International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-7 Issue-1, October 2017

A Review on Data Leakage Detection for Secure Communication

Kishu Gupta, Ashwani Kush

Abstract: ***Data is an important asset for an enterprise. Data must be confined against loss and damage. In IT field massive amount of data is being exchanged among multiple parties at every moment. During data sharing, a great probability of data vulnerability, breach or variation exists. Along with data availability and accessibility data security is also very important. The term Data leakage is expressed as the accidental or unintentional allocation of confidential or sensitive data to a not permitted third party. This paper focuses on the data leakage concept, DLD modules & techniques to identify data leakage. A literature review for data leakage techniques is been presented in this paper. Commonly, water marking technique is used to handle the data leakage and hence causes data alteration. Distributor can allege his rights over the data if this altered watermark copy of data does exist at some not permitted location [1]. Various Data allocation strategies are in use to prevail over disadvantages for using watermark; these techniques enhance the probability of detecting guilty parties. The guilty agent(s) is an individual or a group of malicious users who cause data breach. Finally the algorithms were implemented which enhances the chance to detect guilty agents using fake objects.***

Keywords: Data Leakage, Data Leakage Detection, Data Leakage Prevention, Encryption, Watermarking.

I. INTRODUCTION

The NIST expresses the word computer security as "protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability, and confidentiality of information system resources (includes hardware, software, firmware, information/data, and telecommunications)". Data Leakage is expressed as the accidental or unintentional distribution of confidential or sensitive data to a not permitted third party. Most common examples of confidential data include various intellectual property (IP), financial data, patient data, personal credit card information, & some other information based upon the business and industry [2]. Actually, confidential data is needed to share among a range of stakeholders like human resources working from outside the site (e.g. on laptops), business colleague, & clients. Hence this need to share data enhances the chances that sensitive information, later can be found at some unauthorized location. Reason for this data breach can be either deliberate will or an unintentional mistake performed by any resource person working inside or outside the site.

Manuscript published on 30 October 2017.
* Correspondence Author (s)
Kishu Gupta*, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119, India, Email: kishugupta2@gmail.com
Dr. Ashwani Kush, University College, Kurukshetra University, Kurukshetra-136119, India, Email: akush20@gmail.com

Disclosure of confidential data can create serious damage to any organization. A possible data leakage scenario has been shown in Fig. 1.

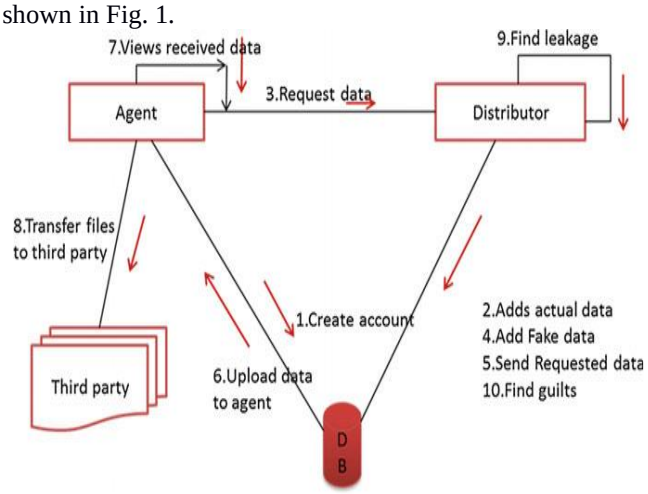

Fig1: A Possible Data Leakage Scenario [3]

A. States of Data in Data leakage: –

1. Data at Rest (DAR) - Data that is stored in file system, various databases or using some other storage methods i.e. the data that resides on an internal server within the organization. At large scale organization, it becomes difficult to identify and manage data and their sensitiveness [4]. This kind of data is less vulnerable to hackers because hackers always desire to attack the system that is not highly secure but consisting of a bulk quantity of valuable data, for example, the end user system.
2. Data in Motion (DIM) - The data which moves outside the site through some network (via internet) to another authorized user. This kind of data is vulnerable to the hackers those who attacks over the network. Data in motion is at a very high risk to get a leak when an employee of the company delivers the sensitive information by mistake, to wrong email address [5].
3. Data in Use (DIU) – The data residing at the network endpoints i.e. the data being used by users located in the laptops, USB storage devices, CD / DVD, iPods and etc. This type of data is highly vulnerable to a data breach [5]. Because end-user devices can easily be lost or stolen, and due to technological advances, these user devices store a huge volume of sensitive data, but these devices don't possess the processing capacity, to provide high-level data security as like a centralized server has, & hence fails in securing sensitive data.

B. Data Leakage Location:-

Data leakage generally occurs at 3 places. First possible location is inside, here data leakage is performed by a source residing within the organizations physical boundary [6];

Next possible location for data leak is outside, here data leakage is performed by some external source residing outside the organizations perimeter; & the last possible location is third-party location, Where data leakage is performed by an authorized third-party. The insider attacks are the most precarious threats challenged by an organization. Dealing with these insider attacks is very difficult as insiders knows each and every detail about the system, and hence cause a great loss to the organization. Comparing to traditional systems, Cloud are impacted more due to these insider attacks [4]. A survey performed by U.S. State of Cybercrime in 2014, illustrates the vindictiveness of insider attack. As per this survey, insider attacks has been faced by about 37% of organizations already & about 32% of the participants involved in this survey said that insider attacks are more precarious as compare to the outsider one. In maximum of the cases say about 82%, confidential information get leaked unintentionally.

C. Types of Losses on data Leakage incidents: –

First is the Direct Losses; these losses cause real harm, are very simple to calculate & to determine quantitatively. These losses occur mainly due to the violation of rules for example disobedience of rules for defending customer privacy cause in fine, settlement or customer compensation fees, a legal action involving lawsuits, harm to future sales, expenses for investigation [2]. Another is Indirect Losses; are very difficult to quantify. These losses proclaim a wide impact in terms of cost, place & time. Because of these losses share price get reduced and hence cause a pessimistic publicity, harm to goodwill & status of company; customer leaving behind; & most risky impact is that competitors come to know about various intellectual properties like business plans, code, financial reports, and meeting agenda.

D. DLP challenges:–

The challenges Data Leakage Detection/Preventions systems are facing currently, are shown in Fig. 2. The challenges to be solved by the DLP are as follows:

1. *Encryption Challenge*- It is difficult to detect and intercept some encrypted confidential data [4]. Though encryption provides integrity, confidentiality & authenticity of the data, but to identify the data leakage happening over encrypted channels becomes difficult.
2. *Access Control Challenge*- It is very difficult to manage employee's access over data repositories [5]. It is very typical to provide access control for data in transit & in data in use; however, data at rest is handled appropriately by these access control. Hence it is a fact that, it becomes very critical to provide access control after data is received once from the storage area. For example, it will become extremely difficult to identify whether a user/programmer is involved in data leak when a system to control access allows complete access to all code storage areas for all users.

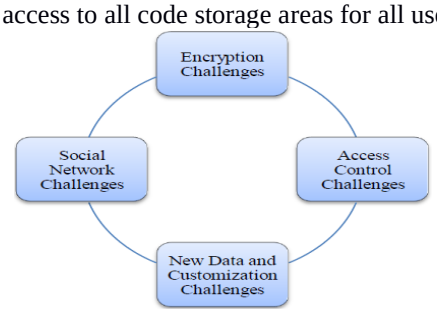

Fig. 2: DLDs challenges

3. *New Data and Customization Challenge*- It is not easy to customize a DLP system for particular an employee if the system utilizes old methods of data protection like regular expressions, keywords or digital fingerprints. A customization process may take long time to create regular expressions, manual keyword analysis and so on.
4. *Social Network Challenge*- It is not adequate to capture diverse communication groups where people belong to multiple groups. At that time it is difficult to reveal a person leaking data (an outsider) in a communication or to detect persons having access to limited access data [5].

In Section II related work has been presented. Data leakage module has been discussed in section III. Furthermore, section IV introduces various data leakage techniques. Conclusion & future work have been discussed in section V.

II. RELATED WORK

In 1996, J. J. K. O Ruanaidh et. Al. authors have discussed a approach for the embedding of robust watermarks in the digital images [7]. Watermarks are developed to be invisible, they are like invisible even to any careful observer, but they consist of enough information to recognize origin & receiver of the image with very less error probability. In transform based methods information bits can be placed adaptively, thus makes watermark robust to attack. A watermark is designed to contest features of image to be protected thus it is imperceptible. Furthermore, it has been proved that Transform-based methods are robust to the image compression & are robust to the operations of standard image processing. In 1998, F. Hartung et. Al. showed additive spread-spectrum watermarking of video scheme [8]. Authors also proposed an MPEG-2 compressed video in bit stream domain as a new scheme. For the practical watermarking applications, working on an encoded video rather than on the un-encoded video is significant. Elementary idea was the embedding of watermark in transform domain as presented in entropy coded DCT-coefficients. Authors implemented the approaches as compatible extension of the watermarking for the uncompressed-video, but in fact it can embed any other additive signals. Even though an existing MPEG-2 bit streams is partially edited, this approach evades visible artifacts by the addition of drift, compensation signal. From the already recorded signals watermark can be fetched deprived of information of original. With the appropriate features, watermarking-scheme in MPEG-2 bit stream domain can attain the data rates for watermark of few bytes\second for the ITU-R601 format video while being robust compared to friendly or the hostile manipulations. MPEG recording complexity is comparable to this scheme. This principal can be applied to the other schemes of hybrid coding, such as MPEG-1, MPEG-4, ITU-T H.261, or ITU-T H.263. In 2001, P. Buneman et. Al. explained a framework for understanding & describing provenance of the data in context of the SPJU queries & views [9]. Data provenance is inspected from the two viewpoints, (1) Why is there a piece of the data in output? (2) From where did a piece of data come? Authors used syntactic approach for understanding of both notions of provenance & authors explained.

system of rewrite rules which answers that why provenance has been preserved over class of well-defined queries & where provenance is well-preserved over class of the traceable queries. Furthermore, authors studied on Data Provenance characterization studied that how extra constraints on input instances, such as functional dependencies, can aid to get a additional whole explanation about where- provenance of the piece of data.

In 2002, R. Agrawal et. Al. proposed an approach in which an exclusively detecting image or text is inserted within the each copy which is then circulated to the authorized agents [10]. This exclusive code will help to detect agent accountable for leak. The main problem in this method is it consists of altering of original data or information set. Furthermore, weakness in this approach is such watermarks can be damaged to adequately mislead the unique detecting code. Sometimes malicious recipient can completely distort it.

In 2003, Y. Cui et. Al. described a whole set of approaches for the data-warehouse lineage tracing when data of warehouse is loaded by graph of the general transformations [11]. This technique is based on a range of *transformation properties* that occurred often in practice & can be stated simply by transformation authors. Authors also presented approaches for optimizing the performance of lineage tracing, comprising of building indexes & merging of transformations. Authors executed all presented algorithms in prototype lineage tracing system & reported preliminary performance results. According to authors, results of them can help to make principles for developing transformations & transformation graphs which are agreeable to the lineage tracing. To confirm that the most efficient tracing procedure is chosen, as first step, the transformation authors can be certain to state most restricting features that transformation satisfies grounded on the author's property hierarchy. As second step, a transformation might be altered a little to optimize its properties, for example making a dispatcher as a backward key-map by keeping key values. As a third step, splitting the transformation into general transformations with improved properties to avoid complex black-box transformations. Furthermore, in general, for lineage tracing it is good to state the smaller/simpler atomic transformations in place of larger ones, since lineage tracing system will merge the transformations automatically anyhow when it is helpful to do so.

In 2007, P. Buneman et. Al. explained that the provenance of data has recently been identified as the central to trust one places in the data [12]. It is also significant to annotation, to data integration & to the probabilistic databases. Several researchers attempted to give an overview of research in provenance the in databases with an attention on the recent database research & technology in the area.

In 2010, A. P. Noble et. Al. discussed DLPs corresponding to management point [13]. They suggested that executing DLPS requires carefully planning & study corresponding to the need, size & the organization's objective. Authors said that un-planned execution can setback aim of using a DLPS in the first place. For e.g., if any organization used Data Leakage Prevention Schemes to evade business damage, but used wrong execution of DLPs then resulted in disruption of business. Delaying workflow by the far-reaching traffic inspection & by the weak collaboration with the other security methods are included within wrong implementation of DLPs. Authors also discussed several contests that need to be discussed or solved before implementing a DLPS. However, these challenges vary among the organizations, according to the type of business & the amount of transactions.

In 2010, R. Mogull et. Al. presented a paper on DLP solutions understanding & selection of a particular DLP solution [14]. Authors in the paper analyzed DLP market and differentiated between a Data Leakage Prevention feature & a Data Leakage Prevention solution. The paper also focused on the confusion corresponding definition of Data Leakage Prevention Schemes & the confusion regarding varied commercial products (corresponding to vendors) which resulted as same output but having different product names. Authors defined Data Leakage Prevention Schemes as 'product which is depended on the central policies, which identifies, monitors & protects the data at rest, data in the motion & data in the use, through the help of deep content analysis'. Furthermore, authors explained differences between the content & the context analysis and suggested that content analysis is better. Overall, paper presented strengths & weaknesses of content analysis approaches, like exact file matching, statistical analysis, rule-based or regular expressions, fingerprinting & partial document matching.

In 2011, P. Papadimitriou et. Al. presented a non-intrusive leakage identification system that can identify guilty agent or party without altering integrity of the original data [15]. Authors proposed main ground on which much work has been based in this area. For improving probability of identifying leakages, authors proposed many data allocation strategies (across agents). The main roles done by authors were that the proposed methods did no change on original data but change the data allocation way to detect guilty agent. In 2011, P. Raman et Al. focused on the significance of Data Leakage Prevention research area & said that this area needs more research [16]. Authors discussed general Data Leakage Prevention approaches & related problems. Furthermore, authors gave ideas corresponding to future work like presented text clustering & social network analysis as possible solution for future problem. In 2012, A. Agarwal et. Al. worked upon issues in the data leakage issues which arise from the common applications such as IM, email & the further Internet channels [17]. The Electronic Mail filtering was done on foundation of fingerprints of the message bodies, the black & white lists of the email addresses & the words exact to spam. Authors said that corresponding to data leakage by the trusted agents, distributor need to estimate odds that disclosed records came from one or from more agents. Corresponding to this aim, authors used the data allocation methods or injecting "realistic but fake" data records for increasing detection of the leakage. In 2012, A. Shabtai et. Al. presented a complete survey on the DLP. According to authors DLPS is a system which is developed to identify & prevent the unauthorized, use, access or transmission of confidential information [2]. Paper included a classification of DLPS with academic & commercial examples.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Misuse identification in the information retrieval systems or database, network or web-based protections, email protection, data hidden in files, encryption and access control are categories of academic Data Leakage Prevention methods. Furthermore, paper presented case studies, Data leakage/misuse scenarios & future trends.

In 2012, N. P. Jagtap et Al. employed a system named as the DW (Data Watcher) & LD (Leakage Detector) to identify & prevent data leakage [18]. There are two models in it - first, when a data leakage happens (due to an employee of an organization uses the confidential data deprived of the approval of the owner), then Data Watcher model is used to detect data leaker. Second, when data leakage occurs (due to an employee who has delivered data exterior the organization), then Leakage Detector is used for tracking the "guilt" among the parties. Authors Guilt model used fake objects as the watermarking approach to increase probability of detecting the guilty third parties.

In 2012, S. A. Kale et. Al. discussed the outcomes of the employment of Data Leakage Detection Model [19]. Authors said that presently watermarking method was used for data protection. But it lacks complete security. Authors differentiate watermarking and data leakage detection model's approach.

In 2013, A. Kumar et. Al. in their work, explained by using many methods that an agent may be accountable for a data leakage [20]. Authors implemented algorithms using four techniques & said that in the real world, the provider can use any of above depending upon requirement. Authors perceived that spreading data carefully may increase chances of identifying agents efficiently particularly when a large overlap in data that agents must receive exists. Author's first objective was verification of result of proposed algorithms [3]. Authors found that for most cases s-max approach performs better than s-overlap & s-random approach. After executing these approaches it has been found that in s-overlap case, all the approaches depicts same performance & any one of above can be utilized on the basis of the distributor's application requirements. However, authors found that in s-max algorithm case, Round-Robin & SRF depicts better than other two. Hence, authors concluded that if the provider needs to fully satisfy an agent before assigning any object to the other agents, then SRF approach must be utilized to increase chances of detecting the leaker.

In 2014, G. Katz et. Al. presented a new context centered model for the unintentional & intentional DLP [21]. This takes advantage by either searching for the exact keywords and phrases or by taking help of many statistical methods. This model has two parts: training & detection. In training, authors formed clusters of the documents. Then corresponding to each cluster authors generated confidential content graph representation. Graph comprises of the key terms & context corresponding to which need to seem in order to be considered as the confidential.

In detection step, document tested is allocated to many clusters. Then matching is done of contents to every cluster's particular graph to evaluate confidentiality of the document. Major benefit of method is that it identifies minor section of the confidential information implanted in non confidential documents. Authors generated a good model that can be evaluated & modified by users.

In 2015, X. Shu et.Al. discussed on the privacy maintaining detection of the sensitive data exposure [22]. Authors proposed a Data Leak Detection method which can outsource & can be implemented in any semi-honest detection surroundings. The benefit of this method is it permits data owner to safely give the detection operation to any semi-honest provider deprived of uncovering private data to provider.

Authors used the fuzzy fingerprint approach that increases the data privacy throughout the DLD operations. The data owners preprocess and prepare fuzzy fingerprints and gives fingerprints to Data Leakage Detection provider. The Data Leakage Detection supplier evaluates fingerprints from network traffic & detects potential leaks in them. To stop Data Leakage Detection supplier from collecting the particular knowledge regarding sensitive data, the pool of potential leaks is combined of the real leaks & the noises. Then, he reports full data leak alerts to data owner. Post processes of potential leaks (sent by Data Leakage detection provider) are done by Data owner and the he decides of any real data leak.

In 2016, M. Backes et. Al. formalized the problem of provably associating guilty agent to the leakages, & also worked on the data lineage methods to resolve information leakage problem in several leakage scenarios [23]. Authors defined LIME, as a generic data or information lineage framework for the flow of data through many entities in the malicious surroundings. Authors involved three characters - owner, consumer, & auditor. Determination of guilty party for data leak is done by Auditor, & it also defines particular properties or characteristics for communication.

The main benefit of this model is that it imposes responsibility by design. It overcomes the current situation where most of the lineage methods are implemented after happening of leakage. For un-trusted sender & un-trusted receiver, the model's protocol enforces a fascinating combination of the oblivious transfer, robust watermarking & the signature primitives. In watermarking, Cox algorithm has been used.

In 2016, S. Alneyadi et. Al. presented a paper that explained that how confidentiality of data is preserved with the help of security methods like information security rules or policies with old security mechanisms like virtual private networks firewalls & the Intrusion Detection Systems [24]. But, these methods lack pro attention or pro activeness & devotion to protection of the confidential data. Furthermore, these methods require already defined rules generally which causes serious results, as appearing of confidential data or information in distinct forms in the distinct leaking channels. Therefore, there is huge requirement to resolve these drawbacks by using efficient methods. Recently, DLPSs have been presented as devoted method for detection & prevention against leakage of the confidential data or information in the use, at the rest & in transit. Data Leakage Prevention systems has different ways for analyzing content & context of the confidential data to detect and prevent leakage. Authors said that up to now many DLPSs have been designed & developed but still the term is ambiguous. Authors presented complete survey of current Data Leakage Prevention Systems mechanism.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Furthermore, authors gave future directions corresponding to more efficient Data Leakage Prevention Systems.

In 2016, S. Peneti et. Al. authors focused on time stamp corresponding to data leakage prevention system [25]. In DLP, time stamp is main, because in this permission is granted to use a particular data, but in a fixed period of time. Within time stamp data is secret & after expiry of time stamp it could be non-secret. There are two phases in time stamped DLP, first is Learning Phase & second is Detection Phase.

In learning phase, gather confidential & non-confidential documents of organization. Make clusters via K-means with cosine similarity function. Identify key terms (on the basis of their frequencies) for each cluster. Corresponding to every key term evaluate the score & allocate time stamp for a document based on organization schedule deadlines. Tested document is compared corresponding to confidential score & time stamp in detection phase. If the tested document's time stamp is equal or greater than the time stamp then that document is considered as confidential & it is made blocked.

In 2016, X. Shu et. Al. authors focused on inadvertent leak detection. Identifying the exposure of sensitive data was difficult because of data transformation in content [26]. Transformations (such as insertion and deletion) outcome in the highly unpredictable leak patterns. In the data leak detection model, they analyzed two types of sequences: sensitive data sequence & content sequence. Content sequence is sequence which is to be examined for leaks. The content may be data extracted from the file systems on workstations or personal computers or payloads extracted from supervised network channels. Sensitive data sequence comprises of information (e.g., customer's records, proprietary documents) that requires to be protected & cannot be uncovered to the unauthorized parties. The sensitive data sequences are known to the analysis system. Here they utilized sequence alignment approaches for detecting the complex data-leak patterns.

In 2017, S. Sodagudi et. Al. presented a paper which describes that in any ad-hoc networks there can be situation when a data provider owns sensitive data destined to set of supposedly trusted agents, some data can be leaked at such authorized place [3]. After the introduction of the fake objects, data allocation methods gives a path in detecting leakages thus ensuring extra security with encryption. Corresponding to this, light weight system has been proposed to describe data loss & to reduce performance degradation in such networks. Light weight system comprises the idea of cryptography & routing protocol execution at various stages of the data transfer. A social network has been taken as challenging issue to achieve proposed scheme.

III. DATA LEAKAGE DETECTION MODULES

Generally, there are five modules corresponding to data leakage detection as shown in Fig. 3. Modules are given below:

A. Data Allocation Module

This module deals with the data allocation problem means in what way a distributor should provide data to the agents so that the chances to identify a guilty agent can be improved [27]. One way to fulfill the same purpose is that admin sends files to authorized users. Agents receives secret key via mail and hence chances to detect guilty agents rise up.

B. Data Distributor Module

In this module sensitive data is provided to a single or group of authenticated agents named as third parties. Later on at some stage if data is available at some unauthorized location for example on a web which means data is leaked.

C. Fake Object Module

In order to enhance effectiveness to detect guilty agents, the distributor sometimes adds fake objects in the data. However, fake objects are not preferred all the time because they may affect the accuracy of what agents do. This idea to perturb data to identify leakage is not latest one [5]. Generally, in maximum circumstances, single objects are perturbed, by adding up a watermark into an image, or by introducing some random noise in sensitive salaries. Adding fake objects may cause problems in some applications, e.g. consider that the data objects to be distributed are medical records & hospitals are the agents. In such situation, a minor alteration in any records of actual patients is undesirable. One thing to consider is that adding up to various fake medical records is acceptable sometimes, as no patient exists in reality who matches these records, & hence, nobody will get treated on bases of these fake records. One common case is using fake objects in mailing lists by means of "trace" records. In such situation, let's consider company A supplies a mailing list to the company B with a constraint that mailing list can be used one time only for example mailing list is supplied to launch advertisements. Company A adds trace records. Hence, every time whenever purchased mailing list is used by company B, company A receives a copy of mail. These trace records, which are a kind of fake objects helps to find out the improper use of data.

D. Optimization Module

Distributor specifies one objective and one constraint to distribute its data among various agents. Distributor's objective is to become capable enough to detect the party who causes data leakage or leakage of any part of data. Constraint for the agent is to get their request fulfilled, by availing the number of objects they needed.

E. Agent Guilt Module

Purpose of this module is to find the guilty agent, and to fulfill this goal, probability Pr $\{G_{ij}/S\}$ is needed to compute. For e.g. consider that S contains the e-mails of individuals as objects. Let's assume a person is here to locate the e-mail of 100 individuals. Let's say the specified person is able to trace 90 e-mails, hence it can be concluded that 0.9 is the probability of finding one e-mail.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

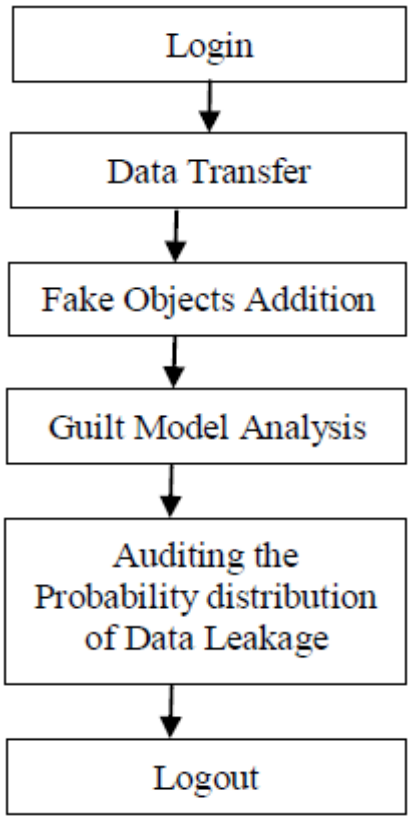

Fig 3: Data Leakage Detection Modules

Algorithm Steps

Step: 1 Distributor select agent to send data. The distributor selects two agents and gives requested data R1, R2 to both agents [28].

Step: 2 Distributor creates fake object and allocates it to the agent. The distributor can create one fake object (B = 1) and both agents can receive one fake object (b1 = b2 = 1). If the distributor is able to create more fake objects, he could further improve the objective.

Step: 3 check number of agents, who have already received data Distributor, checks the number of agents, who have already received data.

Step: 4 Check for remaining agents Distributor chooses the remaining agents to send the data. Distributor can increase the number of possible allocations by adding fake object.

Step: 5 Select fake object again to allocate for remaining agents. Distributor chooses the random fake object to allocate for the remaining agents.

Step: 6 Estimate the probability value for guilt agent. To compute this probability, we need an estimate for the probability that values can be "guessed" by the target.

IV. DLD TECHNIQUES

Some of commonly used data leakage detection techniques are discussed here below:

Some of commonly used data leakage detection techniques are discussed here below:

A. Watermarking

Generally, data leakage is managed by using watermarking. This technique implements a unique code in every copy of data. If later, a distributed copy of data is found at some unauthorized location, the guilty agent can be detected very easily [29]. Watermarks seem to be very useful at some places, but it causes some changes into the real data. Furthermore, watermarks can be even damaged sometimes, if the data receiver is nasty. For example, a hospital supplies its patient records to the researchers who use this data to formulate new more effective treatments. In the same manner, a company running in partnerships with some other companies requires sharing of its client data. Major disadvantages of watermarking technique can be listed as:

1. It causes some variation/changes into data by modifying some of the data attributes and hence makes data less sensitive. This modification of data is known as the perturbation. On the other hand in some situations, real data can't be modified at any level. For example, if an agent requires the exact salary to perform payroll. Salary can't be modified here.
2. Next problem is that, if the recipient is nasty, watermarks can be damaged very easily.

B. Data Allocation Strategy

Another technique used for data leakage detection is named as data allocation strategy. This technique formulates various strategies to allocate the data among the agents so that the probability to detect the guilty agent(s) can be improved. These methods are not based on the modification of the data as watermarks do. In some situations, we need to embed "realistic but fake" data records with real data so that the chances to identify data leakage and hence to detect the guilty agent can be improved. A large number of algorithms are in use to distribute objects to agent effectively. The prime goal is to find that is critical data been leaked by any of the agents, and next target is to detect the guilty agent who caused data leak.

V. CONCLUSION

Data leakage is the most precious crisis in the area of information security. To Shield sensitive data from leakage is a very critical challenge nowadays. The leakage of confidential data from computer systems and network imposes a major threat to the security of an organization. Research highlights that because of improper encryption over files is the most common reason for data leakage. This paper presents a chronological review of various techniques used to identify and avoid data leakage existing in the system. This paper concludes that the industry for data leakage detection is extremely heterogeneous because from the study it has been cleared that the term data leakage detection is come into existence after using a big number of technologies for example firewalls, encryption, access control, identity management, machine learning content/context-based detectors and much more. Watermarking is the most common technique for leakage detection, which implements a unique code in every copy of data so that origin of leakage can be traced with absolute certainty. Sometimes watermarking is not applicable to all the data, in case of such difficulties, a concept of overlap of shared data is used which makes possible to find out the probability that whether an agent caused a leak.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Further, the concept of embedding "fake" objects with the distributed data was introduced. These objects are not real entities but appear like real objects to agents. In simple words, these fake objects work as a kind of watermark with the only difference that it doesn't cause any modification to real data. From this study, it has been concluded that current system based on probability approach to detect data leakage is more practical as compared to the watermarking model. Watermarking uses various algorithms through encryption to offer security, whereas probability-based model provides both the security as well as detection technique to identify guilty. This model is really helping a lot in a range of industries, where data is shared with third parties by means of some public or private.

REFERENCES

1. G. Tuscano, H. Kotadiya, V. Bhat, R. Fernandes, and A. Pancha, "A Survey on Data Leakage Detection," International Journal of Engineering Research and Applications, vol. 5, no. 4, pp. 153-158, April 2015.
2. Shabtai, Y. Elovici, and L. Rokach,. NewYork: Springer, 2012, ch. Introduction to Information Security and Data Leakage, pp. 1-87.
3. S. Sodagudi and R. R. Kurra, "An Approach to Identify Data Leakage in Secure Communication," in 2nd International Conference on Intelligent Computing and Applications, vol. 467, Singapore, 2016, pp. 31-43.
4. N. Rechal and S. Aliyoglu, "A Survey On Data Leakage/Loss Prevention Systems (DLPs)".
5. R. S. Kadu and V. B. Gadicha, "Review on Seuring Data by Using Data Leakage Prevention and Detection," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 5, pp. 731-735, May 2017.
6. Baby and H. Krishnan, "A Literature Survey on Data Leak Detection And Prevention Methods," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, pp. 2416-2418, May-June 2017.
7. J. J. K. O Ruanaidh, W. J Dowling, and F. M Boland, "Watermarking Digital Images for Copyright Protection," IEE Proc. - VIS. Image Signal Processing, vol. 143, no. 4, pp. 250-256, August 1996.
8. F. Hartung and B. Girod, "Watermarking of Uncompressed and Compressed Video," Elsevier , vol. 66, no. 3, pp. 283-301, May 1998.
9. P. Buneman, S. Khanna, and W. C. Tan, "Why and Where: A Characterisation of Data Provenance," in International conference on database theory (ICDT), 2001, pp. 316-330.
10. R Agrawal and J Kiernan, "Watermarking Relational Databases," in 28th Int'l Conf. Very Large Data Bases (VLDB '02), Honkong, China, 2002, pp. 155-166.
11. Y. Cui and J. Widom, "Lineage Tracing for General Data Warehouse Transformation," VLDB Journal, Springer-Verlag, vol. 12, no. 1, pp. 41-58, January 2003.
12. P. Buneman and W. C. Tan, "Provenance in Databases," in SIGMOD ACM, Bejing, China, 2007, pp. 1171-1173.
13. P. Noble, R. Kopaee, A. Melek, and N. Nandy, "Data Leak Prevention," ISACA, USA, White Paper 2010.
14. R. Mogull, "Understanding and Selecting a Data Loss Prevention Solution," SANS, Securosis, LLC., Arizona, White Paper 2010.
15. P. Papadimitriou and H. G. Molina, "Data Leakage Detection," IEEE Transaction on Knowledge and Data Engineering, vol. 23, no. 1, pp. 51-63, January 2011.
16. P. Raman, H. G. Kayacık, and A. Somayaji, "Understanding Data Leak Prevention," in 6th Annual Symposium on Information Assurance(ASIA'11), Albany, NewYork, USA, 2011, pp. 27-31.
17. Agarwal, M. Gaikwad, K. Garg, and V. Inamdar, "Robust Data leakage and Email Filtering System," in International Conference on Computing, Electronics and Electrical Technologies (ICCEET), IEEE, 2012, pp. 1032-1035.
18. N. P. Jagtap, S. J. Patil, and A. K. Bhavsar, "Implementation of data watcher in data leakage detection system," International Journal of Computer & Technology, vol. 3, no. 1, pp. 44-47, August 2012.
19. S. A. Kale and S. V. Kulkarni, "Data Leakage Detection," International Jouranl of Advance Research in Computer and Communication Engineering, vol. 1, no. 9, pp. 668-679, November 2012.
20. Kumar, A. Goyal, A. Kumar, N. K. Chaudhary, and S., S. Kamath, "Comparative Evaluation of Algorithms for Effective Data Leakage Detection," in IEEE Conference on Information and Communication Technologies (ICT 2013), vol. 13, 2013, pp. 177-182.
21. G. Katz, Y. Elovici, and V. Shapira, "CoBAn: A Context Based Model for Data Leakage Prevention," Elsevier, Information Sciences, vol. 262, no. 1, pp. 137-158, October 2014.
22. X. Shu and D. Yao, "Privacy-Preserving Detection of Sensitive Data Exposure," IEEE Transactions on Information forensics and Security, vol. 10, no. 5, pp. 1092-1103, May 2015.
23. M. Backes, N. Grimm, and A. Kate, "Data Lineage in Malicious Environments," IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 2, pp. 178-191, March/April 2016.
24. Sultan, E. Sithirasenan, and V. Muthukkumarasamy, "A Survey on Data Leakage Prevention Systems," Elsevier Journal of Network And Compute rApplications, vol. 62, no. 1, pp. 137-152, January 2016.
25. S. Peneti and B. P. Rani, "Data Leakage Pevention System with Time Stamp," in International Conference on Information Communication and Embedded System (ICICES), 2016, pp. 1-3.
26. X. Shu, J. Zhang, D. Yao, and W. C. Feng, "Fast Detection of Transformed Data Leaks," IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 528-542, March 2016.
27. SHAJ and K. P. KALIYAMURTHIE, "A Review on Data Leakage Detection," International Journal of Computer Science and Mobile Computing, vol. 2, no. 4, pp. 577-581, April 2013.
28. R. Karthik, S. Ramkumar, and K. Sundaram, "Data Leakage Identification and Blocking Fake Agents Using Pattern Discovery Algorithm," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 9, pp. 5660-5667, September 2014.
29. C. Bhatt and R. Sharma, "Data Leakage Detection," International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp. 2556-2558, April 2014.