Big Data Security and Privacy- A Survey

Mrs. Shanmugapriya. E, Dr. R.Kavitha
Assistant professor, Professor
Department of Computer Science and Engineering, Department of Information Technolog)
Anna University Regional Campus, Velammal College of Engineering

Abstract— *Big Data is the term for representing enormous digital data that comes from many resources. Big Data involves not only having access to a large collection of data from various resources but also have the ability to combine them to produce results. Since Big Data is beyond the storage limit offered by the traditional data storage technologies, Cloud provides the concept everything as a service (XaaS).As cloud is a third party to store Big Data, Security and privacy becomes a primary concern to resolve. Even though Elasticity is the attractive feature of cloud, Cloud Service Provider (CSPs) failed to secure the confidential data. This Paper presents the need of security and privacy, security methods and security techniques*

Keywords - *Big Data, Security, Privacy, Encryption, cloud.*

I.INTRODUCTION

This Information era is piloted by digital data and information .The data generation rate is low and can be easily shared in database that becomes accessible from the internet in older days. But, now days, it is growing rapidly. International Data Corporation (IDC) report predicts that the global data volume will grow by a factor of 300, from 130 Exabyte in 2005 to 40,000 Exabyte in 2020 representing double the growth every two years [5]. Hence the information era is changed to Big Data era.. In the era of Big Data, a huge amount of data can be generated quickly from various sources such as smart phones, social networks etc. The emerging big data paradigm has attracted the technology experts and public due to the lower computing cost and it becomes a research hotspot [9].

Big data is the term to represent large and complex dataset whose size is beyond the ability of traditional database software tools to capture, store, manage and analyze. Initially it is characterized by 3V's (Volume, Velocity and Variety) and 2V's(Veracity and Value)are additionally specified to enhance the description. The size of the data is represented by Volume and Velocity is about the speed of data that is created, accumulated, integrated and processed. Variety depicts different format in terms of structured, semi structured and unstructured data. Veracity represents uncertainty or inaccuracy of data. The fifth V(Value)is not about how much data is stored and processed by big data application but it should be of providing the valuable data.

Due to its high Volume, Velocity and Variety, the traditional database tools are not competent to store and process big data.[2].Cloud architecture will support for storing and processing big data. Since cloud is the third party server and it is not trusted by data owner. Moreover, massive amount of data can be computed remotely from the data owner enterprise and accessed from multiple and diverse domain. Security and Privacy becomes a threat. Big Data security implies that the use of big data to implement solution increasing security, reliability and safety of a distributed system.[3] Big Data privacy focuses on the protection of big data from unauthorized use and unwanted inference. This paper focuses on Security and privacy issues in Big Data. The reminder of this paper is divided as follows: Section II portrays Security and Privacy. Section III delineates Security Methods. Section IV discusses about Big Data Security Techniques Section V specially specifies Big Data Privacy and Section VI Concludes the paper.

II.SECURITY AND PRIVACY

Big data are seen to be valuable source of information and it is recorded in the form of photos, video clips, electronic documents and footprints of web surfing by humans [1]. These priceless information can be hacked by using modern technologies and tools such as search engine, social network hacking packages, data mining and machine learning tools [10]. Hence security and privacy becomes a major challenge in big data privacy. Content is varied from person to person and it should be used for any one of the purposes that it was collected not for security [8]. Hence we need to understand large scale collection and storage of data has attracted industry, government and criminals even the privacy regulation is strict.[4]. Therefore, strong security measures should be needed to safeguard big data stores.

2.1 Need of security

Now days, the terabytes of data are handled by various business and other organization. For storing their data, many organizations depend on cloud computing infrastructures. When storing their data to third party, the security and privacy will be a challenging task.[7][3]. Security is measured in terms of data confidentiality and data integrity. The protection of data from unauthorized access is data confidentiality. It is implemented by access control mechanisms. Data integrity refers to the accuracy and consistency of data stored in a database, data warehouse, data mart or other construct. Data integrity is imposed within a database when it is designed and is authenticated through the ongoing use of error checking and validation routines [2]

2.2 Security in large enterprises

Regulatory compliances and post hoc forensic analysis can be maintained by collection of terabyte of security relevant data such as network events. network application events, people 's action events in enterprises[1].Depending on size,10 to 100 billion events generated per day and grow as enterprises enable event logging in more sources, hire more employees, deploy more devices and run more software[4].Due to the data generation rate grows very rapidly; existing analytical techniques wouldn't give better results and leads to false positives, Recently, the enterprises move to cloud architecture for storing and processing, the problem becomes still worse. It has attracted the security community to find the security solution for large scale analysis and processing information. Conventional technologies doesn't support long term, large scale analysis became retaining large quantities of data wasn't economically feasible, inefficient and incomplete in performing analytics and complex queries on large data warehousing has been expensive[8].To resolve the above issues, the modern tools to leverage large quantities of structured and unstructured data have been built.

2.3 Security at various levels

Implementing security controls at the application, operating system and network level is needed to safeguard the entire system using actionable intelligence to detect any malicious activity, emerging threats and vulnerabilities. Elemental Security Platform (ESP) includes passwords, input validation, Role Based Access Control, OS hardening, Persistent control, Responsive no change, In-line Revediction.[7].

2.4 Big data Security & Privacy Challenges

- Secure computation in distributed programming frameworks.
- Security best practices for non-relational data stores
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security/compliance monitoring
- Scalable and compassable privacy-preserving data mining and analytics
- Cryptographically enforced access control and secure communication
- Granular access control
- Granular audits
- Data provenance

III.SECURITY METHODS

3.1 Type based keyword search for security of Big Data

Cryptography is the evergreen technology to secure information. There are number of cryptographic algorithms available such as symmetric key encryption, public key encryption, AES(Advanced Encryption Standard),DES(Data Encryption Standard),etc[5]. These algorithms are different for the users to retrieve desired information from encrypted big data. The novel algorithm (PEKS) has been built to enable searching keywords in the form of fuzzy search, subset search and rank search.

3.2. Achieving Big Data privacy via hybrid cloud

Since information generation rate has been grown rapidly from medical systems, surveillance systems or social networks it makes difficult to store and process. For storing, processing their data to third party server (Cloud server) is good fit with the advantage of low computing cost. [5]But security is a Question because Cloud Service Provider (CSP) can have the full rights to the sensitive data. They may intentionally or unintentionally misuse the data for the other purposes such as advertisement, outsourcing etc[6]. Moreover attackers can steal the data by using modern technological tools.AES like Cryptographic algorithm may be the solution. This algorithm can be suitable for text and not for image since it requires heavy computation overhead [8]. This problem can be eliminated by the following two methods. The first method is by using various image encryption algorithms such as chaos-based approach with substitution diffusion layout. The second method is by separating sensitive data from non-sensitive data and storing them in trusted private cloud and untrusted public cloud respectively [5]. This hybrid

cloud will lead some disadvantages that is requirement of large storage space in private cloud communication overhead between private and public cloud larger delay between private and public cloud.[3] Hence, a novel random one-to-one mapping function is proposed for image encryption and stored in private cloud to minimize the storage.

IV.BIG DATA SECURITY TECHNIQUES

To ensure security & privacy, Organization employs several methods. The first and foremost method is written and oral statement.[2] But this method is not a promising solution in this era. The methods like Authentication, Access Control, and Cryptography can be applied to secure the sensitive data.

4.1 Authentication

Authentication (to valid) is a process in which the credentials provided are compared to a file in a database of authorized users it involves of two steps: Identification step and verification step. Former will give an identifier to the security system, and latter represent authentication information that support the binding between the entity and the identifier.[9] To safeguard the sensitive data, authentication mechanisms should be used. So only the valid users will be able to access such information. Since the level of control needed for different types of data the access control mechanisms must be tuned to realize the security levels needed for authentication. [6].Authentication includes three types such as Password based authentication, Address based authentication and Cryptographic Authentication protocols is more efficient and secure than password based authentication and Address based authentication.

4.2 Access Control

Access control mechanisms strengthen the security measures by providing end users data to only authorized users. Access control will be a good measure to provide privacy and Security. SELinux[Security Enhanced Linux] is the mechanism for supporting access control security policy through the use of Linux security modules in Linux kernels.[9] Traditional access control mechanisms such as access control list doesn't provide a good solution since cloud cover may disclose the data to some authorized users intentionally or unintentionally. The modern access control scheme like Attribute Based Access Control is used to enable the end-users to control the access of their own data. In this scheme, the access policies are defined for their data and encryption can be done by using this access policies. The eligible users can decrypt the data after satisfying the access policy. Sometimes, the access policy may leak privacy because the encrypted data is plaintext form.[7].To overcome this privacy issue, attribute can be hidden in the access policy. But it has another problem of disclosing data not only to unauthorized users and also for authorized users, due to decryption problem [6]. This problem leads to find the solution through partial attribute hiding & whole attribute hiding. In partial attribute hiding method, only the value of attribute name can leak privacy.[4] So , the second method hides the whole attribute by hiding attribute name & attribute values. Privacy leakage due to the personal information is revealed with extended datasets for business operations such as adding values to business, user's shopping habits data collection etc.

4.3 Cryptography

Cryptography is the method of transforming a plaintext into cipher text and again into plaintext. This will maintain the privacy of data. Generally cryptography algorithms classified into four main areas. These are Symmetric Encryption, Asymmetric Encryption, Data integrity and Authentication protocols. Symmetric Encryption is used to protect the contents of data of any size whereas asymmetric Encryption is to prevent small size of data which is used in digital signature. Data integrity includes protection of message from alteration.[8] Authentication Protocols is used to authenticate the identity of entities. Cryptography allows data owners to protect their data proactively instead of relying solely on legal agreements that are difficult to monitor or enforce.[4]. To understand the protection of cryptography we need to consider traditional security goals such as confidentiality, integrity and availability. There are many methods in cryptography which achieve security but homomorphic encryption, verifiable computation and secure multi-party computation seems to the best than the traditional method such as identity based encryption, functional computation and attribute based encryption.

V.BIG DATA PRIVACY

According to Big data life cycle phases (data generation, storage and processing), various privacy mechanisms can be adopted. Privacy protection in data generation phase are access restriction and satisfying data techniques.[4]. Limiting access to individual's private data is the first approach and second approach is to alter the original data before releasing to an untrusted party.[2] The following approaches have been used to protect the sensitive information in data storage phase. They are Attribute

Based Encryption (ABE), Identity Based Encryption (IBE), Storage path encryption and additionally Hybrid Cloud method. In processing phase Privacy Preserving Data Publishing (PPDP) and knowledge extraction from the data are used.

5.1 Privacy in data storage

Access restriction, Anti tracking extensions, advertisement/script blockers, encryption tools are used to preserve private data [3].. In addition anti-malware, antivirus software is also used to protect the digital data. Sometimes, it is not possible to protect sensitive information. In such cases, data should be secured before revealed to third party.

5.2 Techniques

A tool Socket Puppet: Hiding online identity of individual by deceptive actions of false identity of individuals, the online activities and pretending to be someone else. This way of acting multiple socket puppet can deviate the data collectors find it makes difficult to extract the knowledge. Hence, the true activities of users and private information could not be identified.

Mask Me: Security tool to mask individual's identity. This technique creates aliases of personal information such as e-mail id or credit card number. Masking data can be given when the user supported to enter the personal information.

5.3 Privacy in data storage phase

Conventional security mechanisms in data storage phase can be divided into four categories. There are File level data security scheme, Application level encryption scheme,[6] DAS(Direct Attached Storage, NAS(Network Attached Storage).DAS will not applicable for big data due to its scalability hence system is preferred since cloud is a third party security threats leads to privacy violation. and NAS will be a conventional storage architecture.

5.4 Approaches to privacy protection in cloud:

ABE(Attribute Based Encryption)

ABE gives end-to-end big data privacy in cloud computing environment. Data owner defines the access policies and encrypt the data under the access policies. The decryption can be done by users whose attributes satisfies the access policies.[11]. The ABE without policy updating will not give better results. Hence ABE with privacy updating by data owner improves the privacy. But in this method, communication overhead between data owner and cloud server since data owner has to update the policy every time and data should be encrypted.[8] To reduce its overhead, the new method has been adopted. Instead of policy updating by data owner, can send query to the cloud server. The cloud server itself update the policy without decrypting data.

5.5 IBE(Identity Based Encryption)

To preserve the anonymity of senders and receivers.IBE was proposed by using human entities since email address or IP address. Source and destination address protected by applying these primitives [6].

5.6 Drawback of ABE & IBE

Both doesn't support cipher text updating of recipient. Some approaches have been followed to update the ciphertext.But, encryption& decryption is time consuming and costly or big data since it involves communication overhead. In this mode, the data owner has to be online always.[6] This has been resolved by an alternative approach that to update the cipher text receiver by delegating this text to a trusted third party with the details of data owner's decryption key. But it has drawbacks the third party is fully trusted by data owner and the third party should be known about the script to precede new encryption. Hence, the data owner fully depends on third party and it leads to suspicious sometimes.

5.7 IBE privacy –PRRE

To handle the above problem, PRE(Proxy Re Encryption) is proposed. In this scheme data having between different recipients is achieved by transforming a cipher text planned for one user into a cipher text planned for another user for the same message without revealing the knowledge of manage or decryption keys.[8] Here, the proxy handles the workload of data owner and it doesn't have to be online always.

5.8 IBPRE (Identity Base Proxy ReEncryption)

In IBPRE**, t**he identity information of sender and receiver is anonymous. Cipher text receiver can be updated multiple users.

5.9 Homomorphic Encryption

This type of encryption is based on computation as encrypted data[12]. By some functions one can compute function of message directly the encryption can be done. Since, it involves computational complexity; it leads the higher cost, and harder implementation.

5.10 Storage Path Encryption

In this method, before storing the data onto the cloud, it has been separated as public data & confidential data possible data can be accessible by all. The confidential data her to be kept in secure storage environment and it can't be accessible by unauthorized users. To provide secure storage for confidential data, a[3] trapdoor is used to encrypt the storage path instead of encrypting is called cryptographic virtual mapping of big data. Confidential data is also encrypted or some applications. The replications for each piece of data on cloud storage are done to improve the availability and robustness of big data.[6] .In this scheme, the data owner will maintain the storage index information.

5.11 Privacy Preserving Data Processing

To safeguard information from untrusted disclosure because collected data may have sensitive information PPDP is used.[7] To extract meaningful information form data without violating privacy,the sensitive information of original data may have ID(identifier)uniquely identify a person (e.g)name,driving license number of accounts.QID(Query Identifier)linked with the some external dataset may be able to identify the records.

VI CONCLUSION

Several security methods changed the landscape of Big Data. The digitization has changed the world smarter and smarter and led to the Big Data era. But security and privacy is a major threat to this era. This paper portrayed a look back on several security methods. Since Big Data creates a world that maintains the control over our security and privacy, we need to increase our efforts and educate a new generation of computer scientists and engineers on the value of privacy and tools to be developed for designing big data system that follow commonly agreed privacy guidelines.

REFERENCES

[1]. Sophia Yakoubov, Vijay Gadepally, Nabil Schear, Emily Shen, Arkady Yerukhimovich,"A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud",*2013.*

[2].Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Sha," Toward Efficient and Privacy-Preserving Computing in Big Data Era",2014.

[3]. Big data Security and Challenges Available at https://cloudsecurityalliance.org/group/big-data/,2016.

[4].EiEi Mon, Thinn Thu Naing, "The Privacy-Aware Access Control System Using Attribute-And Role-Based Access Control In Private Cloud", *Proceedings of IEEE IC-BNMT,2011*

[5].Kalyani Shirudkar, Dilip Motwani**,"** Big-Data Security**",***vol 5,march 3,2015.*

[6]. Ariel Hamliny Nabil Scheary Emily Sheny Mayank Variaz Sophia Yakoubovy Arkady Yerukhimovichy," Cryptography for Big Data Security",*dec 7,2015.*

[7].Ariel Hamliny Nabil Scheary Emily Sheny Mayank Variaz Sophia Yakoubovy Arkady Yerukhimovichy,"Cryptography for big data privacy",*Sharing,Storage and Security*, 2016.

[8]. S. Srinivasan,"Data Privacy Concerns Involving Cloud", International Conference for Internet Technology and Secured Transactions (ICITST-2016).

[9]. M Aramudhan, R Charanya,"Survey on Access Control in Clooud",2016.

[10]. Anas Ibrahim, Abdelkader Ouda," Innovative Data Authentication Model", The University of Western Ontario,2016.

[11]. Kan Yang, Student Member, IEEE, and Xiaohua Jia, Fellow, IEEE," Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage", *vol 25,no.7,2016*

[12].Yin Hu," Improving the Efficiency of Homomorphic Encryption Schemes",2016.

[13].Paigude, Tejashree, and T. A. Chavan. "A survey on privacy preserving public auditing for data storage security",2013.

[14].Paigude, T., & Chavan, T. A. " A survey on privacy preserving public auditing for data storage security",2013.

[15].Paigude, Tejashree, and T. A. Chavan. "A survey on privacy preserving public auditing for data storage security." 2013.