Vulnerabilities of Data Storage Security in Big Data

Govind Murari Upadhyay*
Harsh Arora**

Abstract

In the era of big data, the massive amount of data needs to be managed, organized and secure over the network for all users. Vulnerabilities and threats make the data insecure and unreliable. In order to keep this bulk of data secure and safe, some security mechanism related to same need to be implemented. Confidentiality and privacy in big data must be maintained. Authentication and data integrity are other related parameters of big data security. Various challenges, issues and problems arise when available data gets interrupted by third party intruders. The intruders can be from external and internal environment from an organizational point of view. These individuals access, view, edit the sensitive data by unauthorized means. There should be some counter measures and access control methods, algorithms and the corresponding techniques like Map Reducer, Data Filtrations methods, Various encryption methods and finally we cannot ignore the best technique "Hadoop" which is used to handle, manage, organize and secure the big data specifically. Hence, monitoring and detection of attacks and prevention of threats to be implemented altogether. Although huge amount of data is present yet the available data must be focused according to cyber security point of view. There should be no compromise with big data storage, security, integrity and reliability. Big data available should be valid and secured using security mechanism and data controlled techniques.

Keywords: Big data, vulnerabilities, filtration, data security, Hadoop, Map Reducer, Intruders, Encryption, threats.

I. Introduction

The word "Big Data" is used to describe the increased and massive volumes of structured and unstructured data which is so large that it is very difficult to process this data using traditional databases and software technologies. Big data storage leads to large amount of data storage and managing and securing the big data is equally important and valuable at the same time. Big data requires bigger responsibilities as companies of all sizes and in virtually every industry are struggling to manage exploding amounts of data. In order to analyze complex data and to identify patterns, it is very important to securely store, manage and share large amount of complex data. When it is the matter of "Big data", we cannot ignore the word or technique "Hadoop"- the term which is used to support the processing of large set of data in distributed computing environment [1]. As both business and IT executives know all too well, managing big data involves far more than just dealing with storage and retrieval challenges- it requires addressing a variety of privacy and security issues as well. So there should some mechanism to protect such a huge amount of data present on the web. For making big data secure, organizations need to employ three key types of security protocols or controls.

(A) Preventive

(B) Detective

(C) Administrative

The techniques such as encryption, logging and honeypot detection must be necessary. In many organizations the deployment of big data for fraud detection is very attractive and useful [1]. As the phenomenon of large data exist almost in every field whether it is physics, biology, ecology, scientific area and others so it is the prime objective to secure the bulk of data present everywhere and big data as an information security problem which has a lot of challenges which have to curbed [2].

Govind Murari Upadhyay*
Department of Information Technology
Institute of Innovation in Technology & Management, New Delhi, India
Harsh Arora**
Department of Information Technology
Institute of Innovation in Technology & Management, New Delhi, India

II. Data Security in Big Data

(A) Confidentiality and privacy in Big data

Computations and database operations are done on massive amount of data, so it's highly important to protect data in order to make it reliable, integrated, and confidential and privacy is maintained. In this era, many organizations are using the technology to protect and secure this bulk of data in order to make it integrated. Not only security but also data privacy challenges exists in the industries and federal organizations [1]. Out of this bulk and huge data called big data it is required to make secure data highly confidential among various companies and stakeholders. There should be a balance between data privacy and national security. Big data contains huge data volume and this requires a new generation of encrypted solutions and on the other hand big data techniques can also be used to address and security changes in network system [2]. Security policies need to apply on big data which refer to set of rules and practices that specify or regulate how a system organization provides security services to protect sensitive and critical system resources [3]. Confidentiality covers two related concepts:

1. Data confidentiality assumes that private or confidential information is not made available or disclosed to unauthorized individuals.
2. Privacy assumes that individuals control or influence what information related to them may be collected and stored and by when and to whom that information may be disclosed.

(B) Authentication and data integrity

Big data needs to be authenticated and integrated that gives the assurance that data received are exactly as sent by an unauthorized entity that contains no modification, insertion, deletion or replay [3]. Big data integrity deals with a stream of messages, assumes that messages are received as sent with no duplication, insertion, modification or replays. Integrity service relates to active allocates, the concern is detection as well as prevention.

(C) Security mechanism and access control related to big data

In this era of big data, in order to manage this huge and massive amount of data on the network is a critical issue that must be handled to keep it safe, secure and integrated so all security mechanism need to be implemented. In the context of network security, access control is the ability to limit and control the access to host systems and applications via communication links. Apart from access control, other security mechanisms are required to be protected and integrate the big data over the network. These security mechanisms are authenticated exchange, traffic padding, routing control and notarization.

III. Problems, challenges, issues related to big data security

Encryption of data is the prime objective if it is the concern of data security in big data. Also appropriate policies are required for managing and sharing of data on the web. For this purpose security algorithms need to apply on big data to make it secure. As per the network access security model, there are many internal security controls to protect and secure the data still there are many threats to the available big data. The existence of vulnerabilities in a network system and database makes the big data insecure and there should be some mechanism to exploit these vulnerabilities in the system as these vulnerabilities has great impact on all the system programs such as utility programs and many more. There can be the possibilities of various threats to big data and various factors associated with network and communications are responsible for these threats. These threats basically allow the system open to intruder mean the third party object or role players which can easily access the data available on network. In other words it would be appropriate to say that the valuable data can be extracted by any of the unauthorized person or hacker during the communication of data from source to destination all over the network. So issues related to confidentiality and privacy comes into the limelight that exposes critical corporate data and related personal information to new security threats. The threats can be related with information access threats and service threats. Now the question in this context is how threats have become the challenges and how they are represented. The insight threat acts as security challenge when intruder or individuals misuse the data and sensitive data is extracted by some internal intruders or individuals by authorized access. The threats can be because of

internal or external intruders. The individuals within the organizations have been reliable to same and have authority to access the data and thus possess the necessary authorization to access propriety or sensitive data. The other types of active intruders are from external environment. These individuals access, view, edit the sensitive data by unauthorized access means, thus act as a big outside threat to the big data in the network. Challenges related to cyber security and of course related with the topic in context "Big data security" can be depicted through various parameters and aspects.

1. **Distributed computation framework challenges**
   When data is accessed in a distributed and parallel fashion than computations need to occur in parallelism and to access and manage the massive amount of data, the distributed computations framework becomes the challenge. Complications regarding this frame need to be resolved by the attacks preventions measures which would describe how to secure the big data in which manner and also at the same time describe how data can be secure in presence of entrusted manner. Untrusted mappers may return wrong results which further generate incorrect aggregate results. Map reduce framework used for secure computations in distributed programming framework where each input file is divided and reduced in the multiple parts or blocks in the first phase of map reducer. Each individual block or chunk is being read by mappers and related calculations are performed and output related to these blocks are combined together later on form of list of key and value pairs [1].

2. **Security for traditional databases challenge**
   Security for traditional databases includes various secure policies that must be evolved with respect to security infrastructure. Here the context is about non relational data storage and its security. In traditional databases where the relativity among data is not so advanced and managing such data in context of big data itself is a big challenge coming on another side nosql database were built to tackle different challenge in analyzing the data as data security was never part of such design at any point [TTBDS]. Hence robustness gets affected such a case which could be processed by clustering aspect of nosql database. Huge volume of data is being handled and processed as a challenged by various companies. Where the criteria is to deal with big bulk of unstructured data sets that originally become the part of traditional relational database but from security and efficiency point of view traditional relational database is being converted into nosql database. For accommodating and accessing big data hence the idea is not to compromised with operational feature of database and traditional database must be handled from the vision of accessing it on the web where the complete data must be secure and companies must review securities policies. For the middleware by enhancing it with addition of security feature to its main counterpart.

3. **Big data storage, Accessing the valuable Transactional Logs Challenges**
   In context of accessing the data and various transactions in available databases, the data has to be travelled through multiple tiers or layers as far as storage media is concern the much tier architecture or framework of database is very helpful whether it is traditional database, relational nosql database management system or big data concern. Multitier storage media helps and allows to move data between tiers and communication of data takes place in effective and efficient manner. But the challenge is how to make data secure rather than any other factor of communication and storage of big data. The vulnerabilities and threats have always been exist as challenges in the network and database security. Also besides normal database storage, the size of datasets has been continuously growing exponentially depending upon the requirements of companies as per their applications and relativity with others. So it has become as necessity to auto apply tiering methodology on scalable and available data in order to make bulk of data secure, manageable, accessible and reliable. But problems and issues lie here also related to auto tiering concept is that it does not keep tract about the place of data storage which definitely

becomes a new challenge it context of data storage security.

For big data security data must be integrated and aligned and place of available data must be tracked and defined and to make available the big data 24X7, the effective security mechanisms should exist to beat the security challenges. With auto tier storage system, critical information should be properly aligned in different tiers however the lower tier has low percentage of security, so the organizations should study and analyzed these multitier strategies and policies and should track how data is passing, communicating and locating in different locations or places and corresponding tier framework and flow of secure data through multiple paths.

4. **Input validation/ Filtering challenges**
Validation of data is highly important for an enterprise to make to make data integrated all together whether it is centralized database or distributed database. In an organizational network, multiple systems devices software applications are connected together and the corresponding data available may be insecure everywhere on the network. Now the question is how can be assure that source of input is valid. Input validation must be recognized. Data in input sources should not be entrusted, malicious and corrupted. So to make data valid and trusted there must be some mechanism of filtering and protecting data during the network transmission and communication of big data in different locations over the web. Validation problems arise while data is input from various sources on net filtering of data is most critical issue and challenge in context of big data security, which completely leads to input data validation. Filtering and validation challenges of data must be analyzed and corrective measures should be taken by following various routing network algorithms and produces along with various firewall mechanisms. But we cannot ignore the concept of filtering data, otherwise it will remain as a big challenge in context of input data validation .There should be no compromise with big data storage, security and reliability so challenges related to validation of data and filtration must be controlled and counter measures are necessary to handle it all.

5. **Monitoring and Real time Security Challenges**
The problem with big data arises when it is the point of real time security. All devices in a network involved in the communication of amiable data need to have some alert mechanism. The alerts are generated by security devices. This aspect of alerting from connecting devices must be taken into account considerably but normally these things are ignored straightway from enterprise or organization. This monitoring and real time security has become a critical issue and challenge.

IV. Counter measure for big data security

From big data security point of view, the reliable security mechanisms are needed to cope up with the existing threats and challenges. In context of network access security model, access channel is used to maintain the flow control of data. The gatekeepers plays very important role for security control of data on the network. Internal security controls are applied on the data and processes to make data secure and protected. There should some privacy measure algorithms and techniques to cope up with inside threats, intruders' interference and hacking of data. Various detection systems are used to detect the unusual pattern of data access. Various encryption techniques are used to protect the online data or big data which is available everywhere on network. The question is how to tackle and control such a big data. Many organizations are facing to implement the control measures and techniques as this is a great challenge. In a big data environment, beside storage of data, integrity, consistency, reliability, availability, accuracy and security factors are highly important and required. Some regulatory and control measures are used to maintain these properties. So a technique such as attribute based encryption is used to protect and secure data. To make a big data secure, detection and prevention of predictable threats are required that becomes the part of big data analysis. These techniques are used to give valid input data that leads to make it reliable for transmission all over. As with these techniques, detection can be done at early stages, so it helps to prevent distribution of errors and analyze the patterns of all mutable data resources.

Another technique which can be used to protect data from various threats is "Feature Extraction". To make data authentic and reliable and when it is all about big data than we should invariably encounter Hadoop. This technology is highly effective to manage are considered problem or challenge which is how to manage and to secure, integrate, authenticate "big data". So the term finally "big data" should be associated completely with the technology meant to handle this massive data is "Hadoop". Which organizes the data online all over the world and makes data extraction feasible everywhere in efficient and effective manner. In cyber space quality data should be extracted and available to the use at any location along with the privacy, secured measures to cope up the challenges reconcile the data protection with data user privacy. As we can see from cyber security point of view related to big data access control mechanism for data security are required and all challenges and issues are tackled effectively with use of database software technologies like Hadoop. For better benefits, and for better organized data in secure environment.

V. Conclusion

The significance for cyber security for big data lies in the fact to make complete bulk of data available on network should be integrated and secured. Multiple users all over the world should be able to access big data under the umbrella of network and database security. The purpose is how to beat the challenges and issues available in the existing system. These challenges are coping up by using some counter measures and techniques to handle big data which cover filtration methods, encryption methods like attribute based encryption, feature extraction are used. Finally Hadoop technology has crossed most of the challenges and issues related to big data includes to prevent the predictable security attacks and threats and how to cope up with these threats. Accordingly proper counter measures should be taken into account by considering all security policies and parameters.

References

1. V. N. Inukollu, S. Arsi and S. R. Ravuri "Securities Issues Associated with Big Data in Cloud Computing" IJNSA, Vol. 6, No. 3, May 2014.
2. K. Zvarevashe, M. Mutandavari, T. Gotora "A Surveyof the security Use Cases in Big Data" IJIRCCE Vol 2, Issue 5, May 2014.
3. W. Stalling "Cryptography and Network Security" Fifth Edition.
4. Harsh Arora "Study of Securing in Cloud, Virtual and Big Data Infrastructure" NCETIT Vol. 5 Issue. 1(A) 22, March 2014 pg. 98-104.
5. Kalyani Shirudkar, Dilip Motwani " Big-Data Security", IJARCSSE, Volume 5, Issue 3, March 2015.
6. G.M. Upadhyay, K. Dhingra " Web Content Mining: Techniques and Algorithms" IJARCSSE, Vol. 5, Issue 4, 2015.
7. Elisa Bertino "Data Security – Challenges and Research Opportunities" Springer International Publishing Switzerland 2014.