Data Breach Detection

Arathi P[1] Pooja G[2] Pooja S[3] Priyadarshini B[4] Sowmya R P[5]
[1]Associate Professor
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]Dr. Ambedkar Institute of Technology Bangalore, India

Abstract— Data Leakage is a serious threat in information security among organizations. Sensitive data is useful, confidential data in organizations which includes personal, client information, finance, economical status of organization, other business information. Data breach happens when such confidential information is let out by insiders of the organization to the outside world which causes immense loss to the organization. It may include sharing of emails, messages, files, forms etc. The company's data is shared to various shareholders as well, thus this data has to maintained such that data loss does not happen. This increases the risk of data leakage. There are a number of methods proposed to detect data leakage. In this paper, we discuss few very significant data leakage detection mechanisms. We discuss the types, causes of data leakage and preventive approaches as well.

Keywords: Data Leakage, Data Leakage Detection, Data Leakage Prevention

I. INTRODUCTION

Data is information. It is useful to an organization or any enterprise and is confidential to that particular entity. Data is measured, collected and analysed . Data might be confidential in several organizations. Hence, data leakage is possible by malicious attackers. There are several data leakage detection techniques .Data breach or leakage is sharing data to the outside world without permission of concerned authorities.it is a crime and is punishable. It can be intentional or accidental.

Nowadays the data leakage is a big challenge. As the data increases, it becomes more vulnerable to attacks. Confidential data is often termed as sensitive data. This sensitive data includes information regarding economical status of the organization, personal information, shareholder information etc. With the advancement in technology, there are plenty of mechanisms to detect data leakage detection in an organization. In this paper, we discuss the types of attacks, detection algorithms and preventive measures that can be undertaken to (avoid) prevent leakage of data(sensitive).

II. DATA LEAKAGE

Data is a sensitive information holds some meaning and is significant. Data Leakage is broad sense in sharing or sending data to unauthorized entities. Nowadays data leakage is a very serious issue to tackle. It is technically defined as an accident or intentional sharing of important or confidential data to the outside entities which are not a part of the organization.

In several organizations the sensitive data is shared to the outside world accidently which is also considered as to data leakage. The potential damage of data leakage can be of types, direct loss and indirect loss. Direct losses refer to the losses that can be measured and estimated. Indirect losses are harder to quantify and include exposure intellectual property organization irrespective of their working fields, size suffer data leakage.

III. CAUSES OF DATA LEAKAGE

1) Data deletion: This is the case when files get accidently deleted in the drive. It may remain unnoticed. Administrative errors also fall into this category.
2) Formatting the drive: Users tend to format their drives which may result in data loss.
3) Natural disasters: Natural disasters like earthquakes, flood, fire can damage computer system. This results in data loss.
4) Power failure: Saved data is lost.
5) Distorted data: If the file system is corrupt or database is is distorted, there is a chance of data loss
6) Software failure: When the software crashes and does not function properly, it leads to data loss. Operating system crash is the most serious type of software failure.
7) Virus attack: Attacks by viruses, spyware etc. causes severe damage to the system which certainly leads to data loss.

IV. TYPES OF DATA LEAKAGE

There are several types of attacks-

1) Hacking or computer intrusion – Which includes phishing, ransomware or malware, skimming.
2) Insider threat- Chances of an insider to leak the data (sensitive) to the outside world.
3) Physical theft- A simple example can be plugging a USB drive into a sensitive system.
4) Data leakage due to failure in web and exposure to internet- The data is moved to applications based on cloud infrastructure as the organizations migrate, thus it increases chances of data exposure on internet
5) Unauthorized access – Due to less secure access control mechanisms, data leakage takes place.

V. DATA LEAKAGE DETECTION TECHNIQUES

There are several mechanisms to detect data leakage. Each of them has its own advantages and disadvantages. Some of the mechanisms can detect any kind of data leakage. The latest hacking techniques target the companies ,organizations and government agencies etc., and attacks on a specific system and analyses the vulnerabilities within the system for a long time, this kind of attack is called Advanced Persistent Threat(APT) and is very hard to detect than traditional attacks[4].

A. Watermarking System

This method of watermarking is difficult to remove by an attacker though many of them conspire with different copies of the watermarking data

1) Embedding and Extraction:

In this type of method the significant portion of the fractional part of the pixel intensity value of cover image is encoded to provide watermark[5].

2) Wavelet Based Watermarking:

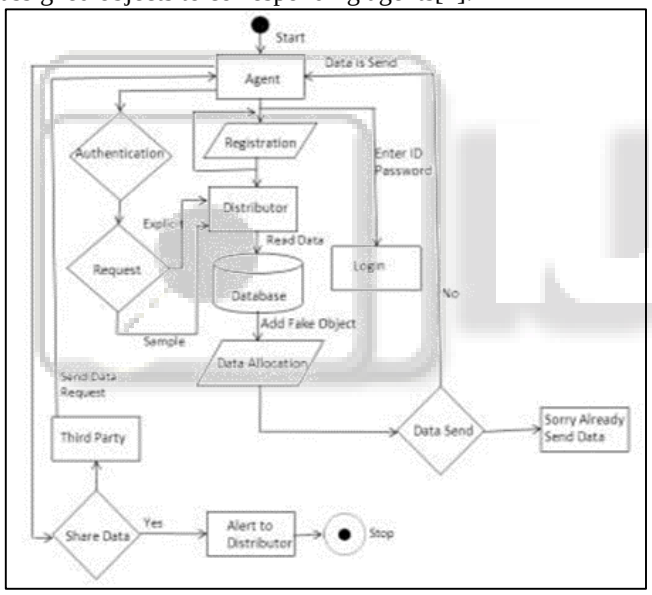
This is carried out by performing multi resolution data fusion for embedding where the image and watermark are transformed into discrete wavelet domain[5].

3) Invisible Water Marking:

This is a robust scheme of invisible watermarking used for embedding and extracting of a digital watermark into image. Insertion of invisible watermark is done only in the most significant parts of the host image[5].

B. Detecting the Guilty Agent:

A guilty agent is an individual or system who breaches the important confidential data to unauthorized systems, people or entities. In this technique use fake objects which are created by the distributor with the intension to detect data leak. These fake objects are designed to make them appear like real objects and are distributed to those who request for data(agents). These fake objects help in the easy detection of leakage. The distributor detects the guilty agents based on the assigned objects to corresponding agents[2].



C. Agent Guilt Model:

In this system, Fake objects are used for detection. The major intent of this system is to detect when the distributor's confidential data has been leaked by some agent and the agent that leaked the data is identified.

Modules are:-

1) Data allocation module
2) Fake Object module
3) Optimization module
4) Data distributor module
5) Agent Guilt module

1) Impact of probability:

Let T contains 16 objects and these objects are given to U1.18 objects are given to U2. Probability of guilt is calculated within [0,1]. As P approaches O1, the target can guess all the 16 values. Now each of the agent has enough leaked data that its individual guilt approaches 1.

D. K-Anonymity Algorithm:

Guilt agents are detected using fake objects. To increase the effectiveness and accuracy, the distributors add fake objects to the distributed data in order to detect the guilty agents[7]. There are two data allocation methods

1) Explicit Data Requests:

In this method, the distributor is prohibited from adding the fake objects to the distributed data. So agent's data request fully defines the data allocation.

Input: $R_1,\ldots, R_n, cond_1,\ldots, cond_n, b_1,\ldots,b_n, B$
Output: $R_1,\ldots, R_n, F_1,\ldots,F_n$

1. $R \leftarrow \Phi$
2. For i = 1,…, n do
3. If $b_i > 0$ then
4. $R \leftarrow R \cup \{i\}$
5. $F_i \leftarrow \Phi$
6. While B > 0 do
7. $i \leftarrow$ SELECTAGENT($R, R_1,\ldots,R_n$)
8. $f \leftarrow$ CREATEFAKEOBJECT($R_i, F_i, cond_i$)
9. $R_i \leftarrow R_i \cup \{f\}$
10. $F_i \leftarrow \cup \{f\}$
11. $b_i \leftarrow b_i - 1$
12. If $b_i = 0$ then
13. $R \leftarrow R \setminus \{R_i\}$
14. $B \leftarrow B - 1$

2) Sample Data Request:

The object of S is defined by the sample data request. It is used to compute the overall system reliability while the probability is used to identify the guessing agents that have been leaked the information. Based on the experiments conducted, the estimation of probabilities are done.

Input: $m_1,\ldots, m_n, |T|$
Output: $R_1,\ldots, R_n$

1. $a \leftarrow 0_{|T|}$
2. $R_1 \leftarrow \Phi,\ldots, R_n \leftarrow \Phi$
3. remaining $\leftarrow \sum_{i=1}^{n} m_i$
4. **while** remaining > 0 **do**
5. for all I = 1,…,n : $|R_i| < m_i$ do
6. $k \leftarrow$ SELECTOBJECT(I, R_i)
7. $R_i \leftarrow R_i \cup \{t_k\}$
8. $a[k] \leftarrow a[k] + 1$
9. remaining $\leftarrow$ remaining - 1

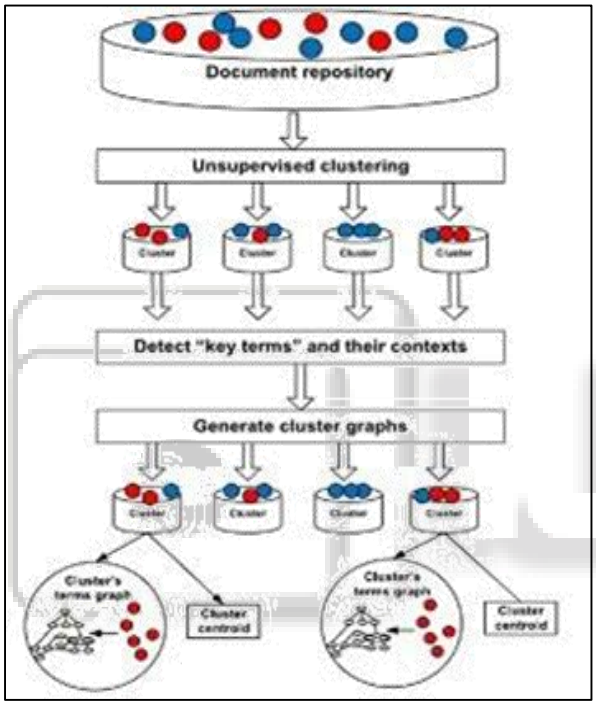
K-anonymity algorithm is used in this method. This provides simple approach is securing the personal and confidential information of an individual via liberating K anonymous views of data sets. A label is constructed and if the table assures K anonymity for some value K, then anyone who sense only the quasi identifier values of an individual cannot identify the documentation. AES algorithm is used in the Agent Guilt Module and it is built on permutation and substitution. Permutation are reordering of data and Substitution is to oust one unit of data by another unit. It has four operations: substitution bytes, shift rows, mix columns, add round key. The numbers of rounds depends on key size whether 128,192 or 256 bytes.

E. CoBAn:

It is a model which is based on unintentional and deliberate prevention of data leakage. There are two phases, namely training and the detection phase. In this course of time i,e in

training, group of documents are prepared and the context of this cluster is represented in graph. In the detection phase, the documents that have been tested is allocated to the group and the context in it are to be matched to the cluster with respect to the graph in order to decide the confidentiality of the context. The potential to decide whether the information is large non-confidential more than the approach [9].

This particular approach has two phase namely learning phase and detection phase .In the first phase i,e in learning phase , a graph is generated based on the sensitive content for each of the particular document that is confidential by using the training set. In the detection, the whole document is analysed and is checked with the graph in order to calculate the confidentiality score. If it exceeds the threshold value, the document is considered to be confidential.



VI. DATA LEAKAGE PREVENTION

Data leakage prevention (DLP) is a method which is used to ensure that the sensitive information is not leaked outside the organization by the end users. The phrase Data leakage prevention is used to find the potential data breaches and prevent these kind of situation by monitoring and supervising the sensitive/critical data while it is in use or even at rest so that it will not be acquired by any unauthorised party. Various terms that are associated with DLP are information leakage prevention (ILP), Extrusion prevention system. The data leakage prevention is managed by some threats and also by privacy loss. The file content will be examined based on the business rule by the data leakage prevention product and it would tag the lightly confidential information or the sensitive information so as the users will not be able to disclose it to anyone here tagging is a process where the data could be classified based on its confidentiality and marked appropriately. Ex; Tagging could prevent spreadsheet in the financial sector from being mailed by any of the employee to another employee within the organization. The data leakage prevention product have the following components, they are

- End points: Supervise and control the activities.
- Network: Filter the stream of data.
- Storage: Protects the data which is at rest.

In most of the organization there would be more number of servers along with the directories and also files that will be stored and particular type of data which has been tagged. This software is used for finding out contents like in terms of social security but fails in terms of intellectual property like graphic components, schematics and formulae. The data leakage prevention technique primarily focus on monitoring as well as location of traffic for matching confidential data that is at rest or also when in motion. In order to implement this solution one has taken up the approach that would risk the address along with the steps of migration and assurance measure. Apart from DLP, CI-750 uses the technique of fingerprints to identify the data that is structured such as social security and data that is unstructured like the document, source code etc. It was done by the code green technologies they create hash values of the data that has to protected and scan the outgoing traffic for matching. This particular technology has built in agents for email transfer and alerts us the data leakage occurring within the e-mail, web or even compressed archive transmission.

REFERENCES

[1] ”Data Leakage Detection and Data Prevention Using Algorithm” Dr. A R. Pon E. Thenmozhi.
[2] “Data Leakage Detection” Ghagare Mahesh1, Yadav Sujit2, Kamble Snehal3, Nangare Jairaj4, Shewale Ramchandra5
[3] ”Data leakage detection“ Ms. N. Bangar Anjali1, Ms. P. Rokade Geetanjali2, Ms. Patil Shivlila3, Ms. R. Shetkar Swati4, Prof. N B Kadu. Vol. 2, Issue. 5, May 2013, pg.283 – 288
[4] ”Data Leakage Detection and Prevention System” Chirag S. Patil , Swapnil S. Nalawade , Vikas D. Natekar , Prof. Neha Saxena .Volume 5 Issue 2, Mar – Apr 2017
[5] “Data Leakage Detection” Sandip A. Kale1, Prof. S.V.Kulkarni Vol. 1, Issue 9, November 2012
[6] “Fast detection of transformed data leaks“V.Prathibha, E.Dilipkumar– Volume 4 Issue 3, May – June 2017
[7] “Data Leakage Detection with KAnonymity Algorithm” Vol. 7 (4) , 2016, 1911-1915
[8] “CoBAn: A context based model for data leakage prevention” Gilad Katz , Yuval Elovici ,Bracha Shapira.

