Efficient Security Approach for Sharing Sensitive Data on Big Data Platform: User Trustworthiness and Multi Factor Authentication

[1]Sona Saxena, [2]Prof. Krunal Vaghela
[1]M.Tech scholar, [2]HOD CE/IT
Computer Science Engineering,
RK University, Rajkot, Gujarat, India.

Abstract **- Big Data is a buzzing word in today's computer science industry. Every organization be it big or small generated enormous amount of data on daily basis. This data that keeps on generating by each and every source on the internet can be termed as Big Data. This data has to be stored and analyzed for various purposes. Since the amount of data is very large and always on increasing basis, its security is major concern for its users. In this paper, first the introduction to big data is given along with the five V's of Big Data. Importance of security in Big Data is then determined which forms the basis of the research here. The two security approaches are then stated which includes user trustworthiness and multi factor authentication.**

Index terms**- Big Data, security, cloud, user trustworthiness, multi factor authentication, Big Data security.**

1. INTRODUCTION

Big Data refers to extremely large data sets that keep on generating by each and every process digitally. Big Data can be both structured and unstructured. The structured data follows fixed format and is relatively easy to process and use by the computer programs whereas the unstructured data does not follow any format and is difficult to process. The examples of structured data are excel spreadsheets, word files etc. and images, videos, online posts come into category of unstructured data.
The examples of generation of Big Data includes social platforms like Facebook, Twitter, Instagram and search engines like Google and Yahoo that generated Petabytes of data every day. For example, each user action action on Facebook that is making a profile, updating it, liking pictures of friends, posting something etc and all related actions are stored by Facebook in their database and is made available to user whenever required. And in case of Google, it stores data related to each user like its email id, actions performed on the email portal, search enquiries made on search engines etc. Then, these companies uses the user data to better understand the behavior of the user and making business plans accordingly.
The amount of data is not important in Big Data, what organizations do with that data is important. Various organizations keep the big data and uses it for various operations. Big Data can be analyzed for insights that lead to better decisions and strategic business moves. Thereby, Big Data is a buzzing field which is used everywhere in today's digital world.
But handling such enormous amount of data with keeping two things in mind i.e. in desirable time and gaining insights into data is very difficult. Traditional database system can not handle Big Data due to its large volume and carried structure. Also, if we use traditional system to process Big Data the processing time will be very large and it will be of no use by then. So, a different processing system needed to be designed for handling Big Data. Doug Cutting and Mike Cafarella invented Hadoop and the name became Hadoop after his son 's toy elephant.

2. 5V'S OF BIG DATA

a. Volume - Volume implies the amount of data that is being generated and the amount stored for analysis of BD.
b. Variety - The type and nature of the data captured is called variety. This helps the analyst to take better decisions for every type of data captured for eg. text, audio, video and images.
c. Velocity - The speed at which data is being generated by the users is called velocity of data.
d. Variability - Inconsistency or missing data can be a problem for analyst. This is called variability of Big Data.
e. Veracity - The quality of captured data can vary in many aspects which can result in false results.

3. LITERATURE REVIEW

B. Saraladevi, N. Pazhaniraja, P. Victor Paul, M.S Saleem Basha, P. Dhavachelvan (2015) - The basic architecture and structure of Hadoop is studied. Main concern lies in the security perspective of hadoop and three approaches- Kerberoes, bull's eye and name node approach is studied.
Pedro H.B. Las-Casas, Vinicius Santos Dias, Wagner Meira Jr. and Dorvigal Guedes (2016) – Attacks on Big Data is studied and its correlation to cyber security is stated here. The problem of phising in emails are studied and then solution is stated.

Ather Sharif, Sarah Coongy, Shengqi Gong, Drew Vitek (2015) - Levels of security of Big Data is first defined and then a new approach called Sticky Policy Framework is defined.
Aayush Gupta, Ketan Pandhi, P V Bindu, P Santhi Thailgam (2015) – The breaches in the cloud security is first studied. Security in an organization using a private cloud is enhanced using role and access based data segregator.
R. Barco, L. Dez, V. Wille, and P. Lzaro (2009) - The concept of self protecting documents is stated. Every document before sending it via cloud are first encrypted using any of the algorithm and then transported.

4. IMPORTANCE OF BIG DATA SECURITY

Various organizations generate data at all aspects of their working hierarchy at regular intervals. The data is being generated every day which keeps on growing larger and larger day by day, which is called BIG Data. But all of this data cannot be stored in traditional systems and at organizations. So, Cloud Technology comes into picture. Organization hire cloud servers to store their Big Data. But, since a single cloud are shared by many organizations at the same time, security of data is of major concern in cloud. Security is a major concern on all Big Data platforms. Organizations loses control of their data once they hire cloud providers for storing and processing their data. One, they are unaware at what physical location of their data stored and Second, they have to fully trust cloud providers for their data. Therefore, security is of major concern while dealing with Big Data.

5. BIG DATA SECURITY APPROACHES

5.1 USER TRUSTWORTHINESS

The users of the data that organizations stores as Big Data on cloud are varied. In a typical hierarchy of any organization, Big Data are being used by each and every employee of that company ranging from the top level employee to the lowest level. Each user uses the data for which he is authorized to (access control). Access control is a mechanism through which it is being stated that which user can access what data of the organizations. Like the top level employee is granted access to highly confidential data whereas the lowest level employee is not granted access to that data.
Here we are stating a mechanism called User Trustworthiness to secure the data in motion (data on the go). There is a trust factor associated with each user of the data. Let the trust factor is denoted by (T). The factors on which user is considered as trustworthy includes-

1. The level (L) of hierarchy of the user in an organization (higher the level of user in hierarchy, more trustworthy the user is)

$$L \alpha T$$

2. The level of confidentiality (C) of data. (more confidential data to be accessed by more trustworthy user)

$$C \alpha T$$

3. Number of times (NT) the user has accessed same data. (more number of access, more trustworthy user is)

$$NT \alpha T$$

4. Number of failed transactions (FT) to access the data. (more number of failed credentials, less trustworthy user is)

$$FT \alpha 1/T$$

By applying these rules on checking the user credibility, the data can be kept more secure in an open environment like that in cloud. Before granting access to any user for the data required, all these above factors should be checked. All the data related to user behavior is stored in one database (like position in hierarchy, security aspect of data, number of times user has accessed the data) and whenever required must be matched with the database and then the access is granted.

5.2 MULTI-FACTOR AUTHENTICATION

It is a method of computer access control in which a user is granted access only after successfully presenting several separate pieces of evidence to an authentication mechanism. MFA is a process through which a user is only granted access to any secured data only when he provides required information. The various information required by a system to grant access to user depends upon the security requirement of the data. Depending upon the security requirements there can be 2 factor or more than 2 factor authentication applied. They are:
a) Something you know, such as a password or passphrase -This method involves verification of information that a user provides, such as a password/passphrase, PIN, or the answers to secret questions (challenge-response).
b) Something you have, such as a token device or smartcard. This method involves verification of a specific item a user has in their possession, such as a physical or logical security token, a one-time password (OTP) token, an employee access card, or a phone's SIM card. For mobile authentication, a smartphone often provides the possession factor in conjunction with an OTP app or a cryptographic material (i.e., certificate or a key) residing on the device.
c) Something you are, such as a biometric- This method involves verification of characteristics inherent to the individual, such as via retina scans, iris scans, fingerprint scans, finger vein scans, facial recognition, voice recognition, hand geometry, and even earlobe geometry.
There may be other types of information also which can contribute in Multi Factor Authentication such as location and time of the user. At least two of the above factors must be used for authentication.

The above mechanism stated can also be used in securing data on big data platforms. The organizations store all of its data on Big Data platforms and in turn they are stored in cloud technology. The cloud providers such as Amazon and Google store the user's data in their various data centers around the world. The data as when required by the user is presented to it b accessing the cloud. Since the data is stored ad a distant place and many users of the same organization can access the data, the data needs to be protected. MFA can play a vital role in securing the Big Data from malicious access.
In this scheme, before the access is provided to any user for the required data the authentication of user is checked on various parameters. The number and type of parameters can be defined by the security analyst of the organization depending upon the data and its security needs.
Currently, in Big Data authentication is provided based on passwords and other traditional systems. But passwords do not provide strong authentication for highly secure data. Two factor authentication is also being used for Big Data authentication. Many organizations use biometrics of their employees to authenticate them to use the data in cloud. But only biometrics does not serve the purpose as it is a slow process and hackers can hack the database containing the biometrics of the employee and then can use them to access the data on their behalf.

How the authentication process will work?

The core principle of authentication is- match user inputs with the available data in the system. The different authentication systems are described below:

a) In the password-based system, the password provided by the user is usually matched with that stored in the database in an encrypted format earlier.
b) In the multi-factor system, the system matches multiple passwords — some of which are stored in the database and the remaining dynamically generated — with the inputs provided during the access request.
c) In the biometric system, the system collects data from a person's voice, fingerprints or iris and uses that data to authenticate the user.
d) In the big-data-based system, the system creates a profile of the user based on the data it regularly collects. It authenticates access requests by matching access inputs with the data in the profile.

6. PROPOSED SYSTEM

The authentication system first created the profile of all the valid users containing their information. The information is collected on regular basis and the database is then updated. Every time the user wants to access the data, firstly its information is first matched with that of the database and then the access is granted based on the successful matching. The various criteria for defining the user's behavior are:

a) Information entering behavior- user using physical or virtual keyboard. (virtual keyboard is more secure than physical)
b) Level of security permissions user have.
c) Number of attempts user takes to enter the correct password (more number of attempts less trustworthy user is)
d) Number of times user access the system in a day.
e) The number of times user have reset the password in past.

7. CONCLUSION

User trustworthiness and Multi Factor Authentication can play a big role in enhancing the security of Big Data stored in public datacenters of cloud. Both these parameters will check the credentials of users before granting them access to data. Both these methods can be applied in any type of cloud and is suitable for every organization. The database of user's information if created properly will enhance the security of big data stored in cloud to a great extent as much of the malicious access is prevented.

8. FUTURE WORK

The above proposed system can be implemented in Big Data. Firstly, a database of records of all the users containing the information about them is created. The database should be itself created with or without intimation to the user. The database should also be self updated every time the user logs into the system. The information in the database contains all the above credentials of users on which the user's trustworthiness and multi factor authentication is being judged. As soon as, the user demands for a particular data from the cloud, the above rules are first applied on the user's credentials and then only after successful matching of information, access is granted. Further, ways to decrease login time of user should be made to decrease by using faster method of authentication.

9. REFERENCES

[1] B. Saraladevi, N. Pazhaniraja, P. Victor Paul, M.S Saleem Basha, P. Dhavachel-van, Big Data and hadoop- a study in security perspective, 2015
[2] Pedro H.B. Las-Casas, Vinicius Santos Dias, Wagner Meira Jr. and Dorvigal Guedes, A big data architecture for security data and its application to phishing characterization, 2016.
[3] Ather Sharif, Sarah Coongy, Shengqi Gong, Drew Vitek, Current security threats and prevention measures relating to cloud servicees, Hadoop concurrent processing and big data,2015.

[4] Aayush Gupta, Ketan Pandhi, P V Bindu, P Santhi Thailgam, International Conference on Emerging Trends in Engineering, Science and Technology- 2015.
[5] R. Barco, L. Dez, V. Wille, and P. Lzaro, "Automatic diagnosis of mobile communication networks under imprecise parameters," Expert Syst. Appl., vol. 36.