



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 2, February 2017

## Detection of Data Leaks on Mail Server Using Lucene Framework

B.H.Swaroop Kumar<sup>1</sup>, A.Sharan<sup>1</sup>, N.Sabarish<sup>1</sup>, Dr.R.Sugumar<sup>2</sup>

B.E Student, Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India<sup>1</sup>

Professor, Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India<sup>2</sup>

**ABSTRACT:** Statistics from security firms, research institutions and government organizations show that the number of data-leak instances has grown rapidly in recent years. The objective of this system is to provide security against sensitive data, using data leak detection technique on the transformed data. Human mistakes are one of the main causes of data loss. According to a report from Risk Based Security (RBS) the number of leaked sensitive data records has increased dramatically during the last few years, i.e., from 412 million in 2012 to 822 million in 2013. Deliberately planned attacks and inadvertent leaks remain the other causes. Detecting and preventing data leaks requires a set of complementary solutions, which uses Lucene framework for detecting data leaks and to know the sender who tries to send the sensitive data.

**KEY WORDS:** Data-leak detection, Data packet inspection.

### I. INTRODUCTION

As the internet grows and network bandwidth continues to increase, administrators are faced with the task of keeping confidential information from leaving their networks. Today's network traffic is so voluminous that manual inspection would be unreasonably expensive. In response, the researchers have created data loss prevention systems that check outgoing traffic for known confidential information. These systems stop naïve adversaries from leaking data, but are fundamentally unable to identify encrypted or obfuscated information leaks. What remains is a high-capacity pipe for tunneling data to the Internet. An approach for quantifying information leak capacity in network traffic is used. Instead of trying to detect the presence of sensitive data—an impossible task in the general case—our goal is to measure and constrain its maximum volume. The advantage of the insight is taken into account that most network traffic is repeated or determined by external information, such as protocol specifications or messages sent by a server. By filtering this data, we can isolate and quantify true information flowing from a computer.

### II. LITERATURE SURVEY

Yeongjin Jang (2014) proposed a way to capture the user's intent through their interactions with an application so that it is easy to verify that the resulting system output can be mapped back to the user's interactions. The disadvantage is that Mapping back to the user interaction is a time consuming task. Domenico Ficara (2010) had used Deterministic Finite Automata (DFA) to match regular expressions so that the outsourcing content similar to the content is matched with the regular expressions. But, Automata construction is required for every information which is stored. Sailesh Kumar (2007) had introduced the Collection of signatures of known security threats and viruses. This system is incapable of efficiently keeping track of counts. Somesh Jha (2008) had proposed the approach of Calculating the edit distance and Smith-Waterman similarity scores between two sequences. The disadvantage of this system is circuit representation requires several Gigabytes of memory. Kevin Borders (2009) had focused on quantifying information leak capacity in network traffic and detecting the presence of sensitive data. Analyzing network traffic leaks would be unreasonably expensive and error-prone which the drawback.



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirccce.com](http://www.ijirccce.com)

Vol. 5, Issue 2, February 2017

## III. EXISTING SYSTEM

In the existing system a straight forward realizations of data-leak detection require the plaintext Sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is compromised, then it may expose the plaintext Sensitive data. In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext. There is no privacy preserving in the existing system, so providers can access the data without data-owners permission.

### A. PROBLEM DEFINITION

1. Inadvertent data leak: The Sensitive data is accidentally leaked in the outbound traffic by a legitimate user. Inadvertent data leak may be due to human errors such as forgetting to use encryption, carelessly forwarding an internal email and attachments to outsiders.
2. Malicious data leak: A rogue insider or a piece of stealthy software may steal Sensitive personal or organizational data from a host.
3. Data Traffic and Time consumption: Data traffic on proxy server and mail server impacts the performance of data leakage detection technique and time delay of common legitimate users.
4. Static filtering of authorized users: Static approaches of authorized user filtering technique affect the efficient of data leakage detection

## IV. PROPOSED SYSTEM

A data-leak detection solution is proposed which can be outsourced from an organization. To avoid data leaks and provide privacy preserving to sensitive data Lucene search engine framework, Levenshtein-distance technique are used. Two most important players in the proposed model is

**Data Owner** owns the Sensitive data and authorizes the DLD provider to inspect the network traffic from the organizational networks for anomalies, namely inadvertent data leak.

**Mail Server - DLD provider** inspects the network traffic for potential data leaks. The focus is on detecting inadvertent data leaks along with the assumption that the content in

file system or network 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. Authority has the threshold for every categorized position of users. In security model the assumption is that the analysis system is secure and trustworthy. Privacy-preserved data-leak detection can be achieved by leveraging special protocols and computation steps. It is another functionality of a detection system.

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirccce.com](http://www.ijirccce.com)

Vol. 5, Issue 2, February 2017

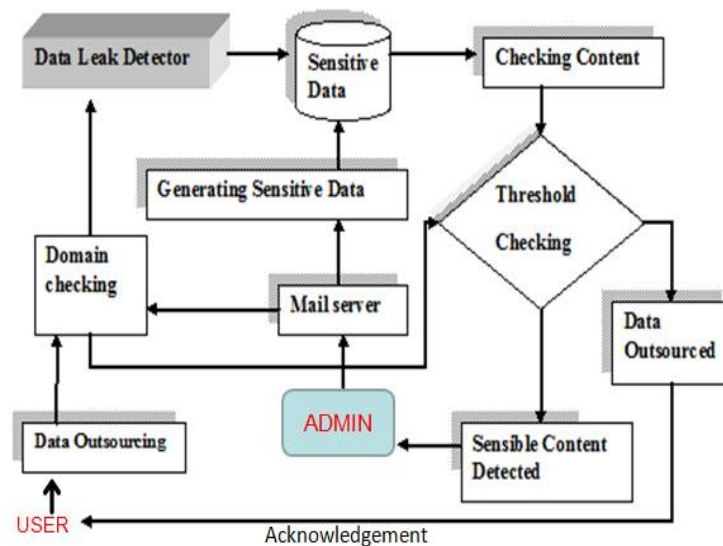


Fig.1 Architecture of Data leak Detection

The web service is implemented to maintain the users and Sensitive content instead of data bases because of static implementation and rough data handling. Even the Sensitive data storage has to be preserved from threats in existing system. For that purpose the Sensitive data is maintained in the cloud

## V. RESULTS

The setup of the portals with their description is as follows

### CONTENT OUTSOURCING WITH DLD

All the outsourced contents are check with sensitive data. The sensitive data are maintained in a index file. Using this index file, DLD identifies the sensitive data concurrently with domain filtering and threshold assigning based on their email domain.

### THRESHOLD ASSIGNMENT

Each user is assigned a threshold for outsourcing the contents based on his designation with the organisation.

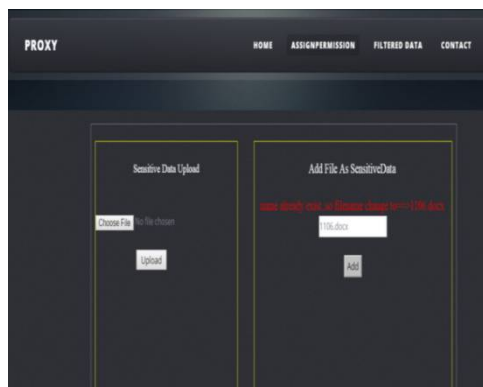


Fig.2 Sensitive Context Upload in proxy server



Fig.3 Threshold Assignment

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirccce.com](http://www.ijirccce.com)

Vol. 5, Issue 2, February 2017

## DATA LEAK DETECTION

If the sensible content percentage of transferred file exceeds the threshold percentage, an alert mail is triggered to the Admin of the proxy mail server. Alert mail consists of entire details about the users even what are the sensible contents are pings from the transferred content by the DLD framework.

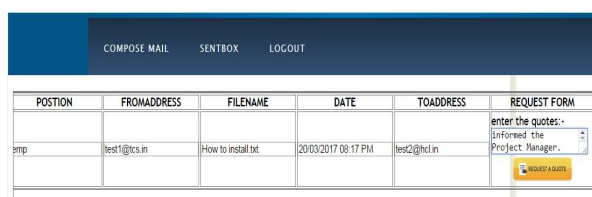


FROM-MAIL-ID	TO-MAIL-ID	FILE NAME	DATE	FILTERED CONTENT	ORIGINAL MAIL CONTENT
lll@hcl.in	ppp@tcs.in	How to install.txt	20/02/2017 07:39 PM	<p>mailcontent = sensitivedata</p> <p>Welcome=Welcome</p> <p>gooyaabitemplates=gooyaabitemplates</p> <p>support=support</p> <p>portal=portal</p> <p>How=How</p> <p>install=install</p> <p>upload=upload</p> <p>blogger=blogger</p> <p>template=gooyaabitemplates</p> <p>your=your</p> <p>Read=Read</p> <p>here=here</p> <p>https=https</p> <p>com=Welcome</p> <p>how=how</p> <p>If=if</p> <p>you=your</p> <p>need=need</p> <p>additional=additional</p> <p>Please=Please</p> <p>visit=visit</p> <p>forum=forum</p> <p>Welcome=Welcome</p> <p>gooyaabitemplates=gooyaabitemplates</p>	<p>Welcome to gooyaabitemplates support portal .....</p> <p>- How to install/upload blogger template in your blog</p> <p><a href="https://gooyaabitemplates.com/how-to-install-blogger-template">https://gooyaabitemplates.com/how-to-install-blogger-template</a></p> <p>additional support, Please visit.. <a href="https://gooyaabitemplates.com/support-forum/">https://gooyaabitemplates.com/support-forum/</a></p>

Fig.4 Filtered Context


## Quote Request

In few situations, the User will need to send the sensitive data but we would not have the permission to do so. He could send a request to the Admin to give permission to outsource the sensitive data. If the Admin gives permission, the user can outsource the data.



POSITION	FROMADDRESS	FILENAME	DATE	TOADDRESS	REQUEST FORM
emp	test1@tcs.in	How to install.txt	20/03/2017 08:17 PM	tes2@hcl.in	<p>enter the quote:-</p> <p>Informed the Project Manager.</p> <p><input type="button" value="REQUEST QUOTE"/></p>

Fig.5 User Quote Request



-ID	FILE NAME	DATE	QUOTE REASON	ORIGINAL MAIL CONTENT	POSITION	ACTION
	How to install.txt	20/03/2017 08:17 PM	This is the data which the	Welcome to gooyaabitemplates support portal .....	emp	<p><input type="button" value="Request a Quote"/></p> <p><input type="button" value="DECLINE"/></p>

Fig.6 Admin Quote Request List



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 2, February 2017

## VI. CONCLUSION

Implementing this system on a large scale can prevent any data using Data leak detection technique. The system can be further expanded or improvised by ensuring guidelines where any private or confidential information are prevented from outsourcing by the users.

## REFERENCES

- [1] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid and parallel content screening for detecting transformed data exposure," in *Proc. 3<sup>rd</sup> Int. Workshop Secur. Privacy Big Data (BigSecurity)*, Apr./May 2015, pp. 191–196.
- [2] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid screening of transformed dataleaks with efficient algorithms and parallel computing," in *Proc. 5<sup>th</sup> ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, San Antonio, TX, USA, Mar. 2015, pp. 147–149.
- [3] (Feb. 2015). *Data Breach QuickView: 2014 Data Breach Trends*. [Online]. Available: <https://www.riskbasedsecurity.com/reports/2014-YEDataBreachQuickView.pdf>, accessed Feb. 2015.
- [4] Kaspersky Lab. (2014). *Global Corporate IT Security Risks*. [Online]. Available: [http://media.kaspersky.com/en/business-security/Kaspersky\\_Global\\_IT\\_Security\\_Risks\\_Survey\\_report\\_Eng\\_final.pdf](http://media.kaspersky.com/en/business-security/Kaspersky_Global_IT_Security_Risks_Survey_report_Eng_final.pdf)
- [5] L. De Carli, R. Sommer, and S. Jha, "Beyond pattern matching: A concurrency model for stateful deep packet inspection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1378–1390.
- [6] A. V. Aho and M. J. Corasick, "Efficient string matching: An aid to bibliographic search," *Commun. ACM*, vol. 18, no. 6, pp. 333–340, Jun. 1975.
- [7] R. S. Boyer and J. S. Moore, "A fast string searching algorithm," *Commun. ACM*, vol. 20, no. 10, pp. 762–772, Oct. 1977.
- [8] S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese, "Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia," in *Proc. 3<sup>rd</sup> ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, 2007, pp. 155–164.
- [9] H. A. Kholidy, F. Baiardi, and S. Hariri, "DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks," *IEEE Trans. Dependable Secure Comput.*, vol. 12, no. 2, pp. 164–178, Mar./Apr. 2015.
- [10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Molecular Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
- [11] Global Velocity Inc. (2015). *Cloud Data Security from the Inside Out— Global Velocity*. [Online]. Available: <http://www.globalvelocity.com/>, accessed Feb. 2015.
- [12] GTB Technologies Inc. (2015). *GoCloudDLP*. [Online]. Available: <http://www.goclouddlp.com/>, accessed Feb. 2015.
- [13] V. Paxson, "Bro: A system for detecting network intruders in real-time," in *Proc. 7<sup>th</sup> Conf. USENIX Secur. Symp. (SSYM)*, vol. 7, 1998, p. 3.
- [14] P.-C. Lin, Y.-D. Lin, Y.-C. Lai, and T.-H. Lee, "Using string matching for deep packet inspection," *Computer*, vol. 41, no. 4, pp. 23–28, Apr. 2008.
- [15] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in *Proc. 23<sup>rd</sup> USENIX Secur. Symp.*, 2014, pp. 79–93.
- [16] K. Borders and A. Prakash, "Quantifying information leaks in outbound Web traffic," in *Proc. 30<sup>th</sup> IEEE Symp. Secur. Privacy (SP)*, May 2009, pp. 129–140.
- [17] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *Proc. IEEE Symp. Secur. Privacy*, May 2008, pp. 216–230.
- [18] D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procissi, and F. Vitucci, "Sampling techniques to accelerate pattern matching in network intrusion detection systems," in *Proc. IEEE Int. Conf. Commun.*, May 2010, pp. 1–5.