DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4338

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A COMPREHENSIVE STUDY AND COMPARISON OF VARIOUS METHODS ON DATA LEAKAGES

B. Raja Koti
Department of IT, GIT
GITAM University,India

Dr. GVS Raj Kumar
Department of IT, GIT
GITAM University,India

Dr. Y. Srinivas
Department of IT, GIT
GITAM University,India

Abstract: Data security is an important asset for any organization. Mainly data leakages will take place when a system is premeditated such that some vital information about the organization is revealed to unauthorized parties. In addition, during the time of data sharing, there may be huge chances for the data exposure, leading to leakage or unauthorized modifications. The protection and prevention of sensitive data from leakages is a vital issue to every organization, as the data is the most valuable source for any organization. Many authors have developed and presented their views of safeguarding the data such that the vital information is not explored or leakage. In this article, a detailed survey many such recent innovations on data leakage techniques aimed at detecting data leakages are highlighted for the researchers to have further directions.

***Keywords*:** Data leakage, Fake Object, Fuzzy Finger Print, Storage Capsules, Information Security, Map Reduce Algorithm, Network Security.

I. INTRODUCTION

The value of the data is more important for any organization and it is customary to protect the data from unauthorized leakages or any other anonymity attacks. In order to achieve these objectives, many models are designed for the data security using different technologies/algorithms, with the central focus on upholding the issues of the reliability to the users of those systems. It is very much difficult for the system administrator to trace out the data leaks/guilt systems among the authorized users, as it is a sensitive issue in any working environment, resulting in a great challenge for the organizations. Data security states that the data or information or resource that is available on the network can only be accessed by authorized user and it also prevents the data or information or resources from an unauthorized user from malicious practice. In order provide authorization for users to access their data, Network administrators play a vital role. There are two types of networks available, viz., Private and Public Networks. In Private Network, Security is provided by an organization or company. In Public Network, Security is providing globally, by means of users name and password. Only the person who has the authorization can access while others are denied.

Data leakage is defined as the accidental or involuntary distribution of the sensitive data to unauthorized parties. Sensitive data of companies and organizations include intellectual property (IP), financial information, patient information, personal credit-card data and other information depending on the business and the industry. In many cases, sensitive data is being distributed among various employees who work from far-away from the organizational premises, business partners and clients working for the organization remotely. Therefore, in such cases, the possibility of private information falling into unauthorized parties increases exponentially. Whether caused by the malicious plan, or an unplanned mistake, by an insider or outsider of an organization, exposed sensitive information can seriously hamper the liabilities and assets of an organization.

The potential damage and adverse consequences of a data leakage incident can be classified into the following two categories, namely, Direct and Indirect loss. If the sensitive data is revealed by an authorized user to some unauthorized parties/ persons, that user is considered to be a guilt agent.
This article is structured as follows, a brief insight is provided about the Data leakage identification, protection, prevention of sensitive data leakages and its techniques. In the next section, a brief review of recent Literature is presented, so as to outline the various techniques that are in usage and the advantages and disadvantages of these techniques, the next section highlights about the various modules to be followed. The techniques for identification are highlighted in the next section and in the final section, the conclusions are presented.

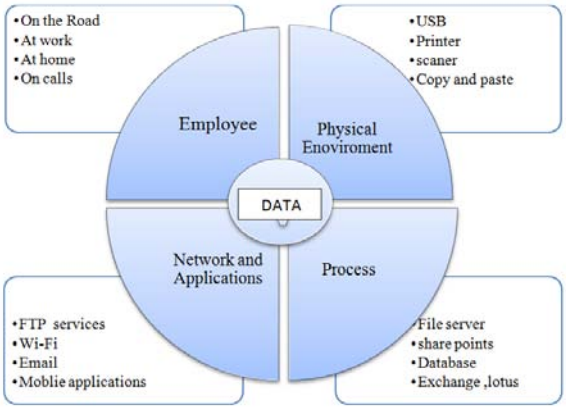

Fig 1: Data leakages possible ways

II. LITERATURE REVIEW

The Data Leakage detection concept was proposed by Papadimitriou and Hector Garcia Molina [1], [2], which can let us detect the guilt agent without changing the reliability of the original data. An investigation of data leak is a scary proposition. Security practitioners should always compact with data leakage issues that take place from email and other Internet channels. But now with the use of mobile technology, it became easier for data loss to happen, whether by chance or maliciously.

To uphold the security within the organization, watermarking Rakesh Agrawal, Jerry Kiernan [3], S. Czerwinski, R. Fromm and T. Hodes [4], techniques are considered, wherein a unique code or a signal is secure, imperceptibly, and robustly embedded into each of the copies that are to be distributed among the third parties. Watermarking was initially used in images J.J.K.O. Ruanaidh, W.J. Dowling, and F.M. Boland [5]. The technique of data provenance is also employed for the purpose of detection. Watermarking technique includes modification in original data which is not acceptable in certain cases. Recently, R. Sion, M. Atallah, and S. Prabhakar [6], Y. Li, V. Swarup, and S. Jajodia [7], and other works have also proposed the techniques of embedding water marks to relational data for safeguarding the data.

X.Shu and D.Yao [8] proposed a Fuzzy Finger print algorithm to minimize the exposure of sensitive data. Data leakage is detected with high efficiency and with a small number of false alarms. As of now, the internet is growing and the network bandwidth is being increased exponentially, and therefore, it is necessary to maintain the confidentiality of data. In such practical cases, the outgoing traffic is monitored; however, it is highly challenging to identify the encrypted data leakage. Generally, during the exploration of sensitive data, Map Reduce algorithms are considered, for identifying the possibilities of data leak. During detection, the exposure of sensitive data is minimized, which is generally done using the help of privacy-preserving data transformation algorithms. Y. Jang, S. P. Chung [14] proposed Fuzzy search scheme rely builds on expanding index that covers possible misspelling words provided with large index file size and higher search complexity.

R. Chen, B. C. M. Mohammed [10] has presented an algorithm for measuring the amount of data leakage. The goal is to calculate the maximum volume of the leaked data Measuring algorithm is provided for the Hyper Text Transfer Protocol, with the only intention to protecting the personal data from falling into wrong hands that lead to a huge loss either to an organization or an individual.

Border and A. Prakash [11] in their article has presented a model which helps to achieve privacy using a tailored model based on the trajectory data anonymization. The aim of this article is to preserve the location time doublets and frequent sequences in a trajectory database. Data utility is highly improved by using this mechanism. Borders [12] used Storage Capsule for encrypting the data file container to Protect personal or confidential information from personal computers, which need to be sent to the trusted parties globally, when the user finishes all their modification it the system is restored to its normal state and the output resumes normally. G. Karjoth and M. Schunter [13] mainly focused on establishing consistency between privacy-policies and its purported implementation. The data can only be used for its intended purpose and accessing purpose.

Most of the works presented by the authors are only to safeguard the system so that the vital information is intact and doesn't leave the system. The unauthorized mechanisms for accessing the data may lead to any number of problems for the larger organizations. The individuals also become victims of these leakages in cases ranging from protecting the bank account details, personal health information and other information, which is deemed to be personal. Therefore, data leakage is not about just damaging the vital data of the organization, but also it is all about safeguarding the individual's personal information.

III. MODULES OF DATA LEAKAGE DETECTION SYSTEMS

A. Data allocation module

The main task of any secured data allocation model is how to identify guilt-agent more intelligently. During the disturbing of data, administrator formulates the authentication rules so that only authenticated users can access that data and have the access to modify the data. The authenticated users will get a secured key by some mail or any other source, in this model, the admin before sending the data to users within a system will generate the random key termed as a secret key to restrict the access of the data file. In the case of an unauthenticated user, trying to access the data without the secret key or entering with a wrong secret key, then this system will send an acknowledgment to invoke the admin so that admin will lock that data file with the desired password to avoid, that access to that sensitive data.

B. Fake object module

The fake objects are created by a distributor and these objects are added to the data while distributing among the agents. These are generally implemented to get information about the leakage of data from the user who is working on data. These fake objects are helpful to the distributors for detecting the leakage chances. It will benefit while tracing the emails and duplicating of data. The fake object will work like an agent and fully operates under the control of the main administrator. Adding to a fake object like, random noise to original data, generates specks in the real time usage, suppose we have to share a data from user1 to user2, the data sent to user1 will be added to a fake object for tracing records during the mailing list of their data so user2 will receive it. Whenever user2 will use that mailing list from user1, a copy of that data will be generated, there by user 1 can be traced when used by the unauthorized user. This model was not suitable for all the data for example: data like financial information, budget and employees salary

all this type of data will not add the fake object because it can lead to financial crises to an organization.

Table I: Comparison

Parameters / *Modules*	*Accuracy*	*Complexity*	*Capacity*	*Detection*	*Robust*	*Strength*
Data Allocation	High	Low	Low	------	Yes	Modification of data not possible
Fake object	Low	Moderate	Depend on the hidden data	Not easy	No	It finds the probability of data leakage by agents Also find the fake data.
Optimization	Low	Low	Its depends on size of the data	Not easy to get it its depend up on the technology that use for it	No	It reduces the information that used for the main user.
Data disturbers	High	Moderate	Moderate	Its Depend on Admin intelligent	Yes	To change the permissions to not to access the data by unauthorized user.
Guilt Agent	High	Moderate	Moderate	Not easy	Yes	To detect the no of users that leaked the data.

C. Optimization module

The Optimization Module is a data allocation module that allocates to agents by distributors with one constraint and one objective. The distributor's requests are satisfied by the agent's constraint thereby providing them with the number of objects that are required or objects aiming to satisfy their circumstances. That object will be able to detect an agent who leaks any segment of his data. After satisfying the agent request, it is crucial to identify the possibility of any sort of leakage in which some of the agents were involved for a particular leaked set S. For the same purpose, we can use a notation to state formally the distributor's objective. As discussed earlier, Pr (G_j | A_i) is the probability that agent G_j is guilty if the distributor discovers a leaked table S that contains all A_i objects [3]. We define the different functions Δ (i, j) as

$$\Delta (i , j) = Pr(G_i |A_i) - Pr(G_j |A_i) \quad where\ i , j = 1,2,\ldots,n$$

Here Δ (i, j) is a metric for assessing the guilt of an agent, assuming that leaked set contains all Ai objects and Agent Ai is at least as likely to be guilty as any other agent. Thus, for every such i, j pair we find Δ (i, j) values. Difference Δ (i, j) is positive for any agent Ai whose set does not contain all data of leaked set. If the agent that leaked the data is to be identified then the minimum value of Δ (i, j) should be maximized.

D. Data distributor

The distributor is the main authorized admin/user of the organization. Admin can only accept the registration request of the user/agent and send the user ID and password for authorized person of the company to access data. It maintains sensitive data in a database and distributes as per agent request by adding the fake object into the original data. If the distributor has received the alert message from the agent notifies as leaked data or the agent is guilt then the fake object can send a warning message to the guilt agent and then admin can take action against that user/agent. The distributor also changes the password for the users; can able to view the file which is leaked along with the fake user's details, he can delete any agent by click on delete agent option. The distributor must assess the likelihood that the leaked data from one or more agents also.

E. Agent guilt module

A guilt agent is a person who transmits authorized data (or information) to an unauthorized organization or person. This guilt agent is a definite employee or a person with access to this sensitive data of the company. In this model agents will find out by using probability of estimation that can be done by guessed by the targeted it can be done by doing some surveys like asking few mailing information, call records, unique ID numbers like that in any case that user fails to provide that information by the analysis then we confirmed the guilt agent.

IV. TECHNIQUES OF DATA LEAKAGE DETECTION SYSTEMS

A. A Privacy Policy Model for Enterprises

In privacy policy model aimed at enterprises that can serve as the basis for an internal access control system to handle received data in accordance with privacy standards [13]. Thus, the enterprise receiving the information will handle it according to the stated privacy policy. The organization as well can verify that its business users are not in conflict with the privacy policies posted on its Web site, which is usually considered a binding contract between the site owner and the people visiting the site. This privacy policy model is

formulated by combining the user consent, obligations, and distributed administration. Conditions impose restrictions on the use of the collected data, such as modeling guardian consent and options, or narrowing the set of accessing principals.

B. Quantifying Information Leaks in Outbound Web Traffic

A new approach for quantifying information leaks in web traffic [11]. Instead of inspecting a message's data, the goal was to quantify its information content. The algorithms achieved precise results by discounting fields that are repeated or constrained by the protocol. It's focused on web traffic, but similar principles can be applied to other protocols. Their analysis was processed on static fields in HTTP, HTML, and JavaScript; where an insight is laid on to create a distribution of expected request content. It also executes dynamic scripts in an emulated browser environment to obtain complex request values in this model.

C. Protecting Confidential Data on Personal Computers with Storage Capsules

Storage Capsules is a new mechanism for securing files on a personal computer. Storage Capsules are similar to existing encrypted file containers, but protect sensitive data from malicious software during decryption and editing [12]. The Capsule system provides this protection by isolating the user's primary operating system in a virtual machine. The Capsule system turns off the primary OS's device output while it is accessing confidential files, and reverts its state to a snapshot taken prior to editing when it is finished. One major benefit of Storage Capsules is that they work with current applications running on commodity operating systems. Covert channels are a serious concern for Storage Capsules.

D. Data Leak Detection as a Service.

A novel privacy-preserving data-leak detection model is using for special digests, the exposure of the sensitive data is kept to a minimum during the detection and conducted extensive experiments to validate the accuracy, privacy, and efficiency of fuzzy fingerprint technique. Later focus on designing a host-assisted mechanism for the complete data-leak detection for large-scale organizations to use the fuzzy fingerprint as a service.

E. Privacy-Preserving Trajectory Data Publishing by Local Suppression.

Special challenges of trajectory data anonymization show that traditional K-anonymity and its extensions are not effective in the context of trajectory data. Based on the practical assumption of L-knowledge, achieve a (K, C) L-privacy model on trajectory data without paying extra utility and computation costs due to over-sanitization. Local suppression introduces to trajectory data anonymization to enhance the resulting data utility. Consequently, they propose an anonymization framework that is able to remove all privacy threats from a trajectory database by both local and global suppressions. This frame work is independent of the underlying data utility metrics and, therefore, is suitable for different trajectory data mining workloads [10].

F. Privacy-Preserving Multi-Keyword Fuzzy Search over Encrypted Data in the Cloud

Multi-keyword fuzzy search problem over the encrypted data, and integrated several innovative designs to solve the multiple keywords search and the fuzzy search problems simultaneously with high efficiency [15]. In this approach of leveraging LSH functions in the Bloom filter to construct the file index is novel and provides an efficient solution to the secure fuzzy search of multiple keywords. In addition, the Euclidean distance is adopted to capture the similarity between the keywords and the secure inner product computation is used to calculate the similarity score so as to enable result ranking. The authors have proposed a basic scheme as well as an improved scheme in order to meet different security requirements.

G. Gyrus: A Framework For User-Intent Monitoring Of Text based Networked Applications

It is a new application which is considered for automation of the analysis, generation the User Intent (**UI)** and the traffic signatures. Extending Gyrus' output monitoring to include disk transactions would allow Gyrus to support no networked applications such as word processors. Integrating with a delegated computation verifier would allow Gyrus to support a broader range of applications. In addition, Gyrus could verify that the input to a computation verifier is actually from the user. Another interesting future direction would be to implement Gyrus on other platforms. Gyrusfills an important gap, enabling security policies that consider user intent in determining the legitimacy of network traffic [14].

H. Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with Map Reduce algorithm.

Map Reduce algorithms have been developed [9]. Map Reduce systems are intended for detecting the occurrences of sensitive data patterns in massive-scale content in data storage or network transmission. The system provides privacy enhancement to minimize the exposure of sensitive data during the outsourced detection. Deployed and evaluated their prototype with the Hadoop platform on Amazon **EC2 (**Amazon Elastic Compute Cloud**).**

Table II: Advantage and disadvantage for techniques.

Techniques	*Advantage*	*Disadvantage*
Privacy Policy Mode Flexible Authorization Framework(FAF)	Protects personal data from privacy violations, Access decisions, Privacy control language, Enterprise provides access to the data, who are the data recipients.	Access control, Increase the risk of exposing sensitive data to outsiders.
Privacy-Preserving Data-Leak Detection	Data –Leak takes place in the network traffic. Hackers can take data from compromised computers, Attackers can easily mingle with the normal activity and preventing of leak becomes difficult, Too large legitimate network traffic.	Repeating fields are discounted by the protocols, For obtaining complex request values it allows execution of dynamic scripts in an emulated browser environment.

Storage Capsules	Protecting confidential files on a personal computer, Edit sensitive files without malware, Security, Usability.	In existing solutions for preserving confidentiality that does not rely on high integrity, Inadequate at preventing malware infection.
A Novel Fuzzy Fingerprint	It is used to detect accidental data leaks due to human errors or application flaws. Sensitive data exposure is highly minimized, Enables the data owners to transfer the sensitive data's with the higher level of safety, Cloud providers are allowed for conducting data-leak detections naturally.	The content of the sensitive data is leaked during data-leak detection. Outsourced data leak-detection can't be realized due to privacy.
Private Sanitization	Can be applied to different trajectory data using sanitization algorithm, Increase the efficiency, Constrained inference using inherent constraints of prefix tree for better utility, A practical solution large amount of trajectory data under various privacy.	Various application domains have generated and collected it with increasing prevalence of location–aware devices and trajectory data, Trajectory data carries much rich information.
Novel Multi Key Word Fuzzy Search Scheme	Effectively supports multiple keyword fuzzy searches without increasing the index, Minimize search complexity.	Data utilization a challenging problem, larger index file size and higher search complexity, Impractical when the data volume is large.
GYRUS	Used to detect accidental data leaks by human or application flaws and highly minimized Sensitive data exposure. Enables higher level safety to owners for transfer the sensitive data are with users.	The content of the sensitive data is leaked during data-leak detection; Outsourced data leak-detection can't be realized due to privacy.
Map Reduce	The data leak detection can scan the content of the data for leakage without learning what the sensitive data is, and Confidentiality is maintained during outsourcing through third party service provider.	The confidentiality of the original and personal data is exposed, Compressed system, The loss of devices takes place frequently, Unencrypted data storage, mostly data leaks happen due to unintentional mistakes of employees or data owners.

From the above Table II, one can identify the advantages and disadvantage of the various techniques and models. To overcome all these disadvantages, one needs to develop new models which can identify the guilt agents and also can protect data, and even in case of the data leakages also it should be able to protect the data such that it can be very useful in various industries, where the sensitive data is to be shared out through any public or private channels with the intrusion of third parties.

V. CONCLUSION

This article elucidates the works presented by eminent authors in uploading the private information and concealing the individuals/ organizational data from unauthorized persons. The works presented by various authors have been highlighted showcasing the advantages and disadvantages associated with each of these models. From the works presented in this study, one can ascertain that the data leakage detection system model is very much useful for any of the organizations. The works presented by the authors help to understand various means of security threats, ranging from, sharing or transmission of the data. This paper gives an insight of the encryption algorithms aiming at providing security together with the detection techniques which are very helpful for various organizations, in which data is distributed through any of the public or private channels and shared with the third party.

VI. REFERENCES

[1]. Panagiotis Papadimitriou, "Data Leakage Detection", IEEE Transactions On Knowledge And Data Engineering, Vol. 23.

[2]. P. Papadimitriou and H. Garcia-Molina, "Data leakage detection", Technical report, Stanford University, 2008.

[3]. Rakesh Agrawal, Jerry Kiernan, "Watermarking Relational Databases", IBM Almaden Research Center.

[4]. S. Czerwinski, R. Fromm, and T. Hodes, "Digital music distribution and audio watermarking".

[5]. J.J.K.O. Ruanaidh, W.J. Dowling, and F.M. Boland, "Watermarking Digital Images for Copyright Protection," IEEE Proc. Vision, Signal and Image Processing, vol. 143, no. 4, pp. 250-256, 1996.

[6]. R. Sion, M. Atallah, and S. Prabhakar, "Rights Protection for Relational Data," Proc. ACM SIGMOD, pp. 98-109, 2003.

[7]. Y. Li, V. Swarup, and S. Jajodia, "Fingerprinting Relational Databases: Schemes and Specialties," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 34-45, Jan.-Mar. 2005.

[8]. X. Shu and D. Yao. "Data Leak Detection as a Service" It was published in the "8th International Conference on Secure Privacy Communication Network"2012.

[9]. F. Liu, X. Shu, D. Yao and A. R. Butt "Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with Map Reduce algorithm". It was published in the "CODASPY '15 Proceeding of the 5th ACM Conference on Data and Application Security and Privacy". 2015.

[10]. R. Chen, B. C. M. Mohammed, B. C. Desai and K. Wang. "Privacy-Preserving Trajectory Data Publishing by Local Suppression" It was introduced by "Journal of Information Science". 2013.

[11]. K. Borders and A. Prakash. "Quantifying Information Leaks in Outbound Web Traffic" It was published in IEEE Symposium on Security and Privacy". 2009.

[12]. K. Borders, E. V. Weele, B. Lau, and A. Prakash. "Protecting Confidential Data on Personal Computers with Storage Capsules" 18th USENIX Security Symposium. 2009.

[13]. Karjoth and M. Schunter. "A privacy policy model for enterprises" Computer Security Foundations Workshop. 2002.

[14]. Y. Jang, S. P. Chung, B. D. Payne, and W. Lee. "Gyrus: A framework for user-intent monitoring of text-based networked applications." 23rd USENIX Security Symposium.2014.
[15]. Bing Wang, Shucheng Yu, Wenjing Lou, Y. Thomas Hou, "Privacy-Preserving Multi-Keyword Fuzzy Search over Encrypted Data in the Cloud", IEEE 2014.