ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Volume 1, Issue 3, May 2013

Data Leakage Detection and E-mail Filtering

Mr. Zarif Shaukat Ansari [1], Ms. Anagha Mahadeo Jagtap [2], Ms. Shilpa Suresh Raut [3]

Student, Dept. Of Computer Engineering, Trinity College of Engineering, Pune, Maharashtra, India[1,2,3]

Abstract: While doing business sometimes it is necessary to hand over company's or organisations sensitive data to supposedly trusted third parties (Agents).If this distributed data is found in an unauthorized place, it is quite possible that the distributed data has been leaked by one or more agents. Data can be leaked through e-mails, instant messaging, databases, spreadsheets and without knowledge of the distributor (owner of data). We proposed the system to find the guilty agent (agent who has leaked data). This project proposes data allocation strategies and adding "realistic but fake records" that improve the probability of identifying leakages. The goal is to detect when the distributor's sensitive data have been leaked by agents, to identify the agent that leaked the data and possibly to filter the e-mails in order to make distributors data secure.

Keywords: Sensitive data, Data leakage, Data allocation strategies, Fake records, E-mail filtering.

I. INTRODUCTION

Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity.In the course of doing business, sometimes it is necessary to give organizations sensitive data to trusted third parties. Sensitive data in companies and organizations includes intellectual property (IP), financial information, patient information, personal credit-card data, and other information depending on the business and the industry. 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. Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase.

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. The owner of the data is the distributor and the supposedly trusted third parties are the agents. It may happen that any of agents gives this sensitive data to any unauthorized user without knowledge of owner. So our goal is to detect when the distributor's sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data and what data has been leaked. Data can be leaked through e-mails. So there is a need to provide security to the data. For this purpose e-mail filtering concept has been implemented in the system. In this, even if any of the agents sends distributors sensitive data to unauthorized person via e-mail, unauthorized person will not be able to see or download that e-mail.

II. EXISTING SYSTEM

Traditionally, Watermarking techniques were used for handling data leakage detection. Watermarks were initially used in images, video or audio data whose digital representation includes considerable redundancy. Watermarking aims to identify a data owner and, hence, is subject to attacks where a pirate claims ownership of the data or weakens a merchant's claims. In watermarking technique, a unique code is embedded in the original data copy. And if that copy of data later found at unauthorized place or with unauthorized person then the leaker can be identified. But the drawback of this technique is that- it requires some modification of data. Also in some cases the watermarks can be destroyed if the data recipient is malicious. Hence there is need to propose efficient technique to find data leakage.

III. PROPOSED SYSTEM

In the proposed system, we develop a model for finding guilty agents. For this purpose different data allocation strategies are used. We are using "Fake records" which are not real but appear as real records in order to find the guilty agent. Here these Fake objects acts as a watermarks like in watermarking technique. It improves over the limitations of watermarking technique as it does not require any modification of original data. If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty. Also in proposed system, we have implemented e-mail filtering technique in which unauthorized users will be unable to see and download the contents of the e-mails which is send by guilty agent. So Distributors sensitive data remains secure.

IV. PROBLEM SETUP AND NOTATIONS

T is a set of data objects.
T= {*t1, t2... tn*};
Set of agents= {*U1, U2... Un*}

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Volume 1, Issue 3, May 2013

The objects in *T* could be of any type and size, e.g., they could be tuples in a relation, or relations in a database.
An agent *Ui* receives a objects *Ri*, where *Ri* is a subset of *T*, determined either by a sample request or an explicit request:

- Sample request *Ri* = SAMPLE (*T, mi*): Any subset of *mi* records from *T* can be given to *Ui*.
- Explicit request *Ri* = EXPLICIT (*T, condi*): Agent *Ui* receives *all* the *T* objects that satisfy *condi*.

We say an agent *Ui* is *guilty* and if it contributes one or more objects to the target. We denote the event that agent *Ui* is guilty as *Gi* and the event that agent *Ui* is guilty for a given leaked set *S* as *Gi*|*S*.

V. MODULE DESCRIPTION

- ***Fake objects:*** Fake objects are objects generated by the distributor in order to increase the chances of detecting agents that leak data. The distributor is capable to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. The use of fake objects by the system is inspired by the use of "trace" records in mailing lists. Fake objects are the objects which look exactly like the real objects. Fake objects are created by the distributor before sending data to agents. In every data which distributor will send to his agents, the position and number of fake objects will differs. Depending on the number of records the number of fake objects will differ so that it will be easy for system to detect the guilty agent.

- ***Data allocation strategies:*** The data allocation problem as how can the distributors "intelligently" gives data to agents in order to improve the chances of detecting a guilty agent. Data allocation depends on the request done by the agent and whether system can add fake object to it. The request done by the agent can be of two types:
  Sample- In sample data request agent receives a subset of distributers data which required by agent.
  Explicit- In explicit data request data satisfying a special condition is given to agent.

- ***E-Mail filtering:*** The Mail is being sent to authorized user and unauthorized user. As the unauthorized user receives the mail, the system detects that the mail has been send to the unauthorized user illegally; the system filters the data and block the contents of the mail. Here, on the user side, if the unauthorized user downloads that mail, the mail does not display the original contents of the mail and the downloaded file will be of size zero.

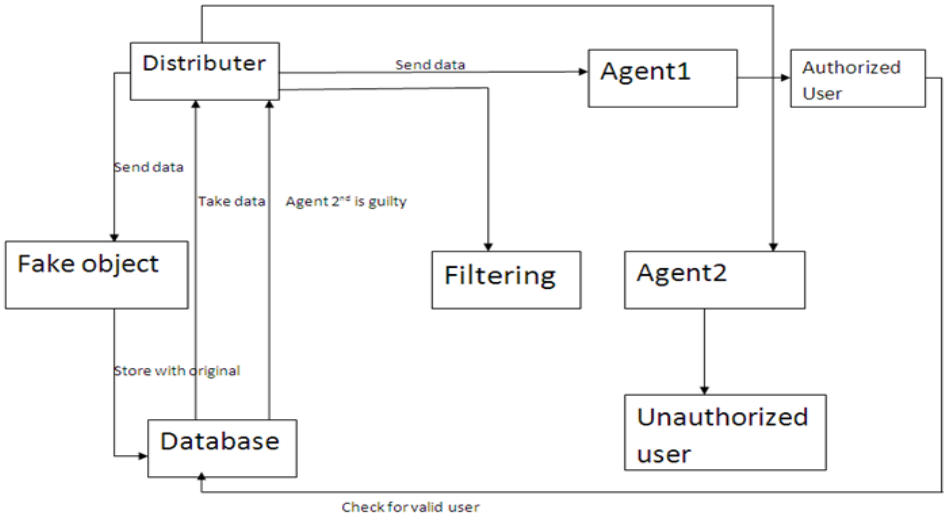

Fig 1: Block Diagram of Data Leakage Detection and E-Mail Filtering

As shown in fig 1, A distributor can insert original as well as fake records in the Database. A new agent can be registered by entering personal details. A registered agent can Login and make a request to the distributor for data. The request can be of two types- Sample or Explicit. The system then extracts the requested data from the main database and performs the addition of fake records to the set of original records. It then provides this data to the agent. The agent may pass on this data to an unauthorized party. The agent or the unauthorized party may leak the sensitive data on the internet, television or other media via email, instant messaging, and webmail or by any other means. Whenever the distributor discovers the leaked set of his data, he will note the objects present in it. He will compare these objects to those present in the data sets handed over to different agents. He will list out the names of the agents whose data set

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Volume 1, Issue 3, May 2013

contains the objects found in the leaked set. The fake records present in the data sets of the agents will be checked and the probability of an agent being guilty will be computed for each agent in the list. The agent having the maximum probability will be the guilty agent.

VI. CONCLUSION

To deal with the problem of Data leakage, we have presented implementation a variety of data distribution strategies that can improve the distributor's chances of identifying a leaker. Also we have implemented the concept of E-mail filtering to secure the distributors data. Scope of this system can be extended by making provisions for the generation of fake records dynamically according to the agent's request. Our future work includes the investigation of agent guilt models that capture leakage scenarios that are not studied in this paper.

ACKNOWLEDGEMENT

The success of any project is never limited to an individual. Similarly, our project is also an outcome of ideas contributed by many and we would like to gratefully acknowledge them here. We express our sincere thanks to all those who have provided us with valuable guidance towards the completion of this report as a part of the syllabus of the degree course.

We deeply thank our HOD Prof. P. Pankhti for her useful guidance. We also thank our Project Guide Prof. Arti Bhore without whom this project would have been a distant reality. We also thank them for giving us moral support, timely comments and discussion in all phases of the project.

We express our sincere thanks and gratitude to the project coordinator Prof. A. Nadaph for inspiring us throughout the completion of our project. We would also like to extend our sincere thanks to all teachers and staff for their valuable suggestions and feedback.

REFERENCES

[1] R. Agrawal and J. Kiernan, "Watermarking Relational Databases,"Proc. 28th Int'I Conf. Very Large Data Bases (VLDB '02), VLDB.Endowment, pp. 155- 166, 2002.
[2] F. Guo, J. Wang, Z. Zhang, X. Ye, and D. Li, "An Improved Algorithm to Watermark Numeric Relational Data," Information Security Applications, pp. 138-149, Springer, 2006.
[3] S. Jajodia, P. Samarati, M.L. Sapino, and V.S. Subrahmanian, "Flexible Support for Multiple Access Control Policies," ACM Trans. Database Systems, vo1.l34
[4] Panagiotis Papadimitriou and Hector Garcia- Molina, "Data Leakage Detection," lEE Transactions on Knowledge and Data Engineering, Vol 23, No.1 january2011
[5] B. Mungamuru and H. Garcia-Molina, "Privacy, Preservation and Performance: The 3 P's of Distributed Data Management," technical report, Stanford Univ.,2008
[6] S. Jajodia, P. Samarati, M.L. Sapino, and V.S.Subrahmanian, "Flexible Support for Multiple Access Control Policies," ACM Trans. Database Systems, vol. 26, no. 2, pp. 214-260,2009

BIOGRAPHY

Mr. Zarif Shaukat Ansari was Born in February 1991 in Pune, currently pursuing Computer Engineering in TCOER, Pune.

Ms. Anagha Mahadeo Jagtap was Born in February 1992 in Pune, currently pursuing Computer Engineering in TCOER, Pune.

Ms. Shilpa Suresh Raut was Born in November 1991 in Pune, currently pursuing Computer Engineering in TCOER, Pune.