IJCTA, 10(03), 2017, pp. 73-79

A Review on Cryptographic Approaches for Secured Processing of Big Data

Melbin J. Reena* and A. Shajin Nargunam**

Abstract : Big Data refers to large datasets and so it is not possible to store, manage and analyze it using commonly used software systems. The emergence of smart phones, social networks and online applications has led to the generation of massive amounts of structured, unstructured and semi structured data. Big data analytics has received sizeable attention since it offers a great opportunity to uncover potentials from heavy amounts of data. The large volume of data necessitates the use of cloud for analytics. Privacy and security in big data processing is the current need as the derived knowledge is used in several domains. This paper present a review about the cryptographic methods used to ensure secured big data processing.

Keywords: Big Data analytics; Privacy and Security; Cloud Storage; Encryption.

1. INTRODUCTION

Recent technological developments results in generation of large amount of data and the data is generated frequently [12]. Various domains like internet, health care, user-generated data, financial companies, social networks and supply chain systems are generating large amount of data. The term Big Data represents large volume of data and also possesses unique characteristics when compared with traditional transaction data. The four important features of big data are volume, velocity, variety, and value.

Usage of analysis algorithms running on powerful supporting platforms for identifying hidden knowledge from massive amounts of unstructured data is called big data analytics. The trending Big Data analytics helps in identifying unknown correlations or hidden patterns [12]. Big Data processing may be either streaming processing or batch processing. Data freshness is given more importance in Streaming processing paradigm. This method analyzes data without any delay after it arrives to derive results. Limited memory is used to store small block of recent data and is being analyzed. In the batch processing paradigm, data are first placed in memory and then analyzed. Nowadays, the most common batch processing model is MapReduce and is widely used [11].

The basic idea of MapReduce is splitting up of data into small units. These chunks are then processed in distributed and parallel manner for deriving intermediate results. All the derived intermediate results are then analyzed and aggregated to achieve final result. This model minimizes communication overhead of data transmission by scheduling computation resources close to the data location. Big data processing, whether streaming or batch, security and privacy in processing and deriving knowledge is inevitable.

The general framework for Big Data analytics involves three main Stages: Collection of Big data from multiple sources, Distributing data over cloud storage, and Analyzing data securely to derive facts. Data can be generated from various distributed sources. The amount of data generated has exploded in the past few years. Usually, the data generated are normally associated with a specific domain such as business, internet, research, etc. A data storage system consists of two parts: hardware infrastructure and data management. Hardware infrastructure refers to utilizing information and communications technology

* Research Scholar, Noorul Islam Centre for Higher Education. *E-mail : mjreena82@gmail.com*

** Director (Academics), Noorul Islam Centre for Higher Education. *E-mail : ashajins@yahoo.com*

resources for various tasks. Data management refers to the set of software deployed on top of hardware infrastructure to manage and query large scale data sets. Data processing phase refers basically to the process of data collection, data transmission, pre-processing and extracting useful information [15].

Data centers play an important role in modern information systems which always perform complex computations and retrieve large amount of datasets from data centers. In a distributed environment, an application may needs several datasets located in different data centers and therefore face some challenges such as data security, privacy preservation and authentication. Generally there are four kinds of conventional security mechanisms to protect data. The first scheme is related to the file-level data security, which can be implemented on the host. The second scheme mainly focused on database-level data security, which can be applied when the data stored in a database.

The third data security scheme for data centers is media level security technology, which is a new method involving static data encryption on storage equipment such as hard disks and tapes. The final scheme is application-level data encryption technology, which is an end-to-end encryption solution. It can ensure that only certain users get to access the data through a particular application. This scheme will be very costly because it must maintain many parameters and data structures [3].

Designing and deploying a big data analytics system is not a trivial or straightforward task. Big data is beyond the capability of current hardware and software platforms. The new hardware and software platforms in turn demand new infrastructure and models to address the wide range of challenges of big data. Data collection and management addresses massive amounts of heterogeneous and complex data. Data representation, Redundancy reduction and data compression, Data life-cycle management, and Data privacy and security are the challenges of big data [16].

2. BIG DATA AND CLOUD

A big data system providing analytics functions is complex and it has to deal with entire digital data life cycle, starting from data introduction to secure destruction of data after analytics [12]. Collecting big data from multiple sources, storing big data in distributed manner, and processing big data in secured way to derive results are the three important phases in Big Data analytics [4]. Privacy preservation is highly recommended in big data collection, big data storage and big data processing. Big Data could be structured, semi-structured, or unstructured, which adds more challenges when performing data storage and processing tasks.

Cloud-based platforms are playing an increasingly significant role in big data analytics and storage applications. Usage of cloud computing ideas in big data analytics provides advantages such as scalability, flexibility, agility, energy efficiency, and cost effectiveness [8]. Enhanced security and privacy mechanisms must be in place to cope up with the rising need of correlating patterns from big data and for maintaining large-scale cloud infrastructures. Because traditional security mechanisms are only suitable for securing small-scale data, they can’t satisfy the needs of big data. Also, the inherent vulnerabilities of a cloud based environment require significant focus on both privacy and security together with risk management procedures. The tremendous growth of cloud computing and cloud data stores is the major reason behind the emergence of big data. Cloud computing has significant benefits over traditional deployment models. It saves computing time and resources by making use of standardized technologies. Cloud Service Providers come in all shapes and sizes and offer many different products for big data.

Cloud computing provides enormous computing resources on demand. Charges acquired based on usage only. Performance and Capacity are the two categories of cloud storage challenges in big data analytics. High performance platforms are necessary for managing and analyzing highly unstructured big data. Also, the cloud storage service for data analytics must be highly available, highly durable, and scalable in size. Thus cloud services become inevitable in Big Data analytics due to the large volume and variety of big data. All domains ranging from healthcare to *e*-commerce generates massive amounts of data frequently [11].

The three major categories of cloud deployment models developed over time are private cloud, public cloud and hybrid cloud. Private clouds are meant for one organization and suitable for organizations where data sensitivity in mainly concerned. As private cloud does not share physical resources secured storage and usage of data is facilitated. Also, the accidental or malicious access through shared resources is avoided. The main drawback of private cloud is that it is not scalable. As big data deluge, the private cloud does not scale well to hold and process data. Also, Analytics in private cloud is hectic as it does not possess compatible analytics frameworks with public cloud. Data Sharing is not possible in private cloud, which prohibits users from sharing data even among their community [14].

Public clouds share physical resources for data transfers, storage, and processing. Hybrid cloud may hold the answer for complicated big data analytics. As hybrid cloud is the combination of private and public cloud, user's sensitive information can be stored in private cloud and other data can be stored in public cloud. This hybrid model is more suitable for big data and it offers benefits like improved performance, cost effectiveness and enhanced security for businesses involved in big data analytics.

The varying nature of Big Data requires Elasticity and Infrastructure Flexibility. Through the use of hybrid cloud, down time and performance degradation due to increased workload can be minimized. Hybrid cloud model helps to allocate new workloads to powerful machines or will add more machines when there is a need for additional resources.

3. SECURITY CHALLENGES

Despite big data could be effectively utilized for us to better understand the world and innovate in various aspects of human endeavors, the exploding amount of data has increased potential privacy breach [15]. To maintain integrity and security of big data stored in cloud servers, sharing user's sensitive data in cloud searching need to be avoided. Security mechanisms must be devised to protect user's sensitive data from migration between cloud servers. At many instances the owner of data does not have direct control over his shared data in cloud. Any third party can verify validity of data in the name of data auditing. So, it is a very big challenge to protect shared big data stored in distributed manner. As hybrid model suits big data processing, security and privacy in cross-cloud big data environment is highly complex. The integrity of data in cloud storage can be identified only by the process of data auditing. To maintain the reputation of the services and to avoid loss most of the cloud service providers are unwilling to notify about data errors to clients [2].

User's privacy may be breached under the following circumstances: Personal information when combined with external datasets may lead to the inference of new facts about the users. Those facts may be secretive and not supposed to be revealed to others. Personal information is sometimes collected and used to add value to business. The sensitive data are stored and processed in a location not secured properly and data leakage may occur during storage and processing phases [15].

Also, vulnerabilities in the cloud's underlying technologies permit intruders to abuse cloud computing services and allowing unauthorized sharing of data [7]. Through research various security schemes such as encryption, authentication, access control, firewalls, intrusion detection systems (IDSs), and data leak prevention systems (DLPSs) are devised to handle security issues [6]. In spite of all these security mechanisms big data security still remains as an issue because of the complex nature of cloud environment and lack of schemes to well protect data.

Secure sensitive data sharing involves four primary safety factors. First, the channel between data owner and cloud platform must be safe to transmit sensitive data. Second, there can be security threat in storing and processing sensitive data on the big data platform. Third, there are security issues in unauthorized usage of user's sensitive data on the cloud platform. Fourth, secure data destruction after analytics is essential [5]. The following section describes existing techniques for privacy-preserving big data analytics.

4. CRYPTOGRAPHIC SCHEMES FOR BIG DATA SECURITY

Identity-Based Cryptographic Schemes (IBCS)

IBE is an alternative to PKE which is proposed to simplify key management in a certificate-based public key infrastructure (PKI) by using human identities like email address or IP address as public keys. To preserve the anonymity of sender and receiver, the IBE scheme was proposed. Encryption scheme like IBE does not support the update of ciphertext receiver. There are some approaches to updating the ciphertext recipient. For instance, data owner can employ the decrypt then re-encrypt mode. However, if data are large as it is mostly the case when dealing with big data, the decryption and re-encryption can be very time consuming and costly because of computation overhead [15].

The two basic flavors of identity-based cryptographic schemes are identity-based encryption (IBE) and identity based signature (IBS). In an identity-based encryption scheme, a trusted party is designated as private key generator (PKG) and is responsible for generating secret master key and public parameter. The public parameter can be distributed to every party involved in the communication. To obtain private keys from PKG, users must authenticate themselves to PKG. The PKG will issue private keys to users associated with their identities. The key transmission channel between private key generator and data user must be well protected to prevent eavesdropping and hence avoiding unauthorized data access.

In identity-based signature scheme, the user submits identity and PKG computes the private key and by using it the signer can sign a message. Any party can verify the validity of the signature provided the message, the signer’s identity, and the signature. The usage of digital certificates is avoided in both IBE and IBS. In IBCS, implicit certification of each user within the system is needed to enforce security [8].

Thus, Identity-based cryptography schemes are slightly more costly than most efficient public key encryption schemes [8]. This method requires a centralized server PKG and channel between PKG and data users must be secure. A common issue in identity-based cryptography is key escrowing problem [8].

Identity-Based Proxy Re-Encryption (IB-PRE)

In Proxy re-encryption, the ciphertext produced under Alice’s public key can be re-encrypted by a designated proxy. The re-encrypted text can be decrypted by another party, say Bob’s private key [1]. In an identity-based proxy re-encryption scheme, a delegator allows a proxy to transform an encryption under Alice’s identity into one encrypted under Bob’s identity. The trusted proxy will utilize re-encryption keys to execute the conversion. It is not possible even for the proxy to gain knowledge of any information about the original text. Also, it is not possible to deduce the private keys of Alice and Bob from the re-encryption keys. A semi-trusted agent with a proxy key can re-encrypt ciphertext. In the authorization process, it is not possible for the proxy to acquire the original text or to deduce decryption key of either party involved in message transfer [8].

The functionalities and concepts of both IBE and proxy re-encryption is combined in identity-based proxy re-encryption without compromising the security [8]. But this scheme is prone to collusion attacks.

Attribute-based Encryption (ABE)

When there is a need for data owner to share information with a number of users, Attribute-based encryption (ABE) can be used in such applications. ABE can attain efficient one-to-many encryption and it proves to be a dominant encryption scheme to achieve both data security and fine-grained access control. ABE comes in two flavors called Key-Policy ABE (KP-ABE) and Ciphertext-Policy ABE (CP-ABE). The key security aspect of ABE is Collusion resistance. The problem with ABE is that to encrypt data the owner needs to use every authorized user’s public key, which results in increased computational complexity.

Key-Policy Attribute Based Encryption

KP-ABE technique associates a set of attributes with the ciphertext and the private key is linked to the access structure [10]. Sets of attributes are used in labeling Ciphertexts and private keys are coupled with monotonic access structures. The role of Access structures is to direct the user in selecting ciphertexts that can be decrypted. KP-ABE scheme is designed for one-to-many communications.

When compared with ABE scheme, the KP-ABE scheme can realize fine-grained access control and more flexible to control users. Key Policy Time specified Attribute based encryption schemes offers great benefits by facilitating user-defined time-specific authorization, fine-grained access control and secure self destruction of data. The problem with this method is high computational cost incurred by time consuming operations such as pairing and exponentiation.

The problem with KP-ABE scheme is the person who encrypts data cannot decide who can decrypt the encrypted data. This technique cannot be adapted to applications where the data owner hesitates to trust key issuer.

Ciphertext-policy Attribute Based Encryption

Unlike KP-ABE, CP-ABE associates the ciphertext with the access structure and the private key contains a set of attributes. In order to facilitate the user to decrypt the data, the set of attributes linked with the user’s private key must satisfy the access policy coupled with the ciphertext. CP-ABE has restrictions in maintaining and managing attributes of users and also specifying policies in encryption.

Timed-Release Encryption

In order to protect sensitive data, shared data can be self-destroyed after the user defined expiration time. An interesting encryption service is provided by Timed-release encryption (TRE) where an encryption key is connected with already set up release time, and a receiver can make and use the corresponding decryption key only at this time instance [10]. At the time of uploading data to cloud itself the data owner can mention the authorization period of data access. The authorization time interval starts from the introduction of data and ends at the expiration time set up by the data owner. The user can construct decryption key only before the expiration time and time instant must be within authorization period. As the shared data will be securely self-destructed after expiration time instance, the shared data can only be accessed by the user before expiration time. Thus, [10] presents full life cycle privacy protection for shared data in cloud computing.

Fully Homomorphic Encryption

Unlike other encryption techniques, in Homomorphic encryption the ciphertext is not decrypted throughout the process. Computations are carried out only on ciphertexts. The computations on encrypted text yield encrypted result. The interesting fact is that when the encrypted results are decrypted, the result matches the result of operations carried out on plain text. Privacy-preserving aggregation, which is based on homomorphic encryption, is a most accepted data collecting technique. [4] Proposes privacy-preserving cosine similarity computation based on homomorphic encryption for big data processing. The Fully Homomorphic Encryption (FHE) method allows an explicit algebraic computation based on ciphertext that yields a still encrypted result. Though the data is not decrypted throughout the entire process, retrieval and comparison of the encrypted data produce correct results [4].

Fully Homomorphic Encryption improves the efficiency of secure multiparty computation. As this technique never decrypts its input, it can be run by an untrusted party without revealing input and internal state. In spite of security, it should be noted that it is difficult to encrypt massive amounts of big data. The FHE scheme entails extremely extensive computation, and so it is very difficult to realize it with existing technology.

Key-policy Time Specified Attribute Based Encryption

The idea behind KP-TSABE is based on the fact that, in today’s practical cloud application scenarios every data item is coupled with number of attributes and also every attribute is linked with time factor. The data owner encrypts his data to share with users in the system. The access tree mechanism is used in which every user’s key is associated with an access tree and each leaf node is associated with a time constant. To enforce security and to avoid malicious data access, the ciphertext cannot be decrypted after expiration of time interval [10].

The computational cost seems to be expensive. KP-TSABE scheme provides advantage by supporting user-defined time-specific authorization, fine-grained access control and secured self-destruction of data.

Table 1
Comparison of Cryptographic Schemes

Parameters *Schemes*	*Fine-grained access control*	*Computational Overhead*	*Communication Overhead*	*Collusion Resistance*	*Flexibility*
IBCS [8]	Better	Moderate	Low	Yes	Good
IB-PRE [1]	Good	High	Average	Yes	Average
ABE [13]	Low	High	____	Average	Average
KP-ABE [10]	Low	High	Low	Good	Average
CP-ABE [5]	Better	Average	Low	Good	Average
TRE [10]	Good	High	High	____	Good
FHE [4]	Low	Complex	____	Average	inflexible
KP-TSABE [10]	Good	High	High	Yes	Flexible

5. CONCLUSION

This paper presents a survey about existing cryptographic solutions to protect big data and thus processing it in secured way. Encryption techniques are mainly used to protect big data. Big data are always enormous in size and so it is impossible to encrypt them as a whole. Most of the existing techniques incur high computation and communication cost. Full life cycle privacy protection is needed for maintaining integrity of big data but most security mechanisms applies only to a single phase. Thus the problem of formulating efficient security mechanism to provide full life cycle privacy and protection of big data with reduced computation and communication overhead is still open.

6. REFERENCES

1. Kaitai Liang, Willy Susilo, and Joseph K. Liu, “Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage,” IEEE Transactions on Information Forensics and Security, vol.10, no.8, 2015.
2. Boyang Wang, Baochun Li, and Hui Li, “Oruta: Privacy-preserving Public Auditing for Shared Data in the Cloud,” IEEE Transactions on Cloud Computing, vol.2, no.1, 2014.
3. Cheng Hongbing, Rong Chunming, Hwang Kai, Wang Weihong, and Li Yanyan, “Secure Big Data Storage and Sharing Scheme for Cloud Tenants,” China Communications, 2015.
4. Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao, “Toward Efficient and Privacy-Preserving Computing in Big Data Era”, IEEE Network, 2014.
5. Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu, “Secure Sensitive Data Sharing on a Big Data Platform,” Tsinghua Science and Technology, ISSN 1007-0214 08/11 ,pp 72-80, volume 20, number 1, 2015.
6. Zhiyuan Tan, Upasana T. Nagar, Xiangjian He, Priyadarsi Nanda, Ren Ping Liu, Song Wang, and Jiankun Hu, “Enhancing Big Data Security with Collaborative Intrusion Detection,” IEEE Cloud Computing, 2014.

7. Chang Liu, Jinjun Chen, Laurence T.Yang, Xuyun Zhang, Chi Yang, Rajiv Ranjan, and Ramamohanarao Kotagiri, “Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates,” IEEE transactions on parallel and distributed systems, 2014.
8. Joonsang Baek, Quang Hieu Vu, Joseph K. Liu, Xinyi Huang, and Yang Xiang, “A Secure Cloud Computing based Framework for Big Data Information Management of Smart Grid,” IEEE transactions on Cloud Computing, vol.3, no.2, 2015.
9. Guiyi Wei, Jun Shao, Yang Xiang, Pingping Zhu, Rongxing Lu, “Obtain Confidentiality or/and authenticity in Big Data by ID-based Generalized Signcryption,” Elsevier Journal on Information Sciences, 2014.
10. Jinbo Xiong, Ximeng Liu, Zhiqiang Yao, Jianfeng Ma, Qi Li, Kui Geng, and Patrick S. Chen, “A Secure Data Self-Destructing Scheme in Cloud Computing”, IEEE transactions on cloud computing, vol.2, no.4, 2014.
11. Matturdi Bardi, Zhou Xianwei, Li Shuai, and Lin Fuhong, “Big Data Security and Privacy: A Review,” China Communications, supplement no.2, 2014.
12. Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial,” IEEE Access, 2014.
13. Minu George, C.Suresh Gnanadhas, Saranya.K, “A Survey on Attribute Based Encryption Scheme in Cloud Computing”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2013.
14. Surya Nepal, Rajiv Ranjan, Kim-Kwang Raymond Choo, “Trustworthy Processing of Healthcare Big Data in Hybrid Clouds”, IEEE Cloud Computing, April 2015.
15. Abid mehmood, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, Song Guo, “Protection of Big Data Privacy”, Special Section on Theoretical foundations for big data applications: Challenges and Opportunities, IEEE Access, Volume 4, 2016.
16. Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li, “Toward Scalable systems for Big Data Analytics: A Technology tutorial”, IEEE Access, Volume 2, 2014.