IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

ENHANCED DATA PRIVACY USING VERTICAL FRAGMENTATION AND DATA ANONYMIZATION TECHNIQUES

[1]G.Kasthuri,[2]Padmapriya Arumugam
[1]M.phil scholar,[2]Professor
[1]Computer science,
[1]Alagappa University,Karaikudi..

ABSTRACT
The use of online banking has significantly increased because of the rapid progress of electronic commerce technologies. Virtual and physical cards are used in online banking to make purchases of goods and services Virtual cards are used for online transactions and actual cards being used for offline transactions. Cashless shopping is made possible through net banking, which is quickly becoming the most common method of payment for both regular and online purchases. It will be the most practical way to do internet business, pay bills, etc. Attackers simply require a small amount of information to conduct fraudulent transactions in the online payment mode (secure code, card number, expiration date etc.). Most transactions in this buying technique will take place over the phone or the Internet. A fraudster only needs to be aware of the card information to make these types of purchases fraudulently. Much of the time, the legitimate cardholder is unaware that his card information has been viewed or stolen. Consequently, there are also rising dangers of fraud transactions using banking information. Fraudulent transactions will be identified in the current cyber security system after the transaction has been completed. It is challenging to identify fraud because the issuing agencies will not allow losses. Therefore, we can use a vertical level server architecture in this project to divide the intermediary gateway and boost security. Transaction details are divided and kept in primary and secondary servers as sensitive attributes. In order to defeat conventional cryptography techniques, you should also design a data suppression scheme that transforms string and number characters into special symbols.

INTRODUCTION

Big data refers to data that are so enormous or intricate that conventional data processing software is unable to handle them. The advanced data analytics techniques are used to extract value from data, such as predictive analytics, user behavior analytics etc. Large data sets are a common source of frustration for researchers, corporate leaders, doctors, advertisers, and government officials working in fields like Internet search, finance, urban informatics, and business informatics.

Datasets are increasing now-a-days because of inexpensive and widespread information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio frequency identification (RFID) readers, and wireless sensor networks are used to collect it. Since the 1980s, the world's technical per-capita capacity for information storage has nearly quadrupled every 40 months; as of 2012, 2.5 Exabytes (2.51018) of data are generated each day. Determining who should be in charge of big data efforts that have an impact on the entire company is a concern for major businesses. Big data is frequently challenging to handle for relational database management systems and desktop statistics and visualization software. Massively parallel software running on tens, hundreds, or even thousands of servers may be necessary for the task. When faced with hundreds of gigabytes of data for the first time, several organizations may feel the need to reevaluate their data management alternatives.

I.RELATED WORK

Nusrat Jabeen, et,.al,..[1] proposed an Enterprise decision support is the foundation of the multidisciplinary field of data mining technologies. The study of small datasets is not the exclusive purpose of data mining. In circumstances when the data is located in databases, data repositories, OLAP, or other information existing in the repository, it is regarded as the duty of uncovering intriguing and hidden patterns/data from big pieces of information. Information retrieval, fuzzy and rough set theory, knowledge representation, inductive logic programming, neural networks, database technology, statistics, machine learning, and other approaches are all combined in data mining.

Finding the common item set in a large database is the fundamental challenge in data mining. A wide range of application domains, including bioinformatics and online usage mining, find value in the mining of frequent item sets. To find the frequent item set, numerous diverse algorithms have been created. The most extensively utilized algorithms in association rule mining are the Apriori and FP-Growth algorithms. The Apriori algorithm is essentially a level-wise search algorithm or bottom-up approach algorithm. The Apriori property states that any subset that contains frequent item sets must also contain frequent item sets; for example, if AB is a frequent item set, then A and B must also be frequent item sets.

For the benefit of the general public, set-valued databases that contain web click log, customer transaction, and trajectory data can be analyzed by Dedigunawan, et,.al…[2]. The ability of private preferences and other sensitive information to be connected to a specific person is an unsolved issue with publishing databases to the general public. Therefore, before publishing the database, a step such as data modification to achieve data anonymity should be taken. One straight forward method of data protection is pseudonym. To create anonymous data, it changes specific characteristics like name or personal identification number into pseudonym. These methods have been shown to be insecure because an adversary can use identity linkage attacks by comparing certain qualities in the disclosed database with those in other publicly accessible data bases, like voter lists, and connecting the two sets of data.

Data alteration to maintain data utility while achieving maximum anonymity is known to be an NP-hard task. Due to its advantages for numerous applications, such as recommendation systems and marketing analyses, set-valued database publication has been receiving a lot of attention. However, exposing the original database directly carries a certain amount of danger since identity linkage attacks allow an unauthorized party to violate individual privacy by identifying relationships between people and groups of published items. Typically, an attack is carried out based on the background information that the attacker has learned from earlier research, and this adversary knowledge should be considered while anonymizing

the data. To stop the identity linkage attack, a number of data anonymization solutions have been developed. However, current data anonymization techniques drastically lower the utility or value of data after extensive database alteration, which makes data recipients mistrust their leased databases. Xinyu Xiong, et,.al...[3] Present with Several industries have amassed substantial amounts of data through numerous routes as a result of the explosive rise of data and the quick development of information technology. Data mining has rapidly advanced to extract meaningful knowledge from massive amounts of data for upper-layer applications. It has had a good effect on a variety of fields, including business and healthcare. Along with the many advantages of these developments, the vast amount of data also contains private information that, if not properly managed, could leak. For instance, GPS-enabled smart phone applications track users' locations and upload the information to their servers. Additionally, potential connections between diseases and various types of data are stored in medical records. Medical record data mining and user location data mining can yield valuable information, but they also run the risk of exposing user privacy. Knowledge mining is therefore strongly anticipated with confident privacy safeguards. This study examines ways to mine popular item sets for large data while maintaining privacy. A business (such an information consulting firm, for example) has a sizable dataset. The company wants to make the dataset available to the public so that anyone can mine frequently occurring item sets for collaboration or financial gain. However, the business is unable to directly disclose the original dataset due to privacy concerns. Because of this, processing the data requires privacy procedures, which is the main topic of this study.

Jinyan Wang,.. et,.al…[4] proposed a data protecting privacy publishing offers strategies and resources for disseminating practical data while protecting individual privacy. It has recently drawn a lot of interest from both academia and industry, and numerous ways have been put forth for various data publication circumstances. Transactional data, also known as set-valued data, is data where each record (or transaction) is made up of a group of objects chosen at random from a larger universe of elements. A data stream is a theoretically infinitely large dataset that is always growing because of the advent of big data, where data may enter continuously at high speed. For many data mining activities, such as frequent pattern mining for a variety of online applications, such as retail chain data analysis, network traffic analysis, and web-server log and click-stream analysis, publishing transactional data streams offers vast opportunity. Data mining over transactional data streams has recently gained popularity as a study topic. Although only explicit identifying information must be deleted, publishing raw transactional data streams could jeopardize people's privacy. Many privacy-preserving methods for disseminating static transactional data have been suggested. Data streams cannot be directly anonymized using static data anonymization techniques due to the features of data streams, which must be handled quickly. The privacy issue with releasing transactional data streams based on a sliding window is the first thing we look at in this study.

H Wang,. et.al…[5] present a clients that use public cloud computing store their enormous amounts of data on distant public cloud servers. The stored data includes security risks in terms of confidentiality, integrity, and service availability because it is not under the clients' control. To persuade cloud clients that their data is preserved intact, remote data integrity verification might be employed as a primitive. In some unique circumstances, the data owner may be prohibited from accessing the public cloud server; in these circumstances, the data owner will choose a third party, such as a proxy, to handle the processing and uploading of the data. On the other hand, a capacity-limited end device requires a remote data integrity checking mechanism that is effective. Therefore, we will examine the ID-PUIC protocol based on identity-based public cryptography and proxy public key cryptography. In a public cloud environment, most clients upload their data to PCS and use the Internet to verify the accuracy of their remote data. Some practical issues will arise when the client is a single manager. The manager will be removed by the police if they suspect him of being complicit in the business fraud.

II.EXISTING METHODOLOGY

A credit network facilitates payments between any two agents and simulates trust between them in a distributed context. Credit networks are the foundation for many Sybil-tolerant social networks, spam-resistant communication protocols, and payment systems due to their flexible architecture and durability against infiltration. However, current systems reveal the existence and number of payment transactions as well as the agents' relationships of trust, which are both regarded as sensitive information in social and financial contexts. This brings up a difficult privacy issue that has largely gone unaddressed in the research on credit networks up to this point. Because sensitive data is increasingly often kept on Internet-connected computers, guidelines for protecting privacy have lately been developed. Additionally, a lot of jobs that were formerly completed by hand are now completed by computers, necessitating the need for information assurance (IA) and security. Protecting your privacy is crucial in the fight against identity theft. Businesses also require security to safeguard their proprietary information and trade secrets. One of the main terrorist risks facing our country today is cyber terrorism. As we've already discussed, the enormous amount of information that is now accessible electronically and online exacerbates this issue. Homomorphism encryption is a type of encryption that enables certain computations to be made on cipher text and provide an encrypted result that, when decrypted, is identical to the outcome of operations made on the plaintext. For instance, without either of them being able to determine the value of the individual numbers, one person could sum two encrypted numbers, and another person could subsequently decrypt the result.

LIMITATION

- Unauthorized person can view the details easily. So, security was less.
- Maintain the details in single server.
- Need large amount of storage space for store the encrypted data.
- Easily hack the details.

III.PROPOSED METHODOLOGY

E-commerce and online payment transactions are growing daily because of communications technology. Financial frauds related to these transactions are likewise getting worse, costing billions of dollars annually throughout the globe. Additionally, a variety of incentives, such as cash back, reward points, interest-free credit, offers for discounts at retailers, and so forth, attract customers to make purchases with credit cards rather than cash. The main issue facing today's e-commerce industry is that fraudulent transactions increasingly resemble legal ones, and basic pattern matching tools are ineffective at spotting fraud. The datasets can be clustered into many levels using the vertical clustering algorithm. The fragments are subsets of attributes (i.e., columns). A tuple identifier must connect rows of the fragments that are related to one another. The projection operations on the table correspond to a vertical fragmentation. The original data set can be created by combining the data from the fragments. To connect the columns from the fragments in a vertical fragmentation, the join operator is used on the tuple identifier; in a horizontal fragmentation, the union operator is used to connect the rows from the fragments. Additionally, use the K-Anonymity algorithm, which is a feature that certain anonymized data have.

Produce a release of the data with scientific assurances that the data's subjects cannot be re-identified while the data are still practically valuable, given person-specific field-structured data. If each person's information in a data release cannot be discriminated from the information of at least k-1 other people whose information also appears in the release, the data release is said to have the k-anonymity property. The numerous methods and software applications for producing anonymized data with k-anonymity protection have proven patentable.

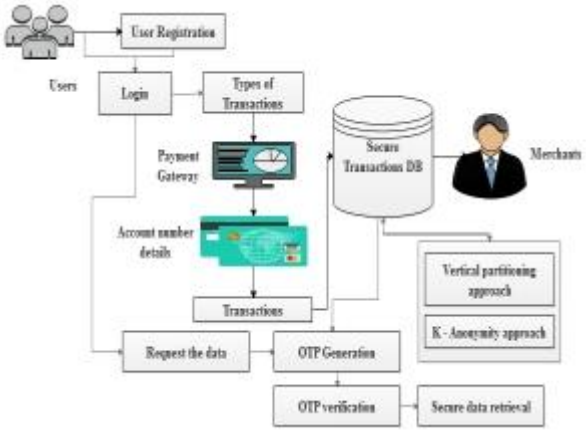

Fig 3.1: Architecture for Proposed Work

3.1BANK INTERFACE CREATION:

You can communicate with the bank in an efficient and automated manner thanks to the bank interface, an electronic information and payment system. You can combine the bank service with the company's accounting system using the bank interface. A web application called the Online Banking System makes sure that a registered user can enjoy doing their banking online. You can send money to other users through this online banking project's web application and keep a close eye on all your transactions. Additionally, we have expanded the security measures in our project for the online banking system. To construct a web-based application specifically for the financial sector, utilize this module. Only the user and bank administrator will utilize this program. The user can conduct online transactions by utilizing this programme.

3.1.1 TRANSACTION DETAILS:

All transaction data is collected using this module. Customers of banks and other financial institutions can carry out a variety of financial transactions online using the transaction detail, an electronic payment system. In contrast to branch banking, which was the conventional method through which customers received banking services, the online banking system will typically connect to or be a component of the main banking system run by a bank. This web application allows for any transactions, including money transfers and withdrawals. Automatic updating of the amount occurs in the savings account. A single gateway receives updates on all transaction data. The gateway oversees securely transmitting the funds to the designated merchant.

3.1.2 VERTICAL PARTITIONING APPROACH:

By enhancing performance, management, and availability, partitioning can be very helpful for a range of applications. Partitioning frequently enhances the performance of queries or maintenance procedures. Partitioning can also make routine administrative duties simpler. Partitioning also enables database administrators and designers to tackle some challenging issues brought on by cutting-edge applications. Building multi-terabyte systems or systems with extremely high availability needs requires the use of partitioning as a major technique. A database can be partitioned to enhance speed and make maintenance easier. To speed up searches that only access a portion of the data, huge databases can be divided into smaller, independent tables. A table that uses vertical partitioning is split up into smaller sub-tables. We can split the properties in this module into various servers, such as personal information, account information, and transaction information.

3.1.3 DATA SUPPRESSION:

K-Anonymity can be used in this module to protect data privacy. Certain anonymized data have the quality of K-anonymity. For some value of k, there are two typical techniques for establishing k-anonymity. Suppression: In this method, a symbol denoting an asterisk (*) is used in place of some attribute values. A column's values may all, some, or neither be substituted by a '*'. Generalization: With this approach, a wider category is used in place of the individual values of characteristics. For one of the re-identification situations, k-Anonymity over-anonymizes data sets on a regular basis, and this over-anonymization is particularly obvious for small sample fractions. Over-anonymization distorts the data in an excessive amount (high information loss), which reduces the usefulness of the data for further study. In comparison to baseline k-anonymity, we discovered that a hypothesis testing strategy offered the best control over the risk of re-identification and minimized the degree of information loss. Every cell could, of course, be replaced with a * to ensure k-anonymity, but it would make the database useless. The quantity of *s added is the cost of a K-Anonymous solution to a database. The fewest amounts of cells are suppressed in a minimum cost k-anonymity solution in order to ensure k-anonymity.

4.1.4 AUTHORIZED ACCESS:

We can create authorized access to bank clients with this module. OTP security allows the user to log in and view the transfer data. OTPs can be sent as SMS alerts and become visible after a set amount of time. Users can securely view their personal information. On a computer system or other digital device, a one-time password (OTP) is a password that is only valid for one login session or transaction. Traditional (static) password-based authentication has a variety of drawbacks that OTPs do not, with their biggest benefit being that OTPs are not susceptible to replay attacks like static passwords are. This ensures that an OTP that has previously been used to log into a service or complete a transaction cannot be abused by a future hacker because it will no longer be valid.

4.1.5 K-ANONYMITY ALGORITHM:

The observed data cannot be linked to fewer than k responders because of k-anonymity. Finding a quasi-identifier, or the collection of attributes in a dataset that may be linked with outside data to re-identify the data owner, is essential to establishing k-anonymity. By making a duplicate copy of a database and utilizing altering techniques like character shuffles, encryption, term or character substitution, data can be made anonymous. It is challenging to identify or do reverse engineering when, for instance, a value character is changed to a symbol like "*" or "x".

IV. RESULTS AND DISCUSSION

The effectiveness of the suggested system is demonstrated by the experimental results. Utilizing ASP.NET as the front end and SQL Server as the back end, enhanced data privacy via vertical fragmentation and data anonymization techniques are implemented here. This will help to increase the security of files.

PERFORMANCE EVALUATION

Evaluate the performance of the system in term of Anonymization time and also memory consumption.

MEMORY CONSUMPTION

Memory Consumption = (Original data size – Anonymized data set size).

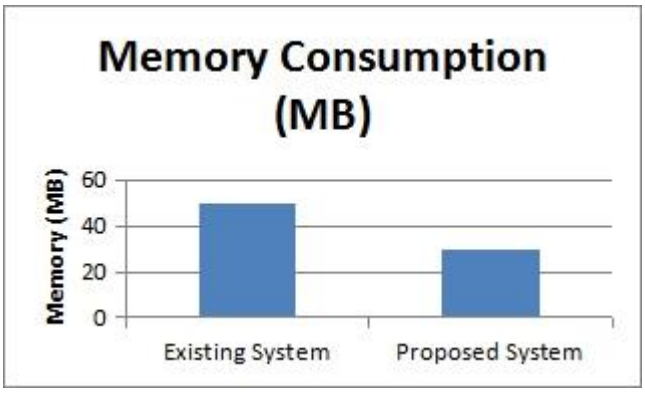

Fig 2. Memory Consumption

From the above graph, proposed system preserves the memory less than the existing encryption schemes. And also time can be reduced at the time of converting data into anonymized data and illustrated in fig 2.

V.CONCLUSION

The protection of personally identifiable information is the main objective of data privacy. The vertical partitioning strategy and the K-Anonymity approach are both used in this paper. A privacy-preserving technique for restricting the disclosure of personal data in data mining is K-Anonymity. A database table is often anonymized by generalizing the table entries, which results in the loss of important data. This drives the hunt for Anonymization algorithms that accomplish the necessary level of Anonymization with the least amount of data loss.

ACKNOWLEDGEMENT:

The authors would like to acknowledge department of science and technology,New Delhi for the financial support in general and infrastructure facilities sponsored under PURSE 2 nd phase programme (order no.SR/PURSE phase2/38(G)dated:21.02.2017)

REFERENCES

[1] Jabeen, T. Nusrat, M. Chidambaram, and G. Suseendran. "Security and privacy concerned association rule mining technique for the accurate frequent pattern identification." International Journal of Engineering & Technology 7.1.1 (2018): 19-24.

[2] Gunawan, Dedi, and Masahiro Mambo. "Set-valued Data Anonymization Maintaining Data Utility and Data Property." Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication.ACM, 2018.

[3] Xiong, Xinyu, et al. "Frequent Itemsets Mining with Differential Privacy over Large-scale Data." IEEE Access (2018).

[4] Wang, Jinyan, Chaoji Deng, and Xianxian Li. "Two Privacy-Preserving Approaches for Publishing Transactional Data Streams." IEEE Access 6 (2018): 23648-23658.

[5] H. Wang, D. He, and S. Tang, "Identity-based proxy-oriented data uploading and remote data integrity checking in public cloud," IEEE Trans. Inf. Foren.Secur., vol. 11, no. 6, pp. 1165–1176, 2016.

[6] K. Liang, X. Huang, F. Guo, and J. K. Liu, "Privacy-preserving and regular language search over encrypted cloud data," IEEE Trans. Inf. Foren.Secur., vol. 11, no. 10, pp. 2365–2376, 2016.

[7] D. Wang and P. Wang, "Two birds with one stone: Two-factor authentication with security beyond conventional bound," IEEE Trans. Depend. Secur.Comput., 2016.

[8] C. GENTRY, "Fully homomorphic encryption using ideal lattice," Proc. ACM STOC 2009, pp. 169–178.

[9] C. Gentry, A. Sahai, and B. Waters, "Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically faster, attribute-based," in Proc. CRYPTO 2013, pp. 75–92.

[10] C. Peikert, V. Vaikuntanathan, and B. Waters, "A framework for efficient and composable oblivious transfer," in Proc. CRYPTO2008, pp. 554–571.