ELSEVIER

Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC

An Efficient Data Chunking Non-Sequential Storage to Defend Against Cyber Attacks

[1]M.Uma and [2]Dr.G.Padmavathi
[1]Ph.D Research Scholar, Department of computer Science,
Avinashilingam university for Home Science and Higher Education for Women, University, Coimbatore – 641043
uma.phdresearch@gmail.com
[2]Professor and Head, Department of computer Science
Avinashilingam university for Home Science and Higher Education for Women, University, Coimbatore – 641043
ganapathi.padmavathi@gmail.com

***Abstract*— The importance of secure data storage in day to day life is increasing rapidly in our society. Data chunking and decentralization plays a significant role in efficiently storing and securing the data. In this research work, server based application is developed using content defined chunking for splitting the content of the files uploaded, markov chain for state transition, Secure Hash Algorithm (SHA-512) is used key generation and Advanced Encryption Standard (AES) algorithm is used for encryption and decryption of the file chunked. The standard way of storing the chunked file into the server in a sequential order is rearranged and introduced a new method say non-sequential way of storing the chunked files into the server. The proposed method outperforms than the existing method and the proposed method is evaluated in terms of performance metrics like False acceptance rate, false rejection rate, and attack detection rate. The proposed method is implemented using Java 1.7 as front end and mysql as back end.**

Index Terms*— *Data chunking, CDC, SHA-512, AES, Markov chain

I. INTRODUCTION

Data chunking is one of the moving target defense mechanisms. Data chunking is the systematic formulation of splitting the data into a predefined size in scientific data set stored as chunks is termed as chunking. For a very large size array which consists of thousands of rows and columns, data chunking will be more valuable for input-output execution [8]. Data chunking is classified into various types based on the necessity of applying chunking for particular applications. Memory based chunking, Frequency based chunking, Content Defined Chunking, Support Vector Machine and Chunk Amalgamation Algorithms are the various data chunking techniques available among all Content Defined Chunking (CDC) are found appropriate for this research work due to its dynamic nature of splitting the content of files according to the size of the file[1][16][19]. The files will be split up into chunks in various sizes smaller than the original file which helps to transmit the data. The chunked file will be given filename with some extension in chronological order say is the filename is *output* then the extended file as "*output.000*", "*output.001*" etc., so that there will not be any complexity in accessing the files in the future[4][22]. No further indexing or maintenance is

required after the files are being chunked. SHA-512 (Secure Hash Algorithm) is a cryptographic hash function's algorithm[5][6] which helps to transform the text file as the value in a fixed length and to compress the message called a message digest. The algorithm acquires 2128 bits as inputs which use 1025 bit blocks to generate 512 bits of the message digest as output. [11][15]Append padding bits, Append length and Initialize hash buffer, 80 rounds and output. SHA-512 compression function and SHA-512 message schedule are the two components of SHA-512. Apart from that, there are few logical functions used for SHA-512 [13]. Advanced Encryption Standard (AES) [2] [7] [14] [17] is standard for symmetric encryption is a block cipher which encrypts of 128-bit block for encryption and 128 bit block for decryption. It also uses a key length as 128, 192, or 256 bits. The transformation is consecutively predefined as 10, 12 and 14 iterations for data block process as correspondingly AES-128, AES-192, and AES-256 which is termed as rounds [9]. Different round keys will be used in every round which functions on two 128 bits from 1 up to 10, 12, 14 iterations [18] [20] [21]. A key expansion algorithm is used to obtain the cipher text, whereas the algorithm will be utilized separately for the encryption / decryption process. All the rounds functions similarly with the following four transformations such as the SubBytes, the ShiftRows, the MixColumns, and the AddRoundKey, in the last round it leave out the MixColumns transformations [10]. The encryption process will have the reverse transformation for decryption as InvSubBytes, InvShiftRows, and InvMixColumns. The structure of the encryption can be formulated precisely from the structure of the decryption.
The organization of the paper is structured as follows: in section 2, background and related work is presented. The overview of the proposed method is described in section 3, and experimental design in section 4. Performance results were presented in section 5 and conclusion in section 6.

II. Background and related work

This section describes about the related work of data chunking. Some of the related works are given below
Diego Perino et al., (2012) proposed a novel framework to eliminate the redundancy and it should be association with the Information-Centric Networking (ICN). They have done a detailed study on Named Data Networking (NDN), Redundancy Elimination (RE), identifying the redundancy, data chunking. Bandwidth, hardware and software suitability is comparatively enhanced with the existing system called vanilla ICN. Complexity is much simpler in ICN-RE when compared to the advanced existing techniques SmartRE and EndRE. Delay is reduced and bandwidth is increased in the proposed than with the existing system.
Punyada M. Deshmukh (2012) developed an application which ensures the data storage security using a distributed scheme. In order to process the user request, a responsible is given to set of master servers. For file recovery and to provide data back up file chunking operation will be executed. The proposed system will benefit android users and chatting applications as well. Also it makes the surroundings of the working place as ease and relaxed.
Mark W. Storer, et al., (2008) proposed a secure data Deduplication for efficient spacing as well as for protecting the data. They have developed two for secure storage such as authenticated and anonymous. Once the data is chunked the key generation is done in order to encrypt the data and also designed map to reconstruct the chunked data to its original form.
Deborah S. Carstens (2006) suggested a technique to evaluate the authentication of password crash. Two levels of experiments were conducted to create password which is pertinent, consequential password and which is not easily reachable by the public. For the development of the password chunking theory this introduced in this work. 7-Character Password Level, Two-Chunk Password Level, Three-Chunk Password Level, Four-Chunk Password Level. Recall rates and paper rates are parameters used to evaluate the technique proposed.
Hong Shen (2004) developed a combinational approach using trigram Hidden Markov Model (HMM) and Data Representation (DR) voting techniques. The analysis is done between Multiple Data Representations and Multiple Learning Models. And suggested that the chunker is faster, simpler method and very accurate in training and decoding they developed. CoNLL-2000 dataset is used.
The related work shows the significance of data chunking.

III. Overview of proposed non-sequential data chunking method

The main idea of this phase is to develop a method which ensures strict verification of users before accessing the files stored in the database [12]. Data chunking provides security system like authentication, secure storage [3] and integrity. The model is developed to defend against cyber attacks say active and passive

attacks. The primary focus of this model is to provide easy access to the legitimate user which ensures data availability to the legitimate users, secure data storage through encryption and decryption. The flowchart of the proposed method is given in figure.1 and proposed algorithm is given in Table.1.

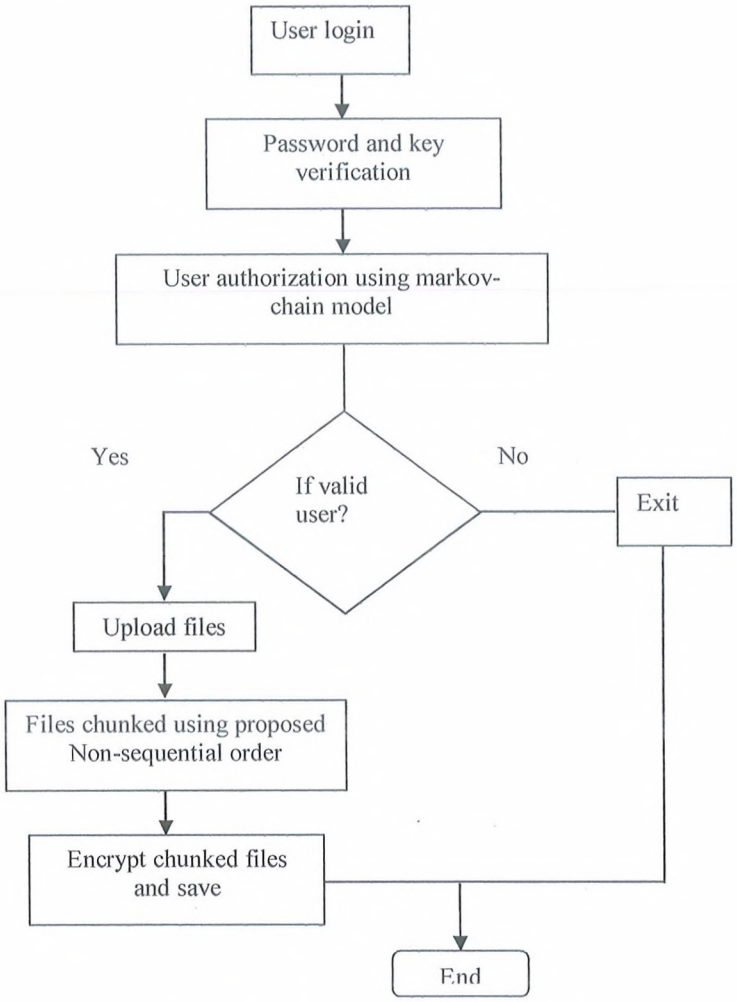

Figure.1 Flow chart of the proposed method

The proposed method consists of the following phases and given in detail below. Data chunking helps to avoid duplication of data storage and also provides security to the data stored. Usually, data chunking will chunk the files in a sequential order. In this research work non-sequential order, i.e., random storage of files is introduced.

A. Registration and key generation

User registration and key generation is the initial process of this proposed method. After the server is initiated, the user can register him/her with the details requested. Once the user is registered he/she will be allotted with unique key. Secure Hash Algorithm-512 (SHA-512) is used to generate key.

B. Attack detection using markov-chain model

In this phase, the cyber attacks are detected using the markov-chain model. Markov-chain model is an anomaly-detection approach. The un-usual event or occurrences from normal profile will be observed by anomaly-detection techniques which are considered as attacks. It is capable of differentiate the known, unknown and new attacks from normal data. Markov chain helps in envision the characteristics of state transition which happens in the future. It is an efficient method in analyzing the probability of state transition.

TABLE I. PROPOSED ALGORITHM

Proposed Method Algorithm
U→User
D→Database
S→Server

Procedure:
Initialize server
for each user logging in
register and generate key
upload file
perform verification
execute markov-chain model

if user authorized
chunk file using proposed non-sequential order
perform encryption
store it is 'S'
else
exit

end if

until final user access

In this proposed method, markov chain model is used to detect processes like network state monitoring and abnormal activity prediction by evaluating the probability of any occurrences which helps in predicting the unauthorized users and attacks.

C. Chunking and proposed non-sequential storage

In Data chunking, the file will be chunked into two or more files and it will be stored in different servers in a sequential order. If there are three servers the files uploaded will be splitted into three files and it will be stored in those three servers in a predefined order like file 1 in server l, file 2 in server 2 and file 3 in server 3. In this proposed method, files will be chunked using content defined chunking which dynamically chunks the file according to the size of the file. But in the proposed method non-sequential order is introduced. Proposed method uses three servers say server1, server2 and server3. The chunked files can be stored in the servers in twelve combinations like

Server1, server 2 and server 3 (1, 2, 3)
Server 1, server 3 and server 2 (1, 3, 2)
Server 1, server 2 and server 3 (1, 2, 3)
Server 1, server 3 and server 3 (1, 3, 3)
Server 2, server 1 and server 3(2, 1, 3)
Server 2, server 3 and server 1(2, 3, 1)
Server 2, server 1 and server 1(2, 1, 1)
Server 2, server 3 and server 3(2, 3, 3)
Server 3, server 1 and server 2(3, 1, 2)
Server 3, server 2 and server 1(3, 2, 1)
Server 3, server 1 and server 1(3, 1, 1)
Server 3, server 2 and server 2 (3, 2, 2)

Among these the unique combinations are selected to avoid the storage of file redundancy in the same server. The five unique combinations used in this research work are:

Server 1, server 3 and server 2
Server 2, server 1 and server 3
Server 2, server 3 and server 1
Server 3, server 1 and server 2
Server 3, server 2 and server 1

The chunked files will be stored in unique combination of servers as given above.

D. Encryption and Storage

In this method, after the chunking process is over, the chunked files will be encrypted and will be stored in different servers. The Advanced Encryption Standard algorithm is used for encryption and decryption process.

IV. Experimental Design

The proposed method is implemented using JAVA 1.7 as front end and mysql as a back end for experimentation. The proposed method is evaluated for performance in terms of attack detection rate, false acceptance rate and false rejection rate. The aim of this section is to evaluate the efficiency of the proposed method in terms of some performance metrics. The performance of the proposed method is evaluated using the following performance metrics:

A. False Acceptance Rate

The false acceptance rate is a fraction of negative entry or unauthorized user was incorrectly identified as positive entry or unauthorized user and it will be calculated using the following formula:

$$\text{False Acceptance Rate} = \frac{number\ of\ false\ rejections}{number\ of\ client\ accesses}$$

B. False Rejection Rate

The false rejection rate is a fraction of positive entry or unauthorized user that was correctly identified as negative entry or unauthorized user and it will be calculated using the following formula:

$$\text{False Rejection Rate} = \frac{number\ of\ false\ acceptances}{number\ of\ client\ accessses}$$

C. Attack Detection Rate

The proposed method is intended to detect the cyber attacks. The attack detection rate is calculated using the following equation:

$$\text{Detection Rate} = \frac{N_{attack}}{N_{total}}$$

V. Results

The results inferred using the proposed method is given in table 2, table 3 and table 4.

TABLE II. RESULTS OF FAR AND FRR FOR EVERY 50 USERS FOR SEQUENTIAL ORDER

Users	False Acceptance Rate	False Rejection Rate
50	0.88	0.34
100	0.84	0.30
150	0.82	0.27
200	0.77	0.22

The results have been taken for every 50 users registering with the proposed method is given in table.2 The above table shows the false acceptance rate and false rejection rate evaluated using the proposed method. The attack detection rate is given below in table.3.

The attack infected files are uploaded along with the normal files to the proposed method. The proposed method detects 57% infected files before chunking process takes place.

TABLE III. CYBER ATTACK DETECTION RATE

Attack Types	Sequential Order	Proposed Non-sequential Order
Active Attacks	84%	88%
Passive Attacks	73%	79%

TABLE IV. DETECTION RATE OF INFECTED FILE BY THE PROPOSED METHOD

File Type	Detection Rate	
	Sequential Order	Proposed Non-sequential Order
Normal Files	100%	100%
Attack Infected Files	50%	57%

VI. DISCUSSION AND CONCLUSION

The main idea of this phase is to develop a model which ensures strict permission and verification in accessing the files stored in the database. This model is developed by rearranging the existing sequential order in saving the chunked files into the server. This proposed method provides security system like user authentication, secure communication and Integrity. The method developed is to defend against cyber attacks say active and passive attacks. The primary focus of this model is to provide easy and secured access to the legitimate user which ensures availability, secure storage. The proposed method is developed to evaluate the performance of the data chunking for security and preventing data or information from cyber attacks. The proposed method outperforms 7% in detecting the cyber attacks than the existing method. Also this model helps in detecting the cyber attacks accurately.

REFERENCES

[1]. Arul Selvan and Dr K Porkumaran, "Two Stage Max Gain Content Defined Chunking for De-duplication" International Journal of Engineering Research and Development, Volume 4, Issue 4, 2012, pp.1-6.

[2]. Bassem Bakhache et al., "Improvement of the Security of ZigBee by a New Chaotic Algorithm" IEEE Systems Journal, 2013, pp.1-10.

[3]. Deborah S.Carstens, et al., "Applying Chunking Theory in Organizational Password Guidelines" Journal of Information, Information Technology, and Organizations Volume 1, 2006, pp.97-113.

[4]. Diego Perino, et al., "ICN-RE: Redundancy Elimination for Information-Centric Networking" Proceedings of ICN'12 ACM, 2012, pp. 91 – 96.

[5]. G. Hanumantha Rao, et al., "Security Providing using Blowfish, RSA and SHA-512 Algorithms" International Journal of Computer Science And Technology Vol. 3, Issue 3,2012, pp.no. 1017 – 1019.

[6]. Imtiaz Ahmad and A. Shoba Das, "Hardware implementation analysis of SHA-256 and SHA-512 algorithms on FPGAs",Elsevier Computers and Electrical Engineering 31 (2005) pp.345–360.

[7]. Jason Van Dyken and Jose G. Delgado-Frias, "FPGA schemes for minimizing the power-throughput trade-off in executing the Advanced Encryption Standard algorithm" Elsevier, Journal of Systems Architecture 56,2010, pp.116–123.

[8]. Jurgen Kaiser, et al., "Design of an Exact Data Deduplication Cluster" IEEE 2013

[9]. Konrad J. Kulikowski et al., "Robust codes and robust, fault-tolerant architectures of the Advanced Encryption Standard" Elsevier Journal of Systems Architecture 53, 2007 139–149.

[10].Ma Jianting, "A Deduplication-based Data Archiving System" 2012 International Conference on Image, Vision and Computing (ICIVC 2012).

[11].Marcio Juliato and Catherine Gebotys, "A Quantitative Analysis of a Novel SEU-Resistant SHA-2 and HMAC Architecture for Space Missions Security" IEEE Transactions on Aerospace and Electronic Systems Vol.49,No.3 July 2013, pp.no.1536 – 1554.

[12].Mark W. Storer, et al.,"Secure Data Deduplication" Proceedings of StorageSS'08, ACM,, 2008.

[13].Prof.V R Kulkarni and Dr. S S Apte, "Alternate Approach for Implementation of SHA-512 Algorithm using Feed forward Neural Network" International Journal of Computer Applications (0975 – 8887) Volume 28– No.5, 2011, pp.no.30-31.

[14].Reham Abdellatif Abouhogail, "New multicast authentication protocol for entrusted members using advanced encryption standard" Elsevier, The Egyptian Journal of Remote Sensing and Space Sciences, 2011 14, pp.no.121–128.

[15].Ryan Glabb et al., "Multi-mode operator for SHA-2 hash functions", Elsevier Journal of Systems Architecture 53, 2007, pp.no.127–138.
[16].Seiichi Ozawa et al., "Incremental Learning of Chunk Data for Online Pattern Classification Systems" IEEE Transactions on Neural Networks 2008, pp.no.1-14.
[17].Simon Heron, "Advanced Encryption Standard (AES)" Elsevier, Network Security Volume 2009, Issue 12, 2009, Pages 8–12.
[18].Tianming Yang, et al., "Alternatives for Eliminating Duplicate in Data Storage" International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013), pp.no.565-568.
[19].Tong Zhang, et al., "Text Chunking based on a Generalization of Winnow" Journal of Machine Learning Research 2, 2002, pp.no.615-637.
[20].Xinmiao Zhang and Keshab K. Parhi "High-Speed VLSI Architectures for the AES Algorithm" IEEE Transactions on very large scale Integration (VLSI), Vol. 12, No.9, 2004, pp.no. 957 – 967.
[21].Xinmiao Zhang,and Keshab K. Parhi, "On the Optimum constructions of Composite Field for the AES Algorithm" IEEE Transactions on circuits and systems – II Express Briefs, Vol.53, No.10,2006, pp.no.1153- 1157.
[22].Yuanjian Xing, et al., "PeerDedupe: Insights into the Peer-assisted Sampling Deduplication" Proceedings of IEEE P2P 2010.

BIOGRAPHIES

M.Uma is a Ph.D. research scholar of Avinashilingam Deemed University, currently doing research on cyber security. Her areas of interest include Information and communication Security. She has 11 publications in her research work. She is currently the principal investigator for one project funded by DST (WOS-A). She is a reviewer for WSEAS, IJSET and TIJCSA.

Dr.G.Padmavathi is the Professor and Head of computer science of Avinashilingam Deemed University for women, Coimbatore. She has 23 years of teaching experience and one year of industrial experience. Her areas of interest include Real Time Communication, Network Security and Cryptography. She has 200 publications in her reascher area. Presently she is guiding M.phil researcher and PhD's Scholar. She has been profiled in various Organizations her academic contributions. She is currently the principal investigator of four projects funded by UGC and DRDO. She is the scientific mentor for one project funded by DST. She is life member of many preferred organizations of CSI, ISTE, WSEAS, AACE, and ACRS.