A Survey on Security Issues and the Existing Solutions in Big Data

Pooja Chaudhary
M.Tech. Scholar
Department Of Computer Science & Engineering
ABES Engineering College, Ghaziabad

Virendra Kumar Yadav
Asst. Professor
Department Of Computer Science & Engineering
ABES Engineering College, Ghaziabad

ABSTRACT
Big data refers different-different type of data, i.e., structured data means Relational Data, Semi-structured data means XML and unstructured data example :-Word, PDF, Text, Audio, Video etc. Here I have describe only one aspect of big data; other attributes are volume, velocity, value, and veracity. Security is technology issues, which seems to be resolvable in the near-term, but represent long term challenges that require research and new paradigms. This technology needs to be more improvement in terms of development because it suffers from different security problems. This paper will discuss about different security problems and define existing solution.

Keywords
Issue in Big data Security, Issue in Hadoop Security, TLS and Mc-TLS, Performance Evaluation

1 INTRODUCTION
"Big Data" meant the a lots of data in terms of 6v's (volume, velocity, variety, veracity, variability and value) that could be processed efficiently by current database methods and tools. It is a collection and analysis of large set data which has information about different-different types of data like:- user data, sensor data, medical data and statistical data etc. Its also collecting information from movies streaming and social media (example:- Facebook, linkedIn and twitter).

Table1. Comparison between Big data and Traditional data

Components	Big Data	Traditional Data
Queries	Largely Abandoned SQL	Traditional SQL
Architecture	Distributed	Centralized
Data Types	Structured, Semi-Structured and Unstructured	Structured
Data Model	No Schema	Fixed Schema
Data Relationship	Unknown or Complex Relationships	Known Relationship
Data volume	Petabytes or Exabyte	Terabytes
Data Traffic	More	Less
Data Integrity	Less	High

1.1 Big data issues

1.1.1 Management Issue
To manage Large set of structure, semi-structure and unstructured data from resources Like: social media, private sector and public sector etc. Big data management means to ensure the quality of data, responsibility, ownership, accessibility and documentation of data set. So that management challenge is volume of big data.

1.1.2 Storage Issue
The volume of large data set is very complex to store. Each time they have invented new storage medium. Many Big Data companies using some storage tools Like: NoSQL, Apache Drill, Horton Works, SAMOA, IKANOW, Hadoop, Map reduce, Grid Gain to handle the Big Data.

1.1.3 Processing Issue
Let's see that data is chunked into blocks of 8 words, so 1 Exabyte equal to 1K petabytes. Now 100 instructions expand by a processor on one block at 5 gigahertz, 20 nanosecond will be spending for end to end processing. Processing time will be required to process 1Kpetabytes is, 635 years.

1.1.4 Security Issue
There are lots of challenges to manage a large data set in security purpose. There are insufficient tools, public and private database having more threats. To secure data in presence of third party is serious problem. Here no more specific tools and technology for converting from homogeneous data to heterogeneous data of large data set with security and policy certificate often developed. Sometimes, publicly available big data set hacked by some hackers who are most of the time aware to copy it and store it in devise like USB, hard disk etc. They are also involve to attack the data by sending some types of attack like:- snoofing attack, brute force attack and denial of services. To overcome this type of security problem some cryptographic techniques and best algorithm must be developed to enhance the security of data.

1.2 Hadoop (Highly Archived Distributed Object-Oriented Programming)
It is a current technology, to store data, manage data and process data. It was developed by Dough Cutting and Mike Caferella in 2005. Hadoop is an open source software framework uses Java programming language with some native code "C" and "Shell Script" for distributed processing of very large data.

The core of apache Hadoop consist two parts one is storage part (HDFS) and another is processing part (Map-Reduce).

Base Apache Hadoop framework is composed of the following module

1.2.1 Hadoop Common

It contains libraries and utilities, which are considered by other modules.

1.2.2 HDFS

HDFS is a distributed file system that store data on commodity machine.

1.2.3 Hadoop YARN

A resource management platform responsible for managing computing resources in clusters and scheduling the user's application.

1.2.4 Map Reduce

A programming model which using for large scale data processing. Hadoop is a collection of additional software package that can be installed. Such as, Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache Zookeeper, Impala, Apache Flume, Apache Sqoop, Apache Oozie, Apache Storm and others.

1.2.5 HDFS Architecture

Its having portability in Java with more scalable, reliable, distributed in the Hadoop framework environment. Hadoop cluster contains single Name Node and group of Data Node, Its perform the operations like: "Write Once, Read many times". HDFS is the base layer of Hadoop Architecture contains different classified data and its more sensitive to security purpose. It has not specific part in the system for security. Most of the public sector and private sector does not use Hadoop framework for storing important data because of less security reason, inside a technology. Due to less security its providing security in outside of Hadoop environment. Some researcher described that the HDFS provide only encrypting the block level and individual file system in Hadoop framework.

2 SECURITY ISSUE

In Big data major security issues are Privacy and Data Provenance problem. Privacy is concern with data will be use for particular purpose that it was collected, It means data is sharing between two party only. No need to access by third party or hackers. But now days, privacy is suffering from technological limitation on the ability to extract, analyze and correlate potentially sensitive dataset[a]. However as well as big data is becoming advance that provide us tools to access data and becoming easier to track identity of data(privacy violation). So that, should develop some application with privacy recommendation to provide more security in terms of identification of data, so data cannot be reused or hacked. Some more privacy preserving techniques are available like shadow coding and de-identification. Need to improve this technique due to large scale of data.Another challenge is Data Provenance. It means authenticity, actually big data is large scale data which expand from multiple sources, and so at the time of communication, receiver of multicast data will verify that the received data is coming from original source.

3 RELATED WORK

In Privacy section can be define some encryption technique which generates privacy code, Like: Blowfish, Link Encryption Technique, Two Fish algorithm etc.

This paper going to represent some related work which already done by some author, so, in this current paper some authors are Amine Rahmani1, Abdelmalek Amine2 and Mohamed Reda Hamou3 represent privacy preserving in big data through De-Identification technique [1]. This technique which provide, deleting and masking the data of identifiable information."Reciprocal of de-identification known as the re-identification". Procedure of de-identification has five steps are: 1.Tokenization, 2.Codification, 3.Detection, 4.Storage, 5.Replacement. Privacy preserving is Data Publishing provided information security of data. That focuses on manipulating a owner data set to create greater anonymity while still manipulating the value of dataset.Using this techniques, A data owner could be anonymize a data set and give that anonymized dataset rather than original dataset to each user. Now receiver of that dataset will be able to use the data for meaningful activities but can not know about the,private information about owner.

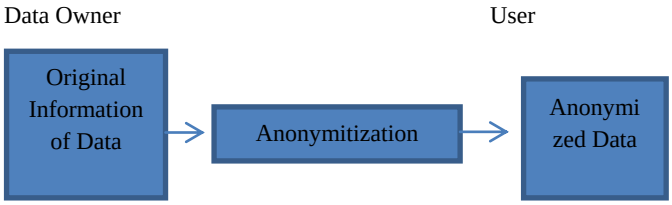


Fig1: Conversion from plaintext into anonymized data

Second paper is Secure Steganography Algorithm Based on Fibonacci Representation which is written by author Nandan Makarand Deval in 2011 [3], so in this paper define two algorithm of steganography one is Spatial domain technique and other is Frequency domain technique. Author used Fibonacci based steganography to hide some data at the time of communication, which follow principle that is, First, Image will be presented in Fibonacci and after that a secret bit will be embedded into the Fibonacci sequence by LSB algorithm.Finally, Fibonacci sequence with secret bit embedded is converted to binary system before being written to stego-image. Steganography algorithm is very simple with very high performance. No need to use all pixel's in an image for hiding data, if embedded secret bit makes the Fibonacci sequences and not satisfying zeckendrof's theorem. So, the robustness of the secret bit is weak, when it is hidden in the LSB.

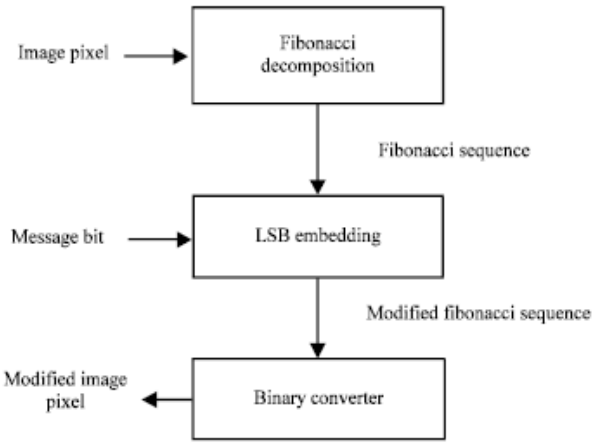


Fig 2: Process to do steganography [3].

and this technique done by three types of analysis one is Perceptual transparency analysis, second is Security analysis and third is robustness analysis.

In 2013, Alvaro A. Cárdenas, Pratyusa K. Manadhata, and Sreeranga P. Rajan have explained some privacy preserving technique for securing Big Data [2], one of them is called apache accumulo is a software project which based on google's big table design and sorted and distributed key/value store is a robust system. It has two features to improve security, Cell Level Security and Server Side Programing.

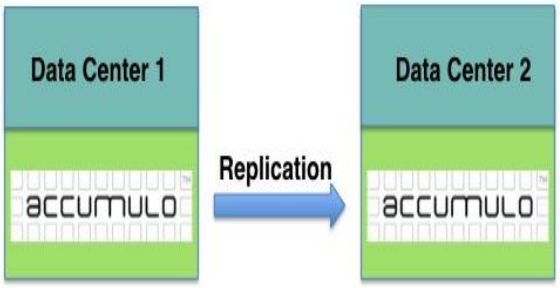


Fig 3: Process to do apache accumulo [2]

In 2015, David Naylor give information about TLS, SSL and HTTPS which provide much better security to Big Data. So this is survey paper about TLS [4], that is simple extension to the OpenSSL library. In this paper author differentiate TLS and mc-TLS. Mc-TLS is extension of TLS with extra features. Transport Layer Security which follow HTTPS protocol, that is standard for end-to-end encryption because its provide some functionality (i) Entity authentication, (ii) Data secrecy, and (iii) Data integrity and authentication. TLS provide us two protocol one is handshake protocol which establishes session and another is record protocol that exchange data with above these three security.

4 EXISTING SOLUTION

This paper describes some technique which uses for Big data security like apache accumulo, De-identification, fibo acci based steganography, TLS and Mc-TLS overview. But now going to describes detail information about Mc-TLS, how its work and how can provide better security than other technique. This paper is going to explain detail information about Mc-TLS, that is a TLS with secure and trusted Middleboxes. It provides secure communication between client and server. Goal of mcTLS is:

1. Encryption

Using HTTPS, provide end-to-end encryption on the web because it ensures

I. Entity authentication

II. Data secrecy

III. Data integrity and authentication, Moreover, it will likely be the default transport protocol for HTTP/2.

2. In-Network Functionality

In-network Functionality is widespread

1. Caching
2. Compression
3. Parental Filter
4. Virus Scanner
5. Packet Pacing

4.1 Mc-TLS Consist:

1. TLS +Middleboxes
2. mcTLS Design Ideas
3. mcTLS Handshake
4. Performance Evaluation

4.1.1 TLS + Middleboxes

It consist two protocol one is **Handshake Protocol** and another is **Record Protocol**. And provide three security property, Entity Authentication, Payload Secrecy and Payload Integrity. This type of security is broken due to its limitation, its designed for two parties, no mechanism to authenticate middlebox, client has no guarantees past middlebox, and middleboxes have full read /write access.

4.1.2 Mc-TLS Design Ideas

mcTLS maintain TLS's security property which is Entity Authentication, Payload Secrecy and Payload Integrity and **also included two** more **property one is visibility control and another is Least Privilege**. Most middleboxes do not need read/write access to all data.Example.

(a) Encrypt the Contexts for access control like send(data, **context**)

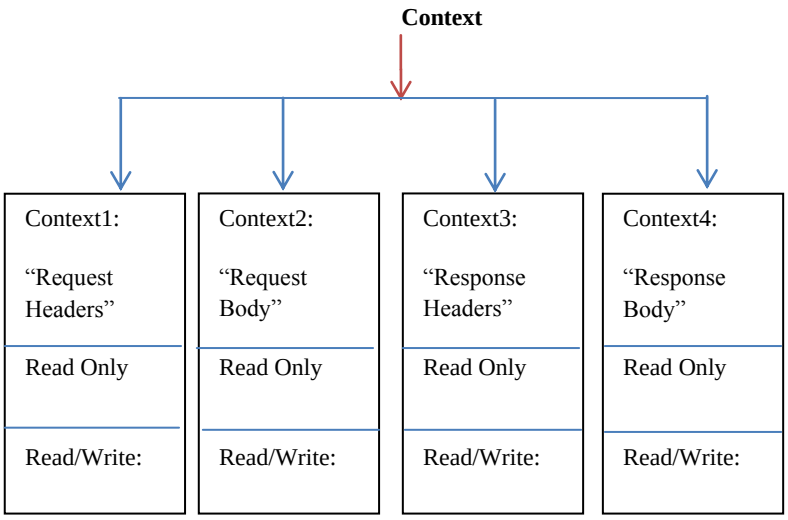
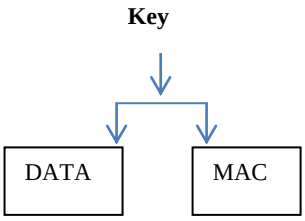


Fig 4: Define encrypted context [4]

In Encryption context TLS uses one key encryption and MAC but mcTLS uses three keys encryption and MAC.

In TLS:



In Mc-TLS:

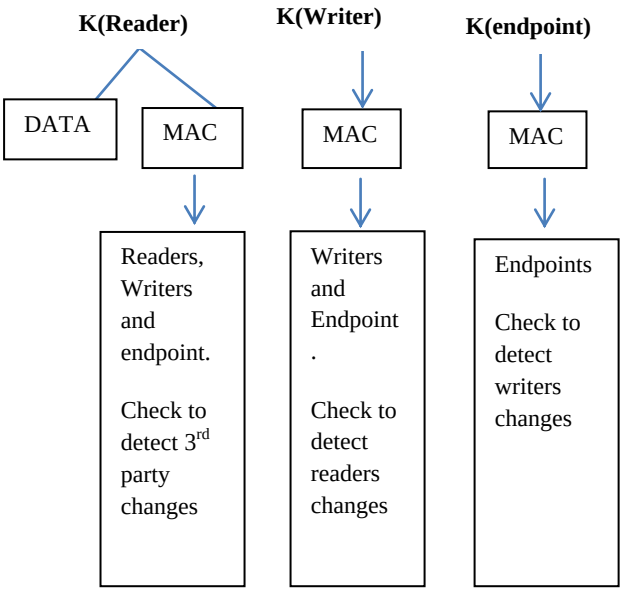


Fig 5: Difference between functionality of TLS and Mc-TLS [4]

Note: Each Context has a read key and a write key

(b). Contributory Context Keys for endpoint agreement: Client and server generate part of each context key.

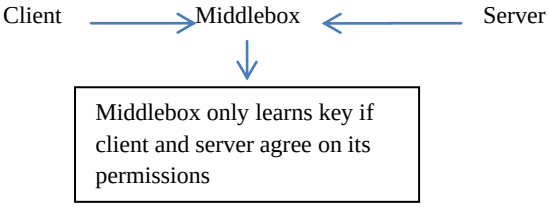


Fig 6: Message passing through authorized middlebox from client to server

4.1.3 Mc-TLS Handshake

Both are Mc-TLS and TLS are same in functionality. Here in this paper going to define purpose of handshake protocol.

- Allow the endpoints to admit on cipher text, a group of encryption context, list of middleboxes with their permission.
- Allow the endpoints for authentication and tota number of middlebox
- Organize shared symmetric key $K_{endpoint}$ between endpoint.
- Organize shared symmetric key $K_{writers}$ for each context in between all of writer and shared symmetric key K_{reader} for each context in between all of reader.

Now this paper going to explain following step which elaborate about Mc-TLS handshake, that has the same 2-RTT "shape" as TLS.

1. Setup
2. Client Hello
3. Certificate and Public Key Exchange
4. Shared Key Computation
5. Context Key Exchange
6. Context Key Computation
7. Finished

4.1.4 Performance Evaluation

Mc-TLS is better access control to creating and distributing keys, computing MACs and sending bigger amount of smaller record than TLS. In this section, define some overhead.

Table2. Evaluate performance in terms of overhead

Overhead	**Description**
Handshake Time	Handshake is not distinctly larger than SplitTLS's or E2E-TLS's
File Transfer Time	File transfer time is not higher than SplitTLS's or E2E-TLS's
Page Load Time	mcTLS has no effect on real world Web page load time
Data Volume	introduces less than 2% additional overhead for web browsing compared to SplitTLS or E2E-TLS.
CPU	mcTLS servers can serve 23%-35% fewer connections per second than SplitTLS, but mcTLS middleboxes can serve 45%-75% more.
Deployement	Enhance an application tomcTLS appears to be straight forward and easy

5. CONCLUSION AND FUTURE SCOPE

This paper describes a lot of security check to establish like apache accumulo, Fibonacci based steganography, De-identification and TLS and mc-TLS. But mc-TLS provide better security rather than other, because mcTLS provide more security between client and server with trusted middleboxes:No Transparent Middleboxes**,** Least Privilege**,** Middlebox Authentication, No Custom Root Certificates.

Currently TLS/SSL used in Hadoop to improve security features. But In this paper has been define some more quality of Mc-TLS rather than TLS/SSL. As per define in this paper, can enhance security after using mc-TLS concept in Hadoop.

6. ACKNOWLEDGMENTS

I am thankful to all faculty who provide me resources and information regarding big data security.

7. REFERENCES

[1] Amine Rahmani1.2015.De-Identification of Textual Data using Immune System for Privacy Preserving in Big Data.

[2] Alvaro A. Cárdenas . 2013. Big Data Analytics for Security., University of Texas at Dallas, Pratyusa K. Manadhata | HP Labs, Sreeranga P. Rajan | Fujitsu Laboratories of America.

[3] Nandan Makarand Deval. 2011. Secure Steganography Algorithm Based on Cellular Automata using Fibonacci Representation and Reverse Circle Cipher Application for Steganography, International Journal of Computer Science and Information Technologies, Vol. 2 (4).

[4] David Naylor?, Kyle Schompy, Matteo Varvelloz, Ilias Leontiadisz, Jeremy Blackburnz,Diego Lopezz, Konstantina Papagiannakiz, Pablo Rodriguez Rodriguezz., Peter Steenkiste.2015. "Multi-Context TLS (mcTLS):Enabling Secure In-Network Functionality in TLS" Carnegie Mellon University Case Western Reserve University Telefónica Research(ACM).

[5] Stephen Kaisler i_SW Corporation. 2013. Big Data: Issues and Challenges Moving Forward. 46th Hawaii International Conference on System Sciences.

[6] John Gantz., David Reinsel.2012. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadow s, and Biggest Grow th in the Far East.

[7] B. Saraladevia., N. Pazhanirajaa., P. Victer Paula., M.S. Saleem Bashab., P. Dhavachelvanc.2015. Big Data and Hadoop-A Study in Security Perspective, 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)(Science Direct).

[8] Roger Schell. 2013. Security – A Big Question for Big Data, University of Southern California, USA, IEEE International Conference on Big Data.

[9] Elisa Bertino., Bharath K. Samanthula.2014. "Security with Privacy - A Research Agenda ", 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaborate Com 2014).

[10] BOUHRIZ Mouniaa., CHAOUI Habibab.2015. Big Data Privacy in Healthcare, The 2nd International Workshop on Privacy and Security in HealthCare (PSCare15).

[11] Koushik Mondal. 2015. Big Data Parallelism: Issues in different X-Information Paradigms, 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[12] Utkarsh Srivastava., Santosh Gopalkrishnan. 2015. Impact of Big Data Analytics on Banking Sector: LearningforIndianBanks. 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).