Analyzing Various Techniques to Safeguard user Sensitive Data

Mrs. K. Pramilarani
Senior Assistant Professor: Department of CSE
New Horizon College of engineering
Bangalore, India

Dr.Vasanthi Kumari P
Associate Professor:
Department of Computer Applications
Dayananda Sagar University
Bangalore, India

***Abstract*— There are several different ways to connect people in and around the world. One such way is the use of internet and interactive communication tools (ICT) such as smart phone, tablet, electronic devices and voice assistant. Each and every second large volume and variety of data will be generated by these devices. This kind of large volume with varying speed and variety of data is called big data. Analyzing and identifying the intelligent pattern from this data is called big data analytics. Several different data analytical methods are used to classify and identify the hidden patterns from the raw data. Data is scattered in different systems in different places. Hadoop is the open source platform for processing such distributed data. Hadoop uses the HDFS and map reduce to do distributed computation in secured way. Even though hadoop and map reduce will impose securityin its own way, the data collected and processed need more secured methods to prevent the data leakage. There are several different methods, tools and techniques available to impose the security and prevent the privacy of any user's identity. This paper will provide the overall view about the challenges faced during the data storage, data generation and processing in distributed and cloud environment. This paper is divided into 4 sections. Introduction part of map reduce, related works and challenges and vulnerability of data security are discussed in first and second sections. Different techniques used to impose the security is analyzed in the third section and fourth section is the conclusion part.**

Keywords—Hadoop; ICT; mapreduce; big data; securit; data leakage; HDFS

I. INTRODUCTION

Data generated by Interactive Communication Tool(ICT) and other devices to communicate with others grow exponentially. Conventional methods will not be useful to handle such large data. The main problem to handle such data is its variety, structure and its speed of transfer. Due to the nature of the data different kinds of analytic tools and techniques are used to identify the hidden unknown patterns from the unstructured and unorganized data set. Processing this different variety of large data is known as big data. Data involved in different stages such as generation, storage and processing is called as big data life cycle. Big data processing will be used in business, healthcare, educational institution, government and private sectors. This large data cannot be stored and processed in a single system due to its distributed nature. So cloud concept is involved here. When the cloud platform is used, the vendor of the data might not have the control over it because data cannot be stored in one place or in one system. There are different kinds of clouds available to store data such as private, public and hybrid. If the data is shared to third party with or without the knowledge of owner of the data then there are more chance to loose or misuse user's confidential data. Many more security layers should be implemented to prevent the data loss or data misuse. Data processing in big data is not the easy one due to its nature. Since the data is having different resources like you tube, twitter and internet, processing the information also needs more care. Big data will be processed by the Apache Hadoop, a open source platform for distributed computing. Hadoop platform includes hadoop distributed file system called as HDFS, map reduce, Hive and Zookeeper with their own security measure. Hadoop is used by Google, Amazon and yahoo.

Even though some analytic methods are used to identify hidden pattern from the information, it may also lead to privacy breach. Consider an example of online teaching during this pandemic situation. All the educational institutions conducted online classes during this covid19 period. Many more organizations conducted the online meet and webinars. This sort of data sharing put a danger to the client's confidential personal data. A recent study showed that more user's personal information is stolen during this online data sharing.

If the sensitive information about the users are stored and processed with less safety measure will lead to the data loss or data misuse by the hackers. Several different mechanisms are developed and used to protect the data against threat.

Depends upon the stages of big data cycle different techniques and procedures are used to prevent the data loss. Few such techniques are falsifying the data, access restriction, access limitation, data distortion and encryption

II. RELATED WORKS

Mohammed Asrar Naveed [9] discussed the challenges and issues related to bigdata in his paper. He also explained about the security in four levels such as network level, authentication, data level and generic issues. He has discussed multiple technologies to improve the security of the data in his work.

Gurjith singh bhathal [5] discussed the hadoop frame work vulnerability and security threats, attacks and he has provided some possible solutions to reduce or eliminate vulnerability in his paper

Katarina Grolinger[9] explained about data processing on a large number of commodity nodes. She also analyzed issues and challenges MapReduce faces when dealing with big data according to four main task types such as data storage, analytics, online processing, and security and privacy. Some concepts to improve and extend MapReduce is also suggested in the paper.

N.Madhusudhana reddy[14] proposed new method to identify the rogue nodes by analyzing the hadoop log files. The rogue nodes are the one which disturbs the functioning of map reduce.

Keerthana rajendran[11] analyzed the principle of anonymization techniques such as k anonymity, l diversity and t closeness. She also discussed the pros and cons of these techniques in her paper along with other authors.

In the literature review many authors proposed many different solution to avoid or reduce the security while transmitting data in the distributed environment in one or other way.

A. Map Reduce(MR)

Google has introduced the map reduce to analyze the huge volume of data in the distributed network. Map reduce has its own security measure to prevent the data loss. There are two tasks in map reduce. One is map task which will split the input and map it and another one is reduce task which will shuffle the input and reduce it to one output . Map reduce uses the key value pair concepts for processing data in parallel. Whenever MR frame work receives the job from distributed network it will divide the given input into several small chunks of input called input split. After the input split the splited input is passed to a map function where the output vales are produced. Map tasks will split the input and map the input in parallel. The map task will be assigned to different nodes in the network. The reduce task will receive input from the map function. The output of the map function will be given as the input to shuffle function to map and consolidate the relevant records. The output from the shuffle function is aggregated. It combines the output from shuffle and returns a single output. The execution of map and reduce function is controlled by Job tracker which acts as a master, one who is responsible for complete execution and multiple task trackers to act like slaves for performing the job assigned to them by the master job tracker. For every submitted job there is one job tracker which will be present in Name node and multiple task trackers which will be present in data node. Task tracker sends the progress report to job tracker to avoid the data loss or node failure. Map reduce function will be disturbed by the hackers by changing the data or by providing the wrong data. Hackers may make the map or reduce function to malfunction. So it is mandatory to identify the malfunctioned node to avoid data loss. There are several methods to identify the malicious mapper. One such way is to monitor the log file and action can be taken according to that.

B. Challenges

Identification of personal information at the time of transmission of data over the internet by the unknown person is the main issue which is to be solved for maintaining the security.

Information privacy is the privilege to have some authority over how the data is collected, processed and utilized. Privacy's main concern is to set up the policy for ensuring the confidentiality of the user sensitive data.

Security safeguards the data from the attackers. Attackers are the one who steal the user's data without their knowledge. Security ensures the confidentiality against unauthorized intervention of the hackers. Security concentrates more on shielding information from malicious attackers. While using the privacy preserving algorithms and techniques one should know the appropriate usage of their data, how safe their data is, where their data is moved or stored. Similarly when security is concerned, users should know the secrecy, uprightness and accessibility of their information. Both security and privacy is compulsory for the data transfer. It is very difficult to have proper protection without security and proper security without protection.

Consider the following cases for information security and data leakage. In the year 2006, [22]American On Line Web portal(AOL) discharged nearly 20 million quest inquiries for around 650 clients by hiding their identity and IP address for some research work. But within two days researchers identified the id's with their analytical skill by correlating with the available data.

In similar way medical information about the patient in healthcare could be identified by combining their data with the externally available data such as voter ID or other data from the government website to reveal the patients confidential information. This can be done by means of some analytical methods or techniques.

Personal shopping habits of the users may be used to expose the user's individual characteristics, habits and taste. These are the few cases where the user's data can be hacked or misused by the third party for their business tricks.

There are several different mechanisms used and developed for saving the user's personal sensitive data. These mechanisms are grouped under three stages of big data life cycle. Protecting data is a difficult process in the distributed environment. Because sensitive information from the users will be hacked with or without the knowledge of users.

One simple way to protect the information is to use the combination of private and public clouds. Non sensitive information can be kept in public clouds where there is no control over data. Sensitive information can be kept in private cloud where only the authorized user will get the access. Data encryption and decryption techniques will be used in cloud to protect the data.

Unauthorized usage of data should be avoided. Users should have the capability to check their reliability of data without retrieving it from the cloud at the data storage stage itself. At the time of data processing user should ensure that their confidential information are not disclosed to unauthorized users. While extracting the meaningful information or hidden patterns from the input data, privacy should not be violated at any cost.

There are several different challenges in big data while preserving privacy and security. [22] Cloud Security Alliance (CSA) highlighted top ten big data-specific security and privacy challenges[4] into four broad category. They are Infrastructure security, data privacy, data management [4], integrity and reactive security. Security in distributed programming frameworks, security in non-relational data stores [4], transactions logs and data storage security [4], input filtering or validation, real-time compliance monitoring, scalable and composable privacy-protecting data mining and analytics, cryptographically enforced access control and secure communication [4], granular access control, granular audits, data provenance and so on are the sub categories.

III. DIFFERENT TECHNIQUES AND SOLUTION FOR PRESERVING AND PROTECTING THE SENSITIVE INFORMATION

Traditional methods are used to maintain the privacy to certain level but full privacy cannot be achieved by the old traditional methods. Several new techniques are developed by overcoming the demerits of traditional methods. Few such methods are listed below.

There are different ways to ensure the security of information in entire life cycle of big data. During the first phase of life cycle data will be generated from various sources. Access restriction and falsifying the data methods will be used to protect the user's privacy from the hackers by giving permission only to authorized users. Some distortion methods can be applied to alter the original data so that the hackers cannot get the real data.

Encryption methods can be included to make the data secure in storage stage. Many different encryption based systems such as attribute based, identity based and storage path encryption and homographic encryption can be used to safe guard the data.

Knowledge will be extracted and privacy will be preserved using privacy preserving data publishing (PPDP) algorithms during the final processing stage of life cycle.

Anonymization techniques are employed to secure the PPDP. By using generalization or suppression methods data can be anonymized. Several different tools and techniques can be employed in knowledge extraction process to mine useful hidden information from large and complex data set.

Association rule mining, clustering, classifications are few example for such mechanisms. Clustering and classification partitions the input data into separate small groups. To identify the relationships and style from the input data association rule mining technique will be used. These are all few different ways to protect the privacy of data in big data life cycle.

A. De identification

In this method sensitive data will be anonymized by generalization and suppression methods. Few values in original data will be replaced with some other values to give generalized data.

After generalization data can be suppressed [2] by without releasing few values. There are three famous techniques to do that. They are K –anonymity, L-diversity and T-closeness for de identification. Some important terms used in de-identification are identifier which directly identifies person by their unique id such as aadhar id, pan number etc., quasi identifier attributes identifies the person partially by their gender, age, ZIP etc., sensitive attributes hold the personal private information such as salary and insensitive attributes are the one which are general and safe in nature. If the data is anonymized with the same values on the quasi-identifiers in all record then it is called equivalence classes.

In k-anonymity technique privacy is preserved by making a single row of information identical to a minimum of (k – 1) rows which makes distinct records in a specific dataset cannot be identified. So k-anonymity technique can be used to prevent database leakages. This is achieved by generalization and suppression. k-anonymity is prone to certain privacy attacks .One such attack is homogeneity attack, which will be addressed by using l-diversity technique.

In l diversity a generalized quasi-identifier (q*)-block will have a minimum of 'l' properly depicted values under the sensitive attribute (S). If every q*-block is l-diverse, then the table is also l diversed table. Entropy l-diversity, recursive l diversity and probablistics l diversity are the extended versions of l diversity. L-diversity method will be vulnerable to skewness attack and similarity attack.

T-closeness is used to prevent privacy by addressing the limitations in the existing k-anonymity and l diversity methods. In t-closeness the distance between the sensitive attribute of an equivalence class is compared with the whole table with certain threshold value to identify whether the table is having t-closeness or not. This reduces the risk of identifying unique information of an individual. The distribution distance between the sensitive attributes is measured using the metric called Earth Mover's Distance (EMD) that considers the semantic proximity of the attribute. The dis advantage of t-closeness is the usage of EMD. It is very hard use EMD to identify the closeness between t-value and the knowledge gained.

B. Data Masking

Data masking is the other way to anonymize the data by hiding data with some other values by character shuffling or encryption or word or character substitution.

C. Data swapping

Data swapping also known as shuffling and permutations which is used to preserve the original records by Swapping attributes that contain identifiers values such as date of birth correspond to original records.

D. Pseudonymization

Pseudonymization is another method to manage data with de identification method. In this method fake or pseudo identifiers are used to replace the data. This modified data can be used for training and testing by preventing the originality of data.

E. Data perturbation

Data perturbation is the another anonymization technique which modifies the original dataset slightly by applying techniques that round numbers by adding some random noise. A small base may lead to weak anonymization at the same time large base could reduce the usage of the dataset.

F. Multidimensional Sensitivity Based Anonymization(MSBA)

MSBA is the improved version of anonymization. This technique will be best suited if the quasi identifier are predefined and the data set is large with less loss of information. The data will be split and kept in different bags with the help of filters in Apache Pig scripting language and based on the probability distribution of the quasi identifiers. Sensitive attributes are represented in class and bottom up generalization will be used on a set of attribute with certain class value. In this approach data distribution will be very effectively done compared to block method. Background knowledge attack will not be possible due vertical partitioning of data in different groups and more over it is very highly impossible to map the data with external sources to identify any person specific information. Due to its scripting nature the code development in pig is very easy and effortless. But at the same time code efficiency is less in Pig when compared to Map Reduce job and due to that Apache Pig script has to be converted into a Map Reduce job to enhance the efficiency. MSBA is best suited if the data is more and at rest and not suited if it is a real time data.

G. *Cryptograhic techniques*

In Cryptographic techniques the data will be encrypted using any encryption technique before transferring it to the cloud. Encryption is the way to modify the original data in source end. There are several ways to use cryptographic technique starting from very simple to very complex. In this technique original data will be replaced by new data by means of some cryptographic algorithm at the source side. This kind of change is called encryption. The other side that is at the destination side again data will be modified back to its original content by decrypting it so that no one can see the original data. But conventional methods cannot be used for large complex data set.

H. *Differential Privacy*

In Differential privacy techniques some calculations will be applied in the form of noise on the data without sharing their sources to preserve privacy. Usually anonymity diminishes the value and meaning of data. But in this technique if only one data is not available out of 100 dataset then also this model will not allow the hackers to deduce the data with the available 99 data. This method adds random noise to the aggregate data to preserve the privacy of data, but it will have very little effect on the pattern

I. *Synthetic data method*

Synthetic data method is the one which creates the artificial data set without any connection with the real data set. Statistical models such as standard deviations, medians, linear regression or other statistical techniques are used to generate the synthetic data which will be applied based on patterns found in the original dataset.

J. *Data distribution*

In Data distribution technique the information is circulated crosswise either in horizontal or vertical over many destinations nodes. If the information is passed crosswise manner over several different destination with or without supervision of variety of associations with same quality depends upon whether they are horizontal or vertical distribution.

IV. CONCLUSION

Large amount of person specific data is collected by the government and private organizations for various purposes. So data extraction and processing is inevitable for the data analytics to identify the hidden patterns for various reasons. Therefore data transmission and storage through clouds is also mandatory for large data set. Active and passive data collection happens during processing. So protecting data and privacy of users are very important. Vulnerability issues should be handled properly by using proper techniques and tools. The data transmitted over internet should reach authorized persons only. There are several levels of security needed to safeguard the data. The challenges faced while transmitting the data and possible ways to prevent the data loss by means of de identification and anonymization techniques with their pros and cons are discussed in this paper. Single method will not be sufficient to protect the sensitive data effectively. Several different layers of security is needed. Security must be imposed in the application level by using firewalls to protect the application from the unwanted users. Similarly security must be implemented in data level by applying techniques such as anonymization , data distribution or de identification and so on. So that the sensitive information will be preserved from hackers. Vulnerability discovery should be done by means of some techniques to identify and avoid the hidden risk. Finally security must be implemented in cloud level also. Users should use the cloud which will provide more security and preserve the user’s information to avoid unnecessary issues.

REFERENCES

[1] Abid Mehmood, "Protection of Big Data Privacy," IEEE ACCESS, pp. 1821 -1834, Volume 4 ,2016

[2] A Mehmood, , Y. Xiang, G. Hua and Song, I. Natgunanathan "Protection of Big Data Privacy," Protection of Big Data Privacy, p. 14, 2016

[3] C S. José Moura, "Security and Privacy Issues of Big Data".

[4] Ibrahim Abaker Targio Hashem, "MapReduce Review and open challenges Springer Scientometrics , p. 389–422, (2016) 109.

[5] Gurjit singh Bhathal,Amardeep Singh,”Big data :Hadoop framework vulnerabilities,issues and attacks,ElsevierArray1-2,JULY 2019,

[6] Katal " Big data: Issues, challenges tools and good Practices," in Proc. of IEEE Int. Conf. on Contemporary Computing , Aug. 2013

[7] Ketaki S. Pathak1, Pratima Bhalekar2,” security issues associated with big data in cloud computing,ICETEMR-16

[8] Katarina Grolinger, Michael Hayes, Wilson A. Higashino,Alexandra L'Heureux,” Challenges for MapReduce in Big Data , Proc. of the IEEE 10th 2014 World Congress on Services (SERVICES 2014), Alaska, USA, June 27-July 2, 1014

[9] Kavya Krishnan and K.Pramilarani ,”A Combination of AES and Key Hash Message Algorithm to Secure Data in Cloud”,International Journal of Scientific Research in Computer Science Applications and Management Studies,2018

[10] Keerthana Rajendran, Manoj Jayabalan, Muhammad Ehsan Rana,” A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data , IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.12, December 2017 pg.172-177

[11] Mohammed Asrar Naveed, Chaitra B,” Security Issues Associated with Big Data,” International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181,Published by, www.ijert.org,NCRTS-2015 Conference Proceedings

[12] Md Maminur Islam1, Subash Poudyal1 and Kishor Datta Gupta,” mapreduce implementation for malicious Websites classification,, International Journal of Network Security & Its Applications (IJNSA) Vol. 11, No.5, September 2019,P 27-35

[13] N. Madhusudhana reddy,dr. C. Nagaraju, Dr. A.Ananda rao." Toward secure computations in distributed programming frameworks: finding rogue nodes through hadoop logs, Journal of Theoretical and Applied Information Technology 15th December 2017. Vol.95. No 23

[14] N. Madhusudhanan Reddy, "protecting privacy of big data in presence of untrusted mapper and reducer," Indian Journal of Computer Science and Engineering IJCSE, pp. 201 -209, 2017

[15] K.Pramilarani ,"Social Media and Big Data"International Journal for Research and Development in Technology,Pg. 39 -42,2017

[16] M. Vijay Prakash, "A New proposal for distributed system security Framework," in AASRI conference on Parallel and Distributed System Security Framework, 2013

[17] Priyanka Jain, "Big data privacy: a technological," Springer open ,Journal of Bigdata, p. 25, 2016.

[18] Philip Deribekoa, "Security and Privacy Aspects in MapReduce on Clouds: A Survey," Elsevier Computer Science Review, p. 42, May 4, 2016

[19] Ram Mohana Rao, "Privacy preservation techniques in big," Springer open ,A journal of Big Data, p. 12 pages, 2018.

[20] Celeste Murnal and K.Pramilarani ,"Secure data Transfer", INTERNATIONAL JOURNAL OF INFORMATION AND COMPUTING SCIENCE,Pg.341-350

[21] Top Ten Big Data Security and Privacy Challenges," Cloud Security Alliance, pp. 1-11, november 2012

[22] Yanfengi Zhang, "i2MapReduce: Incremental MapReduce," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,, pp. 1906-1919, volume 27 ,2015.