International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 5 731 – 735

Review on Seuring Data by Using Data Leakage Prevention and Detection

Rashmi S. Kadu
P. R. Pote (Patil) Welfare & Education Trust's college of Engineering & Management, Amravati
Department of Computer Science & Engineering
Amravati, India
rashmikadu28@gmail.com

Prof. V. B. Gadicha
P. R. Pote (Patil) Welfare & Education Trust's college of Engineering & Management, Amravati
Department of Computer Science & Engineering
Amravati, India
headcse1108@gmail.com

***Abstract*—** Today's life everything including digital economy, data enter and leaves cyberspace at record rates. A typical enterprise sends and receives millions of email messages and downloads, saves, and transfers thousands of files via various channels on a daily basis. Enterprises also hold sensitive data that customers, business partners, regulators, and shareholders expect them to protect. While doing business we need to maintain the sensitive and confidential data. If the confidential data is leaked from the organization then it may influence on the organization heath. So preventing the data many vendors currently offer data leak prevention and detection products; surprisingly, however, there is one technique which is data leak prevention and detection, in this paper review on that Data Leak Prevention and Detection method. Here first term is data leak. Data leaks involve the release of sensitive information to an third party which is unauthorized user intentionally. Data leakage is the unauthorized transmission of data or information within an organization or from an organization to the external destination. The data stored in any device can be leaked in two ways; if the system is hacked or if the internal resources intentionally or unintentionally make the data public. Therefore, organizations should take measures to understand the sensitive data they hold, how it's controlled, and how to prevent it from being leaked or compromised. So that purpose in this review data is preventing by using different technique of data leak prevention and detection.

Keywords- *Data Leakage, Sensitive Data, Watermarking Guilty Agent, Data Leak Prevention*

*****

I. INTRODUCTION

Data Leakage Prevention (DLP) is a computer security term which is used to identify, monitor, and protect data in use, data in motion, and data at rest [1]. DLP is used to identify sensitive content by using deep content analysis to per inside files and with the use if network communications. DLP is mainly designed to protect information assets in minimal interference in business processes. It also enforces protective controls to prevent unwanted incidents. DLP can also business processes. It also enforces protective controls to prevent unwanted incidents. DLP can also be used to reduce risk, and to improve data management practices and even lower compliance cost.

Information loss pertains to both the availability of information for authorized users as well as the unintended access of information by unauthorized users1. Information leakage is thus a subset of information loss with a focus on the security objective of confidentiality. There is using different technique for data leak prevention. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious. In this, we study unobtrusive techniques for detecting leakage of a set of objects or records. Specifically we develop a model for assessing the "guilt" of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker [3]. Finally, we also consider the option of adding "fake" objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects act as a type of watermark for the entire set, without modifying any individual members. If it turns out that an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty. Information leakage is any event, either accidental or malicious, that allows an unauthorized party to access data that is not already public information2. Possible events could involve (but are not limited to) an employee accidentally emailing sensitive information to the wrong person, losing a USB storage device with proprietary business data4, a disgruntled employee walking away with the company's customer list5, or an external hacker stealing a customer's credit card information from a company functions is the data security concerns.

In this paper, we deal with data leakage in analyzing how the DLP technology helps in minimizing the data loss/leakage problem? The study is performed as a case research on DLP technology in organizational perspective. As organization was facing issues with data loss, the objective of our paper is to analyze the evaluation of how well DLP fills security gap in comparison with previously used technology in a motive to solve data leakage problem. This is a very important need for the capability to exchange confidential information securely and easily as the organization is dealing with sensitive payroll data. This is done by doing a detailed study and a case research on Data Leakage Prevention technology in organization.

II. LITERATURE SURVEY

Data loss prevention (DLP) is interested in identifying sensitive data and also is one of the most critical issues facing CIOs, CSOs and CISOs. DLP is now today's strict regulatory and ultra competitive environment. In creating and implementing a DLP strategy, the task can seem to be intimidating. For this the effective solutions are available. This paper presents best practices for preventing leaks,

enforcing compliance, protecting company's brand value and reputation in organization [4]. The guilt detection approach we present is related to the data provenance problem [5]: tracing the lineage of S objects implies essentially the detection of the guilty agents. And assume some prior knowledge on the way a data view is created out of data sources. Our problem formulation with objects and sets is more general As far as the data allocation strategies are concerned; our work is mostly relevant to watermarking that is used as a means of establishing original ownership of distributed objects. Finally, there are also lots of other works on mechanisms that allow only authorized users to access sensitive data through access control policies [4]. Such approaches prevent in some sense data leakage by sharing information only with trusted parties. However, these policies are restrictive and may make it impossible to satisfy agent's requests.

III. EXISTING SYSTEM

There is a large security gap between the existing systems which are used to prevent the data leakage and the real life scenario. Gap is sometimes called, the space between where we are and where we want to be. The gap analysis is undertaken as means of bridging that space. It is a technique for determining the steps that are need to be taken in moving form a current state to desired future state. It begins with questionnaire "what is" and proceeds to "what should be" and finally highlights the gaps" that exist and need to be filled". Here comes what is security gap? Security gaps are nothing but the vulnerabilities or weakness in the organization which is a threat and can be exploited to make an attack. Malware infection, DDOS attack, Man in the middle are few types of attack which are done to gain monetary benefits or to harm the organization assets.
There are two ways of attacks such as External and Internal. External Attacks are those attacks which are done by hackers and other people from the outside of an organization network. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious. *E.g*. A hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents.

IV. STATE USE IN DLP

1) States and Classification of Data

A reasonable classification of data is an integral component of DLP since it has a major impact on the proper handling of data and hence the applied security requirements. For example, data classified as top secret is subject to other restrictions in handling than data classified as restricted or public. Furthermore, the handling of data depends on whether data is present as

- data in motion (DIM)
- data in use (DIU),
- data at rest (DAR)

Data in Motion:
Data in motion refers to data that is leaving the organization through a network to another authorized user28. This type of data is susceptible to hackers that who are attacking the communication network. This type of data is most likely breached when an employee accidentally sends the information to the wrong email address [6].

Data at Rest:
Data at rest refers to data that is being stored in an internal server within the organization29. The difficulty with this type of data is that large organizations with multiple servers and databases do not know where their sensitive data are stored30. More troubling is when the organization might not even realize that sensitive data is being stored in their server completely unprotected. Data at rest residing inside the organization are less susceptible to information leakage by attacks from external parties because hackers prefer to attack the end user systems that are less protected but stills carry a large amount of valuable data[6].

Data in Use:
Data in use refers to data that is being used by the users located in the laptops, USB storage devices, CD / DVD, iPods and etc. This type of data is highly susceptible to a data breach. End user devices can easily be lost or stolen, and due to technological advances, these devices store a large volume of valuable data, but lack the processing power to support the type of protection that a centralized server has. This combination of factors complicates an organization's objective to protect its confidential data [6].

2) Detection of Data

The presented techniques for automated data classification Commonly have to detect data and to recognize the content in order to classify it. All market leaders show detection weaknesses when it comes to unstructured data, cloud support, non-English languages, unsupported data formats, multimedia data, or operating systems other than Microsoft Windows or Apple Macintosh Operating System (Mac). Therefore, methods for automatic content identification and data tagging are interesting applications for DLP, too. Techniques based on watermarking or robust hashes can enhance the chances of detection and identification. Various sophisticated DLP solutions try to introduce behavioral analytics these efforts are in the early stages. Nevertheless, the issue of detecting and identifying data is a major challenge. Due to the existence of encryption, hidden channels, unsupported data formats, as well as large

amounts of multimedia data, DLP solutions can only work within limits [7].

V. DETECTING CHALLENGES

A. Encryption:

Preventing data leaks in transit are hampered due to encryption and the high volume of electronic communications. While encryption provides means to ensure the confidentiality, authenticity and integrity of the data, it also makes it difficult to identify the data leaks occurring over encrypted channels. Encrypted emails and file transfer protocols such as SFTP imply that complementary DLP mechanisms should be employed for greater coverage of leak channels. Employing data leak prevention at the endpoint – outside the encrypted channel – has the potential to detect the leaks before the communication is encrypted.

B. Access Control:

Access control provides the first line of defense in DLP. However, it does not have the proper level of granularity and may be outdated. While access control is suitable for data at rest, it is difficult to implement for data in transit and in use. In other words, once the data is retrieved from the repository, it is difficult to enforce access control. Furthermore, access control systems are not always configured with the least privilege principle in mind. For example, if an access control system grants full access to all code repositories for all programmers, it will not effectively detect data leaks where a programmer accesses a project that he/she is not involved in.

C. Semantic Gap in DLP:

DLP is a multifaceted problem. The definition of a data leak is likely to vary between organizations depending on the sensitive data to be protected, the degree of interaction between the users and the available communication channels. The current state-of-the-art, which is reviewed in Section III, mainly focuses on the use of misuse detection (signatures) and post-mortem analysis (forensics). The common shortcoming of such approaches is that they lack the semantics of the events being monitored. When a data leak is defined by the communicating parties as well as the data exchanged during the communication, a simple pattern matching or access control scheme cannot infer the nature of the communication. Therefore, data leak prevention mechanisms need to keep track of who, what and where to be able to defend against complex data leak scenarios.

VI. PROPOSED SYSTEM

A. Data Leakage Prevention

Data Leakage Prevention (DLP) is a computer security term which is used to identify, monitor, and protect data in use, data in motion, and data at rest. DLP is used to identify sensitive content by using deep content analysis to per inside files and with the use if network communications. To achieve the primary requirement is to scan the whole outbound traffic. We will maintain the DLP (data link prevention) server, which would scan the complete attachment to match the patterns. In case the patter matches, the attachment will be corrupted with the User designed message and an automated response E-mail will be sent out [7].

There give the example of Email Transformation:

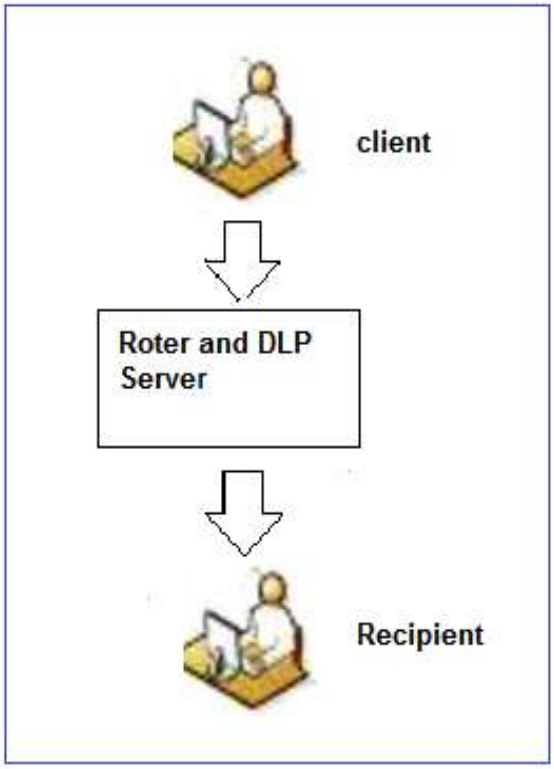

Figure 1: Transmission of an Email between client and DLP server

For the execution of the above DLP solutions we will require the following modules:

1. Email System
2. DLP Server
3. DLP Algorithms for pattern matching
4. Router Programming
5. File corruption system
6. Integration module to integrate all modules.

B. Data Leakage Detection

1. Agent Guilt Model:

To compute this PrfGijSg, we need an estimate for the probability that values in S can be "guessed" by the target. For instance, say that some of the objects in S are e-mails of individuals. We can conduct an experiment and ask a person with approximately the expertise and resources of the target to find the e-mail of, say, 100 individuals. If this person can find, say, 90 e-mails, then we can reasonably guess that the probability of finding one e-mail is 0.9 [8]. On the other hand, if the objects in question are bank account numbers, the person may only discover, say, 20, leading to an estimate of 0.2. We call this estimate pt, the probability that object t can be guessed by the target. Probability pt is analogous to the probabilities used in designing fault-tolerant systems.

That is, to estimate how likely it is that a system will be operational throughout a given period, we need the probabilities that individual components will or will not fail. A component failure in our case is the event that the target guesses an object of S. The component failure is used to compute the overall system reliability, while we use the probability of guessing to identify agents that have leaked information. The component failure probabilities are estimated based on experiments, just as we propose to estimate the pts. Similarly, the component probabilities are usually conservative estimates rather than exact numbers.

2. Fake Objects:

The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. However, fake objects may impact the correctness of what agents do, so they may not always be allowable. The idea of perturbing data to detect leakage is not new, e.g., [8]. However, in most cases, individual objects are perturbed, e.g., by adding random noise to sensitive salaries, or adding a watermark to an image. In our case, we are perturbing the set of distributor objects by adding fake elements. In some applications, fake objects may cause fewer problems that perturbing real objects. For example, say that the distributed data objects are medical records and the agents are hospitals. In this case, even small modifications to the records of actual patients may be undesirable. However, the addition of some fake medical records may be acceptable, since no patient matches these records, and hence, no one will ever be treated based on fake records. Our use of fake objects is inspired by the use of "trace" records in mailing lists. In this case, company A sells to company B a mailing list to be used once (e.g., to send advertisements). Company A adds trace records that contain addresses owned by company A. Thus, each time company B uses the purchased mailing list, A receives copies of the mailing. These records are a type of fake objects that help identify improper use of data. The distributor creates and adds fake objects to the data that he distributes to agents. We let Fi Ri be the subset of fake objects that agent Ui receives. As discussed below, fake objects must be created carefully so that agents cannot distinguish them from real objects. In many cases, the distributor may be limited in how many fake objects he can create [10].

3. Optimization Module:

The Optimization Module is the distributor's data allocation to agents has one constraint and one objective. The agent's constraint is to satisfy distributor's requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His objective is to be able to detect an agent who leaks any portion of his data. User can able to lock and unlock the files for secure.

4. Data Distributor Module:

A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data is leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means Admin can able to view the which file is leaking and fake user's details.

5. Data Allocation Module:

The main focus of our project is the data allocation problem as how can the distributor "intelligently" give data to agents in order to improve the chances of detecting a guilty agent, Admin can send the files to the authenticated user, users can edit their account details etc. Agent views the secret key details through mail. In order to increase the chances of detecting agents that leak data.

Algorithms

Allocation for Explicit Data Requests: In this request the agent will send the request with appropriate condition. Agent gives the input as request with input as well as the condition for the request after processing the data after processing on the data the gives the data to agent by adding fake object with an encrypted format. Allocation for Sample Data Requests: In this request agent request does not have condition. The agent sends the request without condition as agent sends the request without condition as per his query he will get the data. The distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We develop a model for assessing the "guilt" of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding "fake" objects to the distributed set [11].

VII. CURRENT APPROACHES

Various companies have recently started providing data leak prevention solutions. While some solutions secure 'data at rest' by restricting access to it and encrypting it, the state of art relies on robust policies and pattern matching algorithms for data leak detection. On the other hand, related academic work in data leak prevention focused on building policies[12], developing watermarking schemes [13] and identifying the forensic evidence for post-mortem analysis [14]. In order to provide a user-level policy language, hardware enforced policies [12] are proposed to ensure that the sensitive data does not reach the untrusted output channels through network communications, files and shared memory. The proposed runtime information flow security system assigns pre-defined labels to the data and policies are enforced on hardware level to ensure the data flow complies with the policies. Needless to say, such approach involves the labor intensive task of the definition of labels, policies and requires hardware that supports information flow security.

CONCLUSION

In this paper we present that how to prevent the data by using data leak prevention and detection technique.DLP is a multifaceted problem. Determining the sensitive data to be protected, identifying the legitimate use of the data and anticipating data leak channels require the internally. In

addition to data leak detection module the option of adding "fake" objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects act as a type of watermark for the entire set, without modifying any individual members. If it turns out that an agent was given one or more fake objects that Ire leaked, then the distributor can be more confident that agent was guilty. Both data leak prevention and detection share the same common goal, which is to detect potentially harmful activity. Thus, the commercial approach typically employs similar techniques to solve data leak prevention. However, data leak prevention focuses on what is leaked as opposed to detection, which focuses on who is breaking in.

ACKNOWLEGEMENT

The author would like to present their sincere gratitude towards the, Prof. V. B. Gadicha (Guide) and (H.O.D - Department of Computer Science & Engineering) for their extreme support to complete this assignment.

REFERENCES

[1] RichardE.Mackey,Available:http://viewer.media.bitpipe.com/1240246133_118/1258558418_168/sCompliance_sSecurity_Data- Protection_final.pdf 1
[2] Liu, S., and R. Kuhn. "Data Loss Prevention." *IT Professional* 12.2 (2010): 10-13. *Scholars Portal Journals*. Web. 14 June 2011. <http://journals2.scholarsportal.info/details.xqy?uri=/15209202/v12i0002/10_dlp.xml>.
[3] Panagiotis Papadimitriou, Hector Garcia-Molina , IEEE Paper "Data Leakage Detection",2010.
[4] Webspy,Available:http://www.webspy.com/resources/whitepapers/2008%20WebSpy%20Ltd%20%20Information%20Security%20and%20Data%20Loss%20Prevention.pdf
[5] P. Buneman, S. Khanna, and W.C. Tan, "Why andWhere: A Characterization of Data Provenance," Proc. Eighth Int'l Conf. Database Theory (ICDT '01), J.V. den Bussche and V. Vianu, eds., pp. 316-330, Jan. 2001, P. Buneman and W.-C. Tan, "Provenance in Databases," Proc. ACM SIGMOD, pp. 1171-1173, 2007
[6] Takebayashi T, Tsuda H, Hasebe T, and Masuoka R. "Data Loss Prevention Technologies." *FUJITSU SCIENTIFIC & TECHNICAL JOURNAL* 46.1 (2010): 47-55. *ISI Web of Knowledge*. Web. 16 June 2011.
[7] Miller, Ron."PLUGGING Information Leaks. (cover story)." *EContent* 30.1 (2007): 26-30. *Business Source Complete*. EBSCO. Web. 27 May 2011. 29 Takebayashi
[8] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," IEEE Transactions on Knowledge and Data Engineering, pages 51-63, volume 23, 2011.
[9] R. Agrawal and J. Kiernan, "Watermarking Relational Databases," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), VLDB Endowment, pp. 155-166, 2002
[10] Hartung and kutter,‖Watermarking technique for multimedia data‖2003.
[11] P. Bonatti, S.D.C. di Vimercati, and P. Samarati, "An Algebra for Composing Access Control Policies," ACM Trans,Information and System Security, vol. 5, no. 1, pp.35, 2002.
[12] N. Vachharajani, M. J. Bridges, J. Chang, R. Rangan, G. Ottoni, J. A. Blome, G. A. Reis, M. Vachharajani, and D. I. August, "Rifle: An architectural framework for user-centric information-flow security," in *MICRO 37: Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 243–254.
[13] J. White and D. Thompson, "Using synthetic decoys to digitally watermarkpersonally-identifying data and to promote data security," in *2006International Conference on Security and Management, SAM 2006*,June 26-29 2006, pp. 91–99.
[14] S. Lee, K. Lee, A. Savoldi, and S. Lee, "Data leak analysis in a corporate environment," in *ICICIC '09: Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control*. Washington, DC, USA: IEEE Computer Society, 2009, pp.38–43.
[15] Anuja Vasant Kale *et al*, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 513-518