© January 2017 | IJIRT | Volume 3 Issue 8 | ISSN: 2349-6002

Improving the Speed of Data Leak Detection

Pooja A. Dagadkhair, Prof.Vidya Jagtap
Dept. of Computer Engineering,
G. H. Raisoni College of Engineering & Management,
Ahmednagar, India.

***Abstract*— Surveys from many years have shown that many data leakages has been found due different problems like malicious attacks, hacking, different attacks. The leak of sensitive data on computer systems poses a serious threat to organizational security. Statistics show that the lack of proper encryption on files and communications due to human errors is one of the leading causes of data loss. Organizations need tools to identify the exposure of sensitive data by screening the content in storage and transmission, i.e., to detect sensitive information being stored or transmitted in the clear. However, detecting the exposure of sensitive information is challenging due to data transformation in the content. Transformations (such as insertion and deletion) result in highly unpredictable leak patterns. In this paper, we utilize sequence alignment techniques for detection complex data-leak patterns. There algorithm is designed for detecting long and inexact sensitive data patterns. This detection is paired with a comparable sampling algorithm, which allows one to compare the similarity of two separately sampled sequences. There system achieves good detection accuracy in recognizing transformed leaks. The implement a parallelized version of an algorithms in graphics processing unit that achieves high analysis throughput.**

***Index Terms*— Fast data leakage, content inspection, sampling, alignment, parallelism**

I.INTRODUCTION

The number of leaked sensitive data records has grown 10 times in the last 4 years, and it reached a record high of 1.1 billion in 2014. A significant portion of the data leak incidents are due to human errors, for example, a lost or stolen laptop containing unencrypted sensitive files, or transmitting sensitive data without using end to end encryption such as PGP. A recent Kaspersky Lab survey shows that accidental leak by staff is the leading cause for internal data leaks in corporate. The data-leak risks posed by accidents exceed the risks posed by vulnerable software.

In order to minimize the exposure of sensitive data and documents, an organization needs to prevent clear text sensitive data from appearing in the storage or communication. A screening tool can be deployed to scan computer file systems, server storage, and inspect outbound network traffic. The tool searches for the occurrences of plaintext sensitive data in the content of files or network traffic. It alerts users and administrators of the identified data exposure vulnerabilities. For example, an organization's mail server can inspect the content of outbound email messages searching for sensitive data appearing in unencrypted messages.

II. RELATED WORK

Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao and Wu-Chun Feng,[2] proposed utilize sequence alignment techniques for detecting complex data-leak patterns. The algorithm is designed for detecting long and inexact sensitive data patterns. This detection is paired with a comparable sampling algorithm, which allows one to compare the similarity of two separately sampled sequences. There system achieves good detection accuracy in recognizing transformed leaks. Where they are implement a parallelized version of our algorithms in graphics processing unit that achieves high analysis throughput and demonstrate the high multithreading scalability of all data leak detection method required by a sizable organization.

In previous Data leak detection paper, author presented a content inspection technique for detecting leaks of sensitive information in the content of files or network traffic. There detection approach is based on aligning two sampled sequences for similarity comparison. The experimental results suggest that the alignment method is useful for detecting multiple common data leak scenarios. The parallel versions of the prototype provide substantial speedup and indicate high scalability of their design.

X. Shu, D. Yao, and E. Bertino [10] A privacy-preserving data-leak detection (DLD) solution to solve the issue where a special set of sensitive data digests is used in detection. The advantage of their method is that it enables the data owner to safely delegate the detection operation to a semi honest provider without revealing the sensitive data to the provider. They describe how Internet service providers can offer their

customers DLD as an add-on service with strong privacy guarantees. The evaluation results show that this method can support accurate detection with very small number of false alarms under various data-leak scenarios.

They are propose a data-leak detection solution which can be outsourced and be deployed in a semi-honest detection environment. A design, implement, and evaluate their fuzzy fingerprint technique that enhances data privacy during data-leak detection operations. There approach is based on a fast and practical one-way computation on the sensitive data.

F. Liu, X. Shu, D. Yao, and A. R. Butt [8] Data leak detection aims at scanning content (in storage or transmission) for exposed sensitive data. Because of the large content and data volume, such a screening algorithm needs to be scalable for a timely detection. There solution uses the MapReduce framework for detecting exposed sensitive content, because it has the ability to arbitrarily scale and utilize public resources for the task, such as Amazon EC2.

They design new MapReduce algorithms for computing collection intersection for data leak detection. There prototype implemented with the Hadoop system achieves 225 Mbps analysis throughput with 24 nodes. There algorithms support a useful privacy-preserving data transformation. This transformation enables the privacy-preserving technique to minimize the exposure of sensitive data during the detection. This transformation supports the secure out sourcing of the data leak detection to untrusted MapReduce and cloud providers.

Nadkarni and W. Enck [5] Aquifer as a policy framework and system for preventing accidental information disclosure in modern operating systems. In Aquifer, application developers define secrecy restrictions that protect the entire user interface workflow defining the user task. In doing so, Aquifer provides protection beyond simple permission checks and allows applications to retain control of data even after it is shared. Modern operating systems have changed both the way users use software and the underlying security architecture. These two changes make accidental data disclosures easier. To address this problem, they presented the Aquifer security framework that assigns host export restrictions on all data accessed as part of a UI workflow. There key insight was that when applications in modern operating systems share data, it is part of a larger workflow to perform a user task. Each application on the UI workflow is a potential data owner, and therefore can contribute to the security restrictions. The restrictions are retained with data as it is written to storage and propagated to future UI workflows that read it. In doing so, they enable applications to sensibly retain control of their data after it has been shared as part of the user's tasks.

R. Hoyle, S. Patil, D. White, J. Dawson, P. Whalen, and A. Kapadia[12] It can be important and useful to know about accesses to one's information by other parties. Knowing when, how,and by whom, one's information is accessed in practice can facilitate better informed and more effective personal privacy management. A challenge, however, is conveying information about accesses in ways that are concise and unobtrusive yet easily interpretable and noticeable. Toward this end, they built an app called Attire that uses the avatar metaphor to convey information access via changes in the avatar's attire and context. They offered suggestions for further exploration of the design space that include manipulating other aspects of the avatar besides clothing. User evaluation and field trials can shed light on the utility and effectiveness of these approaches and may also contribute further design enhancements.

III. PROPOSED WORK

In our Analytics paper evaluate existing system extends on large-scale Experiments. We proposed evaluate Fast data leak detection system on large scale experiments using a detection is coupled with a comparable sampling algorithm it compares the similarity of two separately sampled sequences.

Statistics from security firms, research institutions and government organizations show that the numbers of data-leakinstances have grown rapidly in recent years. The rising cost of data loss incidents According to a 2010 Ponemon Institute study, the average total cost per data breach has risen to $7.2 million, or $214 per record lost. In addition to the costs of incidents increasing, the number of leaks appears to be increasing every year.

Recently analysis a statistical data leakage prevention (DLP) model is presented to classify data on the basis of semantics. This study contributes to the data leakage prevention field by using data statistical analysis to detect evolved confidential data. The approach was based on using the well-known information retrieval function Term Frequency-Inverse Document Frequency (TF-IDF)[11].

In this proposed system , Existing commercial data leak detection/prevention solutions include Symantec DLP [2], GlobalVelocity [6], and GoCloudDLP [7]. GlobalVelocity uses FPGA to accelerate the system. All solutions are likely based on *n*-gram set intersection. Identity Finder searches file systems for short patterns of numbers that may be sensitive (e.g., 16-digit numbers that might be credit card numbers). It does not provide any in-depth similarity tests. Symantec DLP is based on *n*-grams and Bloom filters. The advantage of Bloom filter is space saving. However, as explained in the

introduction, Bloom filter membership testing is based on unordered *n*-grams, which generates coincidental matches and false alarms. Bloom filter configured with a small number of hash functions has collisions, which introduce additional unwanted false positives.

IV. ARCHITECTURAL VIEW

Sr.no	Paper	Technique	Advantage	Disadvantage	Result
1	Fast Detection of Transformed Data Leaks (2016)[1]	Sampling and allignment Algorithm	Parallelized version to achieves high analysis throughput	Data-movement tracking approache is not used	Good detection accuracy in recognizing transformed leaks
2.	Preventing accidental data disclosure in modern operating systems (2013) [5]	Aquifer as a policy framework and system for preventing accidental information disclosure in modern operating systems	In Aquifer, application developers define secrecy restrictions that protect the entire user interface workflow defining the user task.	Malicious applications are not taken into consideration	Aquifer allows app developers to contribute DIFC-based secrecy restrictions to protect application-specific data objects.
3.	Privacy-preserving scanning of big content for sensitive data exposure with MapReduce (2015) [8]	MapReduce algorithm	Ability to arbitrarily scale and utilize public resources for the task	System is not designed for intentional data exfiltration, which typically uses strong encryption	This transformation enables the privacy-preserving technique to minimize the exposure of sensitive data during the detection. This transformation supports the secure outsourcing of the data leak detection to untrusted MapReduce and cloud providers.
4.	Privacy-Preserving Detection of Sensitive Data Exposure (2015) [10]	A privacy-preserving data-leak detection model for preventing inadvertent data leak in network traffic	It enables the data owner to safely delegate the detection operation to a semihonest provider without revealing the sensitive data to the provider	Noise tolerance property remains an open problem.	Support accurate detection with very small number of false alarms under various data-leak scenarios.
5.	Attire: Conveying information exposure through avatar apparel (2013) [12]	Attire: Mobile app	Attire conveys real-time information exposure in a lightweight and unobtrusive manner via modications to the avatar's clothing .	Attire could also be extended to handle other types of information besides location.	Attire that uses the avatar metaphor to convey information access via changes in the avatar's attire and context.

Table1: Survey Table

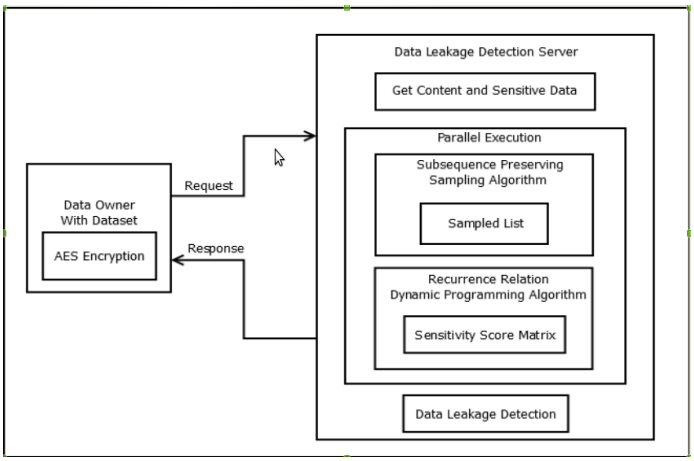

Fig1: Proposed Architecture for Data leakage

The above figure1 shows clearly the architecture view for the proposed system by this we can understand how our project is effective using the AES Encryption and how the data is going to subsequence preserving sampling for the sample list,and recursive relation dynamic programming for the sensitive score matrix which are contain into the parallel execution. After going through this it send back request to the main owner who send the request.

V. CONCLUSION

In this paper different data leakage detection models and techniques are premeditated .thus a content inspection technique for detecting leaks of sensitive information in the content of files or network traffic.the detection approach is based on aligning two sampled sequences for similarity comparison.The result of alignment method is useful for detecting multiple common data leak scenarios.but a parallel versions of the prototype provide substantial speedup and indicate high scalability of there design. The goal of this module is to discover the leakage of confidential data by using a real dataset in public domain and the proposed method try to improve the accuracy and better detection.

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] Symantec. (2015). *Symantec Data Loss Prevention*. [Online]. Available: http://www.symantec.com/data-loss-prevention, accessed Feb. 2015.

[3] Kaspersky Lab. (2014). *Global Corporate IT Security Risks*.[Online]. Available: http://media.kaspersky.com/en/business-security/ Kaspersky_Global_IT_Security_Risks_Survey_report Engf nal.pdf

[4] C. Kalyan and K. Chandrasekaran, "Information leak detection in financial e-mails using mail pattern analysis under partial information," in Proc. 7th WSEAS Int. Conf. Appl. Informat. Commun. (AIC), vol. 7. 2007, pp. 104–109.

[5] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," 2013

[6] Global Velocity Inc. (2015). *Cloud Data Security From the Inside Out—Global Velocity*. [Online]. Available: http://www.globalvelocity.com/, accessed Feb. 2015.

[7] GTB Technologies Inc. (2015). *GoCloudDLP*. [Online]. Available: http://www.goclouddlp.com/, accessed Feb. 2015.

[8] F. Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," 2015

[9] Marecki, Janusz, Mudhakar Srivatsa, and Pradeep Varakantham. "A Decision Theoretic Approach to Data Leakage Prevention." Social Computing (SocialCom), 2010 IEEE Second International Conference on. IEEE, 2010.

[10] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," May 2015.

[11] Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "Detecting Data Semantic: A Data Leakage Prevention Approach." Trustcom/BigDataSE/ISPA, 2015 IEEE. Vol. 1. IEEE, 2015.

[12] R. Hoyle, S. Patil, D. White, J. Dawson, P. Whalen, and A. Kapadia, "Attire: Conveying information exposure through avatar apparel," 2013