Recent Trends in Mathematical Modeling, Simulation Methods, Computations, and Physical Sciences
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

# An Analytical Study of Security and Privacy Issues of Big Data

V. V. Thakare[1], V.S.Tondre[1]
[1]Department of Computer Science, B.B.Sc. College, Amravati, Maharashtra, India

## ABSTRACT

Due to the reasons such as the rapid growth and spread of network services, mobile devices, and online users on the Internet leading to a remarkable increase in the amount of data. Almost every industry is trying to cope with this huge data. Big data phenomenon has begun to gain importance. However, it is not only very difficult to store big data and analyze them with traditional applications, but also it has challenging privacy and security problems. For this reason, this paper discusses the big data, its ecosystem, concerns on big data and presents comparative view of big data privacy and security approaches in literature in terms of infrastructure, application, and data. By grouping these applications an overall perspective of security and privacy issues in big data is suggested.

**Keywords :—** Big Data, Cloud Security, Monitoring, Auditing, Key Management, Anonymization

## I. INTRODUCTION

Data generation and collection quickly surpass the bounds in the digital universe of today. The data has been doubling every 2 years since 2011 [1]. It is predicted that the data will increase 300 times, from 130 exabytes in 2005 to 40,000 exabytes in 2020 [2]. As a result of this technological revolution, the big data is becoming increasingly an important issue in the sciences, governments, and enterprises. Big Data is a data set, which is difficult to capture, store, filter, share, analyse and visualize on it with current technologies [3]. Despite such difficulties, if big data is cope up, it helps to generate revenue, executive efficiency, strategic decisions, better services, defining needs, identifying new trends, and developing new products, all of which is covered in the data science [3].

In addition, data science focus on parallel and distributed processing, similarity search, graph analysis, clustering, stream processing, search ranking, association analysis, dimensionality reduction and machine learning algorithms etc. [4]. However, in this complex computation environment, traditional security and privacy mechanisms are insufficient to analyse big data. These challenges in big data consist of computation in distributed and non-relational environments, cryptography algorithms, data provenance, validation and filtering, secure data storage, granular access control, and real time monitoring [5].

By classifying the sources of problems, results become more efficient with big data. This research paper focus, classify and analyses security and privacy breaches and solutions in big data. This perspective would lead to an understanding of important research

areas and the development of new methods. In addition, the use of big data in analysis would make the systems become safer. It also represents a brief summary of big data. Also contains categorization of big data, concerning security and privacy studies in literature. The results obtained with security and privacy issues in big data are discussed here, and explains how to use big data to maintain security.

## II. DEFINITION AND CHARACTERISTICS OF BIG DATA

Big data refers to large and complex datasets that typical software is inadequate for managing [2]. There are various explanations of big data via Vs. 5Vs are typically used to characterize of Big Data as volume, velocity, variety, veracity and value as shown in Fig. 1 [3]. Volume is the size of data; velocity is the high speed of data; variety indicates heterogeneous data types and sources; veracity describes consistency and trust worthy of data; and value provides outputs for gains from large data sets.

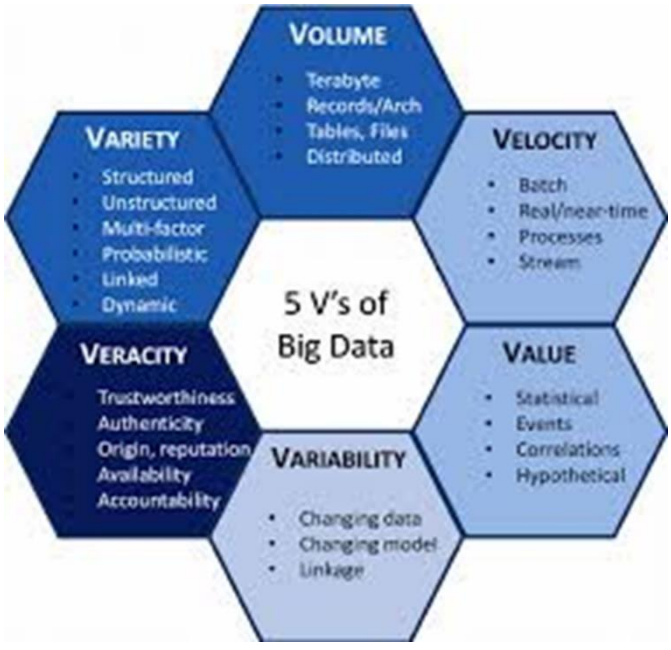


**Fig. 1: 5 V's of Big Data**

Big data is classified into ten categories in terms of data type, data format, data source, data consumer, data usage, data analysis, data store, data frequency, data processing propose, and data processing method as shown in

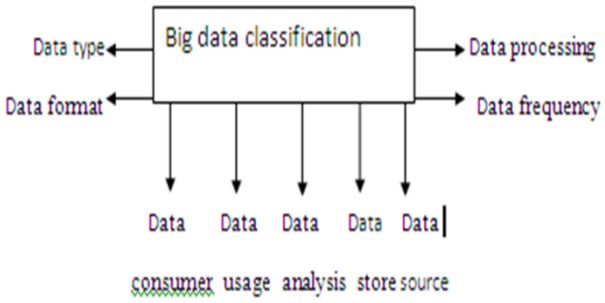


**Fig 2: Big Data Classification**

## III. BIG DATA SECURITY AND PRIVACY APPROACHES

### A. Cloud Security:

The widespread use of cloud computing for such reasons as broad network access, on-demand service, resource pooling and being elastic have made a proper environment for big data [6]. However, cloud hosts traditional threats and new attacks. Data storage on clouds is one of the main problems now a day. Therefore, some precautions must be taken by the service provider. Because of this, a secure way to handle and share big data on cloud platform has been presented [10].

It includes many security methods like authentication, encryption, decryption, and compression etc. to store big data securely. Authentication with email and password has been used for the authorized person. Data has been encrypted and compressed to prevent security issues. It also takes precautions in case of a natural disaster and uses three backup servers for this purpose. In these servers, data has been stored in an encrypted format. If something happens to the server, encrypted data has been decrypted with the secret key.

The classical encrypted technique is not enough for big data security on cloud. Consequently, new scheme to secure big data storage has been proposed [11]. This scheme uses cryptographic virtual mapping to create data path. According to the proposed scheme, big data

has been separated into many parts and each part is located in different storage providers. As a security measure, if all data encryptions are thought to be quite computational and useless, only storage path which shows critical information encryption seems enough, rather than all big data encrypts. The proposed scheme also supports some information encryption to increase the security level. To achieve availability, the scheme holds multiple copies of each part and their accessing index. Thus, if any data part is lost for some reason, information availability is successfully maintained.

### B. Monitoring and Auditing:

Security monitoring is gathering and investigating network events to catch the intrusions. Security audit is a systematic measurable security policy to use different methods. These two elements play an important role in active security.

Intrusion detection and prevention procedures on the whole network traffic is quite difficult. To solve this problem, a security monitoring architecture has been developed via analyzing DNS traffic, IP flow records, and HTTP traffic and honey pot data [11]. The proposed solution includes storing and processing data in distributed sources through data correlation schemes. At this stage, three likelihood metrics have been calculated to identify whether domain name, packet or flow is malicious. According to the score obtained through this calculation, an alert occurs in detection system or process terminates by prevention system. According to performance analysis, open source big data platform on electronic payment activities of a company data, Spark and Shark produce fast and steady results than Hive and Pig. Network security systems for big data should be find abnormalities quickly and identify correct alerts from heterogeneous data. Therefore, a big data security event monitoring system model has been proposed which consists of four modules: data collection, integration, analysis, and interpretation [10]. Data collection includes security and network devices logs and event information. Data integration process is performed by data filtering and classifying. In data analysis module, correlations and association rules are determined to catch events. Finally, data interpretation provides visual and statistical outputs to knowledge database that makes decisions predict network behavior and respond events.

### C. Key Management:

Key generating and sharing between servers and users is another big data security issue. However, using big data centers, quick and dynamic authentication protocols can be suggested.

In [1], a layered model has been proposed for quantum cryptography for strong keys in less complexity and Pair Hand protocol for authentication in mobile or fixed data centers. The model consists of these layers: front end, data reading, quantum key processing, quantum key management and application layers, respectively. This model has been not only increased efficiency but also reduced key search operations and passive attacks.

The big data services consist of multiple groups that need group key transfer protocols for secure communications. For this reason, novel protocol without an online key generation centre based on Diffie-Hellman key agreement and linear secret sharing scheme unlike existing protocols has been offered [5]. The protocol counter attacks via ensured key freshness, key authentication and key confidentiality reducing system overhead.

In more complex systems, conditional proxy re-encryption (CPRE) is used for secure group data sharing. Accordingly, an outsourcing CPRE scheme has been proposed in cloud environment which reduces overhead without downloading all data from the cloud, encrypting them and uploading them to the cloud in a new condition unlike CPRE [6]. When a group membership has been changed, key generation and decryption processes execute on outsourcing server and a condition value changing key has been calculated. Then it is sent to the cloud.

After that, the cloud storage uses this key to transform existing data.

### D. Anonymization:

Data harvesting for analytics causes big privacy concerns. Protecting personally identifiable information (PII) is increasingly difficult because the data are shared too quickly. To eliminate privacy concerns, the agreement between the company and the individual must be determined by policies. Personal data must be anonymized (de-identified) and transferred into secure channels [7]. However, the identity of the person can be uncovered depending on the algorithms and the artificial intelligence analysis of company. The predictions made by this analysis can lead to unethical issues. In [4], PII has been removed from Intel Circuit web portal usage logs to protect users' privacy. The proposed architecture makes anonymization of sensitive fields in log data with AES symmetric key encryption and stores it in HDFS for analysis. When de-anonymization is needed, the logs are moved back and the masking areas are decrypted with the same key. Lastly, the quality of anonymization is measured by k-anonymity based metrics. With the increase of individual and organizational privacy concerns, Privacy Preserving Data Mining (PPDM) has begun to gain tremendous importance. However, these techniques affect the success of applications. To provide privacy protection, an Adaptive Utility based Anonymization (AUA) has been proposed, which depends on association mining [2]. Both native and masked data set has been tested.

## IV. SECURITY AND PRIVACY IN BIG DATA

Seeking new ways to take advantage of big data, organizations need secure mechanisms and regulations to guarantee their systems. It is thought that the traditional techniques are ineffective in big data security and privacy issues. Nevertheless, open source or new technologies (if they are not well understood) also host unknown back doors and default credentials [4]. Therefore, confidentiality, integrity and availability of information must be carefully considered.

### A. Security:

Diversity of data sources, data formats, streaming of data and infrastructures may cause unique security vulnerabilities. The Cloud Security Alliance has divided security and privacy challenges in big data into four categories; infrastructure security, data privacy, data management, integrity, and reactive security [5]. Infrastructure security consists of secure distributed programming and security practices in non-relational data stores. Data privacy refers to privacy preserving analytics, encrypted data centre and granular access control. Data management involves secure data storage and transaction logs, auditing and data provenance. In addition, integrity and reactive security include validation, filtering and real time monitoring. On the basis of these proposed issues, authorization and authentication mechanisms must be constituted for both users and applications, and encryption and data masking must be implemented for both data rest and stream.

### B. Privacy:

The development of systems and applications has led to the sources such as databases of big companies, internet and telecommunication under cover of protecting US citizens [2]. Many big data projects like this indicate the violation of people's privacy. The increasing privacy's concern in big data include knowing new and secret facts about people, combining their personal information with other data sets, adding value to their organizations with collected data from unaware people, treating illiterate people by predictive analysis of social media, tagging discriminated people by law enforcement, conflicting laws in different countries, lastly exchanging datasets between organizations [2]. To cope with such complex issues, laws and regulations must be enforced with clear-cut boundaries in terms of unauthorized

access, data sharing, misuse, and reproduction of personal information termination of the individual control about collection and usage of PII. According to the latest news, National Security Agency (NSA) eaves dropped personal data from heterogeneous data

## V. BIG DATA ANALYTICS FOR SECURITY

Big data analytics aims to obtain beneficial information from large scale and complicated data [3]. The increase of stored or streamed data and development of analysis systems has led to using these activities in information security. The anomaly detection, intrusion detection, fraud detection, advanced persistent threats (APT) detection, and forensics from big data has been accomplished by examining the logs, system events, network traffic, website traffic, security information and event management (SIEM) alerts, cyber attack patterns, business processes and other information sources. To detect these attacks, large volume and variety of data is accumulating and associate with network history. The advantageous uses of big data, such as performing without deletion of logs after a certain period, running complex queries on large and unstructured datasets, and facilitating human-computer interactions via visual interfaces, for security is becoming quicker and cheaper than traditional methods [11].

There is no need to delete the cancelled accounts or old logs as they can be used for the purpose of forensics later. In addition, real time and agile decision support applications, automatic defense and risk reduction systems, prediction of attack, determining of zero-day attack duration and tracking of attackers can be developed by analyzing suspicious and malicious patterns from information security data [10].

## VI. ANALYSIS

**Table 1.: Analysis of Cloud Security Approach and Techniques for Challenges**

| Cloud Security challenges | Cloud Security approach | Cloud Security techniques |
|---|---|---|
| Physical security | 1. Data location<br>2.Server storage<br>3.Network | 1.CCTV<br>2.Security guards<br>3.Protective barriers |
| Organisational | 1.Resource planning<br>2.Organisational change management | 1.Software<br>2.Platform<br>3.Infrastructure as a service via a cloud |
| Data security | 1.Identify access management<br>2.Availability and backup<br>3.Data privacy and security | 1.Authentication with email and password |
| Technological | 1.Application development<br>2.Portability<br>3.Lack of interoperability standards | 1.Store<br>1. 2.Filter<br>2. 3.Share<br>3. 4.Analyse<br>4. 5.Virtualise |
| Auditing | 1.Legal challenges<br>2.Business continuity<br>3.Disaster recovery | 1.DNS traffic<br>2.IP flow records<br>3.HTTP traffic |

## VII. CONCLUSION

Big data needs extra requirements for security and privacy in data gathering, storing, analyzing, and transferring. In this paper, we examined studies on big data security and privacy, comparatively. According to the literature, network traffic should be encrypted with suitable standards; access to devices

should be checked; employees should be authorized to access systems; analysis should be done on anonymised data; communication should be made for the secure channel to prevent leakage, and network should be monitored for threats.

Big data privacy, safety and security are the biggest issues to be discussed more in the future, so new techniques, technologies and solutions need to be developed in terms of human-computer interactions or existing technologies should be improved for accurate results. It is hoped that this study would help understand the big data and its ecosystem better and develop better systems, tools, structures and solutions not only for today but also for the future.

## VIII. REFERENCES

[1]. T. Vijay, A. Aiiad, "Big Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center", Procedia Computer Science, vol. 50, pp. 149–156, 2015.

[2]. B. Matturdi, X. Zhou, S. Li, F. Lin, "Big Data security and privacy: A review", Big Data, Cloud & Mobile Computing, China Communications vol.11, issue: 14, pp. 135 – 145, 2014.

[3]. C.L.P. Chen, C.Y. Zhang, "Data Intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, vol. 275, pp.314-347, 2014.

[4]. N. Miloslavskaya, M. Senatorov, A. Tolstoy, S. Zapechnikov, "Information Security Maintenance Issues for Big Security-Related Data", Future Internet of Things and Cloud (FiCloud), pp. 361 – 366, Barcelona, 2014.

[5]. Cloud Security Alliance Big Data Working Group, "Expanded Top Ten Big Data Security and Privacy Challenges", April 2013.

[6]. A.T.H. Ibrahim, Y. Ibrar, B.A. Nor, M. Salimah, G. Abdullah, U.K. Samee, "The rise of "big data" on cloud computing: Review and open research issues", Information Systems, vol. 47, pp. 98–115, 2015.

[7]. P. Adluru, S.S. Datla, Z. Xiaowen, "Hadoop eco system for big data security and privacy", Systems, Applications and Technology Conference (LISAT), Long Island, Farmingdale, NY, pp. 1 – 6, 2015.

[8]. M. Divakar, K. Shrikant, J. Shweta, IBM, "Big data architecture and patterns, Part 1: Introduction to big data classification and architecture", http://www.ibm.com/developerworks/library/bd-archpatterns1/ (Accessed Date: 1 August, 2015).

[9]. B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan, "Big Data and Hadoop-A Study in Security Perspective", Procedia Computer Science, vol. 50, pp. 596 – 601, 2015.

[10]. A. Kumar, L. HoonJae, R.P. Singh, "Efficient and secure Cloud storage for handling big data", Information Science and Service Science and Data Mining (ISSDM), pp. 162 – 166, Taipei, 2012.

[11]. H. Cheng, C. Rong, K. Hwang, W. Wang, Y. Li, "Secure big data storage and sharing scheme for cloud tenants" Communications, China, vol. 12, issue: 6, pp. 106 - 115, 2015.