International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 629 – 631

Privacy Preservation in data leakage detection of dynamically changing sensitive data

Manish Prabhu
Student,ME second year
SKNSITS,Lonavala
Maharashtra
manishprabhu812@gmail.com

Supriya Sarkar
Assistant Professor
Computer Engineering Department
Maharashtra
supriya.sarkar@rediffmail.com

***Abstract*—**The number of data leakage instances is increasing day by day. Human mistakes are main cause of data leakage. Data leakage caused by human mistakes has some solutions. These solutions provide an alert when there is a data leakage. One of the common approach is to give the contents in the databases and transmitted information over the network to the data leakage detection service provider for getting information about the leakage. But this is an open invitation for an outsider to gain the knowledge about sensitive information about the company or an organization; because the detection is based on third party and contents are available with the data leakage detection service provider. In this report, author has proposed technique for data leakage detection using Rabin fingerprint and RSA algorithm. Using this technique, data owners can perform data leakage detection by preserving the privacy of their sensitive information.

Keywords-*Data leakage detection,shingles,fingerprint,false negatives*

*****

I. INTRODUCTION

Some organizations are frequently involved in the transactions. Those transactions may be financial transactions, business transactions etc. It means that they are continuously transmitting some information to third party. In this case, the organizations are called as data owners because they have their own set of sensitive data.

If somebody gets illegal or unauthorized access to the data, then this scenario is called as data leakage. This may include transferring emails to third parties, assigning wrong privileges to the user etc. These kinds of leakages must get avoided first. If avoidance is not possible then they should get detected or prevented. There are so many techniques to perform data leakage detection like perturbation, watermarking, access control, stealthy malware detection etc.

Sometimes the data is handed over to the trusted third parties called as agents. It may be done for analysis. For example, a company may give the transactional records to the third party for finding profit-loss information and for deciding new business strategy. If these agents are curious, then there is a possibility of leakage. For that purpose, the data is given to the agents by adding some noise. This technique is called as perturbation. Here the original data is changed. Hence it becomes useless for further processing. Watermarking is another technique in which a unique code is inserted in each distributed copy. If it is found at unauthorized place, then it can be easily detected. If there is a transmission of data over the network, then there should be some entity which will identify sensitive data flowing through the network. The entity which performs data leakage detection and it provides data leakage detection as a service is called as Data Leakage Detection provider. This entity may be present on the boundary of the network or it may be present outside the network. This entity is responsible for analyzing the packets flowing through the network and it will generate alert when it detects the sensitive data patterns. For detection purpose, data leakage detection provider must have sensitive data with it.

Again this becomes privacy issue for an organization because the data leakage detection provider is considered as trusted or semi-trusted third party. If the data leakage detection provider server is compromised, then there will not be any security of the data.

Hence the data owner should give their data to the data leakage detection provider by processing it. They must apply some algorithm on the data so that plaintext data can be hidden and there will not be any problem for data leakage detection task. One of the algorithm is Rabin fingerprint algorithm[3] which works similar to hash functions used in security system. The heart of Rabin fingerprint algorithm is Irreducible polynomial. This polynomial is taken as a divisor for getting reminder by dividing the data. In this process, divisor, dividend i.e. data and reminder is represented in binary form. The reminder of the division operation is considered as a fingerprint. The irreducible polynomial is considered as a key for finding the fingerprint. If this key is derived, then this algorithm becomes insecure. Hence we have used Rabin fingerprint along with RSA algorithm[1] to get more security.

II. EXISTING SYSTEM

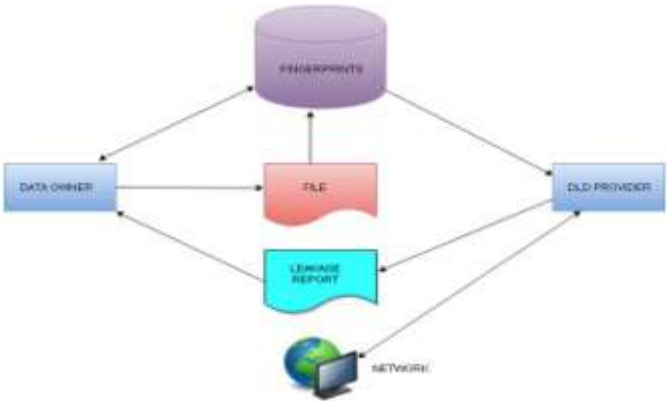


Fig 1. Existing system architecture

In existing system[3], the idea of Rabin fingerprint algorithm is proposed for achieving data leakage detection task. This system works based on shingling of the data. Shingle is the combination of characters present in the data. For example, if abcd is the data and if shingle length is 3 then there are two shingles of the data like {abc, bcd}. Once we get the shingles of the data, a Rabin fingerprint algorithm is applied to get the fingerprint of the shingle. These fingerprints are further processed to get fuzzified fingerprints.

These fuzzified fingerprints are given to the data leakage detection provider along with the parameters used for calculating fingerprints. The data leakage detection provider will analyze the packets and extracts the payload of the packet. Once it gets the payload, it will apply same algorithm on it and does the same processing. The result will be a set of fuzzified fingerprints. The data leakage detection provider will perform set equivalence test on the received fingerprints and calculated fingerprints and it will generate the report of matched fingerprints. The matched fingerprints are treated as a leakage. If two or more data results in same fingerprint, then the leakage report will consist of real leakage and noise. When the data owner receives the leakage report, then it applies post-processing on it to identify real leakages.

III. PROPOSED SYSTEM

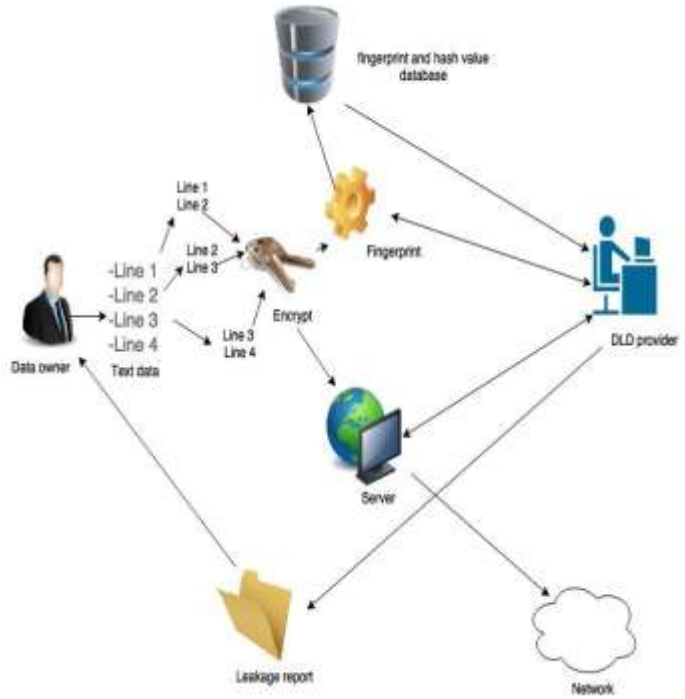


Fig 2. Proposed system architecture

A. *algorithm for proposed system*

for data owner

step 1: find the shingles of the data, encrypt them using RSA algorithm.

Step 2: find the fingerprint of the encrypted shingles

Step 3: assign unique identifier to each fingerprint

Step 4: calculate hash value using unique identifier and fingerprint. Use SHA-1 algorithm[2] for calculating hash value.

Step 5: fuzzify the fingerprints, store the date and time of creating the fingerprint.

Step 6: send set of fingerprints, hash values and unique identifiers to data leakage detection provider.

For data leakage detection provider

Step 1: get encrypted network data.

Step 2: apply rabin fingerprint algorithm on it based on shingling of the data.

Step 3: calculate the hash valuesof the fingerprint using unique identifiers.

Step 4: match the records using calculated hash values and received hash values.

Step 5: prepare the report and send it to the data owner.

B. *Mathematical model*

Let q,p(x),M,fid.pd are the parameters used for the calculation of the fingerprints. q is the length of the shingles,p(x) is the irreducible polynomial. Eaxh fingerprint is having length pf and fuzzy length pd. M is the bitmask which contains pd 0s at random positions. f is the set of fingerprints. fdot is the random binary string which is having same length as that of fingerprint. Fstar is the set of fuzzified fingerpints.

$$Fstar=((NOT\ M)AND\ fdot)XOR\ f \quad \ldots\ldots\ldots\ldots(1)$$

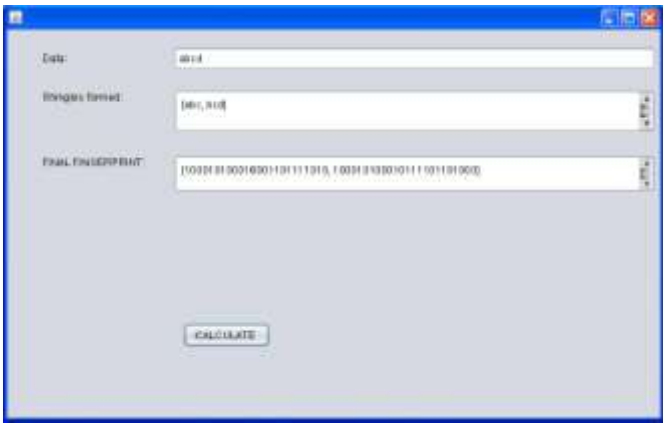
Let h(f,fid) is the hash function used to calculate unique hash value for the fingerprint where f is the fingerprint and fid is the unique value assigned to the fingerprint.

IV. IMPLEMENTATION DETAILS

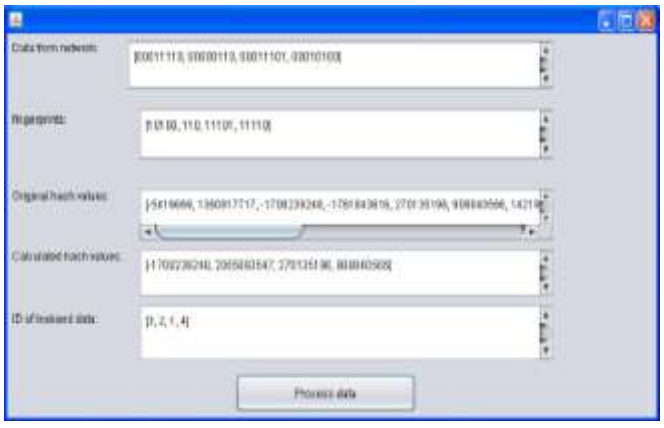
Software requirements: Netbeans IDE,JDK 1.6or higher,MySQL

Hardware requirements: minimum Dual core processor machine with atleast 2GB RAM and 320 GB HDD.

A. *Data owner form*



B. *DLD provider form*



CONCLUSION

The Rabin fingerprint algorithm with RSA provides better privacy as compared to simple Rabin fingerprint algorithm. By using hash value for each fingerprint, one can get exact match for each fingerprint hence the number of false positives will be less. By storing the date and time information for the fingerprints, one can get the idea about the data modification or updates. This will be useful in case of dynamic or frequently updating data.

REFERENCES

[1] Supriya Singh,"Data leakage detection using RSA algorithm", International Journal of Application or Innovation in Engineering and Management,November 2014.

[2] Imran Shoukat Kazi,Zakir Mujeeb Shaikh,"Data alteration detection: A PC to USB pen drive perspective", International Journal of Application or Innovation in Engineering and Management,November 2014.

[3] Xiaokui Shu, Danfeng (Daphne) Yao,"Privacy preserving detection of sensitive data exposure",IEEE transaction on information forensics and security.