INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

Improving the Speed of Data Leakage

Pooja A. Dagadkhair
G. H. Raisoni College of Engineering & Management, Ahmednagar
poojadagadkhair6@gmail.com

Vidya Jagtap
G. H. Raisoni College of Engineering & Management, Ahmednagar
vidya.jagtap@raisoni.net

Abstract: A human mistakes are one of the main causes of data loss.in recent year, the leak of sensitive data has grown rapidly for research institutions. The exposure of sensitive data in storage and transmission poses a serious threat to organizational and personal security. Commonly, information is contained in storage and transmission for exposed sensitive data. Such an approach usually requires the detection operation to be conducted in secret and leakages of the data are not able to predict. However, detecting the data leak by scanning information or files some detectors are available into the market. (e.g. Antivirus(AV),Network Intrusion Detection System(NIDS)). In this paper, for detecting large and inexact sensitive data patterns we use a principal component analysis algorithm and advanced encryption algorithm. Our algorithm is designed for high detection accuracy. This detection is presented by parallel execution of sampling and AES algorithm. Our system is more efficient for detection of accuracy in recognizing transform leaks. We encrypt the important data from owner and receiver side and send the data leak detection result to the owner. We describe the high encryption from both side for getting high security and maintaining the secrecy.

Keywords*: Content Inspection, Data Leakage, Encryption, Parallelism, Privacy, Sampling.*

I. INTRODUCTION

Sensitive data could be of different types, it could be related details about a client, employee, health records of an individual, finance, credit cards details etc. Sensitive data is the information where the disclosure is protected by laws and regulations and mainly by the organization policy. The loss of sensitive data leads to financial damage and the reputation of an organization. This loss of sensitive information affects the organization, customer and the external parties whose information are compromised. Thus the leakage of sensitive information should be prevented from unauthorized transmission of data to the public domain. While we understand the importance of the data security and data loss, it's also necessary to understand the impact of the data losses in Today scenario. Whenever is the data transform from owner to client that time it goes through the server and the attackers are ready to hack that important data at that situation what we do? So we encrypt the data first when it transform from the owner and then send to the next operations. In our system, we analysis data and identify exact data as per data pattern. The sensitive data is transmitted in a plain text without security it will be a hack and may be misuses of that important data. Existing data leak detection approaches are based on the set intersection. Set intersection is performed on two sets of *n*-grams, one from the content and one from sensitive data. The set intersection gives a number of sensitive *n*-grams appearing in the content. The method has been used to detect similar documents on the web [10], shared malicious traffic patterns [11], malware [12], as well as email spam [13]. The advantage of *n*-grams is the extraction of local features of a string, enabling the comparison to tolerate discrepancies. Some advanced versions of the set intersection method utilize Bloom filter, e.g., [14], which trades accuracy for space complexity and speed. Shu and Yao extended the standard use of *n*-grams and introduced data-leak detection as a service. They proposed the first solution for detecting accidental data leak with semi-honest providers [15].

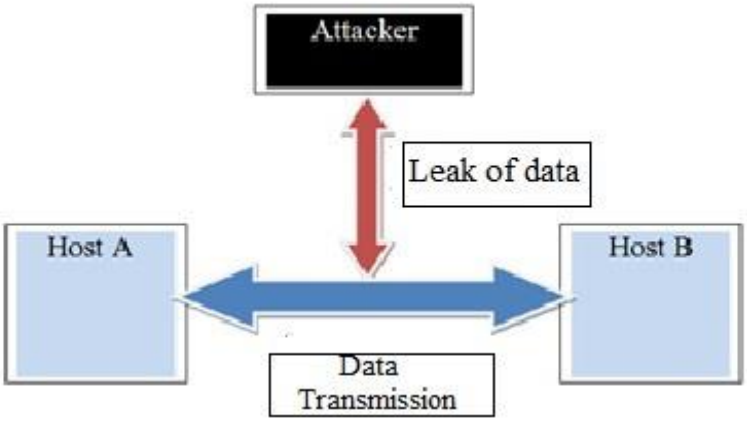

Fig. 1. Data Leakage in Transmission

In this above Figure, 1 shows that the Host A sends and receive the data from Host B. The communication between Host A and B the process which is continuously done. At that time the attacker directly intersects with their communication process and into the middle, it accesses all data of Host A and Host B.

Nowadays for avoiding this type of problem lots of applications available into the market like anti-virus like NPAV, Kaspersky, Avast etc. Which protect the sensitive data of the system. The sensitive data has stored and transmission form, so the attacker who leaks the data get a chance to access it. For protecting the sensitive data while it stored or in transmission from it will be directly detected to the sender and receiver.The detection of where data leak and track of leakage is done and finding out the who access our important data. Our implementation is beneficial for controlling the cybercrime. For example, the rising market in today's world is the consumer market. E-commerce had made anything and everything possible to buy in today's world. While this is nice to have, it has created high-risk areas on information security. Organizations are compelled to provide accessibility by all means like mobile, kiosks, laptops, i pads, etc. And consumer could be anyone. So the information of an organization is available in all the places wherever there is connectivity. This increases susceptibility, sharing of data through social media and accessible to unintended data owners. Like that situation, our detection technique is more efficient and give the accurate result.

II. TECHNOLOGY

The system work presents an efficient sequence comparison technique needed for analyzing a large amount of content for sensitive data exposure. Our detection approach consists of a comparable sampling algorithm and a sampling oblivious alignment algorithm. The pair of algorithms is more useful for managing the content of sensitive data set.

The system design enables the detection of partial data leaks. Our detection runs on continuous sequences of n bytes .i) the subsequence preserving sampling algorithm increases the power to detect distant homologous. We set out to determine whether sub-sampling of sequences for alignment profile searches and the combination of the sub-sample search results, was superior to a single search using the entire alignment. The purpose of our comparable sampling operation is to enhance the analysis throughput. System defines the sampling requirement needed in data leak detection. ii) The system uses Principal Component Analysis as a tool to quickly identify a correlation between futures, helping feature extraction and selection.

III.PROPOSED SYSTEM

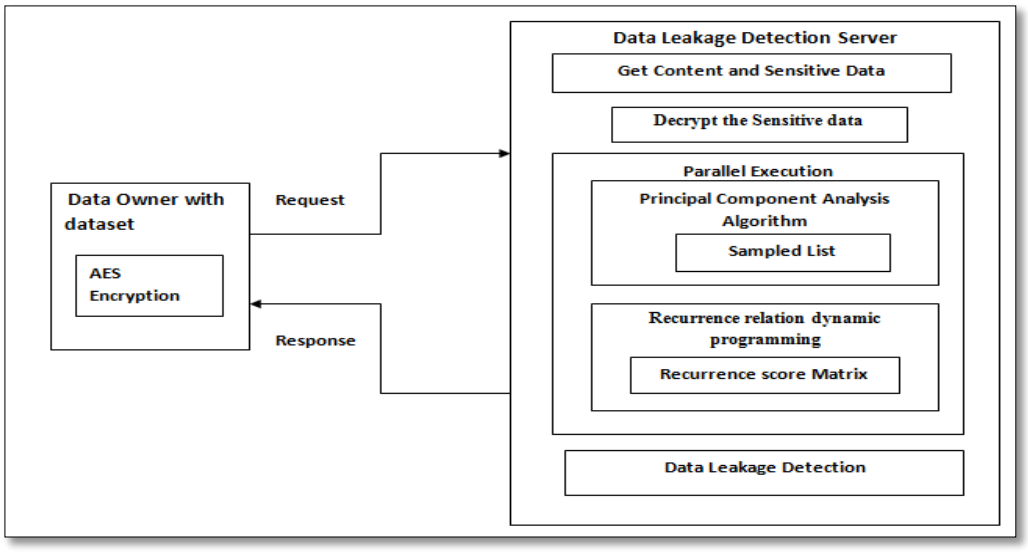

Fig2: Proposed Architecture

The system mainly focuses on detecting inadvertent data leaks, and we assume the content in the file system or network traffic (over supervised network channels) is available to the inspection system. A supervised network channel could be an unencrypted channel or an encrypted channel where the content in it can be extracted and checked by an authority.

The proposed system effective using the AES encryption and how the data is going to PCA for the sample list, and recursive relation dynamic programming for the sensitive score matrix which are contained into the parallel execution. The cryptosystem involves a set of rules for how to encrypt the plain text and how to decrypt the cipher text. The encryption and decryption required a key, which uses for encryption and decryption. The AES encryption having the two type of encryption algorithms. Sometimes the encryption and decryption keys are the same this is called symmetric encryption and because of it having the secret key for encrypting the data from data owner.After sending a request the data leak detection server which having sensitive data it goes through the parallel execution.

IV.SYSTEM ANALYSIS

The parallel execution of Principal Component Analysis algorithm and Recurrence relation dynamic programming algorithm as follows:

1) Principal component analysis Algorithm: Principal component analysis algorithm it is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences. PCA is a powerful tool for analyzing data. The main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. By reducing the number of dimensions, without much loss of information.
2) Recurrence relation dynamic programming algorithm: We present the recurrence relation of our dynamic program alignment algorithm in Algorithm 2. We present the recurrence relation of our dynamic program alignment algorithm in Algorithm 2. For the *i* -the item *Li* in a sampled sequence *L* (the compact form), the field *Li .value* denotes the value of the item and a new field *Li .span* denotes the size of the null region between that item and the preceding non-null item. Our local alignment algorithm takes in two sampled sequences *La* and *Lb*, computes a non-negative score matrix *H* of size |*La*|-by-|*Lb*|, and returns the maximum alignment score with respect to a weight function. Each cell *H(i, j)* has a score field *H(i, j).score* and two extra fields recording sizes of neighboring null regions, namely *null row* and *null cool*. The intermediate solutions are stored in matrix *H*. For each subproblem, three previous subproblems are investigated: *i)* aligning the current elements without a gap, which leads to a *match* or *mismatch*, *ii)* aligning the current element in *La* with a gap, and *iii)* aligning the current element in *Lb* with a gap. A cell candidate *h* is generated for each situation; its score *h.score* is computed via the weight function *few* (lines 1 to 3 in Algorithm 2). The other two fields, *null row* and *nullcol* , are updated for each candidate cell (lines 4 to 9). This update may utilize the null region value stored in the *span* field of an item. All three cell candidates *hup*, *hle f t* , and *hdia* are prepared. The cell candidate having the highest score is chosen as *H(i, j)*, and the score is stored in *H(i, j).score*.

IV.ALGORITHM

Algorithm 1: Recurrence Relation in Dynamic Programming

Input: A weight function *fw*, visited cells in *H* matrix that are adjacent to *H(i, j)*: *H(i* −1, *j* −1*)*, *H(i, j* −1*)*, and *H(i* −1, *j)*, and the *i* -th and *j* -th items *Lai* ,*Lbj* in two sampled sequences *La* and *Lb*, respectively.

Output: *H(i, j)*

(1) *hup.score* ← *fw(Lai*, -, *H(i* − 1, *j))*

2)*hle f t .score* ← *fw(-,Lbj , H(i, j* − 1*))*

(3) *hdia.score* ← *fw(Lai ,Lbj , H(i* − 1, *j* − 1*))*

(4) *hup.nullrow* ← 0

(5) *hup.nullcol* ← 0

(6)*hle f t .nullrow* ← 0

(7) $hdia.nullrow \leftarrow \begin{cases} 0, & \text{if } Lai = Lbj \\ H(i-1,j).nullrow + Lai.span + 1, & else \end{cases}$

(8) $hdia.nullcol \leftarrow \begin{cases} 0, & \text{if } Lai = Lbj \\ H(i,j-1).nullcol + Lbi.span + 1, & else \end{cases}$

(9) $H(i,j) \leftarrow \arg\max h.score \begin{cases} Hup \\ hle\ f\ t \\ hdia \end{cases}$

(10) $H(i,j).score \leftarrow \max \begin{cases} 0 \\ H(i,j).score \end{cases}$

Algorithm 2: Principal Component Analysis Algorithm (PCA)

Input:

Transform an n*d matrix x into an n*m matrix y.

(1) Centralized the data (subtract the mean).

(2) Calculate the d*d covariance matrix:

$$C = \frac{1}{N-1} XTX$$

- $Ci,j = \frac{1}{N-1} \sum_{q=1}^{N} Xq,i.Xq,j$
- Ci,j(diagonal) is the variance of variable i.
- Ci,j(off-diagonal) is the covariance between variables I and j.

(3) Calculate the eigenvectors of the covariance matrix (orthonormal).
(4) Select m eigenvectors that correspond to the largest m
eigenvalues to be the new basis.

V.MATHEMATICAL MODEL

1. Send Encrypted Input text data
   data owner Encrypted send text data to DLD server, to verify whether the data is leaked
   Input-sensitive text data = {t1, t2, …., tn}
   Where,
   tn are the n number of sensitive text data
   For Encryption AES Algorithm is used.
2. At DLD server, two datasets are used D = {D1, D2}
   Where,
   D1 = Enron Dataset{Contain Sensitive Data}
   D2 = HTTPRequest { Contain Content Data}
3. Perform Parallel Execution of Data leakage detection
   A = {A1, A2}
   Where,
   A1 and A2 are the two algorithms required for data leakage detection
   A2= Recurrence Relevance in Dynamic Programming A2 perform number of operations
   A1= {b1, b2}
   b1= Sampling Oblivion
   b2= Recurrence Relation
   Output = Sampling Score in Matrix Form

VI. SYSTEM REQUIREMENT SPECIFICATION

A software requirements specification (SRS) is a comprehensive description of the intended purpose and environment for software under development. The SRS fully describes what the software will do and how it will be expected to perform.

A) Software Requirement Specification

Requirements	Specification
Platform	Java
Tools	JDK 1.5 and above, Eclipse, VMware
Operating System	Windows 7 & above

Table 1: Software requirements specification

1) External Interface Requirements

1.1 User Interfaces

The user will be provided with a simple and user-friendly GUI. The GUI components of the System will be developed using REST Full Web Service , Mobile App is designed using Android SDK, Amzon Cloud Server.

1.2 Hardware Interfaces

Requirements	Specification
Processor	Core 2 and Above
RAM	512 MB to Above
HDD	80 and Above

Table 2: Hardware Requirements Specification

1.3 Safety Requirements

The data handled in the Server system is very vital. The server should always be confirmed to run properly and the data are saved to the database at consecutive intervals.

1.4 Security Requirements

The security system features from having a login for all the user sto access the software.
The login details are encrypted will be used in the system also. AES is used for encryption and decryption.

1.5 Software Quality Attributes Reliability

The system will be designed with reliability as key feature

- The system is guaranteed of providing the services to a user according to his login information.
- This system is guaranteed to be reliable with maximum time.

Maintainability

The system will be developed using the standard software development conventions to help in easy review and redesigning of the system.

Availability

The system is available to continue while transforming data into the network.

Supportability

The system is able to support all type of data.

VII. PERFORMANCE EVALUATION

A) System Flow

a) The data owner sends a request to data leak detection server.
(b) The data leak detection server access a sensitive data in an encrypted form.
(c) Encrypted sensitive data is decrypted into the normal data.

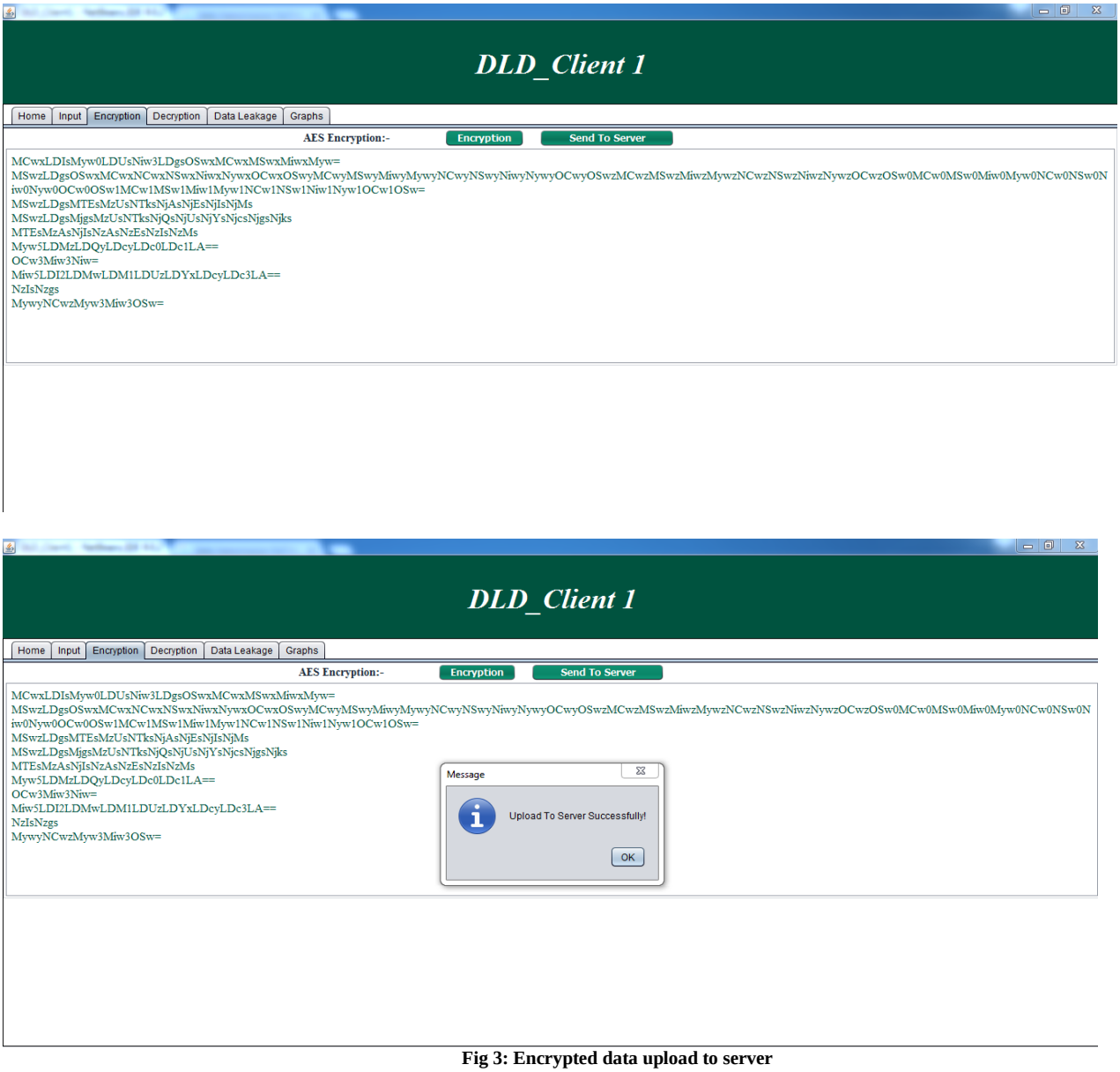

Fig 3: Encrypted data upload to server

(d) Parallel execution of Principal Component Analysis Algorithm and Recurrence relation dynamic programming algorithm.

Fig 4 : Result of PCA Algorithm on server

(e) Principal Component Analysis Algorithm is specified the sampled list of a content data .where it performs some operation like variable mean & variable standard deviation and also identifies the cross product matrix.
(f) The Recurrence relation dynamic programming algorithm which forms the score matrix of stored or transmitted data which shows the alignment and track it.

VIII.COMPARISION OF SIMILAR SYSTEMS

In the existing system, only inspection technique are available for detecting leaks of sensitive information in the content of files or network traffic. AES Encryption Decryption algorithm is used for security purpose. The detecting multiple common data leak scenarios but whereas in the proposed system parallel version of PCA and Recursive relation algorithm which quickly identify the exact data and extract it. In existing system detection approach is based on aligning two sampled sequences for similarity comparison where, as in proposed system enhances the scalability, speed, accuracy, privacy, and efficiency of the system. In proposed system output of the PCA algorithm is getting into direct matrix form which is more time efficient than the existing system.

RESULT

The system is partially ruined for generating time and memory graph & in below figure result is shown:

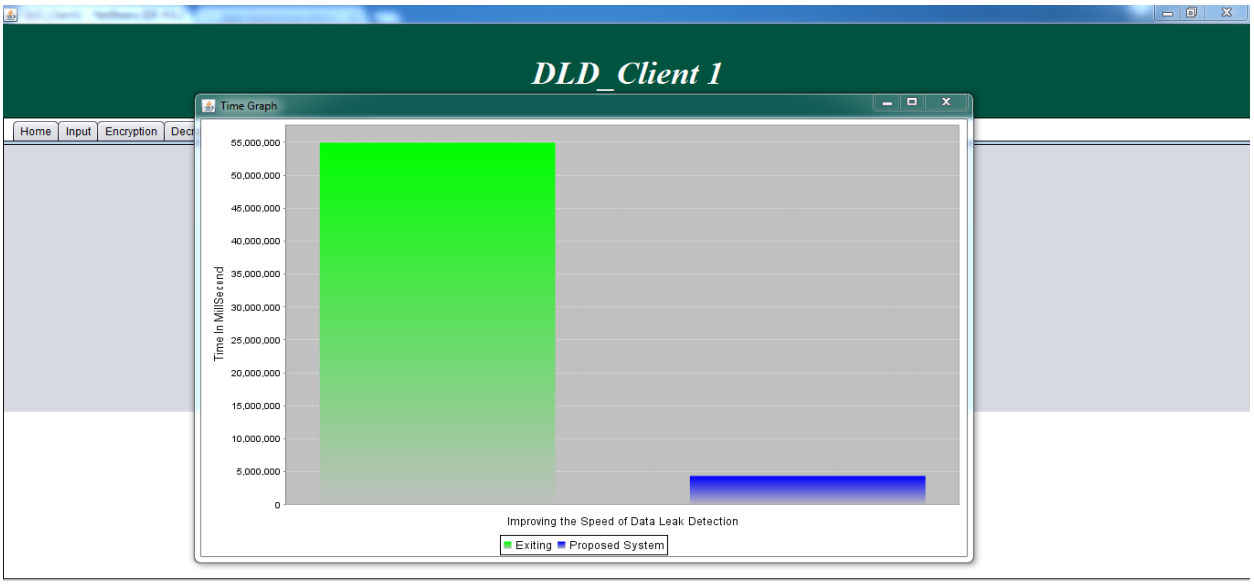

Fig 5: Time Graph

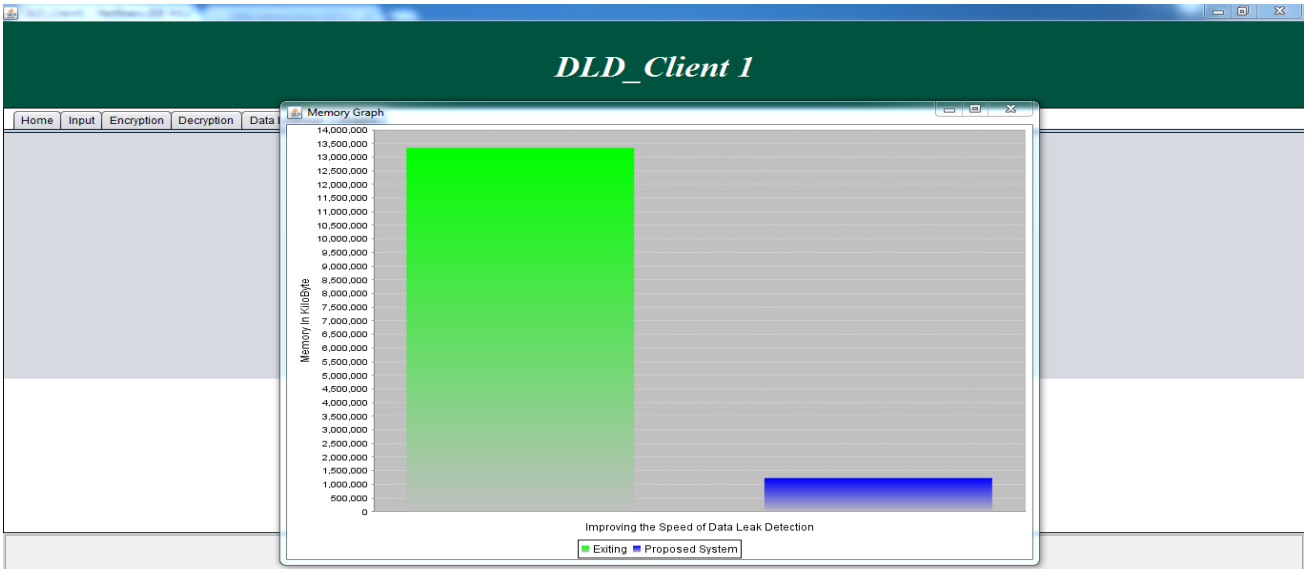

Fig 6: Memory Graph

ACKNOWLEDGMENT

I Miss Pooja Dagadkhair would like to thank Prof. Vidya Jagtap for sharing their pearls of wisdom with me during the course of this research. I am immensely grateful to her comments for designing the system.

CONCLUSIONS

In this work, we have proposed the scalability and speed of the system, we will use parallel approach. We have conducted extensive experiments to validate the accuracy, privacy, and efficiency of our solutions. The System is reducing dimensionality using PCA technique can help us achieve better and quicker results. An AES is more useful for security purpose which maintains the privacy of the data and data should be securely sent to the server. We are evaluating the performance of proposed system with several datasets and prove that the system is accurate and efficient in data leakage detection procedure.

REFERENCES

[1] Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao, Senior Member, IEEE, and wu-chun feng, senior Member, IEEE,"Fast Detection of Transformed Data Leaks", IEEE. Transactions on Information Forensics and Security, vol. 11, no. 3, March 2016.
[2] X. Shu, D. Yao, and E. Bertino, \Privacy-Preserving Detection of Sensitive Data Exposure," May 2015.
[3] F. Liu, X. Shu, D. Yao, and A. R. Butt, \Privacy-Preserving Scanning of Big content for Sensitive Data Exposure with MapReduce," 2015
[4] R. Hoyle, S. Patil, D. White, J. Dawson, P. Whalen, and A. Kapadia, \Attire: Conveying information exposure through avatar apparel," in Pro. Conf. Comput. Supported Cooperative. Work Companion (CSCW), 2013.
[5] A. Nadkarni and W. Enck, \Preventing Accidental Data Disclosure in modern Operating Systems," 2013.
[6] R. Hoyle, S. Patil, D. White, J. Dawson, P. Whalen, And A. Kapadia, \Attire: Conveying Information Exposure Through avatar apparel," 2013.
[7] Wen, Yan, Jinjing Zhao, and Hua Chen. "Towards Thwarting Data Leakage with Memory Page Access Interception." Dependable, Autonomic and Secure Computing (dasc), 2014 IEEE 12th international conference on. IEEE, 2014.
[8] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang,\Appintent: Analyzing Sensitive Data Transmission in Android for Privacy Leakage detection," in proc. 20th ACM CONF. Comput. Commun. Secur, 2013, pp. 1043{1054.
[9] Global Velocity inc. (2015). Cloud data security from the inside out| Global velocity. [Online]. Available: http://www.globalvelocity.com/,Accessed Feb. 2015.
[10] GTB Technologies INC. (2015). Goclouddlp. [Online]. Available: Http://www.goclouddlp.com/, accessed Feb. 2015.
[11] M. Cai, K. Hwang, Y.-K. Kwok, S. Song, and Y. Chen, "Collaborative Internet worm containment," *IEEE Security Privacy*, vol. 3, no. 3, pp. 25–33, May/Jun. 2005.
[12] J. Jang, D. Brumley, and S. Venkataraman, "BitShred: Feature hashing malware for scalable triage and semantic analysis," in *Proc. 18th ACM Conf. Comput. Commun. Secure. (CCS)*, 2011, pp. 309–320.
[13] K. Li, Z. Zhong, and L. Ramaswamy, "Privacy-aware collaborative spam filtering," *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 5, pp. 725– 739, May 2009.
[14] Symantec. (2015). *Symantec Data Loss Prevention*. [Online]. Available: http://www.symantec.com/data-loss-prevention, accessed Feb. 2015.
[15] X. Shu and D. Yao, "Data leak detection as a service," in *Proc. 8th Int. Conf. Secure. Privacy Commun. Netw. (Secure Comm)*, Padua, Italy, Sep. 2012, pp. 222–240.

