A COMPARATIVE EVALUATION OF DATA LEAKAGE/LOSS PREVENTION SYSTEMS (DLPS)

Kamaljeet Kaur[1], Ishu Gupta[2] and Ashutosh Kumar Singh[2]

[1]Govt. Sr. Sec. School, Ambala, Haryana, India
[2]Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India

ABSTRACT

Data is the most valuable assets of an organization that need to be secured. Due to limited computational resources, Customers outsource their workload to cloud and economically enjoy the massive computational power, bandwidth, storage, and even appropriate software that can be shared in a pay-per-use manner. Despite of tremendous benefits of cloud computing, protection of customers' confidential data is a major concern. Data leakage involves the intentional or unintentional release of secure or confidential information to non-trusted environment. Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is enhanced by the fact that transmitted data (both inbound and outbound); including emails, instant messaging, website forms and file transfers. Data leakage prevention system (DLPS) is a strategy for making sure that end users do not send the confidential data or information outside the corporate network. This review paper aims to study data leakage prevention through some challenges and data protection approaches as well as considering some limitations. This survey of DLPS can benefit academics as well as professionals.

KEYWORDS

Data Leakage Prevention (DLP), Insider Attacks, Sensitive Data, Data Access & Protection

1. INTRODUCTION

Data leakage is defined as the accidental or intentional distribution of confidential data to an unauthorized entity. Confidential data of companies and organizations include intellectual property, financial information, personal credit card data scores, information about their sanctions and other information depending upon the business. Data leakage is a serious threat to organizations as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is magnified by the fact that transmitted data are not regulated and monitored on the way to their destination. The diffusion of data can be done through digital media as well as by the company's official person also.

It is more severe when this is carried out by insiders. The researchers discovered that despite the security policies, procedures, and tools currently in place, employees around the world are engaging in risky behaviors that put corporate and personal data at risk [1]. Organizations provide easy access to databases for information sharing, storage and compression technology has allowed for more powerful (high-risk) endpoints. An 80-MB mobile device now holds 6000

Dhinaharan Nagamalai et al. (Eds) : AIS, CSIT, IPPR, IPDCA - 2017
pp. 87– 95, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71008

Microsoft Word documents or 7, 20,000 emails, and new 64-GB removable devices allow an entire hard drive to be copied onto a device same as the size of a pack of gum. These devices make it easier for employees, partners, or data thieves to access, move, or lose intellectual property or customer data. Mitigating data leakage from insider threats is a difficult challenge [2], [3]. Data leakage can occur in many forms and in any place [4], [5]. In survey of United States in 2014, Cybercrime emphasize on the seriousness of attacks caused by insiders of the companies. According to the survey report, companies experienced 37% internal attacks caused by insiders and researchers mentioned that the ratios of insider attacks are more destructive as compared to the attacks that are performed outside of the company. The ratio of private information that accidentally opens to the elements was 82% and in 76% of cases, confidential accounts were stolen [6].

According to statistics, it is stated that insider attacks has a high rate among other attacks that causes data leakage. By using Deep Content Analysis (DCA) techniques such as rule-based, regular expressions, database fingerprinting, exact file matching, statistical analysis, DLPS easily finds out the 'sensitivity' of the information and used to detect 'sensitive' information within traffic. This can be done either to classify the information into categories (e.g. 'confidential', 'secret') or to detect sensitive information within (outgoing) data. When a sensitive piece of information is found leaving the company, DLPS triggers the appropriate alert and action to be taken. There is necessity to implement DLP controls and supporting information security controls on time so that the effectiveness of these controls monitored over time. It helps to improve the management of data with minimum risk. The aim to design and develop DLPSs is to prevent data from breaches.

We can solve the data leakage problem by using Data Leakage/Loss Prevention System (DLPS). Generally DLPS as represented in Fig. 1 is used to discover, monitor, and protect the following type of data [7], [8].

- *Data at Rest*- Inactive data that is stored physically in any digital form like in spreadsheets, mobile devices, laptops and in databases etc. Examples include: - vital corporate files stored on the hard drive of an employee's notebook computer and files on an external backup medium.

- *Data in Motion*- Any data that is moving through the network to the outside via Internet like an email being sent.

- *Data in Use*- Data at the endpoints of the network like data stored in computer's RAM, cache, external drivers and data on USB devices etc. Examples include: - data that is being written, revised, or deleted.

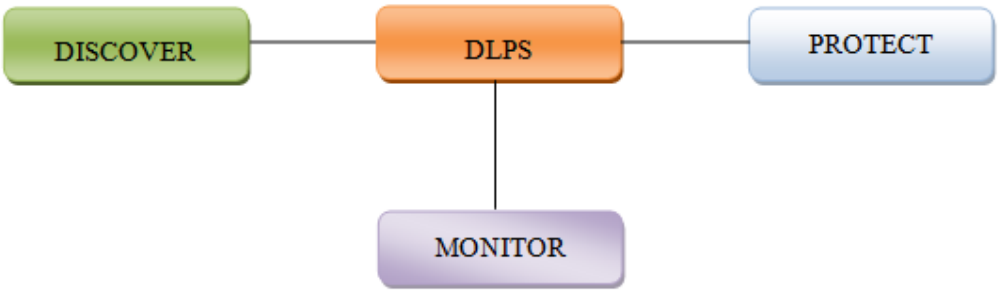

Figure 1. Data leakage prevention system (DLPS)

2. CHALLENGES IN DLPS

There are Common behaviors resulting in potential risk of data leakage like both physical and logical Access control, Accessing unauthorized websites, Leaving passwords unprotected and many more. This section illustrates the current challenges as shown in Fig. 2 to be solved by the DLP as follows:

1. *Encryption Challenge*- encryption is only one approach to secure data and security also requires access control, data integrity, system availability and auditing. So, it is difficult to detect and intercept encrypted confidential data and to recognize the data leakage occurring over encrypted channels [9].

Figure 2. DLPS challenges

2. *Access Control Challenge*- In the field of Information Security, Access control is a way of limiting access to a system, or to physical or virtual resources. In corporate, it is not easy to control employee's access to data repositories. For e.g. An employee of a company want to access data that he/she is not involved into, can steal some information if an access control system grants full access to all code repositories for all employees [10].

3. *New Data and Customization Challenge*- Sometimes, it is difficult to customize a DLP system particular for an employee, if the system utilizes old methods of data protection like regular expressions, keywords, or digital fingerprints. To create regular expressions, manual keywords for new customization process may take longer time. Moreover, this process is meant to be repeated as new type of confidential data appears.

4. *Social Network Challenge*- It is not sufficient to capture heterogeneous communication groups where people belong to more than one group, or even more when new communication groups are formed, old one disappears. In this situation, it is difficult to reveal a person who leaks the data (an outsider) in a communication or to detect persons having limited access to data [11].

3. CURRENT APPROACHES FOR DLP

This section categorizes current approaches for Data Leakage Prevention and identifies their main benefits and shortcomings:

A Learning and specification based system for Data leakage Prevention- This hybrid model combines signature based and anomaly based solutions, enabled on both detection and prevention. Two main dimensions are used to characterize the model: *i) filtering approach*, which describes whether users are permitted or not and *ii) model construction*, which describes how model is constructed. In Filtering, *blacklist* is used for well-known threats or undesired behaviors and *white listing* is used to identify the permissible activities. Only those transactions are considered to be legitimate that will match the model. Two main approaches are used to build the model *i) Specification approach and ii) Learning-approach* [12].

- *Specification-approach:* This approach is based on expert's knowledge and background of the transactions that lead to very accurate models. As, for instance, specification based blacklisting systems, also known as *Signature- based systems* that find the known attacks. A specification-based white listing system is used to detect unknown attacks.

- *Learning-approach*: This approach automatically learns the behavior of model using some techniques like machine learning and statistical modeling.

Shortcomings: These approaches created models that are less accurate as compared to those manually specified. As a consequence, these are inclining to high false positive rate. To check whether transaction is legitimate or not, a large number of alerts are generated and analyzed by human operator that cause to high operational cost [13].

Secure Key Stream Analyzer for Data Leakage Prevention- This approach illustrates that many data leakage prevention solutions depend on scanning file content. Key Stroke Profile not only scans the content of file rather it is capable to parse different file formats. But, risk of data leakage still exists for unsupported file formats. This approach proposed a new DLP model named as *Secure Key stream Analyzer (SKA)* [14].

- *Secure Key Stream Analyzer-* This SKA works on keyboard Application program interface (API). By hooking on keyboard API, it helps to track profile user key stroke behavior and discovers sensitive data. Data creator can be identified according to keystroke behavior.

Shortcomings: There are some issues in keyboard API that needs to be solved: Instead of using a keyboard, if a user uses mouse to make some text modifications like copying text and pick information from auto filled forms, in this situations SKA does not work. It only detects the text typed linearly [15].

A Result based Approach for Data Leakage Prevention- This approach discussed an information flow between one origin and many destinations (receivers) .The *Partially Observable Markov Decision Processes* (POMDPs) method is used over a fixed period called decision epochs where:

- To share a packet is equally important between origin and destination (i.e. a deterministic receiver either leaks all the packets, it receives or none of them).

- Leakage of packets that have been shared is a reward for destination; although disagreeable from the origin (i.e. the receiver is deterministic and considered a foggy receiver who leaks % f of the packets, it always receives).

- Sharing decision from origin is determined by using faulty observations of the accidental leakage of information from the destination, i.e. if packets are shared from origin with multiple foggy receivers and a different percentage of packet leaks occur at each destination [16].

Shortcomings: As the ratio of leak packets increase, it increases the tolerance at origin side, results in effect on the expected incentive of its most favourable strategy.This POMDP requires a huge amount of calculation and it suffers from scalability limitations.

There is a need of DLP solution that will allow secure sharing of confidential information in companies [17].

A Turkish Language Based Data Leakage Prevention System- This approach proposed a data leakage prevention system for Turkish language consisting two phases *i) training phase* and *ii)detection phase*. Two algorithms are used to describe the system: *Boyer Moore (BM Algorithm)* [17] is used to search exact sensitive strings exposed to whitespace attack and *Smith Waterman (SW) sequential alignment algorithm* [18] is used to detect modified string attacks.

- *Training Phase-* during this phase, list of sensitive words are generated from the sensitive document.

- *Detection Phase-* This phase is used to detect the modified sensitive content that attacker used to bypass the security system.

TF-IDF method is used to extract the sensitive words of sensitive documents. Latent Semantic Indexing (LSI) is used to construct the model document topics. This approach used Zemberek tool for extracting and analyzing the Turkish language [19].

Shortcomings: Attacks like adding, deleting and changing characters in 'sensitive' word, deleting white spaces from both sides of 'sensitive' word and adding white space to the middle of the 'sensitive' word were used to design the system. This tool is not only required for Turkish/English, but also for other languages [20].

4. DATA LEAKAGE PROTECTION TECHNIQUES

Data protection for various data states is represented in Table 1. Fig. 3 shows the various activities performed by DLPS to protect the data at various states.

Safety measures for Data-at-Rest: To protect data leakage, content discovery solutions is required. It helps to detect the sensitive data reside in separate locations by performing scanning in laptops, FTP servers, SMTP servers and in database [21]. Techniques for content discovery are as follows:

- *Local scanning of data-* In this technique, an agent is installed on the host machine that regularly scans the content which are stored in the files. It relocates, encrypts and quarantines the content after finding anything malicious in it. During the process, agents are always active, execute a policy even when devices are not placed locally and are not connected to the network.

  Disadvantage: On the target system, agents have low processing power and less memory.

- *Remote Scanning*- Scanning is performed from remotely located computers by maintaining a connection with server and application level protocols.

  Disadvantage: When scanning is performed from a remote computer that results in increased network traffic and low performance.

Table 1. Data leakage protection for different data states.

Type	Description	DLP goal
Data-at-rest	Information stored in an organization like files, servers, document management systems and email servers.	Content discovery
Data-in-motion	Organization data is restricted to network traffic such as web traffic	Block transmission of sensitive data.
Data-in-use	Information currently used at the end points such as http, https, print, file to USB and outlooks.	Prevents unauthorized usage of data (e.g. copying to a thumb drive).

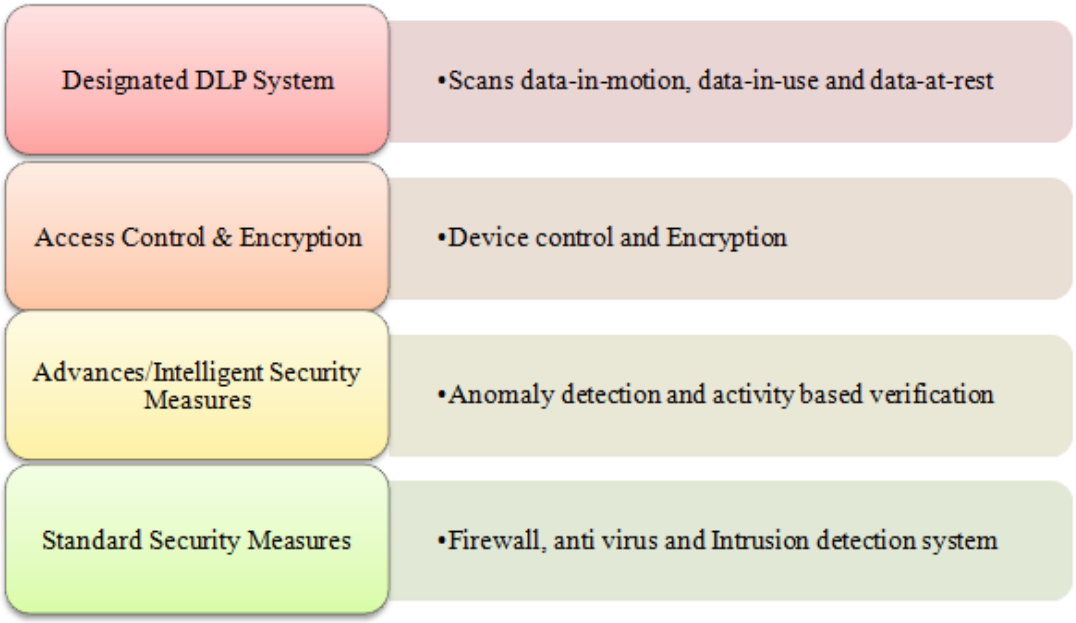

Figure 3. DLPS activities

Safety measures for Data-in-Motion: Network-based solutions are deployed on company's gateway. Gateway computer search the sensitive content and block the malicious activities immediately that violates the policy. These solutions capture the full data and perform the content analysis in real time [22], [23].

Safety measures for Data-in-Use: Local agents and host machines regularly check sensitive data such as data copied from one location and pasted into another location, data from print screen, unauthorized data transmission and copying data to a USB/CD/DVD [5].

5. LIMITATIONS

A DLP solution help organization to control sensitive data, but it has some pretty significant limitations also.

- *Graphics*- Graphics files consist sensitive information of companies like company's design documents, academic records, credit card scores etc. that demands scanning to prevent data leakage from unauthorized users. Scrutinize a file manually and then blocking the information shows that there is a gap exists in company's control. Companies having considerable IP for scanning graphics format should expand strong policies that administrate the use and allotment of data.

- *Third-party service providers*- While sending the company's sensitive information to third party, there should be mirrors of same level to control over the information. A vigorous third-party should comprise effective convention speeches over data leakage prevention and a supporting audit program that will help to moderate the risk.

- *Cross-application support*- DLPs have limited application level type functionalities. If DLP agent monitors data manipulation in any application and at the same time, it wants to perform same operation on another file then it is not able to do so. Companies must be ensured about DLP solutions that will prevent data leakage and identify applications which manipulate company's sensitive data.

- *Limited client OS support*— Many DLP solutions do not support data leakage prevention solutions for operating systems such as Linux and Mac operating systems because their usage as clients are fewer in companies [24].

6. FUTURE ANALYSIS FOR DLPS

In future, following activities will be followed to prevent company's data from leakage.

- *System Isolate*- To prevent data from leakage, companies should isolate their departments. They should close FTP port, TELNET port. Only HTTP ports should work but with some protection policies. Companies should ensure that traffic will pass through HTTP port.

- *E-mail Security*- In companies, grouping can be performed to prevent data leakage. There can be a group of 10 persons who can exchange emails within the group only. There should be some restrictions for sending emails. Companies should enforce some policies while sending an email from one department to another department so that the person of another department could not send an email to outside the network.

- *System Specific*- To prevent data leakage, each employee of the company must be restricted to their system. Administrator of the company should ensure that employees will use their allotted system only.

- *Smart Phones*- The employees of the companies uses smart phones and it cannot be stopped completely. Smart phones are enabled with new functionalities that handles as much data as you need. For companies, these smart phones are the main cause of sensitive data leakage such as transferring of e-mails and important documents accidently or intentionally.

These are the major factors that contribute to grow Data Leakage market. DLP solution focuses on organizations towards meeting regulatory and compliance requirements and data saved on public and private cloud.

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we discussed the challenges in DLPs and current approaches for data leakage prevention. We described how company's confidential information can be protected from unauthorized user's access. We explained various techniques like learning and specification, secure key stream analyzer, result based approach for data leakage prevention, but still there are various elements that leak the company's data. As we know data leakage happens through social media, cybercrimes and with the help of insider attacks. All these factors have a great impact on the company's reputation. Companies know which data is important to their business, where it is located and how it is sent to the outside network. Companies should enforce some policies, rule & regulations to prevent their data from unauthorized user's access.

Data Leakage Prevention System is a solution for all these problems that helps to discover, monitor and project the company's important data. There are some challenges that need to be solved. Cluster analysis algorithm has the ability to group data into cluster for further analysis that will help to cope with access control challenge and social network challenge.

Hence, there is necessity of research that will take a balanced approach for cloud computing data leakage and incorporate not only to end-users, but also with cloud provider and the cloud customers.

REFERENCES

[1] Ernst & Young, "Data loss prevention: Keeping your sensitive data out of the public domain," Insights on governance, risk and compliance, October 2011.

[2] S. Alneyadi, E. Sithirasenan and V. Muthukkumarasamy, "Detecting Data Semantic: A Data Leakage Prevention Approach," 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, 2015, pp. 910-917.

[3] S. Alneyadi, E. Sithirasenan and V. Muthukkumarasamy, "Discovery of potential data leaks in email communications," 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), Gold Coast, QLD, 2016, pp. 1-10.

[4] B. M. Babu and M. S. Bhanu, "Prevention of Insider Attacks by Integrating Behavior Analysis with Risk based Access Control Model to Protect Cloud," Procedia Computer Science, Vol. 54, pp. 157-166, 2015.

[5] D. Kolevski and K. Michael, "Cloud computing data breaches a socio-technical review of literature," 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015, pp. 1486-1495.

[6] S. Mathew and M. Petropoulos, "A data-centric approach to insider attack detection in database systems, " in Recent Advances in Intrusion Detection, ser. LNCS 6307, Springer, pp. 382–401, 2010.

[7] Frost and Sullivan, "World Data Leakage Prevention Market," Technical Report ND34D-74, United States, 2008.

[8] B. Hauer, "Data and Information Leakage Prevention Within the Scope of Information Security," in IEEE Access, vol. 3, no., pp. 2554-2565, 2015.

[9] P. Raman, H. G. Kayacık, and A. Somayaji, "Understanding Data Leak Prevention," in 6th Annual Symposium on Information Assurance (ASIA'11), pp. 27, 2011.

[10] S. Alneyadi, E. Sithirasenan,V. Muthukkumarasamy, "A survey on data leakage prevention systems," Journal of Network and Computer Applications, Vol. 62, pp. 137-152, February 2016.

[11] DLP Technologies, Challenges and Future Directions 268462340_ [accessed Jun 23, 2017].

[12] E. Costante, D. Fauri, S. Etalle, J. D. Hartog and N. Zannone, "A Hybrid Framework for Data Loss Prevention and Detection," 2016 IEEE Security and Privacy Workshops (SPW), San Jose, CA, 2016, pp. 324-333.

[13] A. Shabtai, Y. Elovici and L. Rokach, "A survey of data leakage detection and prevention solutions", ser. Springer Briefs in Computer Science, Springer-Verlag, New York, 2012.

[14] J. S. Wu, Y. J. Lee, S. K. Chong, C. T. Lin and J. L. Hsu, "Key Stroke Profiling for Data Loss Prevention," 2013 Conference on Technologies and Applications of Artificial Intelligence, Taipei, 2013, pp. 7-12, 2013.

[15] K. Revett, F. Gorunescu, M. Gorunescu, M. Ene, S. T. de Magalh˜aes and H. M. D. Santos, "A machine learning approach to keystroke dynamics based user authentication," Int. J. Electronic Security and Digital Forensics, Vol. 1, No. 1, pp. 55–70, 2007.

[16] J. Marecki, M. Srivatsa and P. Varakantham, "A Decision Theoretic Approach to Data Leakage Prevention," 2010 IEEE Second International Conference on Social Computing, Minneapolis, MN, 2010, pp. 776-784.

[17] M. Srivatsa, P. Rohatgi, S. Balfe and S. Reidt, "Securing information flows: A metadata framework," in Proceedings of 1st IEEE Workshop on Quality of Information for Sensor Networks (QoISN), 2008.

[18] Y. Jeong, M. Lee, D. Nam, J.-S. Kim, and S. Hwang, "High performance parallelization of Boyer–Moore algorithm on many-core accelerators," Cluster Computing, vol. 18, pp. 1087-1098, 2015.

[19] Y. Canbay, H. Yazici and S. Sagiroglu, "A Turkish language based data leakage prevention system," 2017 5th International Symposium on Digital Forensic and Security (ISDFS), Tirgu Mures, 2017, pp. 1-6.

[20] Y. Liu, C. Corbett, K. Chiang, R. Archibald, B. Mukherjee, and D. Ghosal, "Detecting sensitive data exfiltration by an insider attack," in Proceedings of the 4th annual workshop on Cyber security and information intelligence research: developing strategies to meet the cyber security and information intelligence challenges ahead, pp. 16, 2008.

[21] R. Tahboub and Y. Saleh, "Data Leakage/Loss Prevention Systems (DLP)," 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, 2014, pp. 1-6.

[22] S. Liu and R. Kuhn, "Data Loss Prevention," in IT Professional, vol. 12, no. 2, pp. 10-13, March-April 2010.

[23] G. Lawton, "New Technology Prevents Data Leakage," in Computer, vol. 41, no. 9, pp. 14-17, Sept. 2008.

[24] "Data leak prevention," Information Systems Audit and Control Association, Technical Report, 2010.