An Overview of Data Protection by Data Leakage Detection

Ms Neha S. Barley[1], Prof. R. R. Keole[2]

[1]P.G. Department of Computer Science & Information Technology, HVPM, COET Amravati, Maharashtra, India.
[2]Asst. Professor , P.G. Department of Computer Science & Information Technology, HVPM, COET Amravati, Maharashtra, India.

Abstract—*In today's business world, the owner of the data are called as distributors and the trusted third parties are called as agents. Data leakage happens every day when confidential business information such as customer or patient data, company secrets, budget information etc are leaked out. This paper contains the results of implementation of Data Leakage Detection Model. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker can be identified. Currently watermarking technology is being used for the data protection. But this technology doesn't provide the complete security against date leakage. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious. This paper includes the difference between the watermarking & data leakage detection model's technology.*

Keywords— agents, distributor, data leakage, data security, fake records.

I. INTRODUCTION

In a company, sensitive data must be handed over to supposedly trusted third parties. Also the security depends on employees by learning the rules through training and awareness-building sessions. However, security must go beyond employee knowledge and cover the following areas such as a physical and logical security mechanism that is adapted to the needs of the company and to employee use then the procedure for managing updates and finally it needs an up to date documented system.

Data leakage happens every day when confidential business information such as customer or patient data, source code or design specifications, price lists, intellectual property and trade secrets, and forecasts and budgets in spreadsheets are leaked out. When these are leaked out it leaves the company unprotected and goes outside the jurisdiction of the corporation. This uncontrolled data leakage puts business in a vulnerable position. Once this data is no longer within the domain, then the company is at serious risk.

When cyber criminals "cash out" or sell this data for profit it costs our organization money, damages the competitive advantage, brand, and reputation and destroys customer trust. To address this problem, we develop a model for assessing the"guilt" of agents. The distributor will "intelligently" give data to agents in order to improve the chances of detecting a guilty agent like adding the fake objects to distributed sets. At this point the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. Mainly it has one constraints and one objective. The Distributor's constraint satisfies the agent, by providing number of object they request that satisfy their conditions.

II. SYSTEM IMPLEMENTATION MODELS

2.1. Data Allocation Module

In this module, administrator has to login with his id and password. Administrator has all the agent information, user data inside his database. Administrator is now able to view the database consisting of the original data as well as the fake data. Administrator can also list the agents here. He will be able to add additional information to the database. Agent's information can be added here.

2.2. Fake Object Module

The distributor creates and adds fake objects to the data that he distributes to agents. Fake objects are objects generated by the distributor in order to increase the chances of detecting agents that leak data. The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. Our use of fake objects is inspired by the use of "trace" records in mailing lists. In case we give the wrong secret key to

download the file, the duplicate file is opened, and that fake details also send the mail.

2.3. Data Distributor Module

Once the agent has been added by the administrator, he can create one username and password for that particular agent, in fact registering. After the agent has been successfully registered we now want to send the data to agent according to their request. Administrator will now select a requested amount of data and then export these data into an excel file in byte format. After the file is created, the administrator will send the data to agent. Sending the data includes transferring the data through the network (LAN).At the same time the administrator will keep the record of the agent with his id.

III. SYSTEM REQUIREMENTS SPECIFICATION

3.1 Aim of the project

Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data.

3.2 Description of the project in short

In this project we are finding out the data is leaked or not. The age will give the information to broadcast the data on the news media. We will check whether the authorized user leaked the data to another news media channel.

3.3 Algorithms

Allocation for Explicit Data Requests

In this request the agent will send the request with appropriate condition. Agent gives the input as request with input as well as the condition for the request after processing the data after processing on the data the gives the data to agent by adding fake object with an encrypted format.

Allocation for Sample Data Requests

In this request agent request does not have condition. The agent sends the request without condition as per his query he will get the data.

3.4 Operating Environment

3.4.1 Hardware

- Processor : Intel (R) Core(TM) i3 CPU
- Installed RAM : 1 GB
- System type : 32 bit operating systems

3.4.2 Software

- Java 1.6
- My SQL.

3.5 Project Plan

Plan of Execution

- Identification – Searching for different project ideas. Identifying and finalizing one of them for further implementation.
- Conceptualization and design – The concepts required for building the project are studied in detail.
- Also the high level designing is done at this stage. Preliminary presentation given for more clarification of project.
- Detailed design – At this stage, low level designing is done. User Interface is designed to give better visualization of the project idea.
- Coding – Actual implementation of the project starts at this stage. Coding for each module of the four modules will be done. Coding and testing will take approximately 8 to 10 weeks.
- System Testing – The product was tested in the context of the entire system. Different Linux systems will be used for system testing and the performance will be monitored.
- Documentation – A detailed document about the project shall be prepared at this stage.

IV. WATERMARKING THE DATA

A **Watermark** is a signal that is securely, imperceptibly, and robustly embedded into original content such as an image, video, or audio signal, producing a watermarked signal and it describes information that can be used for proof of ownership or tamper proofing. It provides an effective watermarking technique geared for the relational data. This technique ensures that some bit positions of some of the attributes of some of the tuples contain specific values. The tuples, attributes within a tuple, bit positions in an attribute, and specific bit values are all algorithmically determined under the control of a private key known only to the owner of the data. This bit pattern constitutes the watermark. Only if one has access to the private key then it is possible to detect the watermark with some high probability

Detecting the watermark neither requires access to the original data or the watermark. The watermark can be detected even in a small subset of a watermarked relation as long as the sample contains some of the marks. Protection of these assets is usually based upon the insertion of digital watermarks into the data. The watermarking software introduces small errors into the object being watermarked. These intentional errors are called marks and all the marks together constitute the watermark. The marks must not have

a significant impact on the usefulness of the data and they should be placed in such a way that a malicious user cannot destroy them without making the data less useful.

In Digital Media such as video, audio, images, text the information are easily copied and easily distributed via the web. While sharing secured information as provided some traditional data like Stock market data, Consumer Behavior data (Wal-Mart), Power Consumption data, Weather data the Database outsourcing is a common practice. So the Watermarking provides an effective means for proof of authorship by signature and the data as the same object and also it provides an effective means of tamper proofing by integrity information is used and embedded in the data.

4.1 Disadvantage

This data is vulnerable to attacks. There are several techniques by which the watermark can be removed. Thus the data will be vulnerable to attacks.

4.2 Advantage

This system includes the data hiding along with the provisional software with which only the data can be accessed. This system gives privileged access to the administrator (data distributor) as well as the agents registered by the distributors. Only registered agents can access the system. The user accounts can be activated as well as cancelled. The exported file will be accessed only by the system. The agent has given only the permission to access the software and view the data. The data can be copied by our software. If the data is copied to the agent' system the path and agent information will be sent to the distributors email id thereby the identity of the leaked user can be traced.

V. DATA LEAKAGE DETECTION USING CLOUD COMPUTING

In the virtual and widely distributed network, the process of handover sensitive data from the distributor to the trusted third parties always occurs regularly in this modern world.

Generally, the sensitive data are leaked by the agents, and the specific agent is responsible for the leaked data should always be detected at an early stage. Thus, the detection of data from the distributor to agents is mandatory. This project presents a data leakage detection system using various allocation strategies and which assess the likelihood that the leaked data came from one or more agents. For secure transactions, allowing only authorized users to access sensitive data through access control policies shall prevent data leakage by sharing information only with trusted parties and also the data should be detected from leaking by means of adding fake records in the data set and which improves probability of identifying leakages in the system. Then, finally it is decided to implement this mechanism on a cloud server Key to the definition of cloud computing is the "cloud" itself. For our purposes, the cloud is a large group of interconnected computers. These computers can be personal computers or network servers; they can be public or private. For example, Google hosts a cloud that consists of both smallish PCs and larger servers. Google's cloud is a private on (that is, Google owns it) that is publicly accessible (by Google's users).This cloud of computers extends beyond a single company or enterprise. The applications and data served by the cloud are available to broad group of users, cross-enterprise and cross-platform. Access is via the Internet. Any authorized user can access these docs and apps from any computer over any Internet connection. And, to the user, the technology and infrastructure behind the cloud is invisible. From Google's perspective, there are six key properties of cloud computing.

Today, because to cloud computing technology, the data of a database can be stored in the cloud, on collections of web server instead of housed in a single physical location. This enables users both inside and outside the company to access the same data, day or night, which increases the usefulness of the data. It's a way to make data universal.

VI. CONCLUSION

From this study we conclude that when the occurrence of handover sensitive data takes place it should always watermarks each object so that it could able to trace its origins with absolute certainty, however the data leakage detection system model is very useful as compare to the existing watermarking model. We can provide security to our data during its distribution or transmission and even we can detect if that gets leaked. Thus, using this model security as well as tracking system is developed. Our model is relatively simple, but we believe it captures the essential tradeoffs.

REFERENCES

[1] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," IEEE Transactions on Knowledge and Data Engineering, pages 51- 63, volume 23, 2011.

[2] R. Agrawal and J. Kiernan, "Watermarking Relational Databases," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), VLDB Endowment, pp. 155-166, 2002

[3] *Hartung and kutter*,‖Watermarking technique for multimedia data‖2003.

[4] *Chun-Shien Lu*, Member, IEEE, and *Hong-Yuan Mark Liao*, Member, IEEEMultipurpose Watermarking for Image Authentication and Protection

[5] Edward P. Holden, Jai W. Kang, Geoffrey R. Anderson, Dianne P. Bills, Databases in the Cloud: A Work in Progress,2012.

[6] Michael Miller, ‖Cloud Computing‖ Web-Based Applications that change the way you work and Collaborate Online,Pearson Education,2012.

[7] Data Leakage Detection, an IEEE paper by Panagiotis Papadimitriou, Member, IEEE, Hector Garcia-Molina, Member, IEEE NOV-2010.