EFFICIENT PRIVACY PRESERVATION TECHNIQUE FOR HEALTHCARE RECORDS USING BIG DATA

S.Sathya[1], Dr T.Sethukarasi[2]

Master of Engineering in Computer science and Engineering[1], Professor[2]

Department of Computer Science and Engineering

RMK Engineering College, Chennai

sathyasmecse@gmail.com, tsk.cse@rmkec.ac.in
Ph no : 9791419820

ABSTRACT

Big data, because it can mine new knowledge for economic growth and technical innovation, has recently received considerable attention, and many research efforts have been directed to big data processing due to its high volume, velocity, and variety (referred to as "3V") challenges. However, in addition to the 3V challenges, the flourishing of big data also hinges on fully understanding and managing newly arising security and privacy challenges. If data are not authentic, new mined knowledge will be unconvincing; while if privacy is not well addressed, people may be reluctant to share their data. Because security has been investigated as a new dimension, "veracity," in big data, in this article, we aim to exploit new challenges of big data in terms of privacy, and devote our attention toward efficient and privacy-preserving computing in the big data era. Specifically, we first formalize the general architecture of big data analytics, identify the corresponding privacy requirements, and introduce an efficient and privacy **Triple DES** as an example in response to data mining's efficiency and privacy requirements in the big data era.

Keywords- big data, privacy, triple DES.

I. INTRODUCTION

Big data, because it can mine new knowledge for economic growth and technical innovation, has recently received considerable attention, and many research efforts have been directed to big data processing due to its high volume, velocity, and variety (referred to as "3V") challenges.

There is no hard and fast rule about exactly what size a database needs to be in order for the data inside of it to be considered "big." Instead, what typically defines big data is the need for new techniques and tools in order to be able to process it. In order to use big data, you need programs which span multiple physical and/or virtual machines working together in concert in order to process all of the data in a reasonable span of time.

Getting programs on multiple machines to work together in an efficient way, so that each program knows which components of the data to process, and then being able to put the results from all of the machines together to make sense of a large pool of data takes special programming techniques. Since it is typically much faster for programs to access data stored locally instead of over a network, the distribution of data across a cluster and how those machines are networked together are also

important considerations which must be made when thinking about big data problems.

However, in addition to the 3Vchallenges, the flourishing of big data also hinges on fully understanding and managing newly arising security and privacy challenges. If data are not authentic, new mined knowledge will be unconvincing; while if privacy is not well addressed, people may be reluctant to share their data.

Because security has been investigated as a new dimension, "veracity," in big data, in this article, we aim to exploit new challenges of big data in terms of privacy, and devote our attention toward efficient and privacy-preserving computing in the big data era. Specifically, we first formalize the general architecture of big data analytics, identify the corresponding privacy requirements, and introduce an efficient and privacy **Triple DES** as an example in response to data mining's efficiency and privacy requirements in the big data era.

II. RELATED WORKS

2.1 Privacy-Preserving Cosine Similarity computing Protocol

The privacy-preserving cosine similarity computing (PCSC)[10]can efficiently calculate the cosine similarity of two vectors without disclosing the vectors to each other. The protocol describes that we can directly calculate the cosine similarity in an efficient way. When we consider inter big data processing the direct cosine similarity computation (DCSC)would disclose each other's privacy. Hence we can apply homomorphic encryption (HE), such as Pallier encryption (PE) to provide privacy but this requires time-consuming exponentiation operations. So, PCSC protocol for big data processing is used based on the lightweight multi-party random masking and polynomial aggregation techniques which does not require time consuming operations.

The issue with this protocol is that it has computational overheads with the increase in length of the vector.Compared to DCSC protocol. This protocol does not address the unique privacy which becomes another issue. But it is efficient towards time.

2.2 Cryptographic Approaches for Big-Data Analytics

There are many widespread techniques in cryptography which are used to provide the security for the data. Here we consider some of the cryptographic approaches to securing big data analytics in the cloud.Homomorphicencryption (HE), Verifiable Computation (VC), Multi-Party Computation (MPC) are the three cryptographic techniques which can be deployed on trusted, semi-trusted and untrusted clouds [19].Homomorphic Encryption (HE) allows functions to be computed on encrypted data without decrypting it first. Given only the encryption of a message, one can obtain an encryption of a function of that message by computing directly on the encryption. The cloud nodes are not trusted to protect confidentiality. Input holders encrypt data before it enters the cloud, and data receivers decrypt the data after it leaves the cloud. In Verifiable Computation cloud nodes are not trusted to protect integrity. The compute nodes provide proofs of correct computation, and the data receiver verifies the proof. The dashed line denotes physical isolation from outside networks. The secured Multi-Party Computation (MPC) is deployed in semi-trusted cloud. The input holders secret-share the data among the compute nodes who perform multi-party computation on the shares. The data receiver reconstructs the output.

Of the three MPC provides confidentiality, integrity and even authenticity. But MPC is much suited for semi-trusted cloud in the presence of honest parties.

2.3 Data Mining with Big Data

Big data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, big data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences.

This presents a HACE theorem that characterizes the features of the big data revolution, and proposes a big data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the big data revolution.

2.4 An Efficient and Privacy-Preserving Aggregation Scheme for Secure Smart Grid Communications

The concept of smart grid has emerged as a convergence of traditional power system engineering and information and communication technology. It is vital to the success of next generation of power grid, which is expected to be featuring reliable, efficient, flexible, clean, friendly, and secure characteristics.

EPPA uses a superincreasing sequence to structure multidimensional data and encrypt the structured data by the homomorphic Paillier cryptosystem technique. For data communications from user to smart grid operation center, data aggregation is performed directly on ciphertext at local gateways without decryption, and the aggregation result of the original data can be obtained at the operation center.

EPPA also adopts the batch verification technique to reduce authentication cost. Through extensive analysis, we demonstrate that EPPA resists various security threats and preserve user privacy, and has significantly less computation and communication overhead than existing competing approaches.

2.5 Toward Privacy-Assured and Searchable Cloud Data Storage Services

Cloud computing is envisioned as the next generation architecture of IT enterprises, providing convenient remote access to massively scalable data storage and application services. While this outsourced storage and computing paradigm can potentially bring great economical savings for data owners and users, its benefits may not be fully realized due to wide concerns of data owners that their private data may be involuntarily exposed or handled by cloud providers.

Although end-to-end encryption techniques have been proposed as promising solutions for secure cloud data storage, a primary challenge toward building a full-fledged cloud data service remains: how to effectively support flexible data utilization services such as search over the data in a privacy-preserving manner. In this project, identify the system requirements and challenges toward achieving privacy-assured searchable outsourced cloud data services, especially, how to design usable and practically efficient search schemes for encrypted cloud storage.

This present a general methodology for this using searchable encryption techniques, which allows encrypted data to be searched by users without leaking information about the data itself and users queries. In particular, discuss three desirable functionalities of usable search operations: supporting result ranking, similarity search, and search over structured data.

For each of them, we describe approaches to design efficient privacy-assured searchable encryption schemes, which are based on several recent symmetric-key encryption primitives. This analyze their advantages and limitations, and outline the future challenges that need to be solved to make such secure searchable cloud data service a reality.

2.6 A Secure and Privacy-Preserving Opportunistic Computing Framework for Mobile-Healthcare Emergency

With the pervasiveness of smart phones and the advance of wireless body sensor networks (BSNs), mobile Healthcare (m-Healthcare), which extends the operation of Healthcare provider into a pervasive environment for better health monitoring, has attracted considerable interest recently.

However, the flourish of m-Healthcare still faces many challenges including information security and privacy preservation. In this, we use a secure and privacy-preserving opportunistic computing framework, called SPOC, for m-healthcare emergency. With SPOC, smart phone resources including computing power and energy can be opportunistically gathered to process the computing-intensive personal health information (PHI) during m-Healthcare emergency with minimal privacy disclosure.

In specific, to leverage the PHI privacy disclosure and the high reliability of PHI process and transmission in m-Healthcare emergency, we introduce an efficient user-centric privacy access control in SPOC framework, which is based on an attribute-based access control and a new privacy-preserving scalar product computation (PPSPC) technique, and allows a medical user to decide who can participate in the opportunistic computing to assist in processing his overwhelming PHI data. Detailed security analysis shows that the proposed SPOC framework can efficiently achieve user-centric privacy access control in m-Healthcare emergency.

In addition, performance evaluations via extensive simulations demonstrate the SPOC's effectiveness in term of providing high-reliable-PHI process and transmission while minimizing the privacy disclosure during m-Healthcare emergency.

III. EXISTING SYSTEM

In existing system using cosing similarity check protocol for search over the encrypted data ,using this protocol ,we can take long time to search over encrypted data, because all the Stored document must decrypted before start comparison process ,all stored patient information must encrypted same public key so, lesssecure .

We can also run operations over encrypted data to protect individual privacy in big data analytics. However, as operations over encrypted data are usually complex and time-consuming, while big data is high-volume and needs us to mine new knowledge in a reasonable timeframe, running operations over encrypted data is inefficient in big data analystics.

Compared to privacy preserving aggregation and operations over encrypted data, deidentification can make data analytics and mining more effective and flexible. However, many real examples indicate that data which may look anonymous is actually not after deidentification; for example, only (5-digit zip code, birth date, gender) can uniquely identify 80 percent of the population in the United States. Therefore, to mitigate the threats from reidentification.

IV. PROPOSED SYSTEM

To overcome the existing , implement different encryption key for each stored document, further for fast way to performing

search process ,all documents searchable keyword are encrypted using same key , so native data maintain secrecy de-identification is a crucial tool in privacy protection, and can be migrated to privacy preserving big data analytics.

However, as an attacker can possibly get more external information assistance for de-identification in the big data era, we have to be aware that big data can also increase the risk of re-identification. As a result, de-identification is not sufficient for protecting big data privacy.

Research work on big data privacy should be directed toward efficient and privacy preserving computing algorithms in the big data era, and these algorithms should be efficiently implemented and output correct results while hiding raw individual data. In such a way, they can reduce the re-identification risk in big data analytics and mining

V. PERFORMANCE EVALUATION

Server Authentication:

In this module we design new client side page for actively interact with server ,for improve secure purpose we generate unique Authentication scheme for each and every user(patient).

Once Server Authentication process completed , we design patient info web port ,for get all necessary information about patient health report.

Each patient uploaded document must be stored in own encryption key ,that key info only patient can known.

Database Encryption's

Triple DES was designed to replace the original Data Encryption Standard (DES) algorithm, which hackers eventually learned to defeat with relative ease. At one time, Triple DES was the recommended standard and the most widely used symmetric algorithm in the industry.

Triple DES uses three individual keys with 56 bits each. The total key length adds up to 168 bits, but experts would argue that 112-bits in key strength is more like it.

Despite slowly being phased out, Triple DES still manages to make a dependable hardware encryption solution for financial services and other industries.

Improve De-identification

De-identification is a traditional technique for privacy-preserving data mining, where in order to protect individual privacy, data should be first sanitized with generalization (replacing quasi-identifers with less specific but semantically consistent values) and suppression (not releasing some values at all) before the release for data mining.

Compared to privacypreserving aggregation and operations over encrypted data, deidentification can make data analytics and mining more effective and flexible.
However, many real examples indicate that data which may look anonymous is actually not after deidentification; for example, only (5-digit zip code, birth date, gender) can uniquely identify 80 percent of the population in the united states.

Therefore, to mitigate the threats from reidentification, the concepts of *k* anonymity, *l*-diversity, and *t*closeness have been introduced to enhance traditional privacy-preserving data mining. Obviously, de-identification is a crucial tool in privacy protection, and can be migrated to privacy- preserving big data analytics.

However, as an attacker can possibly get more external information assistance for de-identification in the big data era, we have to be aware that big data can also increase the risk of re-identification. As a result, deidentification is not sufficient for protecting big data privacy.

However, compared with privacy-preserving aggregation and operations over encrypted data, de-identification is more feasible for privacy-preserving big data analysitcs if we can develop efficient and privacy-preserving algorithms to help mitigate the risk of re-identification.

VI. CONCLUSION

In this project, this investigated the privacy challenges in the big data era by first identifying big data privacy requirements and then discussing whether existing privacy-preserving techniques are sufficient for big data processing. This have also introduced an efficient and privacy-preserving cosine similarity computing protocol in response to the efficiency and privacy requirements of data mining in the big data era. Although have analyzed the privacy and efficiency challenges in general big data analytics to shed light on the privacy research in big data, significant research efforts should be further put into addressing unique privacy issues in some specific big data analytics.

VII. REFERENCES

[1] X. Wu et al., "Data Mining with Big Data," IEEE Trans. Knowledge Data Eng., vol. 26, no. 1, 2014, pp. 97–107.

[2] R. Lu et al., "EPPA: An Efficient and Privacy-Preserving Aggregation Sheme for Secure Smart Grid Communications," IEEE Trans. Parallel Distrib. Sys.,vol. 23, no. 9, 2012, pp. 1621–31.

[3] M. Li et al., "Toward Privacy-Assured and Searchable Cloud Data Storage Services," IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.

[4] R. Lu, X. Lin, and X. Shen, "SPOC: A Secure and Privacy-Preserving Opportunistic Computing Framework for Mobile-Healthcare Emergency," IEEE Trans. Parallel Distrib. Sys., vol. 24, no. 3, 2013, pp. 614–24.

[5] "Big Data at CSAIL," http://bigdata.csail.mit.edu/.

[6] IBM, "Big Data at the Speed of Business," http://www-01.ibm.com/software/data/bigdata/, 2012.

[7] P. Paillier, "Public-Key Cryptosystems based on Composite Degree Residuosity Classes," EUROCRYPT, 1999, pp. 223–38.

[8] Oracle, "Oracle Big Data for the Enterprise," http://www.oracle.com/caen/ technologies/big-data, 2012.

[9] A. Cavoukian and J. Jonas, "Privacy by Design in the Age of Big Data,"Office of the Information and Privacy Commissioner, 2012.

[10] S. Liu, "Exploring the Future of Computing," IT Professional, vol. 15, no.1, 2013, pp. 2–3.