A Technical Review on Data Leakage Detection and Prevention Approaches

C. Mercy Praba
M.Phil Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India.

Dr.G. Satyavathy
Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India.

Abstract – Data security is the major concern of every application in the distributed environment. The sensitive data protection from being leaked to the others is the ultimate aim of all organization. Many security procedures are followed to maintain the data confidentiality, which preserves the data according to the security policies and rules. However, due to the distributed nature, the confidential and sensitive data protection lack pro-activeness and has many complications. These results in serious consequences and the data can be leaked in various leaking channels. So the analysis and mitigation of these drawbacks using effective mechanism is important. This paper carried out a comprehensive survey on different data leakage detection and prevention techniques and suggests future direction to overcome the weakness of the current data leakage detection and prevention schemes.

Index Terms – Data Leakage Detection, Data Leakage Prevention, Sensitive data management, Information security.

1. INTRODUCTION

In the recent network Data Leakage is an important concern for the business organizations. Prevention of sensitive data from unauthorized entities and monitoring the data flow to avoid more security risks are the main goals of the security domain. Unauthorized disclosure may have serious consequences for an organization in both long term and short term. To prevent from the unwanted access and transaction from happening, an organized effort is needed to control the information flow inside and outside the organization. Data leakage detection and prevention process are the important research issue, which is not always possible because several reasons. Recent news and reports indicates 50 % of data's are leaked in the business sector either partially or fully [1]. This is very difficult to identify the exact details of leaked data and the leaker. However, the data leakage has many channels to leak. So monitoring every channel is an impossible task, and thus creates many serious issues. There is numerous detection and prevention schemes like Intrusion Detection System (IDS), firewall, and virtual private networks are the common security systems used to detect or prevent some unwanted access. These schemes can perform well if the rules are properly defined. However, the rules can be violated from different accessible channels like email, instant messaging, and via other social media attachments. To overcome this problem, Data Leakage Detection(DLD) systems and Data Leakage Prevention(DLP) systems are deployed. There are less adequate researches introduced to thwart the DLP issue, so there is a need and challenge to design and develop a new DLP mechanism with detection ability. Motivated by the DLP field of study, a survey on the Data Leakage Detection and Prevention approaches are presented in this paper. The paper provides the basic process of DLD and DLP along with the recent techniques under the data leakage process. The paper finally contributes the problem and challenges of the recent techniques with future work.

This paper is prepared as follows. Section 2 discusses the DLD and DLP standard. Section 3 describes the challenges facing DLDs and DLPs. Section 4 categorizes the current DLP methods and discusses the advantages and disadvantages of each method. Section 5 concludes the survey paper.

2. DATA LEAKAGE DETECTION AND PREVENTION STANDARDS

Numerous studies conducted to define the area of data leakage detection and prevention in the literature. But the definition of the data leak or information leak prevention is the process of content monitoring and protecting them from the misuse [2]. Although researches on data leakage prevention are rising, there is little research on the detection of data leakage from the perspective of user behavior [3]. Authors in [4] reviewed the DLP approaches and its problems with the appropriate definition. The DLD and DLP process contains three phases such as the data collection phase, analysis phase and the remedial action phase shown in Fig 1.0. The data collection is beginning with the user internet or intranet logs and the database sources. The collected data's are imported in the DLD and DLP analysis phase, which performs rule matching, policy verification, content and context verification processes. The context verification extracts the sender, source id, timings of the data access, format and size from the header information etc., the content is the pre-processed data from regular expression and tagging process. The detection and classifying the data into the predefined class always utilizes the security policies and training samples. Finally the DLD and DLP schemes resolves the issue by selection appropriate remedial actions like alerting, blocking, allowing and doing some other actions in the security policy rule set.

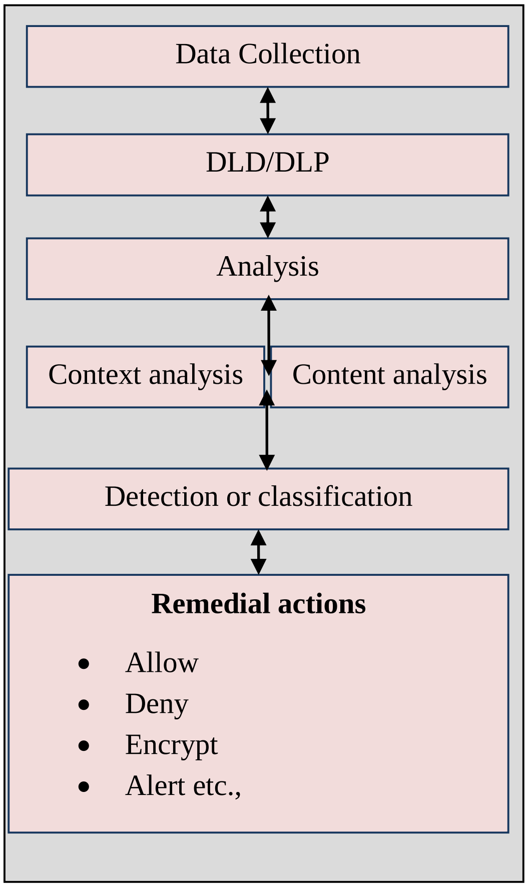


Fig 1.0 Deployment of DLD and DLP

Fig. 1 shows the simplified deployment of a DLD and DLP, where the detection and prevention attributes are analysis, detection and remedial actions. The dataset process includes the internet online data's or in transit data's, user access rights and offline historical data from the database. These three types of data's are generally collected for the DLP process. The fig 2.0 shows the data types used for the DLP. Data in transit is the data being transmitted form one node to another within the same network or different networks. Data in use is the data, which is accessible to the users in document format or email formats within the applications. The data in use format is not encrypted and this can be easily interpreted. The third database data's are normally structured and protected with strong access controls.

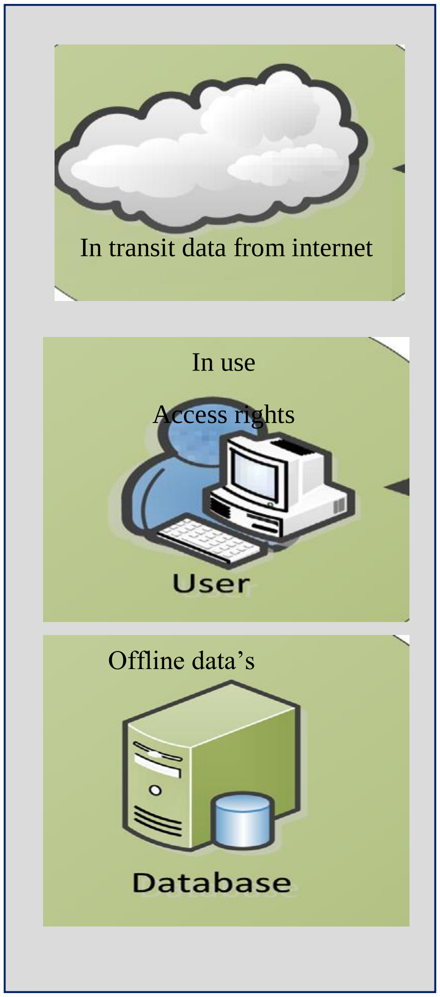
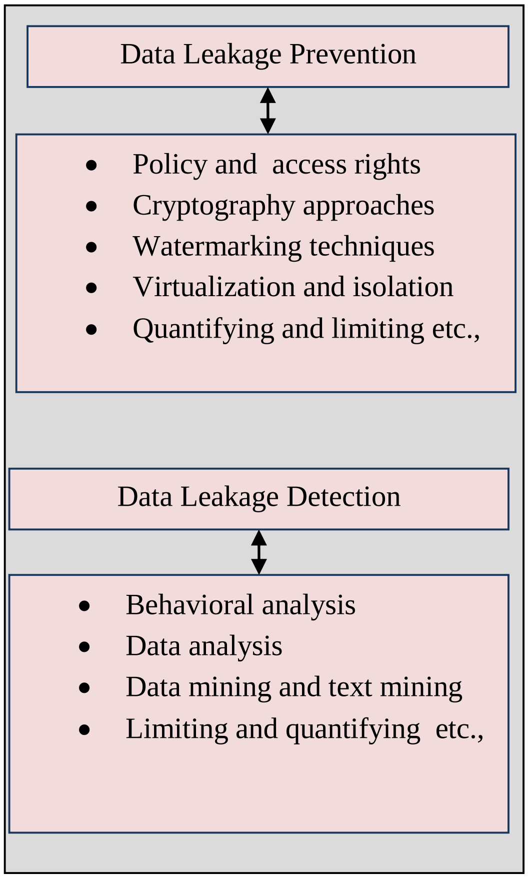


Fig 2.0 Data Formats in DLP

The content based DLP monitors sensitive data using regular expressions by identifying the structure. For example account number, phone number and other sensitive details can be monitored. According to this type, authors in [5] proposed a DLP with regular expression. But the technique was not successful and creates high false positive rates.

The DLDs and DLPs are performed by agent who has the ability to change the accessibility for the confidential data. Author in [6] identified different leaking channels, in which the data can be transfer. This includes the portable Medias like USB, memory cards and many. Author's audits the activities related to the sensitive data access and restricts according to the audit report.

The categories of DLD and DLP techniques are summarized in fig 3.0, which has different types of methods to prevent data leakage and maintain the ownership and detection using behavioral analysis etc., the most popular approaches under DLP is the use of cryptographic and watermarking techniques, this avoids the data from the unauthorized users. In the detection process, the data and behavioral analysis with text mining has several developments.



3. CHALLENGES FACING DLDS AND DLPS

Similar to the other data protection and security techniques, the DLD and DLP face several issues while detecting and preventing the data leakage. There are seven main challenges were identified from the earlier work [7] such as listed below.

- The first challenge is the data transaction may happen in many channels and many ways. If the data transmitted via the desired application then it can be detected or protected. But, if the data channel is other than the specific application like email, USB and other format then it will be a challenging task.
- The second challenge is the data modification, where the data can be modified and that can be partially leaked to the users. Sometimes many approaches find the whole pattern of the sensitive content. So it is failed to detect the partial and full data leakage with or without modification.
- The third challenge is determining and providing appropriate access rights to the specific users according to their security level is more complicate. The improper guidelines and policies may affect the DLP accuracy. The access controls should be properly configured.
- The fourth challenge is the process of encryption and steganography process, this technique can protect the data from the unauthorized user, but it difficult to analyze the data content when the strong encryption techniques are used.
- The use of watermarking concepts preserves the ownership and thus avoids the data leakage. But the watermarking contents are unreliable for all type of data. And moreover, the watermarked contents are easily recoverable. This creates many challenging issues.
- The scalability and integration for vast domain is certainly impossible. This creates a scalability issue. When the network size is huge, then the policy matching, monitoring and access specifications are difficult one.
- Finally the detection of data leakage from the log needs a complete supervised learning process. This creates many uncertainties and cause several issues over sensitive data. These challenges are commonly noted from the literature. Based on these challengers many researches made and that is described in the following section.

4. LITERATURE REVIEW

In paper [8], authors proposed a method named as CoBAn a Context-based model for accidental and intentional Data Leakage Detection (DLD) and Data Leakage Prevention (DLP). The technique has the ability to detect small sections of confidential information with high success rate. This fully deal with the problem of rephrased texts, and it can be detected the modified contents. Authors contributed a novel approach for classification using the context based approach. The graph based schemes are used to match the key nodes. The CoBAn approach has several advantages like; this has the ability to detect the confidential information's in large scale documents. The visual graph idea provides easy understanding of the confidential terms and policies. The configuration process is simple and this can be customized according to the domain. The main demerits of the approach are the features used for the prevention. Additional and important features can improve the performance more. This is also suffers from Long running time. And it needs huge training samples to improve the accuracy.

In paper [9], authors concentrated on dynamic leakage detection scheme and implemented for video streaming application. There are several techniques were used to detect documental data's. This paper utilizes the video content leakage detection system using traffic pattern analysis. Using the video length and traffic patterns the authors detects the leakage. The proposed method allows flexible and accurate streaming content leakage detection. This improves the security and trust of the video content streaming networks. However, the technique is not adequate for the real time contents like documents, email etc.

In the paper [10], authors designed two new algorithms to detect the transformed data leaks. Transformations such as insertion, deletion processes result in the unpredictable leakage patterns. This may affect the sensitive information's. To solve this issue, efficient sequence comparison techniques are used. This consists of a special sampling algorithm and alignment algorithm. This calculates the similarity score between the sensitive or confidential data and the content. This improves the accuracy in data leak detection with low false rate. The main drawback of this paper is, it is not scalable and has several integrity issues. Based on the results of this paper, the sampling and alignment algorithms are enhanced in several researchers later.

In the paper [11], authors studied the effectiveness of statistical analysis techniques in confidential data semantics detection. Authors proposed a data leakage prevention classification technique, which is based on the term frequency (TFIDF). The TFIDF finds the terms and its frequency counts. The classification was based on computing the similarity between the documents and the category values. This model was tested against different scenarios in which the DLPS dealt with known, partially known and unknown data. The overall classification indicates encouraging outcomes across all scenarios. Further, a graphical representation of the classification results was applied using SVD abstraction. The visualization provided a very useful analytical tool for studying the semantics of documents in relation to category centroids. The technique in this paper achieved a high score of 0.99 for both precision and recall. This technique is also suitable for the modified documents. This paper creates several issues and future directions for effective DLP. The techniques completely rely on the training samples. If the class distribution is not evenly distributed, then the result will be invalid. The algorithms creates class imbalance and classification speed related issues.

In the paper [12], authors developed Data leakage Prevention technique with time stamp approach. This is very important for giving permission to access a particular data, because in a particular period of time the data is confidential after the time stamp the same data could be non confidential, here authors developed an algorithm for data leakage prevention with time stamp. This technique collects the confidential and non-confidential documents and creates a cluster using k-means algorithm with cosine similarity function. For each cluster the key terms are identified using TFIDF. Then finally assigns the time stamp foe each document. This timestamp gives the deadlines of the document access. This approach can prevent data leakage within the time period, however the data leakage prevention for only fully leaked contents are developed. So partial data leakages are cannot be detected.

In the paper [13], sequence alignment techniques for detecting complex data-leak patterns are proposed. The algorithm is designed for detecting long and inexact sensitive data patterns. The technique proposed in the paper can only perform the data leakage detection in the network. It failed to detect the data leakage on a host. The sequence alignment techniques are based on aligning two sampled sequences for similarity comparison.

TABLE 1.0 DLD AND DLP TECHNIQUE SUMMARY

Paper Id	Year	Abstract	Techniques	Merits	Demerits
13	2016	In this paper, sequence alignment techniques for detecting complex data leak patterns are proposed. The algorithm is designed for detecting long and inexact sensitive data patterns.	sampling and alignment techniques	good detection accuracy in recognizing transformed leaks	Only data leak detection is performed
12	2016	In Data leakage Prevention, time stamp is very important for giving permission to access a particular data, because in a particular period of time the data is confidential after the time stamp the same data could be non confidential,	DLP with timestamp	Accuracy is high	Only suitable for full data leakage detection

		here we developed an algorithm for data leakage prevention with time stamp.			
11	2015	This paper designs two new algorithms for detecting long and transformed data leaks. The system achieves high detection accuracy in recognizing transformed leaks compared to the state-of-the-art inspection methods.	Alignment based algorithms	high detection accuracy	Scalability and integrity issues
10	2015	Statistical data leakage prevention (DLP) model is presented to classify data on the basis of semantics. This study contributes to the data leakage prevention field by using data statistical analysis to detect evolved confidential data. The approach was based on using the well-known information retrieval function Term Frequency-Inverse Document Frequency (TF-IDF) to classify documents under certain topics. A Singular Value Decomposition (SVD) matrix was also used to visualize the classification results. The results showed that the proposed statistical DLP approach could correctly classify documents even in cases of extreme modification. It also had a high level of precision and recall scores	Statistical Data leakage prevention (DLP) classification approach. Singular Value Decomposition (SVD) matrix was also used to visualize the classification results	Data leakage prevention.	Only 60 percent of the modified documents were able to identify.
9	2014	This focused on overcoming this issue by proposing a novel content-leakage detection scheme that is robust to the variation of length of the video. By comparing the length of different videos, it determined a relation between video length to be compared and the similarity between the videos which are compared.	Novel content-leakage detection scheme. Traffic Pattern	the detection performance improved	Only suitable for video content
8	2013	A new context-based model (CoBAn) for accidental and intentional data leakage prevention (DLP) is proposed. The new model consists of two phases: training and detection. During the training phase, clusters of documents are generated and a graph representation of the confidential content of each cluster is created. During the detection phase, each tested document is assigned	context based model (CoBAn), data mining algorithms	The ability to detect confidential information ‘‘hidden’’ in large non-confidential documents.	Long running time. Need training samples

		to several clusters and its contents are then matched to each cluster's respective Graph in an attempt to determine the confidentiality of the document.			

Table 1.0 shows the overall summary of recent DLD and DLP works. From the recent studies on the data leakage detection and prevention techniques, there are several future directions are found. The DLD and DLP should be performed with and without training samples and policy set. Detection accuracy should not be compromised for partial and full document leakage. By considering the above analysis and issues, an efficient DLD and DLP technique can be developed.

5. CONCLUSION

Data leakage is an ongoing problem in the field of information security. There are numerous researches from various domains are continuously working towards developing data leakage detection and prevention methods to mitigate this problem. Protecting confidential and sensitive information is more important. There are inadequate methods to perform the DLD and DLP with the considerations of all the research challenges. The aim of this survey was to summarize the recent researches and its demerits in data leakage detection and prevention. This paper gives the merits and demerits of the recent techniques and its capabilities are studied. This paper concludes that there is not enough method to concentrate on the partial and full data leakage prevention and as well as detection with high accuracy and document mobility within the host. So, further approaches should overcome all the above issues.

REFERENCES

[1]. Data loss db. Data loss statistics. Retrieved from ⟨http://datalossdb.org/⟩; 2015

[2]. Mogull R.Understandingandselectingadatalosspreventionsolution.Retrieved from ⟨https://securosis.com/assets/library/reports/DLP-Whitepaper.pdf⟩; 2010.

[3]. Boehmer, Wolfgang. "Analyzing Human Behavior Using Case-Based Reasoning with the Help of Forensic Questions." In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pp. 1189-1194. IEEE, 2010.

[4]. Shabtai, Asaf, Yuval Elovici, and Lior Rokach. "*A survey of data leakage detection and prevention solutions*". Springer Science & Business Media, 2012.

[5]. Yu, Fang, Zhifeng Chen, Yanlei Diao, T. V. Lakshman, and Randy H. Katz. "Fast and memory-efficient regular expression matching for deep packet inspection." In *Architecture for Networking and Communications systems, 2006. ANCS 2006. ACM/IEEE Symposium on*, pp. 93-102. IEEE, 2006.

[6]. Hackl, Andreas, and Barbara Hauer. "State of the art in network-related extrusion prevention systems." *Proceedings, 7th international symposuim on database engineering and applications* (2009): 329-35.

[7]. Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "A survey on data leakage prevention systems." *Journal of Network and Computer Applications* 62 (2016): 137-152.

[8]. Katz, Gilad, Yuval Elovici, and Bracha Shapira. "CoBAn: A context based model for data leakage prevention." *Information sciences* 262 (2014): 137-158.

[9]. Sudumbare, Pune. "Content Leakage Detection by Using Traffic Pattern for Trusted Content Delivery Networks." International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7909-7913

[10]. Shu, Xiaokui, Jing Zhang, Danfeng Yao, and Wu-Chun Feng. "Rapid screening of transformed data leaks with efficient algorithms and parallel computing." In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, pp. 147-149. ACM, 2015.

[11]. Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "Detecting data semantic: a data leakage prevention approach." In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, vol. 1, pp. 910-917. IEEE, 2015.

[12]. Peneti, Subhashini, and B. Padmaja Rani. "Data leakage prevention system with time stamp." In *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*, pp. 1-4. IEEE, 2016.

[13]. Shu, Xiaokui, Jing Zhang, Danfeng Daphne Yao, and Wu-Chun Feng. "Fast detection of transformed data leaks." *IEEE Transactions on Information Forensics and Security* 11, no. 3 (2016): 528-542.

Authors

Mercy Praba.C is a Research Scholar in Computer Science Department, Sri Ramakrishna College of Arts and Science for women, Affiliated to Bharathiar University. She received Master of Computer Application (MCA) degree in 2015 from Sri Ramakrishna College of Arts and Science for women, Affiliated to Bharathiar University, India. Her research interests are Information Security, Computer Networks etc.

Dr.G.Satyavathy completed her Ph.D in Computer Applications under Anna University in 2014.She holds M.Phil in Computer Science under Bharathidasan University in 2005.She is Mastered in Computer Applications under Bharathiar University in 1999.She holds an Bachelors Degree in Physics from Bharathiar University.

She has an teaching experience of 17 years from various institutions.She has her **11** research papers published in reputed reviewed International and National Journals.She has presented papers close to **9** papers in International and **7** National Conferences.

Her thrust on research and teaching interests include Networking,Image Processing,Information Security and Mobile Computing.