International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
376 – 382

# Enabling Public Audit Ability and Data Dynamics for Storage Security in Data Mining

[1] N. Nandhini
M. Phil Scholar,
Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

[2]Mrs. K. K. Kavitha
M.C.A., M.Phil., SET., (Ph.D).,
Vice Principal,
Head of the Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

**Abstract:** Data mining has been envisioned as the next-generation architecture of IT Enterprise. It moves the application software and databases to the centralized large data centers, where the management of the data and services may not be fully trustworthy. This unique paradigm brings about many new security challenges, which have not been well understood. This work studies the problem of ensuring the integrity of data storage in Data mining. In particular, we consider the task of allowing a third party auditor (TPA), on behalf of the cloud client, to verify the integrity of the dynamic data stored in the cloud. The introduction of TPA eliminates the involvement of the client through the auditing of whether his data stored in the cloud are indeed intact, which can be important in achieving economies of scale for Data mining. The support for data dynamics via the most general forms of data operation, such as block modification, insertion, and deletion, is also a significant step toward practicality, since services in Data mining are not limited to archive or backup data only.

*****

## I. INTRODUCTION

Data mining is a technique that provides specific information that can detect weaknesses in controls. Furthermore, an objective of data mining techniques is to uncover patterns indicating a broken process and/or develop predictive patterns in business information. The first objective is for the auditors to know the purpose of each data element, including how collective data patterns play a role in business decision making. Typically, there may be hundreds or thousands of data elements or variations that require a great deal of auditors' time in developing an understanding through a partnership with the business owner. The new data storage paradigm in "Data Mining" brings about many challenging design issues which have profound influence on the security and performance of the overall system.

One of the biggest concerns with cloud data storage is that of data integrity verification at untrusted servers. For example, the storage service provider, which experiences Byzantine failures occasionally, may decide to hide the data errors from the clients for the benefit of their own. What is more serious is that for saving money and storage space the service provider might neglect to keep or deliberately delete rarely accessed data files which belong to an ordinary client. Consider the large size of the electronic data and the client's constrained capability, the core of the problem can be generalized as how can the client find an efficient way to perform periodical integrity verifications without the local copy of data files.

## FINANCIAL BENEFITS OF USING DATA MINING TECHNIQUES

Depending upon the organization, there are numerous methods that can be used to reduce the cost of external and internal audits. There are significant benefits to all parties impacted by audit. For example, to reduce external audit fees, the IT internal auditor may use data mining to validate interface software that performs data transfers between systems. The successful comparison of data extraction from each system used in data transfer can validate balancing routines that can occur between systems. Using data mining techniques is especially important when validating data transfers between noncore systems, which are created internally, and an enterprise planning (ERP) system (e.g., SAP) used to record financial statement entries. In addition, data mining can validate the data transfer between the ERP (e.g., Lawson, Oracle) and a financial statement reporting package (e.g., ESSBASE), which is essential to financial statement integrity. At the request of management, data mining can be used to validate a known control such as a preventive and detective duplicate payment control within the accounts payable system (disbursement process). If properly established, the use of data mining appears to be limitless.

## INTRUSION DETECTION MODELS IN DATA MINING

As network-based computer systems play increasingly vital roles in modern society, they have become the target of this enemies and criminals. Therefore, we need

 ISSN: 2454-4248
376 – 382

to find the best ways possible to protect this systems. Intrusion prevention techniques, such as user authentication (e.g. using passwords or biometrics), are not sufficient because as systems become ever more complex, there are always system design flaws and programming errors that can lead to security holes (Bellovin 1989). Intrusion detection is therefore needed as an-other wall to protect computer systems.

## AUDITING TO FIND PROBLEMS IN DATA WAREHOUSES AND DATA MINING

Throughout this lives, we search for things. As children, games such as "hide and seek" or "Where's Waldo?" taught us search-and-find skills that we now use to strategically find this reading glasses or car keys. Those of us who do need reading glasses might also remember looking for the hidden pictures in Highlights magazine, by scanning the pictures back and forth, or in a haphazard way. As an auditor searching for fraud in today's sophisticated business systems, we must take those search-and-find skills to the next level. Fraud auditing is a proactive approach to detecting fraud. There are two key components to fraud auditing: 1.) using a fraud data mining plan, 2.) using fraud audit procedures. Both work closely together in that if the sample does not include a fraudulent transaction, the audit procedure cannot reveal the fraudulent transaction.

### Fraud Data Mining Methodology

Fraud data mining methodology is a structured step-by-step approach to identifying transactions consistent with a fraud scenario, as described through the fraud data profile.

### Identify the inherent fraud scheme

The first step is to establish the scope of the fraud audit. Each business system has five to seven inherent fraud schemes. The audit plan should identify which inherent fraud schemes are within the fraud audit scope. This article focuses on the fictitious vendor and the false billing scheme to illustrate the methodology. A false billing is paying for goods or services not provided; and the vendor is serving as a front company.

### Build the fraud scenario

The fraud scenarios are how an inherent fraud scheme would occur in a specific company. The auditor must consider the variations of the fraud scheme based on the variations of the entity, opportunity and the transactional consideration. Data mining must be driven by the fraud scenario versus the data mining routine. A good example is when the auditor matches the vendor master file to the employee master file. The purpose of the match was to identify fictitious vendors. However the only fictitious vendors identified are those vendors where the perpetrator was obtuse enough to use their home address. The design of the routine excludes all fictitious vendors with a concealed address.

### Obtain the data

In a sense, the concept sounds easy. However, one of the greatest impediments to fraud data mining is the identification and extraction of the data from the IT environment. While IDEA® – Data Analysis Software may be used to convert the data format, you must also consider storage capacity, table identification, data location and IT cooperation to build an effective data mining environment.

### Identify and link the data to the fraud scenario

One of the most critical stages of the data mining plan is to understand the available data and how to use the data to identify fraud scenarios, often referred to as the data mapping phase of the plan. Data mapping is the process of starting with each field in the database, understanding how the data correlates to the fraud scenario and how to search the data for indicators that link to the scheme. In essence, data mapping is the process of drawing a picture of a fraud scenario with data. Auditors should focus on both master file data and the transactional data associated with the business system.

### Developing the data interrogation procedures

The data interrogation plan starts with a fraud scenario. The second step is to determine the extent of data interpretation within the audit process to search for fraudulent transactions. In other words, will the data interrogation look for a fraud scenario with a low sophistication concealment strategy or high level of sophistication? This plan focuses on developing data interrogation based on the following concepts:

### Pattern and frequency

The analysis creates statistical reports by vendor, customer,employee, and transaction type in an attempt to identify an anomaly within the data. You can create these reports using IDEA by utilizing the "Summarize," "Sort" and "Data Extraction" features. Circumvention strategies - The analysis searches for transactions that exhibit a pattern or frequency which suggests someone was processing transactions below the control threshold.

### Duplicate analysis

The search routine searches for duplicate information within the data file, or external to the data file, that should not exist. Proceed with caution as this analysis often produces false positives. The duplicate search routines

 ISSN: 2454-4248
376 – 382

within a file or comparing two files will identify these transactions.

**Changes**

The analysis searches for changes in data that would be consistent with a fraud scenario. Changes such as new, delete, update, and void may all be signs of changes. These transactions maybe located via transaction codes or comparisons to files and two points in time.

Illogical - The analysis searches for transactions that do not fit the normal frequency or pattern that would be expected in the data file.

**Normalize the data**

The intent of this step is to shrink the population through use of the exclusion and inclusion theory. The exclusion theory is intended to create data files that have a high degree of commonality. In this way, an anomaly becomes more obvious. In the inclusion theory, the search routines are designed to search for data characteristics or red flags consistent with the fraud scenario.

**AUDITING MANAGEMENTTECHNIQUE**

Audit management staff members are constantly challenged to cut time in completing testing. They evaluate automated controls constantly by reviewing system options, edit logs, etc. They ask themselves: Are information technology (IT) auditors getting the most out of the available technology that can enable financial/operational auditors to effectively perform their duties to detect inefficient and ineffective processes including identifying fraud, waste and abuse of company. The answer can be at management's fingertips if it uses well-known data mining techniques, which are part of continuous auditing. The automated tools available today, when compared with 20 years ago, are well beyond sorting techniques. These tools are capable of analyzing terabytes of information and searching for patterns that may not be identified easily by manual means. In addition, over the past five years, numerous articles and large consulting practices have been created to assist companies in understanding their data, so they can make the most use of data mining.

**Types of Cloud Computing**

**Public Cloud**

An IT capability as a service that providers offer to consumers via the public Internet.

**Private Cloud**

An IT capability as a service that providers offer to a select group of customers.

**Internal Cloud**

An IT capability as a service that an IT organization to its own business (subset of private cloud).

**External Cloud**

An IT capability as a service offered to a business that is not hosted by its own IT organization.

**Hybrid Cloud**

IT capabilities that are spread between internal and external clouds While there are many roads to cloud computing from existing client-server infrastructures, there are at least three major Fig. 1: Cloud Computing Architecture In this thesis, we address the security issue from information assurance and security point of view. That is, we take holistic view of securing cloud computing by using the third party auditor (TPA), vehicle. Earlier works performs auditing only for static data. We enhance the system with dynamic operations on data locks i.e.) data update, append and delete.

## II. LITERATURE SURVEY

**PRIVACY-PRESERVING PUBLIC AUDITING FOR DATA STORAGESECURITY IN CLOUD COMPUTING**

Cloud computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing. By data, users can be relieved from the burden of local data storage and maintenance. Thus, enabling public auditability for cloud data storage security is of critical importance so that users can resort to an external audit party to check the integrity of data when needed. To securely introduce an effective third party auditor (TPA), the following two fundamental requirements have to be met: TPA should be able to efficiently audit the cloud data storage without demanding the local copy of data, and introduce no additional on-line burden to the cloud user. Specifically, this contribution in this work can be summarized as the following three aspects:

- Motivate the public auditing system of data storage security in Cloud Computing and provide a privacy-preserving auditing protocol, i.e., this scheme supports an external auditor to audit user's data in the cloud without learning knowledge on the data content.
- This scheme is the first to support scalable and efficient public auditing in the Cloud Computing. In particular, this scheme achieves batch auditing where multiple delegated auditing tasks from

different users can be performed simultaneously by the TPA.

- Prove the security and justify the performance of this proposed schemes through concrete experiments and comparisons with the state-of-the-art.

## DYNAMIC PROVABLE DATA POSSESSION

As storage services and sharing networks have become popular, the problem of efficiently proving the integrity of data stored at untrusted servers has received increased attention. In the provable data possession (PDP) model, the client preprocesses the data and then sends it to an untrusted server for storage, while keeping a small amount of meta-data.

The client later asks the server to prove that the stored data has not been tampered with or deleted (without downloading the actual data). However, the original PDP scheme applies only to static (or append-only) files. Present a definitional framework and efficient constructions for dynamic provable data possession (DPDP), which extends the PDP model to support provable updates to stored data. Use a new version of authenticated dictionaries based on rank information.

## STATEMENT OF PROBLEM

### System Model

Three different network entities can be identified as follows:

### Client

Anentity, which has large data files to be stored in the cloud and relies on thecloud for data maintenance and computation, can be either individual consumersor organizations; Cloud Storage Server (CSS): an entity, which is managed byCloud Service Provider (CSP), has significant storage space and computationto maintain clients' data; Third Party Auditor (TPA): a TPA, whichhas expertise and capabilities that clients do not have, is trusted to assess andexpose risk of cloud storage services on behalf of the clients upon request.In the cloud paradigm, by putting the large data files on the remote servers,the clients can be relieved of the burden of storage and computation. As clientsno longer possess their data locally, it is of critical importance for the clientsto ensure that their data are being correctly stored and maintained. That is,clients should be equipped with certain security means so that they can periodically verify the correctness of the remote data even without the existenceof local copies. In case that clients do not necessarily have the time, feasibilityor to monitor their data, they can delegate the monitoring task to atrusted TPA.

## III. RESEARCH METHODOLOGY

### DRAWBACKS OF EXISTING METHOD

There was lot of existing techniques presented for the public audit ability and data dynamics. In all these works, great efforts are made to design solutions that meet various requirements: high scheme efficiency, stateless verification, unbounded use of queries and retrievability of data, etc. Considering the role of the verifier in the model, all the schemes presented before fall into two categories:

- private audit ability and
- Public audit ability..

## IV. PROPOSED METHOD

Consider a cloud storage system in which there are a client and an untrusted server. The client stores their data in the server without keeping a local copy. Hence, it is of critical importance that the client should be able to verify the integrity of the data stored in the remote untrusted server. If the server modifies any part of the client's data, the client should be able to detect it; furthermore, *any* third party verifier should also be able to detect it. In case a third party verifier verifies the integrity of the client's data, the data should be kept private against the third party verifier.

### Advantages

Proposed system has the following main contributions:

- Remote data integrity checking protocol for cloud storage. The proposed system inherits the support of data dynamics, and supports public verifiability and privacy against third-party verifiers, while at the same time it doesn't need to use a third-party auditor.
- Security analysis of the proposed system, which shows that it is secure against the untrusted server and private against third party verifiers.

The proposed system has three parties involved. They are namely clients or data owners, cloud service providers and third party auditors.The cloud server provider maintains required storage space for outsourced data. The clients are responsible to store and retrieve data as and when required while the third party auditor is responsible to verify the integrity of data which is being flown between data owner and service provider. The third party auditor is a trusted entity that is responsible to audit data being flown for verification of integrity.

### The audit process

The audit process consists of checking the established control system. The general objectives --of auditing [ISACF, 1998b] are:

- ❖ To provide management with reasonable assurance that control objectives are being fulfilled.

 ISSN: 2454-4248
376 – 382

- To substantiate the risks caused by a control weakness.
- To advise management on corrective actions.

The methodology we propose in order to carry out this process is based on assessment of risks. Risk is defined as a chance for injury, damage or loss due to using data inappropriately [Curtis and Joshi, 1997]. Starting from the inherent risks that threaten the project, the control objectives that minimize those menaces will be settled. The objectives have been defined in the previous section, although they should not be interpreted in a strict way and should be adapted by the auditor to each project and organization. The goal of control objectives is to have information that satisfies the business requirements. This goal is achieved through seven criteria:

**Effectiveness**

The information has to be relevant and pertinent to the business process, as well as delivered in a timely, correct, consistent and usable manner.

**Efficiency**

The provision of information through the optimal use.

**Confidentiality**

The protection of sensitive information from unauthorized disclosure.

**Integrity**

Relates to the accuracy and completeness of information, as well as to its validity in accordance with business expectations.

**Availability**

The information has to be available when required by the business process. It also involves the safeguarding of necessary and associated capabilities.

**Compliance**

Deals with complying with laws, regulations, and contractual arrangements the business process is subjected to.

**Reliability**

The provision of appropriate information for management to operate the entity.Given a data warehouse project, a specific audit plan should be drawn to achieve the described audit goals. It is necessary to take into account both the drawn control objectives and the available for the audit process so that the plan is realistic and workable. The reached detail level in the audit process depends on the available to execute it.

**Identify and document the current control procedures**

The purpose of this stage is the comprehension of the business; including its requirements, risks and policies, the organization structure, the different roles and positions, ... Through this activity, the auditor will identify and document the detected control measures. This is achieved by both interviewing the adequate staff and reviewing the related documentation.

**Evaluate the controls and their operation**

This phase tries to settle whether the current measures of control are effective in order to get the established control objectives. The result will be the degree to which the control objectives are met. It is necessary to check if: · There are enough controls where it is necessary. Controls are adequate to assure the business against potential risks.Controls work in a continuous way along the project. The auditor has to take into account this time consideration, because the whole project is under revision and not only the delivered products.

## V. RESEARCH METHODOLOGY DRAWBACKS OF EXISTING METHOD

There was lot of existing techniques presented for the public audit ability and data dynamics. In all these works, great efforts are made to design solutions that meet various requirements: high scheme efficiency, stateless verification, unbounded use of queries and retrievability of data, etc. Considering the role of the verifier in the model, all the schemes presented before fall into two categories:

- private audit ability and
- Public audit ability.

## VI. PROPOSED METHOD

Consider a cloud storage system in which there are a client and an untrusted server. The client stores their data in the server without keeping a local copy. Hence, it is of critical importance that the client should be able to verify the integrity of the data stored in the remote untrusted server. If the server modifies any part of the client's data, the client should be able to detect it; furthermore, *any* third party verifier should also be able to detect it. In case a third party verifier verifies the integrity of the client's data, the data should be kept private against the third party verifier.

**Advantages**

Proposed system has the following main contributions:

- Remote data integrity checking protocol for cloud storage. The proposed system inherits

 ISSN: 2454-4248
376 – 382

the support of data dynamics, and supports public verifiability and privacy against third-party verifiers, while at the same time it doesn't need to use a third-party auditor.

- Security analysis of the proposed system, which shows that it is secure against the untrusted server and private against third party verifiers.

**The audit process**

The audit process consists of checking the established control system. The general objectives --of auditing [ISACF, 1998b] are:

- ❖ To provide management with reasonable assurance that control objectives are being fulfilled.
- ❖ To substantiate the risks caused by a control weakness.
- ❖ To advise management on corrective actions.

The methodology we propose in order to carry out this process is based on assessment of risks. Risk is defined as a chance for injury, damage or loss due to using data inappropriately [Curtis and Joshi, 1997]. Starting from the inherent risks that threaten the project, the control objectives that minimize those menaces will be settled. The objectives have been defined in the previous section, although they should not be interpreted in a strict way and should be adapted by the auditor to each project and organization. The goal of control objectives is to have information that satisfies the business requirements. This goal is achieved through seven criteria:

**Effectiveness**

The information has to be relevant and pertinent to the business process, as well as delivered in a timely, correct, consistent and usable manner.

**Efficiency**

The provision of information through the optimal use.

**Confidentiality**

The protection of sensitive information from unauthorized disclosure.

**Integrity**

Relates to the accuracy and completeness of information, as well as to its validity in accordance with business expectations.

**Availability**

The information has to be available when required by the business process. It also involves the safeguarding of necessary and associated capabilities.

**Compliance**

Deals with complying with laws, regulations, and contractual arrangements the business process is subjected to.

**Reliability**

The provision of appropriate information for management to operate the entity.Given a data warehouse project, a specific audit plan should be drawn to achieve the described audit goals. It is necessary to take into account both the drawn control objectives and the available for the audit process so that the plan is realistic and workable. The reached detail level in the audit process depends on the available to execute it.

**Identify and document the current control procedures**

The purpose of this stage is the comprehension of the business; including its requirements, risks and policies, the organization structure, the different roles and positions, ... Through this activity, the auditor will identify and document the detected control measures. This is achieved by both interviewing the adequate staff and reviewing the related documentation.

## DATA DYNAMICS

Data dynamics means after clients store their data at the remote server, they can dynamically update their data at later times. At the block level, the main operations are block insertion, block modification and block deletion.

i. Block Insertion:
The Server can insert anything on the client's file.

ii. Block Deletion:
The Server can delete anything on the client's file.

iii. Block Modification:
The Server can modify anything on the client's file.

## PUBLIC VERIFIABILITY

Each and every time the secret key sent to the client's email and can perform the integrity checking operation. In this definition, it has two entities: a challenger that stands for either the client or any third party verifier, and an adversary that stands for the untrusted server. Client doesn't ask any secret key from third party.

## METADATA KEY GENERATION

Let the verifier V wishes to the store the file F. Let this file F consist of n file blocks. initially preprocess the file and create metadata to be appended to the file. Let each of the n data blocks have m bits in them. A typical data file F which the client wishes to store in the cloud.

 ISSN: 2454-4248
376 – 382

Each of the Meta data from the data blocks $m_i$is encrypted by using a RSA algorithm to give a new modified Meta data $M_i$. Without loss of generality Show this process. The encryption method can be improvised to provide still stronger protection for Client's data. All the Meta data bit blocks that are generated using the procedure are to be concatenated together. This concatenated Meta data should be appended to the file F before storing it at the cloud server. The file F along with the appended Meta data with the cloud.

## SYSTEM TESTING AND IMPLEMENTATION EXPERIMENTAL APPROACH

Experiment supports an External Auditor for processing the auditing mechanisms that happens between a dedicated Cloud Service Provider and cloud user. The TPA may concurrently handle multiple auditing upon different Cloud Service Providers request. In this Experiment we are using Simulation software which is based on process of real phenomenon with a set of mathematical formulae's. This software provides the simulated environment that is similar to real world environment. Simulation software is designed in a manner so that result should be close to real world.

## VII. CONCLUSION

To ensure data storage security, it is critical to enable a TPA to evaluate the service quality from an objective and independent perspective. Public audit ability also allows clients to delegate the integrity verification tasks to TPA while they themselves can be unreliable or not be able to commit necessary computation performing continuous verifications. Another major concern is how to construct verification protocols that can accommodate dynamic data files. In this dissertation, we explored the problem of providing simultaneous public auditability and data dynamics for remote data integrity too. This construction is deliberately designed to meet these two important goals while efficiency being kept closely in mind. To achieve efficient data dynamics, we improve the existing proof of storage models by manipulating the classic Merkle Hash Tree construction for block tag authentication.

## FUTURE WORK

The client file has been modified to clients does not show what modification is done in client file by server, if the user need to know the modification only way to download the corresponding file. In future will show what modification is done in the client file by server to the client. The user can view their file details such as upload files, download files. Modification files can viewedthrough accessing with the help of mobile.

## REFERENCES

[1] BERRY, M.J.A. and GORDON, L. (1997). "Data Mining techniques". John Wiley & Sons, New York.

[2] CHAUDHURI, S. and DAYAL, U. (1997) "An Overview of Data Warehousing and OLAPTechnology". ACM SIGMOD Record 26 (1), pp. 65-74.

[3] CURTIS, MARY and JOSHI, KAILASH (1997) "Internal Control Issues for Data Warehousing". IS Audit & Control Jthisnal, Volume IV.

[4] DARUWALA, A., GOH, C., HOFMEISTER, S., MADNICK, S. and SIEGEL, M. (1995)"The Context Interchange Network Prototype".

[5] Proceedings of the Sixth IFIP TC-2 Conference on Data Semantic (DS-6).

[6] HALEY, B., and WATSON, H. (1998) "Managerial Considerations". Comm. of the ACM.September, pp 32-37.

[7] "Security in the Data Warehouse/Internet Environment". IS Audit & Control Jthisnal, vol. IV, pp. 8-11.

[8] Juels and B.S. Kaliski Jr., "Pors: Proofs of Retrievability for Large Files," Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07), pp. 584-597, 2007.

[9] H. Shacham and B. Waters, "Compact Proofs of Retrievability," Proc. 14th Int'l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT '08), pp. 90-107, 2008.

[10] M.A. Shah, R. Swaminathan, and M. Baker, "Privacy-Preserving Audit and Extraction of Digital Contents," Report 2008/186, Cryptology ePrint Archive, 2008.