# A Review on Security Challenges and Issues of Big Data

**C. Immaculate Mary**
*Department of Computer Science*
*Sri Sarada College for Women (Autonomous)*
*Salem-16*
*cimmaculatemary@gmail.com*

**P. Roshni Mol**
*Department of Computer Science*
*Sri Sarada College for Women (Autonomous)*
*Salem-16*
*roshnimphil@gmail.com*

***Abstract-*** **Lack of Security is the major problem all over the world in each sector. The term big data is due to the terrific usage of data in our day to day life. Among these data, highly sensitive data such as Payment Card Information (PCI), Personally Identifiable Information (PII), and Protected Health Information (PHI) must be handled more securely. Routine steps such as Prevention, Detection, Remediation and Investigation should be followed in order to get a safe and secure big data environment. In this paper, we provide an extensive survey of major and minor security issues and challenges of Big Data, while highlighting the specific concerns in big data security. We have compared the traditional Security Information and Event Management System (SIEM) with the current Security Information and Event Management System (SIEM) for big data. We have discussed the tools and techniques available for securing big data. We have provided a survey of existing solutions, identified research gaps, and suggested future research areas.**

Keywords: Big data, security, issues, SIEM, Encryption, sensitive data, hadoop, honeypots, cyber attacks, security analytics

## 1. INTRODUCTION

Due to the increasing use of technology, the data produced has also been increased leads to big data. (Big Data, 2016) Every day we create 2.5 quintillion bytes of data. For every minute, amazon get 4,310 unique visitors, Netflix subscribers stream 77,610 hours of video, people sends 347,222 tweets. As the technology grows, the security level of the technology needs to be concerned. Always there will be vulnerabilities and security gap when the amount of data is huge. In October 2016, 32 lakh debit cards issued by SBI, HDFC Bank, ICICI Bank, Axis Bank and Yes Bank were compromised in the largest-yet cyber attack on the Indian banking system. Due to Demonetisation of Rs.500 and Rs. 1000 bank notes there is increase in Digital Payments. Though Digital Payments may be useful for transactions, but on the other side there are many possibilities for cyber attacks. An article published by The New Indian Express states that till November 28 2016, top intelligence sources had observed an average of 2 lakh threats & Vulnerabilities per day. This increased to 5 lakh after note ban & further went up to 6 lakh threats by first week of December. According to Vormetric Survey, 35 % of cyber attacks today occur without the knowledge of enterprise. In 30 seconds, a cyber criminal can steal most sensitive data. Android and IOS platform-based smart phones are known to have multiple vulnerabilities which are being widely exploited by the attackers and adversaries. Data such as Payment Card Information (PCI), Personally Identifiable Information (PII), and Protected Health Information (PHI) are highly sensitive privacy related data which should be more secured. In this paper, we provide a wide survey of major, minor security issues and challenges of Big Data. Also we have analyzed various security solutions available for big data security.

## 2. SECURITY SOLUTIONS FOR TRADITIONAL SYSTEMS Vs BIG DATA SYSTEMS

(Rama Roa, 2015) Comparison of traditional and big data security systems can be done based on big data’s Volume, Velocity and Variety. There are many advantages and disadvantages in both the systems. In traditional systems syslog is used to store, report and analyze logs. Big data systems make use of cloud for storage of data. Data logs are difficult to be traced since it is stored in different tiers of cloud environment. Storing & retaining large amount of data was not economically feasible in traditional security systems. Detecting the malicious data in big data is difficult. Auditing can be easily done in traditional systems whereas due to its quantity in big data, it is difficult to perform auditing. Security analytics in traditional systems is quite simple when compared to big data. Performing analytics and complex queries on large, structured data set is inefficient. Big data systems handle unstructured data, whereas traditional systems were not designed to analyze unstructured data. Encryption of big data may take time and there may be performance related issues due to its high volume. Further research can be done on how to handle data which is stored in different tiers of cloud environment securely and how to increase the processing speed of big data while encrypting and decrypting.

## 3. LITERATURE SURVEY

### 3.1. Hadoop and Big Data Ecosystems

Traditional perimeter security provided by the big data Vendor or other traditional infrastructure tools Protecting the data itself, using data-centric security provides a way to protect data against attacks. Sateless key management in combination with field- level, format preserving encryption, enables the secure portability of data throughout Hadoop and big data ecosystems. (Kappenberger, 2016)Transparent Data Encryption (TDE) in hadoop provides file/folder level encryption. Data-at-rest encryption method is to protect sensitive data received in big data environment. Server protection is essential to protect data outside big data environment.

### 3.2. Challenges of Security in Big Data Environment

The challenges of security in big data environment can be categorized into authentication level, data level and network level .If malicious node gets administrative priority then it will steal or manipulate user data. To overcome this authentication level issue (Savant, 2015) logging can be used to record the activity of the nodes. Analysis of Logs such as log collection server, log parser and log transition can be used for security. (Kyung Sik jeon, 2016) PCRE (Perl Compatible Regular Expressions) is a library that supports special separator technique & regular expression for unstructured data normalization technique. For data level challenge it must be encrypted in big data environment. Network level issues are more complicated in big data environment as the data is stored anywhere among the node in cluster. Remote procedure call (RPC) used for communication in network must be encrypted.

### 3.3. Security Controls for Big Data

Security Controls for Big Data such as (Filkins, 2015) Host based application firewalls/IDS, Network based IDS, Encryption, centralized SIEM ( Security Information and Event Management), Security controls within our big data management system, user-activity monitoring , secure development and life cycle practices, database activity monitoring, unified authorization mechanism, data de-identification, data redaction, digital rights management, automated audit aggregation are used to perform survey on effectiveness of security controls. Among these controls 40% of people ranked encryption as very effective.

### 3.4. Big Data Security Framework

Big data security framework core components such as (Gaddam, 2014) data management, identity & access management, data protection & Privacy, Network security, Infrastructure security & integrity must be concentrated.Packet Level Encryption using TLS from the client to Hadoop cluster and within the cluster itself is one way for data protection. For network level security, end users must be allowed to access name nodes only. To control the network traffic out of Hadoop, Apache Knox gateway can be used. For Infrastructure security, SELinux (Secure Enhanced Linux) has a set of patches for linux kernel for linux security. Tools such as OSS Apache Falcon, Cloudera, Navigator or Zettaset orchestrator can be used for data management.

### 3.5. Security of Cloud Environment

Security measures needed to be considered for the (Inukollu, 2014) security of big data in cloud environment are File encryption, Network encryption, logging, software format, node maintenance & authentication, accurate system testing of jobs, honeypot nodes, third party secure data publication to cloud, access control.

### 3.6. Security of IOT

Mirai (PC Quest) malware is a serious cause of concern for the internet of things ecosystem. It scans for IOT (Internet of Things) devices that are still using their default passwords and then bind those devices into botnets which is then used to launch DDOS (Distributed Denial of Service) attacks. Mirai makes use of brute force technique for guessing passwords. Users should use strong passwords, occasionally changing the passwords, regularly check for software updates and implement appropriate security software to make a secure environment.

From the literature survey we came to know more about security issues within and outside big data environment, cloud environment and IOT. Further research can be done on different levels such as authentication level, data level and network level. There are many tools available for protecting these levels of big data environment.

## 4. TOOLS & SOLUTIONS FOR BIG DATA SECURITY

Data fusion & analysis, data privacy & security, data collaboration & Sharing plays major role in cyber intelligence. (Informatica, 2013) Informatics platform directly supports the gathering of intelligence by discovering, distilling & delivering the information needed to detect & prevent new threats.

RSA (RSA,EMC2) security Management portfolio provides customers with comprehensive visibility, agile analytics, actionable intelligence and optimized incident management. RSA security prevents, detects, remediates and investigates threats.

VSS (Monitoring, 2014) Visibility plane provide both line rate network packet preprocessing from 100 Mbps to 100 Gbps as well as deliver packets raw encapsulated indexed and in the form of metadata, manage, monitor and secure the network.

To secure big data life cycle organization requires infrastructure security, data privacy, management, integrity, authentication , authorization of databases, transport security. Oracle (Custom, 2015) enables organizations to separate roles & responsibilities and protect sensitive data. It also provides monitoring, auditing & compliance reporting across big data system and traditional data systems.

Apache Hadoop (Mattson, 2014) library itself is designed to detect and handle failures at the application layer. Solution Providers such as Cloudera, Gazzang, IBM, Intel (open source), MIT (open source), Protegrity and Zettaset provide solutions to access control, authentication, volume encryption, field/column encryption, masking and / or monitoring.MIT use secret-key cryptography to provide authentication for client-server applications.

Though there are many tools and solutions available, it is not sufficient to handle big data thoroughly. Only application layer is protected by apache hadoop , so we have to concentrate on other layers for security.

## 5. CONCLUSION & FUTURE ENHANCEMENTS

In this paper, we reviewed various security challenges and issues in big data environment. Big data plays major role in Internet of Things (IOT). Attackers are now focusing on IOT to locate and compromise IOT devices to further grow botnet. Security measures should be taken to handle IOT devices. Personal Information must be secured in order to avoid privacy related issues. Various encryption techniques can be implemented to make to data secured. Machine learning algorithms can also be implemented for security purpose and to handle big data. Further research can be done in the following four areas of big data environment as shown in Figure 1.

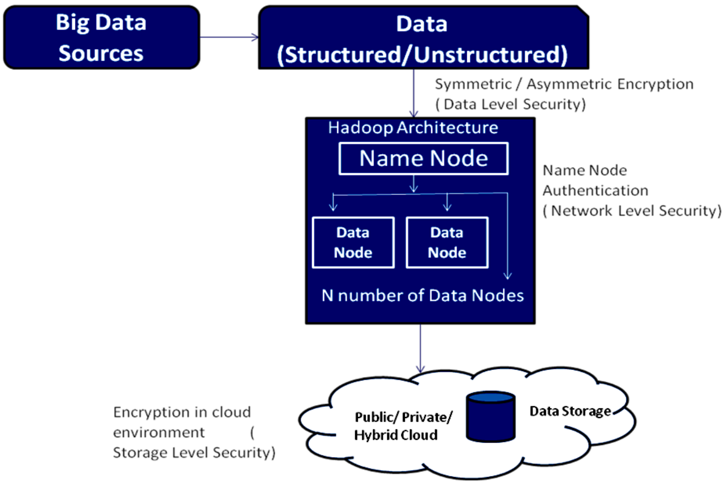
Big Data Sources
Data (Structured/Unstructured)
Symmetric / Asymmetric Encryption ( Data Level Security)
Hadoop Architecture
Name Node
Data Node
Data Node
N number of Data Nodes
Name Node Authentication ( Network Level Security)
Encryption in cloud environment ( Storage Level Security)
Public/ Private/ Hybrid Cloud
Data Storage

Figure 1. Big Data Environment

First step is to encrypt data which is received in big data environment using symmetric or asymmetric encryption. For this encryption existing algorithms as well as new machine learning algorithms can be implemented. Second step is to monitor the network where the data is travelling through. Firewalls and apache knox gateway can be used to control the traffic. Third step is to secure the cloud infrastructure. Big data makes use of Public, private as well as hybrid cloud to store data. We must identify where the data resides in different tires of cloud environment and provide security to those data. Fourth step is to ensure the security within the hadoop architecture. The name node must be more secured because it contains the details of data nodes. Various solution providers are available and we can also develop algorithms for securing hadoop architecture.

## REFERENCES

Bureau (2016). Security breach: Banks block over 32 lakh debit cards; Finance Ministry seeks report. Mumbai/ New Delhi, India.

Custom (2015). Securing the Big Data Life Cycle. Oracle.

Filkins (2015, April). Enabling Big Data by removing security & compliance barriers.

Gaddam (2014), A. Securing your big data environment.

Informatica (2013). Data Fusion for cyber intelligence. Ponemon Institute LLC, Big Data Analytics in Cyber Defense. Informatica.

Inukollu (2014). Security issues associated with big data in cloud Computing. 6 (3).

Rama Roa (2015). Effects of Big Data Characteristics on Security-Leveraging existing security mechanisms for protection. Asian Research Publishing Network, 10 (5).

Kappenberger (2016). Protecting Your Data against cyber attacks in Big Data environments. ISSA, 14 ( 2).

Kyung Sik jeon (2016). A study on the big data log analysis for security (Vol. 10). International Journal of Security and its applications.

Mattson (2014). Bridging the Gap between Access and Security in Big Data. ISACA, 6.

Savant (2015). Approaches to solve big data security issues and comparative study of cryptographic algorithms for data encryption. IJICA , 1 (1).