International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 010 | Oct 2018 www.irjet.net p-ISSN: 2395-0072

An Approach Towards Data Security in Organizations by Avoiding Data Breaches through Standards of DLP

Anil Kumar N[1], Althaf Rahaman Sk[2], Girija K[3]

[1]*Student, Dept. of CS, GIS, GITAM (Deemed to be University), Visakhapatnam*
[2]*Assistant Professor, Dept. of CS, GIS, GITAM (Deemed to be University), Visakhapatnam*
[3]*Reserach Scholar, Dept. of CS, GITAM (Deemed to be University), Visakhapatnam*

Abstract - *Data leakage involves the loss of data due to unauthorized transfer and access by third-party applications. Sensitive and confidential data are a requisite for most companies' top management, administrators and IT managers. Data leakage can be accomplished by simply mentally remembering what was seen (or) by physical removal of storage devices such as tapes and disk, and reports or by subtle means such as data hiding. Since many years incidents involving data breaches have been present showing various reports in the media. In order to control and monitor the data access and usage, this paper is an attempt to survey and study DLP systems resulting analysis that illustrate the DLP and information leakage prevention(ILP) demanding various security measures to reduce the risk of technologies based on data breaches.*

***Key Words*: Data leakage, Data leakage prevention, Data Breaches, Information leakage Prevention, Data Hiding**

1. INTRODUCTION

This article focuses on data leakage prevention within the scope of information security (IS) which exceeds information technology (IT) security. Data in each company is one of the most important assets: therefore the protection of the data must take the first priority. Although the companies have security measurements and technical parries such as firewalls, gateways and etc, still the data leakage occurs. There are many paths for the data to travel and they can travel in many forms such as e-mail message, word-processing documents, spreadsheets, database flat files and instant messaging are a few examples.

In such variety of ways data leaks can cause harm to the data. Improper handling of confidential data can violate government regulations, resulting in fines and other sanctions. More loss of valuable information to competitors can result an outcome in loss of sales and may even threaten the existence of an organization. Detection and prevention of data loss can help preventing brand damage, competitive disadvantage, and/or legal proceedings. The DLP program is the mechanism by which an organization identifies their most sensitive data where the data is authorized to store or process, who or what applications should have access to the data, and to protect from the sensitive data loss.

1.1Need of DLP

DLP is widely required in the today world as large amount of data preceding takes place in every corner of the world from time to time. It makes the data more valuable to be prevented from any leaks. If the implementation of a DLP program is not managed properly, it can present a number of risks to the enterprise.

Most of these risks can have a direct and indirect impact on business operations, so it is important to take appropriate steps. There is a wide range of threats which lead to data and information leakage incidents.

In order to enhance security mechanisms and to prevent data leakage as effectively as possible, the major objective is to analyze and understand past incidents and attacks.

There are different number of ways the sensitive data can be revealed to untrusted third parties.

It is important to identify sensitive data leakage repositories within an organizations. Since, suitable selection of prevention techniques naturally depends on the repositories. Customer record, proprietary source code and sensitive documents on shares are a few examples of repositories.

There are ddifferent prevention techniques that may be appropriate for different data states:

1) Data repository

2) Data in Motion (Over the network)

3) Data in Use (At the end-point)

At the time when the data is at rest, the repository can be protected by controlling the access. However, when the data is in motion or in use, prevention using access control becomes increasingly difficult.

For the data in motion and data in use scenarios, the data leak prevention mechanism should be sufficiently context aware to interfere with the semantics of communication.

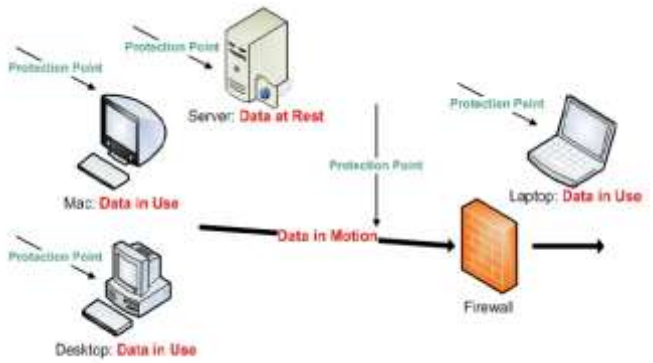

Fig-1: Data states in DLP

DLP is originally needed to alert the organization to the unintended misuse of internal data by an organizational employee, an identifying broken process during the discovery. As time progressing and technology advanced, the use cases for DLP is evolving. The System and network compromise the lead to data breaches by malicious users increasing significantly. This increase prompts a re-evaluation of the threat actors intended in accessing, stealing or destroying the data.

1.2 History of DLP

Data leakage prevention comes from one of the terms that should come to mind whenever there is the use of data. It is about reducing the risk of loss of data to an acceptable level when the frequency and cost of data breaches outweigh the security concerns.

If we take a look 2008 we can clearly see that it was a year of unprecedented events. From a security perspective look a data breaching, the number of records containing sensitive personal information that were involved in data breaches (in the U.S) in the last three years also falls under the "unpreceded" category – approximately 250 million records last year alone 42million records accounted for part of that number. In 200 there were over 127 million records involved in data breaches. The below image represents the data loss amount detected through various regions.

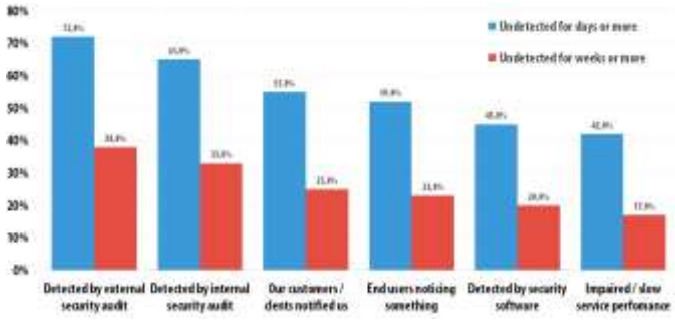

Chart-1: Data Leakage detection in volumes

A point to keep in mind is that the number of records involved in data breaches is either under-reported or in some cases not reported at all. A trend seen is that corporations are facing greater financial risk from insufficient controls and unclear policies.

2. WORKING WITH DLP

DLP may be defined as technologies which perform both content inspection and contextual analysis of data sent via messaging applications such as email and instant messaging over the network in use on a managed endpoint device and at rest, on-premises file servers or in cloud applications and cloud storage. These solutions execute high responses based the on policy and rules which can be defined to address the risk of accidental leaks or exposure of sensitive data within and outside authorized channels.

Fig-2: Abstract Process Flow of DLP

DLP technologies are broadly divided into two categories-

- Enterprise DLP
- Integrated DLP

Integrated DLP provides following features

Strengthens Data Protection and Control:

It Empowers the IT to restrict the use of USB drives, USB attached mobile devices, cloud storage, CD/DVD writers, and other removable media with maximum device control and DLP policies. Detects and reacts to the improper data usage based on regular expressions, keywords and file attributes. Educates employees on corporate sectors of data usage rules and policies through alerts, blocking and reporting.

Supports Compliance:

Simplifies regulatory compliance with out-of-the-box compliance templates. Speeds the audits and enforcement with real-time reporting and forensic data capture.

Streamlines Administration, Lowers Costs:

Simplifies the deploy and maintenance with a lightweight DLP plug-ins. Improves the visibility and control access with a fully-integrated, centrally-managed solution.

Reduces the demand of resources and impacts on performance with a single agent for endpoint security, device control and content DLP.

While, the Enterprise DLP solutions are comprehensive and packaged in an agent software for servers and desktops, physical and virtual appliances taken for monitoring networks and email traffic or soft appliances are used for data discovery.

Understanding the basic differences between contextual analysis and content awareness of data is essential to comprehend any DLP solution in its entirety. A useful way to think of the difference is if the content is a letter, context is the envelope.

While content awareness capturing the envelope and peering inside it to analyze the content includes external factors such as header, size, format, etc. anything that doesn't include the content awareness is that although we want to use the context to gain more intelligence in the content, we don't want to be restricted to a single context.

3. CHALLENGES IN DLP

Organizations are likely looking to reduce the risk of data breaches face several challenges.

1. Protection of Information: Sensitive information must be protected wherever it is stored sent or used. Do not reveal personal information inadvertently.

2. Reduction of data transfer: The organization should ban shifting data from one device to another external device. Losing removable information will put the data under risk.

3. Download restrictions: Any media that may serve as an allegiance to the hackers should be restricted to download. This restrictions could reduce the risk of transferring the downloadable media to external sources.

4. Secure transfer: The use of secure courier services and tamper proof packaging while transporting bulk data will help in preventing a breach.

5. A good password: The password for any access must be unpredictable and hard to crack. Change of password from time to time

6. Identifying threats: The security team should be able to identify suspicious network activity and should be prepared if there is an attack from the network.

7. Monitor data leakage: Periodically checking security controls will allow the security team to have a control on the network. Regular checks on the internet contents to identify if any private data is available for public viewing is also a good measure for monitoring data.

8. Tracking data: Tracking the motion of data within the organizational network will prevent any unintentional use of sensitive information.

9. Define accessibility: Defining accessibility to those who are working on company's sensitive data will bring down the risk of malicious users.

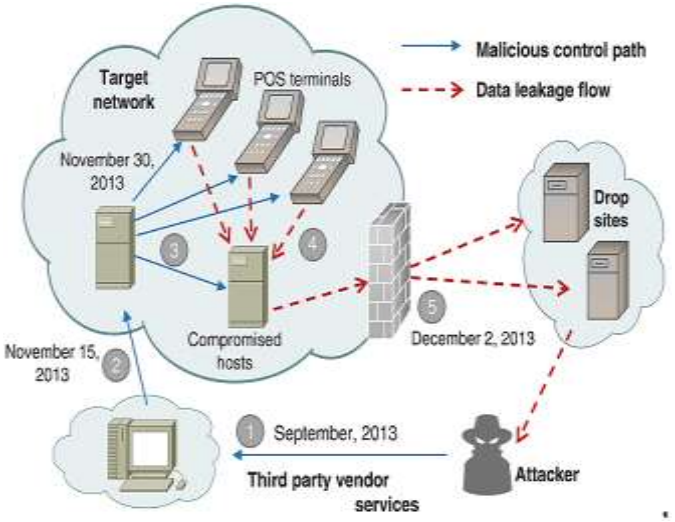

Fig-3: Process of Data Leakage

4. FUTURE RECOMMENDATIONS:

As discussed earlier, there are still many research issues and opportunities where further research efforts are required, especially as the enterprise data volumes are rapidly increasing. We highlight few of them in this section.

Deep data Learning for Insider Threat Detection: In big data settings, where a large volume of data from heterogeneous sources are generated, data mining and machine learning techniques will be increasingly used in DLPD. Deep learning techniques such as the Deep Neural Network have been used nowadays to detect anomalies in different applications. Such advanced techniques can be applied to both content and context analyses in DLPD, which will be able to both reveal stealthy data leaks, and also improve accuracy and achieve timely protection. Deep learning may also help close the semantic gap often encountered in the insider threat detection. The semantic gap is between the high-level user intentions and the low-level machine events. User intentions are the most pertinent to detecting insiders; however, they are not directly measurable. On the contrary, machine events are directly observable; unfortunately, they are not meaningful and need to be mapped to corresponding user intentions. Similar systematic gaps exist in many other research problems, e.g., capturing of image semantics based on pixels. Deep learning techniques have recently shown promises in solving complex sequence-to-sequence translation problems in natural languages. Training of the deep learners to infer sequences of user intentions basing on sequences of machine events is an extremely interesting direction.

DLPD taken as a Cloud Service: The use of cloud computing offers a new option for conducting data leak detection. Enterprises may outsource their data processing to third-party service providers, which brings about data privacy concerns. Collection and intersection approach is based on the similarity of two sets with their element frequent information. Therefore, it might be vulnerable to frequency analysis if the sensitive data is outsourced to a

third party and the third party has enough background frequency information of the n-gram. Privacy preserving the data leak detection algorithms are to be needed to resist strong attacks. An important research direction is for the cloud service provider is to achieve scalability without reducing detection accuracy and incurring significant delay when processing large-scale datasets. Spark is able to process streaming dataset by trucking data streams into small data segments. It is compatible with the MapReduce-based detection approach. However, the small content segments may miss the real leaks if the leak happens across multiple data segments, where increasing the size of data segments also increases transmission delay. Flink is another stream data processing platform that may be used to build enterprise-scale data leak detection.

Monitoring of Encrypted Channels: The Most existing DLPD approaches which are discussed, are vulnerable to large alteration of the original data, and thus are inapplicable to be evolved, obfuscated, or encrypted data. Encrypted traffic inevitably renders existing content-based detection useless. While deploying monitors outside, the encrypted channel can partially mitigate the problem, future DLPD solution needs a way to monitor encrypted channels in order to effectively detect stealthy data leaks. A possible direction is the use of data flow tracking or differential analysis. For example, various researchers have recently leveraged differential analysis techniques to achieve obfuscation resilient privacy leak detection on smartphone platforms. String matching on encrypted data has been one of the hot research areas in the last decade. Techniques in this area may also be used in future DLPD to detect the transfer of sensitive information on encrypted channels.

Benchmarks for DLPD: Sommer et al pointed out that the lack of training data is one of the challenges for applying machine learning to network anomaly detection, which also applies to DLPD. As machine learning techniques are being increasingly used, academic research in DLPD lacks common datasets for testing and evaluation, making it hard to compare with the state-of-the-art solutions and perform sound evaluation. The research community needs to provide mechanisms to modify data sharing and benchmark preparation effort.

Text clustering

Cluster analysis is one of the convenient method for identifying homogenous groups of objects called clusters. Objects in a cluster share many characteristics, but are very different to objects which does not belong to that cluster. In cluster analysis challenges we face are access control, Social networking by grouping textual data into clusters by — "flagging", for example, any fluctuation in usual data usage or labeling some data for further DLP purposes.

Social Network Analysis

Social network analysis (SNA) is measuring and mapping of the relationships and the flows between people, organizations, groups, computers, URLs, and other connected information/knowledge entities. The connection joints in the network are the people and the groups while the links show relationships flow between the nodes. SNA provides both the visual and mathematical analysis of the human relationships. Applying of social network analysis to data leak prevention involves in the monitoring of online communications (such as email, document and code repositories) to discover the social networks. The vast social networks play vital in identifying collaborators such as a team of developers working on the same or reconstructed code repository or a group of employees exchanging the emails to perform a task. Social network analysis or SNA has the potential to discover various types of communication that are not documented as a part of company policy or access controlling documents. Social networks can be easily visualized for manual or automatic validations.

Machine learning algorithms for classification

In the modern growing world a company may obtain new confidential or valuable data frequently. Current DLP systems experience many problems in detecting such data — technologies like regular and common expressions or keywords that cannot be customized to protect a large flow of diverse information. That is why various new types of algorithms have emerged that enables the organizations to use software that learns to detect the types of confidential data that is required for protection. Through the training, this approach will simultaneously improve the accuracy and reliability of finding protected information.

For DLP purposes, it is necessary to classify information into at least two of classes (for example, confidential and non-confidential data) and most of companies have sample data that may be used to train machine learning and AI algorithms to classify and detect the different types of data. That is why we will discuss supervised learning classification algorithms.

5. CONCLUSION

This paper describes the importance of the information or the data used by the organizations and by the local people regarding the seriousness of data leakage. This papers gives reference to current systems used to protect data and the DLP in terms of components and methods to solve the specific problems. This paper covers issues that are to be solved with DPL mechanisms. Each bit of data that is created or transmitted that carries values is precious, so it is recommended to process the data in secure ways by following DLP mechanisms.

REFERENCES

[1] Preeti Raman, HilmiGunes Kayacik, Anil Somayaji,"Understanding Data Leak Prevention", NY: Carleton University, 2011.

[2] Radwan Tahboub, Yousef Saleh, "Data leakage/Loss prevention systems(DLP)", Palestine Polytechnic University, 2014.

[3] Barbara Hauer, "Data and information leakage prevention within the scope of Information Security", Johannes Kepler University Linz, 2015.

[4] Randy Devlin,"Data loss prevention", The SANS institute, 2016.

[5] Open Security Foundation. DataLossDB. [Online]. Available: http://www.datalossdb.org, accessed Sep. 15, 2015.

[6] J. Rowley, ''The wisdom hierarchy: Representations of the DIKW hierarchy,'' J. Inf. Sci., vol. 33, no. 2, pp. 163–180, 2007.

[7] A. Giani, V. H. Berk, and G. V. Cybenko, ''Data exfiltration and covert channels,'' Proc. SPIE, vol. 6201, p. 620103, May 2006.

[8] H. G. Rice, ''Classes of recursively enumerable sets and their decision problems,'' Trans. Amer. Math. Soc., vol. 74, no. 2, pp. 358–366, 1953.

[9] M. Hart, P. Manadhata, and R. Johnson, ''Text classification for data loss prevention,'' in Privacy Enhancing Technologies (Lecture Notes in Computer Science), vol. 6794. Berlin, Germany: Springer, 2011, pp. 18–37.

[10] E. Ouellet, ''Magic quadrant for content-aware data loss prevention,'' Gartner, Inc., Stamford, CT, USA, Tech. Rep., Sep. 2013

BIOGRAPHIES

N Anil Kumar is currently pursuing his Bachelor Degree in Computer Applications at GITAM Institute of Science, GITAM (Deemed to be University), Visakhapatnam,A.P,India. His area of interest includes Computer Networks, Information Security.

SK.Althaf Rahaman is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University) Visakhapatnam, A.P, India. His main area of research includes Data Mining Mobile Computing, Software Engineering, and Information Security.

K.Girija is currently working as Research Scholar in the Department of Computer Science, GITAM (Deemed to be University) Visakhapatnam, A.P, India. Her main area of research includes Big Data, Information Security.