Data Leakage Detection and File Monitoring in Cloud Computing

Niketa Kirdat
kirdatniketa@gmail.com
Terna Engineering College, Navi Mumbai, Maharashtra

Nilima Mokal
digisol16@gmail.com
Terna Engineering College, Navi Mumbai, Maharashtra

Juhi Mokal
juhimokal@gmail.com
Terna Engineering College, Navi Mumbai, Maharashtra

Aishwarya Parkar
aishwaryaparkar19@gmail.com
Terna Engineering College, Navi Mumbai, Maharashtra

R. V. Shahabade
rameshshahabade@ternaengg.ac.in
Terna Engineering College, Navi Mumbai, Maharashtra

ABSTRACT

In the virtual and cosmopolitan network, the method of redeeming sensitive information from the distributor to the trusty third parties continually happens frequently during this times. It is important to safeguard the security and durability of service based on the demand of users. In the present scenario, the transfer of data is an important thing between one users to another. Data is mainly sent by the distributors which are generally the owner of data to the user which wants the information. The information sent by the distributor must be confidential and must be shared by a secure way. Sometimes during sharing of data, multiple copies of information is generated by different parties which cause a huge amount of loss, this is known as data leakage. To prevent this escape of information one should place measures so as to discover the escape in associate early stage. When user downloads a file, shares or views the file notification will be provided to the owner through email. When a user downloads or shares the file the date and time will also be recorded and log files will be maintained.

Keywords: Watermarking, Data leakage, Fake Object, Blowfish Algorithm, AES Algorithm.

1. INTRODUCTION

Data Leakage can occur through a variety of methods - some are simple, some complex. As such, there is no single “silver bullet” to control Data Leakage. Data leakage detection is an increasingly important part of any organization’s ability to manage and protect critical and confidential information. Examples of critical and confidential data that applications can access include: Intellectual Property, Corporate Data, and Customer Data. Watermarks are very useful in a relational database, which involves some modification of data. The goal of our paper is to detect when the distributor’s sensitive data has been leaked by agents, and show the probability for identifying the agent that leaked the data using encrypted fake objects.

Encryption is that the strategy of cryptography messages (or information) in such however that eavesdroppers or hackers cannot scan it, but that licenced parties can. Encrypting data enables confidentiality; this means that if the data falls into unauthorized hands the data is unreadable.

In this paper, we tend to present algorithms for distributing objects to agents, in a very manner that improves our possibilities of distinguishing a source. At last, we additionally take into account the option of adding “fake” objects to the distributed set. Such objects don’t correspond to real entities however seem realistic to the agents. In a way, the fake objects act as a sort of watermark for...
the complete set, while not modifying any person members. If it turns out that an was given one or a lot of fake objects that were leaked, then the distributor may be more assured that agent was guilty.

Cloud computing has been considered as a new model of enterprise IT infrastructure, which can organize huge resource of computing, storage and applications, and enable users to enjoy ubiquitous, convenient and on-demand network access to a shared pool of configurable computing resources with great efficiency and minimal economic overhead. Attracted by these appealing features, both individuals and enterprises are motivated to outsource their data to the cloud, instead of purchasing software and hardware to manage the data themselves.

2. LITERATURE SURVEY

Our work is generally relevant to watermarking that's used as a way of creating original ownership of distributed objects. Watermarks were initially used in images video and audio data whose digital representation includes considerable redundancy. Recently and other works have also studied marks insertion to relational data. Our approach and watermarking are similar within the sense of providing agents with some kind of receiver identifying data. However, by its very nature, a watermark modifies the item being watermarked. If the object to be watermarked cannot be modified, then a watermark cannot be inserted. In such cases, ways that attach watermarks to the distributed information aren’t applicable. Finally, there are also lots of other works on mechanisms that allow only authorized users to access sensitive data through access control policies. Such approaches prevent in some sense information leak by sharing data solely with trustworthy parties. However, these policies are restrictive and may make it impossible to satisfy agents’ requests. Many efficient data leakage detection method had been developed in the last decades; some of the prominent studies are given below.

Shreyta Raj discusses a correct technique for representing and police work inconsistencies of combined secrecy models is to notice once the pc distributor’s sensitive data has been leaked by their agents and if possible to identify the agent that leaked the data. Data leakage is a silent type of threat. This sensitive info are often electronically distributed via e-mail, Web sites, FTP, instant electronic communication, spreadsheets, databases, and the other electronic means that obtainable – all while not your data. Data allocation strategies (across the agents) are proposed that improve the probability of identifying leakages. These methods do not rely on alterations of the released data (e.g., watermarks). “Realistic but fake” data records can sometimes be added by the distributor to further improve our chances of detecting leakage and identifying the guilty party. A model for assessing the “guilt” of agents using C# dot net technologies with MS sql server as backend is proposed to develop. Algorithms for distributing objects to agents, in an exceedingly manner that improves our possibilities of distinguishing a source is aloes presented. Finally, the option of adding “fake” objects to the distributed set is also considered. Such objects do not correspond to real entities but appear.

Sandip A. Kale describes the results of implementation of knowledge leakage Detection Model. Presently watermarking technology is being used for the information protection. But this technology doesn’t provide the complete security against date leak age.

Different Techniques of Watermarking

A. Watermarking by DCT (Discrete Cosine Transforms)

Discrete cosine transform (DCT) is a method for converting a signal into elementary frequency components. In this method, the image is first converted into 8x8 blocks of pixels. After DCT conversion, the mid-frequency range are selected which is based on Gaussian network classifier. Now the mid-frequency DCT coefficients are used for embedding.

B. Watermarking by DWT (Discrete wavelet transform)

This is a modern method which is widely used for watermarking, image compressions etc. This technique uses wavelet filters to transform the image. Wavelets are small waves of varying frequency and limited duration. The wavelet transform decomposes the image into three spatial directions horizontal, vertical and diagonal.

C. LSB (Least Significant bit)

In this method, the watermark is embedded in the LSB of pixels. This method is easy to implement but it is not very secure against attacks i.e. the watermark can be destroyed easily. The watermarking is done by choosing a subset of image pixels and then substituting the LSB of each of the chosen pixels with watermark bits.

D. Watermarking by Embedding and Extraction

In this method, the insignificant part of the fractional part of the pixel intensity value of the main cover image is encoded to provide watermark. The watermark in the insignificant maintains the accuracy of the image. In this method the watermark is unnoticeable. A large amount of watermark can be easily embedded and extracted using this method which will help companies and firms involved in digital information security products. It is an added advantage of this method. Various algorithms for embedding and extraction are used in this technique.
E. Wavelet Based Watermarking
In this method, the multi-resolution data fusion is embedded where the image and watermark are both used and transformed into discrete wavelet form. The watermark is embedded into each wavelet level. The average of the estimates from each resolution level of wavelet decomposition is taken to detect the watermark. This algorithm works for JPEG compression, additive noise and filtering operations.

F. Secure Spread Spectrum Watermarking
In order to save multimedia data such as audio, video and image, a watermark must be added in significant components of a signal if it is to be robust to common signal distortions and malicious attacks. But, the modifications of these components may lead to degradation of the data signal. So, watermark must be added to the spectral components of the data using techniques similar to spread spectrum communications, hiding a narrow band of the signal in a wide band signal. This watermark is very difficult for an outsider to remove, even many outsiders collude with different copies of the watermarked data.

G. Robust Watermarking Technique
This watermarking technique can not only survive general operations such as compressions, adding noise, filtering and so forth but also geometric attacks such as rotation, scaling.

Previously, 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 involve some modification of the original information but they can be extremely helpful in some cases. Also, watermarks will generally be demolished if the information recipient is malicious. E.g. A hospital may give patient records to research who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise could outsource its processing, thus information should tend to numerous alternative corporations. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Disadvantages: However, there are two major disadvantages of the above algorithm: 1. It involves some modification of data i.e making the information less sensitive by altering attributes of the information. This alteration of data is called perturbation. But in some cases, it's necessary to not alter the original distributed information. For example, if an agent is doing the payroll, he must have the exact salary. We cannot modify the salary in this case. 2.

3. PROBLEM DEFINITION

3.1. Agent Guilt Model
An agent Ui is guilty if it contributes one or a lot of objects to the target. The event that agent Ui is guilty for a given leaked set S is denoted by Gi | S. "The next step is to estimate Pr{Gi|S} . i.e., the possibility that agent Gi is guilty given verification S. To compute the Pr{Gi | S}, estimate the probability that values in S can be "guessed" by the target. For illustration, say some of the objects in t are electronic mails of individuals. Conduct an experiment and ask a person to find the email of say 100 individuals, the person may only discover say 20, leading to an estimate of 0.2. Call this estimate as pt, the probability that object t can be estimated by the target.

3.2. Data Allocation Problem
The distributor “intelligently” gives data to agents in order to improve the chances of detecting a guilty agent. Depending on the type of data requests made by agents and whether “fake objects” are allowed, there are four instances of this problem. Agent makes two types of requests, called sample and explicit. Based on the requests the fakes objects are added to data list. Fake objects are objects produced by the distributor that are not in set T with help of encryption algorithm. The objects are designed to look like real objects, and are distributed to agents together with the T objects, in order to increase the chances of detecting agents that leak data.

3.3. Optimization Problem
The distributor’s information allocation to agents has one limitation and one aim. The distributor’s constraint is to satisfy agents’ requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His aim is to be able to detect an agent who leaks any section of his data.

We define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted data in cloud computing (MRSE). We initiate a set of strict privacy demands for such a secure cloud data utilization system. Among various multi-keyword semantics, we choose the efficient similarity measure of “coordinate matching,” i.e., as many matches as possible, to capture the relevance of data documents to the search query. We further use “inner product similarity” to evaluate such similarity measure. As a special case of modification, the operation of deleting existing documents introduces less computation and communication cost since it only requires updating the document frequency of all the keywords contained by these documents.

4. PROPOSED SYSTEM
Our main aim and objective is to detect when the distributor’s personal and sensitive information has been disclosed by agents, and if possible to detect the agent that leaked the info. In this section we develop a model for assessing the “guilt” of agents. We also show algorithms for distributing objects to agents, in a way that enhances 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 acts as a type of watermark for the entire set, without reconstructing any individual members. 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.
In propose system owner can upload and share file in cloud computing get notification the owner and also user can download and share file sent notification to owner.

4.1. Advantage Proposed System:-

- We will give protection to our information throughout its distribution or transmission and even we will observe if that gets leaked.
- We have presented the implementation of a variety of data distribution strategies that can improve the distributor’s chances of identifying a leaker.
- Quick response time
- Customized processing
- Small memory factor
- Highly secure
- Replication in Heterogenic Database
- Easy updating.

4.2. MODULES

A. Data Allocation Module:
Data allocation drawback is that the main focus of our project as to however will the distributor “intelligently” provide information to agents in order to enhance the possibilities of detection.
Example:- Admin can send the files to the authenticated user, users can edit their account details etc. Agent views the secret key details through mail. In order to enlarge the chances of detecting agents that are responsible for data leakage.

B. Fake Object Module:
The distributor creates & adds fake objects to the data that he distributes to agents. The distributor generates the fake objects in order to increase the chances of detecting agents that leak data the distributor is will also be able to add. The distributor is will also be able to add fake objects to the distributed info so as to boost his efficiency to find out guilty agents. We have used the fake objects by taking the inspiration by the use of “trace” records in mailing lists. In case we have a tendency to provide the incorrect secret key to download the file, the duplicate file is opened, and that fake facts additionally send the mail. Ex: The fake object details shall be displayed.

C. Optimization Module:
The distributor’s allocation of the data to agents has one limitation and one objective. The distributor’s restrictions to satisfy agents’ requests, by providing them with the number of objects they request or with all obtainable objects that satisfy their conditions. His objective is to be able to find an agent who leaks any portion of his information.

D. Data Distributor:
A set of supposedly trusted agents (third parties) were given sensitive data by the distributor. In an unauthorized place, some of the data are leaked and found. (e.g., on the web or somebody’s laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means.

4.3. AES Algorithm

The AES algorithm relies on permutations and substitutions. Permutations are the rearrangements of information, and substitutions replace one unit of information with another. Using several different techniques, AES performs permutations & substitutions. Inside loop there are four operations SubBytes, ShiftRows, MixColumns, and AddRoundKey which are called, that executes Nr times—the number of rounds for a given key size, less 1. The number of rounds that the encryption algorithm uses is 10, 12, or 14 and depends on whether the seed key size is 128, 192, or 256 bits.

In this example, because Nr equals 12, the four operations are called 11 times. After the completion of this iteration, before copying the State matrix to the output parameter, the encryption algorithm finishes by calling SubBytes, ShiftRows, and AddRoundKey. In summary, there are four operations that are at the heart of the AES encryption algorithm. From the seed key value the AddRoundKey substitutes groups of 4 bytes using round keys generated. SubBytes substitute’s individual bytes using a substitution table. SubBytes substitutes individual bytes using a substitution table. By rotating 4-byte rows, ShiftRows permutes groups of four bytes. Using a combination of both field addition and multiplication, MixColumns substitutes bytes.

Allocation for Explicit Data Requests: In this request the agent is sending the request with correct and proper 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 information Requests: during this request agent request doesn’t have condition. The agent sends the request without condition as agent sends the request without condition as per his query he will get the data. The distributor must estimate the likelihood that the leaked data came from one or more agents, as contradictory to having been independently gathered by other means. We develop a model for finding out the “guilt” of agents. We tend to additionally present algorithms for distributing objects to agents, in a very method that improves our possibilities of identifying leaker. Finally, we also consider the option of addition of the “fake” objects to the set.
4.4 Blowfish Algorithm

Blowfish is the symmetric block encryption algorithm designed in thought with.

**Fast**: It encrypts information on massive 32-bit microprocessors at a rate of twenty six clock cycles per computer memory unit.

**Compact**: It will run in less than 5K of memory. Simple: It uses addition, XOR, operation table with 32-bit operands.

Simple: It uses addition, XOR, lookup table with 32-bit operands.

Secure: The key length is variable, it will be within the range of 32~448 bits: default 128 bits key length.

It is appropriate for applications wherever the key doesn't modify usually, like communication link or an automatic file encryptor. Unpatented and royalty-free.

**Description of Algorithm:**

Blowfish symmetrical block cipher algorithm encrypts block information of 64-bits at a time.it will follows the feistel network and this algorithm is split into 2 components.

1) Key-expansion

2) Data Encryption

1) Key-expansion:

It can convert a key of at the most 448 bits into many sub key arrays totaling 4168 bytes. Blowfish uses sizable amount of sub keys. These keys are generated earlier to any encryption or decryption. The p-array consists of eighteen, 32-bit sub keys:

P1,P2,...........,P18

Four 32-bit S-Boxes consists of 256 entries each:

S1,0, S1,1,..........., S1,255
S2,0, S2,1,..........., S2,255
S3,0, S3,1,..........., S3,255
S4,0, S4,1,..........., S4,255

Generating the Sub keys:

The blowfish algorithm is used to calculate the subkeys.
1. Initialize 1st the P-array & next the four S-boxes, in order, with the stable string.
   The hexadecimal digits of pi (less the initial 3): P1 = 0x243f6a88, P2 = 0x85a308d3, P3 = 0x13198a2e, P4 = 0x03707344, etc are included in this string.
2. XOR P1 with the first 32 bits of the key, XOR P2 with the second 32-bits of the key, and so on for all bits of the key (possibly up to P14). Repeatedly cycle through the key bits till the whole P-array has been XORed with key bits. (For each short key, there's a minimum of one equivalent longer key; as an example, if A could be a 64-bit key, then AA, AAA, etc., are equivalent keys.)
3. Using the sub keys expressed in steps (1) and (2).The all-zero string should be encrypted with the Blowfish algorithm.
4. Replace P1 and P2 with the output of step (3).
5. Encrypt the output of step (3) using the Blowfish algorithm with the modified sub keys.
6. Replace P3 and P4 with the output of step (5).
7. Continue the strategy, substituting all entries of the P array, then all four S-boxes so as, with the output of the endlessly modifying Blowfish algorithm.

2) Data Encryption:

It has a function to reiterate sixteen times of network. every round consists of key-dependent permutation and a key and data-dependent substitution. XORs and additions are all the operations on 32-bit words. The only other different operations are four indexed array data lookup tables for each round.
Algorithm: Blowfish Encryption

Divide x into two 32-bit halves: xL, xR

For i = 1 to 16:

xL = xL XOR Pi

xR = F(xL) XOR xR

Swap xL and xR

Swap xL and xR (Undo the last swap.)

xR = xR XOR P17

xL = xL XOR P18

Recombine xL and xR

Figure 2: Blowfish Encryption
5. CONCLUSION

Data leakage is a silent type of danger. Sensitive and confidential information can be leaked by an employee of an organization as an insider. This confidential information can be electronically distributed via e-mail, Web sites, FTP, instant messaging, spreadsheets, databases, and any other electronic means available – all without your knowledge. To estimate the danger of distributing information 2 things are necessary, where first one is information allocation technique that helps to distribute the tuples among customers with minimum overlap and second one is calculating guilt probability which is based on overlapping of his data set with the leaked data set. Our model is relatively simple. The algorithms we’ve given implement a range of information distribution methods which will improve the distributor’s possibilities of identifying a leaker. We have shown that distributing objects can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive. Another and most drawback to be solved is protecting information before obtaining leaked.

6. REFERENCES