Security and privacy challenges in big data healthcare

[1]T. Papitha Christobel, [2] Dr.A. Sasi Kumar
[1]PhD Research Scholar, Department of Information Technology,
School of Computing Sciences, Vels Institute of Science, Technology
and Advanced Studies (VISTAS), Pallavaram, Chennai, India.
[2]Professor, Department of Information Technology,
School of Computing Sciences, Vels Institute of Science,
Technology and Advanced Studies(VISTAS),
Pallavaram, Chennai, India.

Abstract: Security of Big Data is a big concern. We know, big data contains structured and unstructured and semi structured data. To provide security to unstructured data is more difficult than the structured data. In emerging IT industry, big data is one. And the database is huge so difficult to manage the data. We need to protect these data against unauthorized person's updating and stolen the data, for this data security is very important one. Big data comes more vital to all industries the challenges also increased high, because it's a heterogeneous data. The private healthcare data are accessed and stored securely by implementing a decoy technique. Telemedicine is an emerging healthcare service. Healthcare is produce big data because now a days, healthcare switches paper based medical records into electronic platform to store, manage, analysis and process in the form of Electronic Medical Record (EMR) or Electronic Healthcare Record (EHR) with the help of internet. Patient agreement is necessary to provide privacy. Agreement includes, who can access patient's particular record with valid purpose. Security can be described as physical and technological measures that can be used to secure healthcare data from unauthorized disclosure or illegal access of any restricted data. Here access control policies are violated that must be prevented.

Key words: Big data, Security, Privacy, Public key infrastructure (PKI), Electronic Medical Record (EMR), Electronic Healthcare Record (EHR), Medical Big Data (MBD), Decoy Medical Big Data (DMBD), Original Medical Big Data (OMBD).

1. Introduction

If we want to protect the data means, we should do the following things, Protection against targeted threads, high performance search and accurate real time reputation. More advanced technological solutions include cryptography and encryption. Encryption is intended to encode data or information such that access is permitted only to authorized individuals who hold the "key" to unlock the encryption code. In 1970's three encryption algorithms there: the symmetric cipher- Data Encryption Standard (DES), the asymmetric cipher-Rivest Shamir Adleman (RSA), and the Diffie-Hellman key exchange.

The most widely used encryption scheme is based on DES. It is referred to as the Data Encryption Algorithm (DEA). Data are encrypted in 64-bit blocks using a 56-bit key. Algorithm transforms 64-bit input in a series of steps into a 64-bit output. The same steps with the same key are used to reverse the encryption. The widely accepted and implemented general purpose approach to public-key encryption is used in RSA algorithm. The RSA scheme is a block cipher in which the plain text and cipher text are integers between 0 and n-1 for some n. The size for n is 1024 bits or 309 decimal digits, that is, n is less than 2^{1024} . The purpose of the Diffie Hellman Key Exchange algorithm is to enable two users to securely exchange a key. That can be used for subsequent encryption of messages. The algorithm itself is limited to the exchange of secret values.

2. Security And Privacy

2.1. Meaning of Big Data Security

Data security is used to prevent unauthorized access to computers, databases and websites. Data security protects data from corruption. It is an important aspect of IT organizations of every size and type. It is also known as Information Security (IS) or Computer Security. Backups, data masking and data erasure are the examples of data security technologies. A key data security technology measure is encryption, where

digital data, software/hardware, and hard drives are encrypted and therefore rendered unreadable to unauthorized users and hackers. One of the most commonly used methods of practicing data security is the use of authentication. In authentication, users must provide a password, code, biometric data, or some other form of data to verify identity before access to a system or data is granted. Data security is also very important for health care records, so health advocates and medical practitioners in the U.S. and other countries are working toward implementing electronic medical record (EMR) privacy by creating awareness about patient rights related to the release of data to laboratories, physicians, hospitals and other medical facilities. Information security and privacy is the most prominent challenge which can be directly attributed to the growth of computing and network infrastructure. The ability of malicious actors to intrude into networks and access restricted data and resources increased exponentially. The challenges of managing information security are well documented, and they are not just limited to firms and individuals. The lack of security led to the growth of a billion dollar industry with a variety of firms providing security products and services.

2.2. Difference between Security and Privacy

Data confidentiality, integrity and availability are called as security. The appropriate use of user's information is called as privacy. The organizations network are used various techniques like encryption, firewall etc. in order to prevent data compromise from technology or vulnerabilities. The organization can't sell its patient/user's information to a third party without prior consent of the user. It may provide for confidentiality or protect an enterprise or agency. It concerns with patient's right to safeguard their information from any other parties. A security offers the ability to be confident that decisions are respected. Privacy is the ability to decide what information of an individual goes and where to.

3. Technologies And Tools

3.1. HADOOP

The most popular big data processor is hadoop. The data sensitive arguments used in hadoop's are data security, data loss, data privacy, data maintenance and data extraction. This means, secured the data without loss of data and use a privacy key to maintain the data without unnecessary extraction. Hadoop is a free, Java-based programming frame work that aids in the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach reduces the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, fault tolerant and flexible. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects namely MapReduce and Hadoop Distributed File System (HDFS).

3.2. Map Reduce

Hadoop MapReduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A MapReduce first divides the data into individual chunks which in turn are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Usually the input and the output of the job are both stored in a file -system. Scheduling, Monitoring and re-executing failed tasks are taken care of by the framework.

3.3. Use Of Cryptographic Encryption Techniques

Encryption is the process of transforming plain text message to some form that can be readable only by communication entities involved in data transmission. Levels of encryption can be performed both software as well as hardware. There are various encryption algorithms such as Triple DES (Data Encryption Standard), AES (Advance Encryption Standard), IDEA (International Data Encryption Standard) etc.

3.4. Authentication Algorithms

There are a number of authentication algorithms used for e.g. Digital signature, Password mechanism, in sensor networks hash functions can be used for authentication. Purpose of information authentication is to protect two parties who exchange information's from any third party. It does not protect the two parties against each other. Several forms of dispute between the two. In this situation there is no complete trust between sender and receiver. More authentications are needed. The most attractive solution to this problem is the digital signature. It is analogous to the hand written signature. It must verify the author and the date and time of the signature. It must verifiable by third parties to resolve disputes. It must be computationally infeasible to forge either by constructing new information for an existing digital signature. It must be practical to retain a copy of the digital signature in storage. There are two types of digital signatures are there: Direct Digital Signature and Arbitrated Digital Signature.

3.5. Big Data Technical Issues And Challenges

Fault Tolerance:
Fault tolerant computing requires extremely complex algorithms. To solve this issue, divide the entire computation to be done into tasks and assign these tasks to different nodes for computation. At the same time, keep a node as a supervising node and look over all the other assigned nodes as to whether they are working properly or not. If an interruption occurs the particular task then the system will be restarted.
Data Heterogeneity:
In big data, 75% of data are unstructured. Handling unstructured data is inconvenient and also expensive. Converting the unstructured data to structured data is unfeasible.
Data Quality:
Storing of big data is expensive; it is a tiff between business peoples and IT peoples, to store the huge amount of data.
Scalability:
In cloud computing, to manage big data the workloads and performance into very large clusters.

4. Role Of Security And Privacy In Healthcare

The concerns over the big healthcare data security and privacy are increased year-by-year. While the automations have led to improve patient care workflow and reduce costs, it is also raising healthcare data to increase probability of security and privacy breaches. In fact, attackers can use data mining methods and procedures to find out sensitive data and release it to the public and thus data breach happens. Privacy of medical data is an important factor which must be seriously considered. Different countries have different policies and laws for data privacy.

4.1. Healthcare Big Data Security And Privacy Issues

Use of technologies such as: Electronic Patient Record (EPR) systems or we can say Electronic Health Record (EHR), Remote patient monitoring using sensor networks and the combination of electronic record with sensor networks as a hybrid can help to enhance the quality of medical process. In these processes protection of patient's data is a very big challenge.

4.2. Recent Approaches Used In Big Data Privacy

Proposed the privacy preserving data mining techniques in Hadoop, that is solve privacy violation without utility degradation. But, its execution time is affected by noise size. Introduced an efficient and privacy preserving cosine similarity computing protocol. But, Need significant research efforts for addressing unique privacy issues in some specific big data analytics. Proposed a scalable two phase top down specialization (TDS) approach to anonymize large scale data sets using the Map Reduce framework on cloud. But it uses anonymization technique which is vulnerable to correlation attack. Proposed various privacy issues dealing with big data applications. But customer segmentation and profiling can easily lead to discrimination based on age gender, ethnic background, health condition, social background etc.

4.3. PKI For Big Data Security

It includes client software, authority server, smart cards and operational procedures. In IT security services, Public key cryptography is very popular and advance. Because it uses the digital certificate (DC), it contains only the name and email address. DC uses to components namely, private and public keys. For example the user wishes to send an email to his business associates; they wants to sign digitally the email with his private key. The email sent to his business associates. Then they will decrypt the message from the user's sending public key. In this example DC provides secret information can be shared with user authentication and also without exchange the secrets in advance. PKI is also used in medical application systems. The uses of PKI are, to encrypt email message and document, smart card login system authenticate the user's applications and bootstrapping is secured communication protocols like internet key exchange. PKI's solution is very popular in e-commerce and business to business companies. And also used in financial, health care sectors, NASA, DoD and e-governance application for authenticate and integrity purpose. The advantages for PKI are no limit to access and exchange of data between client and server only. The characteristics of PKI are suitable for big data. So, PKI is used for big data security.

5. Proposed System

Telemedicine is one of the emerging fields for e-health research. In this service, EMRs including MBD, images and multimedia medical data are transmitted on the fly over insecure internet connections as they are required by the remote doctors. The decoy files are used to retrieve all files from the beginning to ensure better security. The specialty of decoy file is a double security technique by encrypting the original file when an attacker recognizes that they dealing with a decoy gallery; they would need to figure out how to decode the original gallery. Finally, this methodology ensures the users MBD are 100% secure and shortens the process. Suppose if the user is an attacker, by default it offers the decoy big data gallery directly to any user and keeps the original one hidden, which is only made available to a legitimate user after successful verification.

The basic idea behind this technique is to limit the damage caused by stolen data by decreasing the value of the stolen information. To achieve this, the decoy should have certain features. First, it should be believable. In the absence of any additional information, a perfectly believable decoy should make it impossible for an attacker to figure out that the data are not real. Thus, the decoy should seem authentic and trustworthy. Second, the decoy should be enticing enough to attract the attention of the attacker and make them open the file. Third, the decoy should be conspicuous, which is closely related to being enticing. Whereas enticing is related to how curious an attacker is about a decoy, conspicuousness has to do with how easy a decoy is to access. Therefore, the decoy should be easily located by search queries. Fourth, the decoy should be differentiable so that the real user can distinguish between the real and the decoy file. Balancing different ability for authentic users with believability for attackers is one of the critical aspects of any decoy deployment system. Fifth, the decoy should be non-interfering so that the real user will not accidentally

misuse the bogus information contained in the decoy. Finally, the decoy should be detectable; this feature refers to the ability of decoys to alert their owners once they have been accessed.

5.1. DMBD Algorithm

It is used as a trap gallery that makes it not of direct relevance to the legitimate user but it is used to secure his/her OMBD by distracting the attacker. The DMBD is placed in the fog computing as a honeypot to secure the original one, which is located in the cloud. A number of anomaly detection systems are provided by fog computing such as user proling and a decoy file system. For each newly uploaded MBD in the OMBD, a decoy one will be placed on the DMBD.

5.2. User Profiling Algorithm

User profiling can help to determine whether a user is legitimate or not based on certain parameters, such as the user search behavior, amount of downloaded data, nature of operations, division of tasks, and IP address. There are three different types of user profiling, each with different advantages and disadvantages based on the techniques used. The type that will use in our system is the hybrid user profile, which is a combination of explicit and implicit user profiles. The explicit user prole usually contains high quality information because it is gathered from the user, but it requires a lot of effort from the user to update their profile information. Another way, the implicit user profile is automatically updated with minimal user effort; a large amount of interaction between the user and the content is required before an accurate user prole can be created. Thus, combining the two types into a hybrid user profile should reduce the weak points and advance the strong points of each technique used to monitor the cloud data access and detect any unusual data access pattern.

5.3. Key Exchange Algorithm

In our proposed system, the OMBD and the DMBD need to communicate in different situations, for example, when the user uploads a new photo/image, the OMBD is supposed to communicate with the DMBD to inform it to add a new decoy photo. These communications between three parties (the user, the OMBD, and the DMBD) need to be secure.

5.4. Photo Encryption Algorithm

Photo encryption is a technique used to secure a photo by changing it to an understandable one. Different photo encryption algorithms with different properties and different levels of security are available. In our proposed system, using the Blowfish algorithm. Blowfish is a symmetric key cryptography where the key does not change, such as an automatic file encryption. The reasons behind choosing this algorithm are the following: (1) Blowfish has a longer key length, making it the most secure algorithm; (2) it can encrypt any photo file format of any size, black and white or even a color photo. Based on the Blowfish algorithm length, the photo will be divided. The beginning of the array will be directly after the photo header since the header would not be encrypted. The array elements will be stored in rows, left to right ordered, with every photo scan line represented by one row, and the photo rows will be encrypted from top to bottom.

5.5. Photo Decryption Algorithm

Photo decryption is the reverse of photo encryption. This process will restore the encrypted photo to what it was originally. In this process, the same photo encryption key will be used. The only difference between decryption and encryption is that supplying the sub keys (P-array) with photo decryption is in reverse order. To encrypt a photo, the two inputs are the plain photo and the key. Then, the photo will be converted into a cipher photo. For the reverse process, which is decryption, the inputs are the cipher photo and the key, and this process will convert the cipher photo to its original form, which is the plain photo.

5.6. Original Mbd Algorithm

The OMBD contains the real legitimate user's photos, for which the whole system was built to secure them. This gallery is located in the cloud. Each time the user needs to access it, they needs to pass the security challenge first. Each time the user uploads a new photo into the gallery,
A decoy one will be added in the DMBD and the original photo will be encrypted. This was designed to make the system more secure since it has two levels of security: one is the honeypot DMBD and the other keeps the original photos encrypted while stored in the cloud.

5.7. The Purpose Of Blowfish Algorithm

Many different competitive studies have been done regarding encryption algorithms to help us pick the best one. Thus, we used results from studies that were about the competitive analysis on different symmetric encryption algorithms, which are DES, 3DES, AES, and Blowfish. The comparison between the algorithms was based on seven criteria listed below in detail:
1) *Block size*: The relation between security and block size is positive, so the larger the block size, the more secure it is. The block size used for all of the algorithms is 64 bits except for AES, which uses 128 bits. So, it is clear that, based on block size, AES is more secure than others, but at the same time it costs more to implement.
2) *Number of rounds*: This is the same as block size, so the higher the number of rounds, the more secure they are. The number of rounds for AES might be 10, 12, or 14 depending on the key length used. For DES and Blowfish, they have 16 rounds. On the other hand, 3DES has the highest number of rounds, which is 48, and it’s named triple DES since its number of rounds is three times more than DES.
3) *Key length*: The longest key length means a decreased likelihood of successful attacks. Blowfish is the most secure algorithm since its key length is in the range of 32 bits to 448 bits, while DES's key length is 56 bits, 3DES's key length is either 112 or 168 bits, and the AES key length might be 128, 192, or 256 bits.
4) *Encryption/Decryption time*: This is the time needed to convert plaintext into cipher text, and the reverse regarding decryption time. The algorithm that consumes the longest encryption/decryption time is 3DES, while the one that consumes the shortest time is Blowfish.
5) *Power consumption*: Regarding this criterion, 3DES consumes more power than the others, while Blowfish consumes the least.
6) *Memory usage*: 3DES uses a very high memory while the memory usage of Blowfish is very low.
7) *Confidentiality:* The confidentiality of 3DES and AES is high while DES is low, but Blowfish has the highest confidentiality among them Based on the previous comparisons, Blowfish gives a better performance than the others based on its encryption/ decryption time. It is also more secure than the rest based on the key size used. It consumes less memory and power than the others. Thus, Blowfish is the best algorithm and is thus the best candidate to be used in our proposed system.

5.8. Fog Computing

Fog computing can be considered as an alternative name for the Decoy Document Distributor (D3), which is a tool for generating and monitoring decoys. This strategy is used to protect the real, sensitive data by providing a ``fog" of misinformation. Decoy information, such as decoy documents, honey files, and honeypots, among others, can be generated when unauthorized access is detected. This confuses the attackers and makes them believe that they have the real, useful data when they actually do not. Decoys can be created manually by the user ; for example, when the user creates a new document, they can create a fake document that will appear as a mirror document but contains bogus information. Such manual creation of decoys is obviously very tiring for the user, especially if we are talking about a large organization with multiple users and files. For this reason, fog computing is used to create decoys with minimal user intervention.

6. Conclusion

Thus, Privacy and Security are two important factors that must be considered when developing a patient centric EHR. Existing works have focused largely on security and less on how patient mastermind privacy control with the help of the healthcare providers. And, outside threats were of more concern, but more often than expected insiders who have legitimate access to EHR are often overlooked. This research implements cryptographic techniques with biometrics despite its key security challenges. At the end, two photo galleries are generated. The OMBD is kept secretly in the cloud and the DMBD is used as a honeypot and is kept in the fog. Therefore, instead of retrieving the DMBD only when any unauthorized access is discovered, the user, by default, accesses the DMBD. The OMBD is only accessible by a user after verifying the authenticity of the user. The original multimedia data become more secure by setting the default value of the DMBD, while the OMBD is kept in a hidden gallery. To facilitate the above process, an efficient tri party authenticated key agreement protocol has been proposed among the user, the DPG, and the OPG based on paring cryptography. There are two approaches to store data centralized which store data in central database or it can be stored in local database connected to each database resides inside a network.

7. References

[1] M. Koushikaa,S.Habipriya, Mr.S.S.Aravinth, Mr.T. Karthikeyan, Dr. V. Kumar, A Public Key Cryptography Security System For Big Data, Volume 1 | Issue 6 | November 2014
ISSN (online): 2349-6010
[2] S. Indirakumari, A. Thilagavathy, A Secure Verifiable Storage Deduplication Scheme on Bigdata in Cloud, Volume 2 | Issue 2 | ISSN : 2456-3307
[3] Hadel Abdulaziz Al Hamid, S.K. Md. Mizanur Rahman,M.Shamim Hossain,Ahmad Almogren, and Atif Alamri, A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography, September 28,2017.
[4] S.Castro and R.Pushpalakshmi, A Survey on Big Data Security and Related Techniques to Improve Security, Volume 1, Issue 5, Pages 113-116, June 2017
[5] Isabel de la Torre, Begoña García-Zapirain, Miguel López-Coronado, Analysis of Security in Big Data Related to Healthcare,Volume 3, Number 3, Article 5, September 2017.
[6] Kaustav Ghosh,Asoke Nath, Big Data: Security Issues, Challenges and Future Scope, Volume 3, Issue 3, 2016, PP 1-11.
[7] Karim Abouelmehdi,Abderrahim Beni Hessane and Hayat Khalouf, Big healthcare data: preserving security and privacy, 2018.
[8] Cornelia L. Hammer, Diane C. Kostroch,
Gabriel Quirós, and STA Internal Group, Big Data: Potential, Challenges,
and Statistical Implications, September-2017.
[9] Harsh Kupwade Patil and Ravi Seshadri
Nanthealth, Big data security and privacy issues in healthcare, 2014.
[10] Renu Kesharwani, Enhancing Information Security in Big Data, Vol. 5, Issue 8, August 2016.