Data Leakage Detection System

Aman Lamsal[1], Aniket Chakraverty[2], Aniket Paul[3], Chakri Venkat Katam[4], Kiran Bidua[5]
[1]B Tech (TY), Computer Science and Engineering, MIT ADT University, Pune
[2]B Tech (TY), Computer Science and Engineering, MIT ADT University, Pune
[3]B Tech (TY), Computer Science and Engineering, MIT ADT University, Pune
[4]B Tech (TY), Computer Science and Engineering, MIT ADT University, Pune
[5]Professor, Department of Computer Science and Engineering, MIT ADT University, Pune

DOI: 10.29322/DLDS.X.X.2020.pXXXX

----------***----------

Abstract – *Data leakage in an organization is a very important concern that leads to the ex-filtration of data, The work in this paper addresses a novel concept for prevention of data leakage for data in transit.We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party.Encryption is applied only to selected data and not the entire data in transit, ensuring that the hardware resources are efficiently utilized.*

***Key Words*:** DATA LEAKAGE PREVENTION (DLP), ENCRYPTION , WATERMARKS

1. INTRODUCTION

Data leakage is a condition where the confidentiality of the data is compromised. There are several means for data leakage to happen, major causes include 1) Intentional data leakage by an adversary internal to the organization, 2) Data leakage by a person external to the organization but has got temporary access rights to the victim organization's resources, 3)Unintentional leakage by Internal user or administrator. It is said that data is an asset of an organization hence protecting this asset becomes a very important aspect for an organization's growth and development. Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorised entity.

Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase.Data leakage is enhanced by the fact that transmitted data including emails, instant messaging, website forms, and file transfers among others, are largely unregulated and unmonitored on their way to their destinations.

Some of the techniques that can be used as a counter measure for data leakage detection are: 1) Scrambling: Where the sequence of the data is changed so that the meaning is preserved but the sequence is altered. 2) Perturbation: Is a very well known technique where the confidential content of the document is modified and made less sensitive, here the noise is introduced in the text so that the overall Signal to Noise ratio is reduced and hence the data leakage detection capability comes down. 3) Interleaving: Is a technique where the sequence of occurrence of the words is changed resulting in under weighing the confidentiality of the document. Traditional approach for data leakage prevention include 1) Watermarking: where in a unique code is embedded in each document which is confidential, 2) Finger Printing: Here the documents or the text is represented as a set of strings and a hash value is generated. Hash value's for each document is generated in a sliding window method, Now that the Database is available, each document is analyzed and its hash values are then compared against the hash values in the database, if sufficient number of matches in the hash values are found then the document is considered confidential.

In this project the module is providing complete information about the data/content that is accessed by the users within the website. Forms Authentication technique is used to provide security to the website in order to prevent the leakage of the data. Continuous observation is made automatically and the information is send to the administrator so that he can identify whenever the data is leaked.

1.1 Literature Survey

During the detailed reference for the literature research, we came across some words that include:

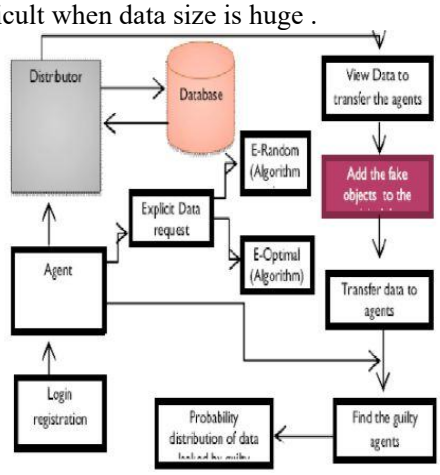
- Statistical(DLP) • Fake Object Algorithm in Sample Data Request • A D-SeGATE architecture • IDA-Algorithm • X-mark watermarking approach

(A) **Statistical Data Leakage Prevention(SDLP)**

Prabha,C.M and Satyavathy,G [1] presented a model to classify the data on the basis of semantics. The algorithms that have presented implement a variety of data distribution strategies like Singular Value Decomposition matrix (SVD) that can improve the distributors chances of identifying a leaker. The Authors have shown that only 60 percent of the modified documents were able too be identified upon transmission. So there exists scalability and integrity issues.

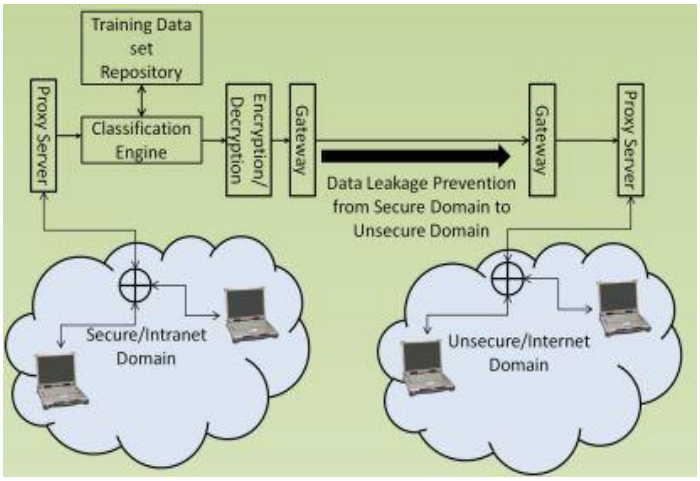
(B) Fake Object Algorithm in Sample Data Request

Verma, R., Gautam, V., Yadav, C.P., Gupta, I. and Singh, A.K., 2020, May[2] presented a model of adding fake object to the distribution set Such objects do not correspond to real entities but appear realistic to the agents. So if it turns out that an agent was given one or more fake objects, that were leaked, then the distributor can be more confident that agent was guilty. The Fake Object Algorithms in Sample & Explicit Data Request is Used. The authors have also identified some issues like analyzing a data file is difficult. Monitoring,matching and accessing their data becomes difficult when data size is huge .



(C) A D-SeGATE Architecture

Mohammed Ghouse DM,Manisha J. Nene Dept. of Computer Science and Engineering,VembuSelvi C MRS, Cyber Security Central Research Laboratory [3]presented a novel concept for prevention of data leakage(DLP) for data in transit. The text under consideration is classified to confidential or non-confidential category based on the content and context using Machine Learning technique. The Algorithms used are Scrambling,Perturbation&Interleaving of the datas. Understand on clustering and classification of texts/documents, and also various approaches are used eg: attribute reduction method, graph representation of texts/documents is presented.



(D) Information Dispersal Algorithm

Garay JA, Gennaro R, Jutla C, Rabin T.[4] presented a model on how to spread information in parts among servers in such a way that recovery is achievable even in the presence of up to inactive servers in his well-known Information Dispersal Algorithm work. Krawczyk later presented an improved mechanism for enabling creation in the presence of malevolent errors that might intentionally modify their pieces of information. These methods, on the other hand, believe that malicious defects only occur during reconstruction. The Algorithm used here is Information Dispersal Algorithm. The Authors have also identified a problem on simply storing a key in the clear would violate the confidentiality requirement.

(E) X-mark Watermarking Approach

Aiswarya, K. K., et al. "Application of Secret Sharing Scheme in Software Watermarking [5] presented a model Xmark watermarking approach to embed the watermark in source code utilising Shamir's secret sharing system and the collatz function. The watermark value is split using Shamir's secret sharing and fed into the collatz conjecture as input. Shamir's secret sharing approach has the advantage of being adaptable to any number of embedding locations. For instrumenting the source code, She utilise the sbt-instrumentation tool. She used the sbt-instrumentation tool to load the config file, which provides the rules needed to instrument the LLVM. The Algorithm used here is watermarking appoarch and the author have also identified the problem that An unscrupulous thief may even put his own water by removing yours.

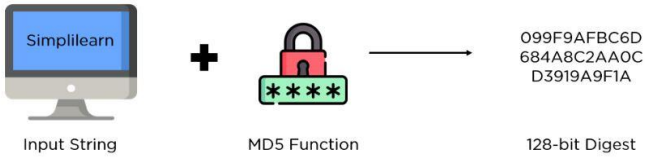
2. Objectives of the Project:

Our Project propose data allocation strategies that improve the probability of identifying leakages. In some cases, we can also inject “realistic but fake” data records to further improve our chances of detecting leakage and identifying the guilty party.

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. Another enterprise may out source its data processing, so data must be given to various other companies. There always remains a risk of data getting leaked from the agent. 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. But again it requires code modification. Watermarks can sometimes be destroyed if the data recipient is malicious. 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. 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. In this paper, we study unobtrusive techniques for detecting leakage of a set of objects or records. Specifically, we study the following scenario: After giving a set of objects to agents, the distributor discovers some of those same objects in an unauthorized place. (For example, the data may be found on a website, or may be obtained through a legal discovery process.) 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. Using an analogy with cookies stolen from a cookie jar, if we catch Freddie with a single cookie, he can argue that a friend gave him the cookie. But if we catch Freddie with five cookies, it will be much harder for him to argue that his hands were not in the cookie jar. If the distributor sees “enough evidence” that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings. In this paper, we develop a model for assessing the “guilt” of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding “fake” objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects act as a type of watermark for the entire set, without modifying any individual members. If it turns out that an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.

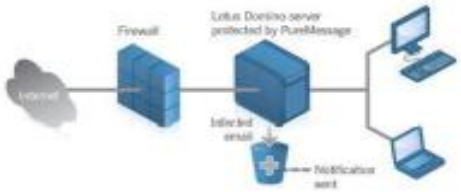
3. PROPOSED ALGORITHM

1.We have used MD5 (Message Digest Method 5) as our cryptographic hashing algorithm used to generate a 128-bit digest from our uploaded txt file.



There are basically four major sections of the algorithm : Padding bits , Padding length, Initialize MD buffer,&Process Each Block .

2. Forms Authentication technique is used to provide security to the website in order to prevent the leakage of the data.



3. We have used K-anonymity algorithm for the data privacy . K-anonymity means that the observed data cannot be related to fewer than k respondents. Key to achieving k-anonymity is the identification of a quasi-identifier, which is the set of attributes in a dataset that can be linked with external information to reidentify the data owner. We have used this algorithm to check between the data-sets and identify the leaker

4. CONCLUSION AND FUTURE WORK

In our project of Data leakage detection, we presented that whenever someone user share files with others there will be a secret key and if someone want to download or access file, they need a secret key which will be request to sender who share this file, after sharing secret key user can download that file, if someone download that file without asking secret key using guess method that will be notified as leaker.

Future Work: In future, we will aim for the experimental results of our proposed Data leakage Detection solution for a system designed for an organizational point of view.We will aim to use the techniques and algorithms which we have not used in this project thus we can improve this project by using

strong symmetric key algoithm's like AES,3DES etc. Also As this project is currently an OFFLINE PROJECT, we can buy domain or host our database (making it online) for sharing the secret key via Email or any other SMS platform. for eg: domain and sms system for transaction otp etc. We wish to explore Data leakage detection for data in different states (i.e data at rest, data in use,encrypted/compressed/zipped data). We will also explore DLD solution for Hand held endpoint devices such as tab, smart phone etc.

5. REFERENCES

[1] Sandip A.Kale, Prof. Kulkarni S.V. (Department Of Computer Sci.
&Engg,MIT College of Engg, Dr.B.A.M.University, Aurangabad(M.S),
India, Data Leakage Detection: A Survey, (IOSR Journal of Computer
Engineering (IOSRJCE)ISSN : 2278-0661 Volume 1, Issue 6 (July-Aug
2012), PP 32-35 www.iosrjournals.org

[2] IEEE Transactions On Knowledge And Data Engineering, Vol. 22,
No. 3, March 2011 Data Leakage Detection Panagiotis Papadimitriou,
Member, IEEE, Hector Garcia-Molina, Member, IEEE P.P (2,4-5)

[3] Faith M. Heikkila, Data Leakage: What You Need to Know, Pivot
Group Information Security Consultant. P.P (1-3)

[4] Rudragouda G Patil Dept Of CSE, The Oxford College Of Engg,
Bangalore.International Journal Of Computer Applications In
Engineering Sciences [VOL I, ISSUE II, JUNE 2011] [ISSN: 2231-
4946] P.P (1, 4) Development Of Data Leakage Detection Using Data
Allocation Strategies

[5] Chun-Shien Lu, Member, IEEE, and Hong-Yuan Mark Liao,
Member, IEEE MultipurposeWatermarking for Image Authentication
and Protection

[6] A. Shabtai, a. Gershman, M. Kopeetsky, y. Elovici Deutsche
Telekom Laboratories at Ben-Gurion University, Israel. Technical
Report TR-BGU-2409-2010 24 Sept. 2010 1 A Survey of Data Leakage
Detection and Prevention Solutions P.P (1-5, 24-25)