Review of Different Approaches and Analysis of Big Data Security Issues

Nikhat Azhar¹, Mohammad Suaib² and Faiyaz Ahmad³
¹,²,³Computer science and Engineering Department, Integral University, Lucknow, INDIA

ABSTRACT
In this paper, we mainly discussed about the security issues related to big data. The basic issues related to the security of big data are mainly considered as Authentication level, Data level, Network level and Generic level issues. We also discuss many approaches to solve these security issues as Data encryption, Network encryption and Logging. For data encryption we use many algorithms as Data Encryption Standard, Advanced Encryption Standard, Rivest, Shamir and Adleman algorithm and Elliptic Curve Cryptography.

Keywords:- Big data, Encryption, RSA algorithm, DES Algorithm, AES Algorithm, IPSec, SSL.

I. INTRODUCTION
Coming century will be Digital, means everything will be digital, Internet of Things (IoT) is an example of it. Kevin Ashton (Father of the IOT), he believed IOT could “turn the world into a data”. A lot of data is generated every second. In the past time it was very easy to handle little data, with the help of the traditional relational data base management system (RDBMS), but very tough to handle such huge amount of data. In the early days of computer, let see the year 2003 the ratio of connected device per person is very less (0.08) and the record changes are noticed in year 2010 (1.84) as in year 2015(3.47) and if talk about the future, it will be 6.58(2020), it means every person will use approximately 6 devices at a time, so we can think huge amount of data will increase. To overcome the drawbacks of the traditional data processing methods, we use a new concept of Big Data, which refer to the large data configuration, storage and analysis of that data.

A. Characteristics of Big Data
The study of big data is incomplete, without the discussion of three main characteristics of Big Data known as 3 V’s [1].
1) Volume (Lots of data)
Volume indicates large amount of data. Every second terabytes to zettabytes of data is collected from various resources.
2) Variety (Different form of data)
Data comes in the structured or in the unstructured form just like image, video, sounds etc
3) Velocity (Analysis of streaming of data)
Data comes at very high speed and with high frequency from social sites just like Gmail, Face book, Twitter and WhatsApp.

To overcome the drawbacks of the traditional data processing methods, we use a new concept of Big Data, which refer to the large data configuration, storage and analysis of that data.

B. Big Data technologies and tools
Due to the variety of data that is encompasses, big data always facing a large number of challenges related to its large amount and complexity in storage and the less security. Complexity refer to the unstructured nature of data, it is very tough to handle unstructured data. There are some technologies & tools to analyze and store Big Data.

- Hadoop
- MapReduce
- NoSQL
- HIVE
- PolyBase

Big Data mainly concern with storing, process the data and extract large amount of data. We use many environments to employ this Big Data as Cloud Computing, Distributed File system and many more [10]. A large amount of data is generated by various resources such as internet,
sensor data, social media due to cloud computing and other ways the data can be easily accessible to everyone, but content of data is belongs to any organization, so it should be secured. That’s why security is main issue after storage of big data.

C. Challenges in big data security

Nowadays every organization has huge amount of data on internet, as much as data is store, more security is required. Any organization will face very big financial loses and other loses due to lack of proper security mechanisms. Now discussing some challenges in Big Data.

1) As we know big data is spread here and there on internet. An additional security is required, which is not available all the time and everywhere.

2) File distributed method, having only one level of data security, which is not enough.

3) A large amount of data is spread everywhere. If anyone wants to use it, the data should be easily Accessible, authentic and having right information otherwise user will be misguided.

4) Unauthorized data access should be a crime.

5) Due to large amount of data and location transparency, it is very tough to detect any changes, occurred in any data.

The generic name for the collection of tools designed to protect data and to thwart is Data Security.

Data Security means employing security operations to achieve these major goals:

1) Privacy/Secrecy/Confidentiality
The assurance that the data is not examined or accessed by unauthorized parties.

2) Integrity
The assurance that the data, we received has not been modified. Once it sent by the sender.

3) Authentication:-
The assurance that the data, we received, is sent by the legitimate party or parties or data authentication is a procedure to verify that received data come from the alleged source and have not been altered. Data authentication may also verify sequencing and timeliness.

II. SECURITY ISSUES

The four important security issues of big data are Authentication level, Data level, Network level and Generic level issues.

A. Authentication level issues:
Authentication is the technique by which a process verifies that its communication partner is who it is supposed to be and not an imposter.

B. Data level issues:
In big data, data is everything. Our main goal is to secure data. Data may be anything which is some important and useful personal information [11]. Big Data is not magic, it doesn’t matter how much data you have if can’t make sense of it”

C. Network level issues:
Data go from one place to other place through network. Network should be secure, otherwise anyone can access confidential data and no one detect it. Network level security problems can be divided into four intertwined areas-

Secrecy, authentication, non-repudiation and integrity control. Hadoop data is not secure.

D. Generic level issues

The generic level security issues specially concerned about the privacy and security of the electronically transmitted and stored data. Secure Electronic Transaction (SET) is an encryption and security technique designed to protect credit card and other money related transactions over the internet. Any sensitive personal information as medical records, property documents related to anyone should be needed high level of security on the internet. In the same way, high level of security is needed in legal matters also over internet [1].

III. APPROACHES TO SOLVE SECURITY ISSUES

To solve these Authentication level, Data level and Network level issues, we use many approaches as Authentication mechanism, Data encryption and Network encryption.

A. Data Encryption

Encryption is the process of converting plaintext into cipher text with the help of algorithms and keys. Encryption is based on algorithms that scramble information into unreadable or non-discernible form. Decryption is just opposite of encryption means we decrypt cipher text with the help of algorithms and keys to get plaintext again. Encryption is done at sender side and decryption is done at receiver side.

- Symmetric Key cryptography:
Symmetric key encryption is also referred to as conventional encryption or single-key encryption. In symmetric key cryptography, we use same secret key for both encryption and decryption. Examples of symmetric key algorithms are Data Encryption Standard (DES), double & triple DES, Advanced Encryption Standard (AES) and IDEA (International Data Encryption Algorithm) [1].

- Asymmetric Key cryptography:
Asymmetric key encryption is also referred to as modern encryption or public key encryption. In symmetric key cryptography, secure key exchange is a problem.In asymmetric key encryption there is no need of key exchange.
We uses two keys, one private and one public in asymmetric key cryptography. Rivest-Shamir-Adleman (RSA), Elliptic curve cryptography (ECC) are examples of asymmetric key cryptography.

DES(Data Encryption Standard)

DES is the best known and most widely used symmetric algorithm in the world. DES is the name of the federal Information Processing Standard (FIPS), which
describes the Data Encryption Algorithm (DEA) created by IBM. The algorithm transforms 64-bit input in a series of steps into a 64-bit output. The same steps with the same key used in reverse order for decryption. Plaintext is encrypted in blocks of 64 bits, giving 64-bit cipher text. The algorithm has 19 distinct stages. The first stage is transposition apply on the 64-bit bit plaintext. The exact inverse of this transposition will take place in the last stages. The second last stage exchanges the leftmost 32 bits with the rightmost 32 bits. The rest of the 16 stages are functionally identical but are operated by different functions of the key. To extend the life of DES, Tuchman proposed a triple encryption method that uses only two keys. The function follows and encrypt-decrypt-encrypt (EDE) sequence. As it is an alternative to DES, so it is known as 3DES or triple DES.

Security provided by DES:

Timing attack is the only attack that is possible on DES. In timing attack, the attacker check how much time is use to decrypt a encrypted message.

1) AES(Advanced Encryption Standard)

AES must be a stream block cipher with a block length of 128 bits and use the key of 128,192 and 256 bits lengths. The input to the encryption and decryption algorithms is a single 128-bit block. The key size depends on the no of rounds [2,3].

It uses 10, 12, or 14 rounds. Depending on the number of rounds, the key size may be 128, 192, or 256 bits. This block is copied into the State array, which is modified at each stages of encryption or decryption. After the final stage, State is copied to an output matrix. Similarly, the 128-bit key is depicted as a square matrix of bytes. This key is then expanded into an array of key schedule words; each word is four bytes and total key schedule is 44 words for 128-bit key [5,7].

Security provided by AES:

AES algorithms not only provided high level of security but provided high speed also. Speed is also very important as well as security. For both efficiency and security, a larger block size is desirable [6]. The Advanced Encryption Standard (AES,) that uses 128 bit block size as well as 128 bits key size was introduced by NIST. So, security, efficiency and speed all three core element of secure system is provided by AES for symmetric key algorithms.

2) IDEA(International Data Encryption Algorithm)

IDEA, originally called Improved Proposed Encryption Algorithm (IPES), was developed by James Massey and Xuejia Lai in 1992. IDEA is a symmetric key, block-cipher algorithm with a key length of 128 bits, a block size of 64-bits. Same key is used for encryption and decryption. IDEA uses 52 subkeys in 8 rounds. Each round uses 6 subkeys, with the remaining four being used for the output transformation.

Security provided by IDEA:

Speed of IDEA is approximately twice of the DES and it is more secure then DES. We can understand the security mechanism of IDEA, if any hacker using brute-force approaches, there are 2^{128} possible keys to decrypt an IDEA encrypted message. It may be in the future attackers use new techniques with the help of powerful machines can decrypt IDEA-encrypted message. However, IDEA will be an extremely secure cipher for a long time.

3) RSA (Rivest, Shamir and Adleman)

RSA is the first known public key algorithm. RSA based on the fact that multiplication of two prime number is easy but it is very tough and time consuming to get these numbers from the answer. The RSA scheme is a block cipher in which the plaintext and cipher text are integers between 0 and n−1 for some n, whose typical size is 1024 bits. Let the multiplication of two prime numbers is n and t=(p−1)(q−1) then We take a integer m

Where gcd (t, m)=1 & 1<m<t;

s=m^{−1}mod t.

now The Public key, Pk={m, n} and, The Private key, Pr={s, n}

Security provided by RSA:

There may be three attacks are possible on RSA encrypted messages, brute-force attack, factoring attack and last one is timing attack. We can avoid brute-force attack by using large key, but this will increase mathematical burden on the user and also slower the speed of the system. It is clear that strength of RSA is mainly depending on the factorization capability of the attacker. That’s why to make it more secure a larger key size is required and a research is going on to increase the key size from 1024 to higher bits key size. By keeping track of how long a computer takes to decrypt messages attacker can determine a private key [2].

4) Elliptic Curve Cryptography (ECC)

Elliptic Curve Cryptography is an approach to asymmetric key cryptography based on the algebraic structure of elliptic curves within finite limit. This method has this name because the cubic equations we use to calculate the circumference of an ellipse are same as the cubic equation use to describe the elliptic curves. Elliptic curves in cryptography was suggested by Neal Koblitz and V.S.Miller. Elliptic Curve Cryptography algorithms entered wide use in 2004 to 2005. It is clear that, considerable a smaller key size can be used for ECC compared to RSA, for example 256-bit ECC public key should provide comparable security to a 3072-bit RSA public key. An elliptic curve is a plane curve over a finite field which consists of the points satisfying the equation.

General form

$qy^2 = rx^3 + ax + b + sx^2 + ty$

We want the set of solutions (x,y) to the equation $y^2 = x^3 + ax + b$

where 4a^3 + 27b^2 ≠ 0.

where a and b are real numbers and x and y are true variables. Elliptic curves are applicable for key agreement, digital signatures and other security related tasks.

Security provided by ECC:

The security of ECC mainly depends on difficulty to get the value of k if kS and S are given. This is known as
elliptic curve logarithm problem [7]. It is known that, if we take smaller and equal key size for both RSA and ECC, the computational efforts required for ECC and RSA is approximately equal.

5) **Diffie Hellman Key Exchange Algorithm**

Diffie-hellmen key exchange, also called exponential key exchange technique, is the first published public-key encryption method. Discovered by Whitefield Diffie and Martine Hellman. It permits two users to exchange a secret key. This algorithm mainly based on two numbers, A prime number and an integer that is a primitive root of that prime number [7].

**Security provided by Diffie-Hellman key exchange Algorithm:**

The security of the diffie-hellman key exchange is depend on the fact that, it is easy to calculate exponential modulo prime, it is very tough to calculate discrete algorithms.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>DES</th>
<th>IDEA</th>
<th>AES</th>
<th>RSA, ECC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>IBM 75</td>
<td>Xuejia, james</td>
<td>Rijman, joan</td>
<td>Rivest, Shamir, Adleman, Neal, Vector</td>
</tr>
<tr>
<td>Block size</td>
<td>64 bits</td>
<td>64 bits</td>
<td>128 bits</td>
<td>Varies, Varies</td>
</tr>
<tr>
<td>Key size</td>
<td>56 bits</td>
<td>128 bits</td>
<td>128,198 or 256 bits</td>
<td>Based on no. of bits, 135 bits</td>
</tr>
<tr>
<td>Security Strength</td>
<td>Not enough</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Good, Less</td>
</tr>
<tr>
<td>Execution speed</td>
<td>Slow</td>
<td>Twice faster than DES</td>
<td>More fast</td>
<td>Slowest, Fastest</td>
</tr>
</tbody>
</table>

By observing above table we get that AES cryptographic algorithm is best algorithm among the all algorithms. As our main concern is Big data’s security analysis, so size of data for encryption is also important and in AES algorithm we are using 128 bits block, so more data will encrypt in less time.

**B. Network encryption**

Network encryption or network level encryption is a network security process that applies encryption services over the network. The main layers which are more responsible for data transmission are data link layer and network layer and transport layer. In the data link layer, big data on a point-to-point line can be encoded as they leave one machine and decoded as they enter other. In the network layer, firewalls can be installed to keep data in or keep data out. In the transport layer, entire connection can be encrypted end-to-end, that is, process to process. Network encryption is implemented through Internet Protocol Security (IPSec), a set of open Internet Engineering Task Force (IETF) standards that, used in conjunction, create a framework for private communication over IP networks. IPSec works through the network architecture, which means that end users and applications don't need to be altered in any way. Encrypted data appear to be identical to unencrypted data and are easily routed through any IP network. Network encryption products and services are offered by a number of companies, including Cisco, Motorola, and Oracle [7]. Data is encrypted only while in transit, existing as plaintext on the originating and receiving hosts. There are many types of security threats arises, when Web is in use. We can divide them into two groups. In first group, these threats known Passive attack and the other one is Active attack. In the second group, we can differentiate the security threats, based on location of the threat: Web browser, Web server and network traffic between browser and server (network Security). For the sake of Web Security, we uses many approaches as SSL and TSL. The main concept behind the designing of SSI to make use of TCP to provide a reliable end-to-end secure services. Tangleper layer security (TSL), TSL is an IETF standardization initiative. The main work of TSL is to generate an Internet standard version of SSL. An SSL certificate is required to make an SSL connection. You would need to give all details about the identity of your website and your company as and when choose to activate SSL on your web server. Now creation of two cryptographic keys will take place a Private key and a Public key. Then Certificate Signing Request (CSR) will submit, this is a data file that have your details along with your Private key. The CA (Certification Authority), would then validate your details. Following successful authentication of all details, they will issue a SSL certificate. The newly-issued SSL would be matched to your Private Key. After all security check have done, our Web Server created a secure (encrypted) link between your website and the customer’s web browsers. An SSL Certificate comprises of your domain name, the name of your company and other things like your address, your city, your state and your country. It would also show the expiration date of the SSL plus details of the issuing CA.[5,7]

**C. Logging**

Authentication logs show you where and how users authenticate, with usernames, location, time, type of auth factor and more. Normalize user patterns so you can identify abnormal activity.
IV. CONCLUSION

In this paper we discussed about big data and its security issues. To solve data level security issues we discussed many encryption algorithms and their comparative study. And in the last part we studied about the approaches to solve the network level and authentication level security issues.

REFERENCES

[2]. Sreenivasa B.L1, Manish Kumar2, Mohammed Nueed Shaikh3 and Dr. S Sathyaranayana4.”A STUDY ON ENCRYPTION DECRYPTION ALGORITHM FOR BIG-DATA ANALYTICS IN CLOUD”, International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM 2016, pp. 323-329, e-ISSN:2278-621X.
[4]. Sophia Yakoubov, Vijay Gadepally, Nabil Schear, Emily Shen, ArkadyYerukhimovich.”A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud”.