Analyzing and Overcoming Challenges in Big Data Security

Pramila Joshi

Computer Science Department, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India
Ext Centre Noida, India

Abstract: **This is the era of Big Data. Everywhere in almost all the organizations huge volumes of data is being gathered and processed to generate information. As the amount of data is growing so is the need to keep it secure. Any ignorance in this part may result in damaging the reputation of the organization and incurring a huge financial loss. It can bring down a company to its worse state. We will discuss various big data security challenges in this paper. We will also see how we can provide more security to big data and also discuss the future course of action as far as data security is concerned.**

I. INTRODUCTION

Era of Big Data has come. Everywhere in almost all the organizations huge volumes of data is being gathered and processed to generate information. As the data in the structured and unstructured form increases in volume, it becomes more and more complex and this complexity is compounded by addition of new applications almost on daily basis.[2, 3, 4].

With data volume is growing exponentially day by day need for big data repositories is also growing, where this massive amount of data can be kept and processed. One cannot imagine having a competitive edge in the business market without the help of big data. [1] Big data helps them gain greater insight by highlighting patterns which usually is impossible to come out from smaller data sets.

Lot of new information tools and techniques are being introduced in that enable organizations to have insights into the massive volume of big data and gives them an edge over others. But this in turn faces challenges of safety and security of the data. Keeping big data secure is becoming more and more challenging.

As the amount of data is growing so is the need to keep it secure. Any ignorance in this part may result in damaging the reputation of the organization and incurring a huge financial loss. It can bring down a company to its worse state.

Cyber attackers are the ones who are being benefitted by the rise in big data phenomenon. Any compromise in IT security in an organization becomes a tempting target from cyber attackers. Breaching into the security of big data repositories may benefit the cyber attackers for they might get big recognitions and other monetary gains. At the same time it might have catastrophic effects on the companies as their data may be valuable in the sense that it may be highly confidential and private and important.

We will discuss various big data security challenges in this paper. We will also see how we can provide more security to big data and also discuss the future course of action as far as data security is concerned.

II. BIG DATA SECURITY CHALLENGES

Hadoop offers solutions to all major big data problems like where to store and how to process the enormous amount of data. Its biggest advantage lies in its scalability feature which it incorporates by using commodity hardware. Using Hadoop as Big data platform may also introduce new threats to the security of the very same data being stored in Hadoop.

Entire data from various sources is collected, collated and stored in one huge data pool from where users from various organizations can access it in real time. The stored data can have sensitive private and confidential information of the customers which if leaked to wrong hands may damage the privacy of users.

Although companies do install various security measures such as antivirus programs and firewalls, still they face the problem of Big data security breach. The reason is all the said security measure were developed keeping in mind that the data resides on the hard disk. In this case of big data, our data crosses the boundaries of hard disk and other stand alone systems, rather it travels much farther such as in memory data grids and various cloud storage systems. Moreover Hadoop continues to evolve and so are the big data security challenges.

Following points have been analyzed.

2.1. The Data

The enormous size of big data itself poses threat to its security. The biggest challenge has come in terms of excessive information overload out of diverse and multiple data sources, different data formats thus resulting in massiveness of data volume. This can be analogous to stove piping where improvements can be considered under the limited scope of present technology. Although lot of research work is going on in the fields of cyber security , data security , counter-terrorism, and crime-fighting applications and big data still a consistent framework is missing to address all these challenges. [6]

Other Big data challenges come in terms of its other attributes like velocity, variety, veracity and volume. Since the sources of big data are not tracked, validated and are not known many times, it is a big danger to security. Therefore it is advisable to make it a practice that whenever at the time of receiving the data its authenticity and accuracy is validated. Another challenge comes in terms of multiple resources of big data, which gives rise to secure access management challenge. Various heterogeneous data resources follow their own security policies and have their own set of access restrictions. When these data sources come together for big data repositories it becomes difficult to manage and balance the various security issues. Protecting privacy of the users from unethical data miners who gather personal information without permission is another huge challenge in big data security. Therefore it is recommended to have routine detailed audits to protect the unauthorized access to private information of users without any proper notification.

2.2. NoSQL

Non-relational databases (NoSQL) are great as far as big data is concerned but they lack security feature. NoSQL offers solutions to manage large volumes of data and their no-cost attribute and high performance has compelled many organizations to adopt them as a solution for their big data. However it is observed that switching from RDBMS to NoSQL without taking into consideration proper security implications is not good. NoSQL being an important tool though is still in its evolutionary stage and is highly prone to data attacks. Having a scalable feature with the use of commodity hardware also adds to the security woes. Adding a new security layer to the NoSQL application software of the organization is highly recommended.

2.3. Data Anonymity

At the time of procurement of big data organizations should be able to decide the importance and utility of the data which is to be kept in the data repository. Also maintaining the privacy of data is equally important. Data must be anonymised adequately by removing its unique identifiers. Still data security is a huge challenge as de-anonymization techniques and crass reference procedures can be used by data burglars to capture the data in its original form.

2.4. Data Encryption

Query processing on encrypted data is another strong challenge for big data security. The current technology doesn’t support data access in the encrypted form. Queries, in the encrypted data whether structured or unstructured have to be decrypted first, which due to vast volumes of data involved is time consuming thus delaying the processing. [10]

Data procurement also comes with the challenge of encrypting the data. In the situation where operations are to be performed by the cloud over the data for generating results, data cannot be transmitted in encrypted format. “Fully Homomorphic Encryption” (FHE) provides a solution for this problem. Here cloud performs operations on encrypted data and generates encrypted output only. At the time of using the output the result is decrypted back.

2.5. Data ownership and access control

The term access control refers to restriction access to a place or property or data. The term can be used in the same meaning as being used in day today life. The way we secure our property or house by employing guards or use mechanical means like locks to achieve physical access control, in the same way access control mechanism is applied to big data. Not much has been done in this area so far to protect big data. Although Access control and network security mechanisms are adopted but they become outdated very soon. Sometimes applying access control strictly results in dividing confidentiality levels in the organization thus discouraging the practice. Nowadays new concepts based on Attribute Based Access Control (ABAC) in large applications are being presented which can be used with Hadoop and NoSQL systems. The Attribute Based Access Control mechanism is implemented in two steps. In first step a list is made which specifies the data to be retrieved for user with sequence of permissions. In second step Query modification is done with ABAC controls applied. [5]

Data can be encrypted and decryption should be allowed only when the user trying to access the big data is genuinely authorized by an access control policy. One more challenge is that Hadoop which is a solution to almost all the problems of big data doesn’t follow default user authentication. It usually relies on conventional security solutions like firewalls and application layer safety standards.

Another significant challenge is determining the ownership of data. The biggest question is who owns the Big data. May be the cloud where the big data is stored? Here one feels the need to have adequate access control mechanisms so that data can be protected.

The recent case of Google Spain and Google Inc v Agencia Espanola de Proteccion de Datos of Mario Costeja Gonzalez decided by the Court of Justice of the European Union in May 2014, made it clear that an individual has the right to request the removal from an internet search engine of information which relates to that individual, and that the individual's right to data protection "will override the economic interest" (i.e. the property right) "of the operator of the search engine". [11] The judgment makes it clear that an individual’s right over his personal data is more than a search engine’s right over that data. It is interesting to see how little control organizations have over their big data.

2.6. The Infrastructure

The distributed nature of Big Data environments also poses a huge problem of infrastructure. It is easy to initialize attacks on distributed data as compared to single high end database server which is heavily protected. Adding to woes is the geographically distributed nature of big data. In such cases there is a need to standardize physical security measure all across the locations. The scalability feature of Hadoop which works by adding more servers for managing massiveness of data, it is a possibility that all those servers may not match in configuration and lack consistency, thus making them prone to attack.

2.7. The Technology

When big data programming tools like Hadoop and NoSQL were initially designed, the main focus was on the enormity of data and not the security. A big disadvantage with Hadoop is that it doesn't authenticate users who communicate with it. Also the data transfer between its nodes also does not happen in encrypted mode. It is a big danger to security and authentication. At the same time some of the security features offered by conventional databases are missing in NoSQL databases. The only advantage of NoSQL was its flexibility to support all new data types which are not supported by traditional databases. But at the same time supporting all new data types still doesn't have straightforward security policies. Additional security measures required for automated data transfer are also not up to the mark. Most of the distributed systems still follow single level of protection which is making data security weak.

III. HOW TO SECURE BIG DATA

Now let us discuss what security measures to take when one decides to move from traditional database management system to Big data. Various points can be considered:

3.1. Application Software Security

When the Big Data Technology stared spreading its wings, the concept of security wasn't anywhere in the picture. But with time as it is gaining momentum, the security issues are also becoming grave. In order to make Big Data secure, It is recommended that more and more open source technologies are used. Secure versions of open source software are address the issue well.

There are a few technologies like Cloudera Sentry in which application level security features are very good. In case of NoSQL databases using Sentry and Accumulo technology enhances access control security.

3.2. Protect the data and communication:

Big data analytics is about operations of huge volumes of data and then analyzing the outcomes. It is imperative that both data and the outcome to be used for analysis must be well protected. Care should be taken to ensure that sensitive and private data is not leaked. Also one must ensure the safety of data when it is being transmitted so that it remains as confidential and integral as was at the origin.

Emergence of computer aided research methods is transforming the way how research are done and scientific data are used. The following types of scientific data are defined [7]

Raw data collected from observation and from experiment (according to an initial research model) Structured data and datasets that went through data

- filtering and processing (supporting some particular formal model) Published data that supports one or another scientific
- hypothesis, research result or statement Data linked to publications to support the wide research
- consolidation, integration, and openness

3.3. Maintenance, Monitoring, and Analysis of Audit Logs.

Big data clusters need to be monitored and audited on a regular basis to ensure proper safety of data. There are various technologies like Apache Oozie which support this feature.

3.4. Secure Configurations for Hardware and Software.

Big data architecture plays a vital role in safety of its data. Server configuration needs to be proper in terms of having secure images of all the systems of the organization. Administrative privileges should be given to less number of people who are authorized in the task.

3.5. Account Monitoring and Control.

All the accounts of big data users need to be monitored and managed regularly. Inactive accounts should be deactivated instantly. Other security measures like strong passwords, detecting repetitive failed login attempts, monitoring account access, etc must be reinforced.

3.6. Continual expansion of the antivirus industry :

Installing a good antivirus software as a protection mechanism is advisable. Therefore antivirus industry also has to evolve and come up with better defence mechanism. Since the volume of big data is growing exponentially, new advancements in antivirus techniques is also sought for.

3.7. Plan before you deploy –

In order to have strong security of big data all the policies must be planned and deployed well before the initial phase of setting up Hadoop and moving Big data into it. Care should be taken to identify and protect any sensitive and critical data. In addition all government policies and regulations should also be understood clearly while the planning phase is going on.

3.8. Don't overlook basic security measures –

Basic security measures if implemented correctly at the time of initial planning will surely provide long term security of big data. Many times when data is private and confidential, care should be taken to monitor the user groups who are going to access that data. As the data is sensitive mapping potent users to groups will help. Basic safety precautions like permission granting and use of strong passwords must be encouraged.

3.9. Choose the right remediation technique –

Desensitization is the process of identifying sensitive data where random dummy data which is carefully formatted is substituted in place of real data. It is used in areas where one doesn’t want to use real factual data of the clients. Problem arises when big data analytics need access to real data, as it has been already desensitized. Two techniques come as an aid for them, masking or encryption. Whatever technique is used for remediation, it has to be integrated with organization’s protection policies and various security standards. Also care should be taken that both masked and unmasked versions of data are kept safely in separate Hadoop files.

Data masking creates a duplicate inauthentic version of data while keeping the structure same. This new copy of data can be used for testing and training purpose.

At the same time data encryption completely transforms data into a different form from its original one enabling only those users who have access to a secret key read the data.

3.10. Examine the cloud providers:

Cloud Computing is the latest technology which offers innovative solutions to most of the Big Data Problems especially storage. The technology is growing rapidly and has carved a niche in the current and upcoming generations of IT Industry and various IT based business companies. Benefits of Cloud technology can be seen in terms of its reliability in providing software, hardware and other necessary infrastructures called IAAS. [8][9]

While storing big data into cloud organizations need to carefully examine the security policies of the cloud provider. One must emphasize the cloud provider to have regular security audits and agreement terms on penalties in case of any security breach of data concerned.

IV. CONCLUSION AND FUTURE DIRECTIONS

Big data security is important not only because of the criticality and sensitivity of data but also because any kind of breach in security will impact other issues also. These can be in terms of damage to the organization’s reputation resulting in legal percussions or intruding customer’s privacy. It is recommended to emphasize more on application security. As far as device security is concerned, if the devices hold sensitive data they must be kept under security. Objective should be to provide reactive and proactive protection and for that event management and real time safety measures should be incorporated. Although Big Data phenomenon is largely seen everywhere, yet best practices still need to be widely implemented all across the organizations. Organizations need to work on a proper access control policy to enable access to authorized users only. Big data users also need to undergo proper training and procedures in terms of Hadoop data security. Although benefits of using Big Data with Hadoop are many but they come at a cost of security. A better understanding of Hadoop system with all security considerations will help organizations keep their precious data safe.

REFERENCES

[1]. Miss. Debalina Nandy, Mr. Renish J Padariya, Miss Tosal Bhalodia (2017). Security Challenges in Big Data. *IJIRST, National Conference on Latest Trends in Networking and Cyber Security*

[2]. Ali Kalantari, Amirrudin Kamsin, Halim Shukri Kamaruddin, Nader Ale Ebrahim, Abdullah Gani, Ali Ebrahimi and Shahaboddin Shamshirband, (2017). A bibliometric approach to tracking bigdata research trends. *Journal of Big Data*

[3]. Hashem IAT. The rise of “big data” on cloud computing (2015). *Review and open research issues. Info Syst. 47:98–115.*

[4]. Diaz M. et al. (2012). Big data on the internet of things, *In 2012 sixth international conference on innovative mobile and internet services in ubiquitous computing.*

[5]. Jim Longstaff,Joanne Noble. Attribute Based Access Control for Big Data Applications by Query Modification (April2016). *IEEE Explore, Electronic ISBN: 978-1-5090-2251*

[6]. Fadoua Badaoui, Amine Amar, Laila Ait Hassou, Abdelhak Zoglat and Cyrille Guei Okou (2017). Dimensionality reduction and class prediction algorithm with application. *Badaoui et al. J Big Data 4:32 DOI 10.1186/s40537-017-0093-4 to microarray Big Data*

[7]. Study on AAA Platforms For Scientific data/information Resources in Europe. *Final report [online] https://confluence.terena.org/download/ attachments/30474266/AAAStudy-Report-0907.pdf*

[8]. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, (2010). A view of cloud computing. *Commun. ACM 53 50–58.*

[9]. Ibrahim Abaker Targio Hashem , Ibrar Yaqoob , Nor Badrul Anuar, Salimah Mokhtar a , Abdullah Gani a , Samee Ullah Khan b (2015). The rise of big data on cloud computing: Review and open research issues. *Elsevier, Information Systems 47 98–115*

[10]. Raghav Toshniwal, Kanishka Ghosh, Dastidar, Asoke Nath (February 2015). Big Data Security Issues and Challenges”. *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2*

[11]. Taylor Wessing (Jul 2014). *The ownership of Big Data, Global Data Hub*