a lot of time to achieve trusting, that effect in performance.	Authentication
Youngseok Lee : Encrypted HDFS **[16] 2013**	Block Cipher	Implement AES algorithm with ECB mode to encode and decode the data in HDFS.	Encrypt HDFS generates an affordable computation Overhead less than 7%.	Different blocks used independently of each other because the paper used ECB mode. The size of encrypted file increased to approximately 50%.	Confidentiality and Authentication.
Mingqi LIU : Triple Encryption Scheme for Hadoop- Data Security **[17] 2013**	Block Cipher	Implement three algorithms, RSA DEA, and IDEA to encrypt data and private key by RSA.	Data Encryption with high secure data.	Encryption and Decryption that effect in the performance using triple encryption scheme.	onfidentiality, Integrity, and Authentication.
Pardeep A Pragmatic Study on Different Stream Ciphers **[14] 2012**	Stream Cipher	Implement RC4 algorithm.	RC4 has faster encryption and decryption among the other stream cipher algorithms.	The RC4 algorithm needs improvement to be more secure.	Confidentiality and Integrity.
Wen-Guey Tzeng Integrating Hybrid Encryption Schemes and HDFS **[20] 2012**	Block Cipher With Stream Cipher	Implement HDFS - Pairing and HDFS-RSA integrations.	Provide Data confidentiality in HDFS by using data encryption.	Asymmetric used by RSA and pairing algorithms are slower and not suitable for encryption and decryption data in Hadoop.	Confidentiality and Integrity.

References

[1] Sun Microsystems, “Introduction to Cloud Computing architecture,” Sun Microsystems, 2009.

[2] I. Nwobodo, “A Comparison of Cloud Computing Platforms,” Int. Conf. Circuits Syst., no. CAS 2015, pp. 283–289, 2015.

[3] D. Borthakur, “The Hadoop distributed file system: Architecture and design,” Hadoop Proj. Website, pp. 1–14, 2007.

[4] Sarhan, E., Ghalwash, A., Mohamed H. Khafagy.” Queue weighting load-balancing technique for database replication in dynamic content web sites “,Proceedings of the 9th WSEAS International Conference on Applied Computer Science, ACS '09, pp. 50-55,2009

[5] Abdel Azez, H.S.H., Mohamed H. Khafagy, Omara, F.A.,” Optimizing Join in HIVE Star Schema Using Key/Facts Indexing”, IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India),35(2), pp. 132-144,2018.

[6] Kaseb, M.R., Mohamed H. Khafagy, Ali, I.A., Saad, E.S.M.” Redundant independent files (RIF): A technique for reducing storage and resources in big data replication “,Advances in Intelligent Systems and Computing,745, pp. 182-193, 2018

[7] D. Namiot, “On Big Data Stream Processing,” Int. J. Open Inf. Technol., vol. 3, no. 8, pp. 48–51, 2015.

[8] R. Srinivas, “Managing Large Data Sets Using Support Vector Machines,” pp. 1–93, 2010.

[9] A. A. Tole, “Big Data Challenges,” Database Syst. J. vol., vol. IV, no. 3, pp. 31–40, 2013.

[10] Sahal, R., Mohamed H. Khafagy, Omara, F.A.,”Comparative study of multi-query optimization techniques using shared predicate-based for big data “, International Journal of Grid and Distributed Computing 9(5), pp. 229-240,2016

[11] Sahal, R., Mohamed H. Khafagy, Omara, F.A.,”Exploiting coarse-grained reused-based opportunities in Big Data multi-query optimization”, Journal of Computational Science, 26, pp. 432-452, 2018

[12] Shanoda, M.S., Senbel, S.A., Mohamed H. Khafagy,”JOMR: Multi-join optimizer technique to enhance map-reduce job”, 9th International Conference on Informatics and Systems, INFOS 2014, 7036682, pp. PDC80-PDC87, 2014

[13] G. Wang, P. She, and F. Mahzabeen, “‘ Big Data Security Analysis .’”

[14] R. Focardi, “Classification of Security Properties (Part II: Network Security),” Network, no. September, 2001.

[15] O. O. Malley, “Integrating Kerberos into Apache Hadoop,” 2010.

[16] P. K. Peteriya, “A Pragmatic Study on Different Stream Ciphers And On Different Flavors of RC4 Stream Cipher,” vol. 12, no. 3, pp. 37–42, 2012.

[17] A. Jayan, “RC4 in Hadoop Security using MapReduce,” 2017.

[18] S. Park and Y. Lee, “Secure Hadoop with encrypted HDFS,” Lect. Notes Comput. Sci.

[19] C. Yang, W. Lin, and M. Liu, “A novel triple encryption scheme for Hadoop-based cloud data security,” Proc. - 4th Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2013, no. September 2013, pp. 437–442, 2013.

[20] P. Adluru, S. S. Datla, and X. Zhang, “Hadoop Eco-System for Big Data Security and Privacy,” 2015.

[21] H. Mahmoud, A. Hegazy, and M. H. Khafagy, “An approach for Big Data Security based on Hadoop Distributed File system,” no. Itce, pp. 109–114, 2018.

[22] H. Y. Lin, S. T. Shen, W. G. Tzeng, and B. S. P. Lin, “Toward data confidentiality via integrating hybrid encryption schemes and Hadoop distributed file system,” Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA, pp. 740–747, 2012.

[23] www.tpc.org/2018