International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Information Security and Privacy Challenges of Big Data

Mohd Shoaib
Zakir Husain College of Engineering and Technology,
Aligarh Muslim University,
Aligarh, India.
md.shoaibs@gmail.com

Abstract— Digitalization has generated huge volumes of data. This volume of data is from heterogeneous type and stored over public cloud infrastructure. The risk of storing data over cloud and also collection of it from multiple sources has given rise to security and privacy risks. In this paper, security and privacy challenges for Big Data are discussed.

Keywords—Big Data, Data Analytics, Security, privacy.

1. INTRODUCTION

Digitalization has generated enormous volume of data content by adopting the use of internet, social media, and smart phones, among many others. This high volumes of stored chunks of data is called as Big Data. Big Data is defined as the collection of data collected from multiple sources, and analyzed to find knowledge. Analytics on this large volumes of data is done to extract valuable knowledge helpful to individual or business, which in turn help to resolve the objectives of an individual's needs or help in businesses in developing and reforming valuable services for customers. As this huge amount of data is collected from heterogeneous sources, a simple processing and analyzing of it is hard to do using conventional methods, therefore it requires massive parallel processing. These processing tools and techniques to be performed on large volumes of data is called as Big Data Analytics [1].

Big Data Analytics is application of advanced analytic and parallel techniques on massive volume of mixed unstructured, semi-structured and structured data to retrieve enormous valuable information and valuable insights. Big Data Analytics provides opportunity to industries and decision makers to review and better apply their strategies of service for stakeholders. However, massive collection of data also represent a major security and privacy challenge, as the data analytic tools work to store, manage, analyze, visualize, share the collected data and explore more data over internet. Various analytics tools are discussed in [2][3]. The flow of all information over internet turns security and privacy vulnerable. In this paper, an overview of Big Data and their Major security challenges surrounding it will be discussed.
Due to the buildup of collection of data to form Big Data from multiple sources and in massive amounts, data has certain characteristics as shown in Fig. 1, also defined as below:

A. Volume
It is the size of data available for processing to extract knowledge, stored in distributed stores and scaled to Zettabytes in storage volumes.

B. Variety
It defines the heterogeneity present of data terms whether data is structured, semi-structured or unstructured data. Also, different techniques and architecture have to be applied according to variety of data.

C. Veracity
Performing analysis on data and getting accurate results is necessary. Therefore, accuracy and integrity of collected data is highly important.

D. Value
Performing analysis on huge volumes, variety, veracity of data at high velocity to extract a hidden knowledge depicts the value of Big data.

E. Velocity
Velocity of big data defines the speed at with real time streams of data and also new data being stored in storage spaces. Big data should be processed, extract and visualize the results in minimal latency.

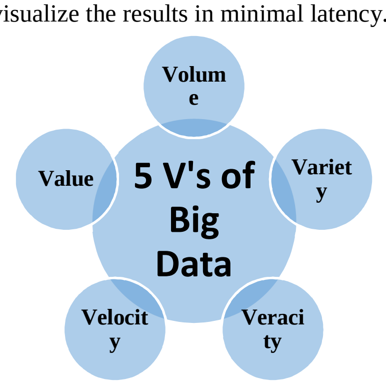

Figure 1. Characteristics of Big Data

2. Security and Privacy Challenges

Big Data is coupled with the on-demand availability of computing power on public cloud infrastructure. The security challenges over cloud arise due to heterogeneous hardware, operating system and software infrastructures for storing and computing on data [4][5].

Cloud Secure Alliance (CSA), a non-profit organization with a mission to promote the use of best practices for providing security assurance within Cloud Computing, has created a Big Data Working Group that has focused on the major challenges to implement secure Big Data services (Cloud Security Alliance, 2013). Each of these aspects faces the following security challenges, according to CSA [6]:

A. Infrastructure Security
- Secure Distributed Processing of Data
- Security Best Actions for Non-Relational Data-Stores

B. Data Privacy
- Data Analysis through Data Mining Preserving Data Privacy
- Cryptographic Solutions for Data Security
- Granular Access Control

C. Data Management
- Secure Data Storage and Transaction Logs
- Granular Audits
- Data Provenance

D. Integrity and Reactive Security
- End-to-End Filtering & Validation
- Supervising the Security Level in Real-Time

Also, Cloud Secure Alliance (CSA) [7] has identified and listed best practices to be followed for securing privacy and security of data. Table I, Table II, Table III and Table IV provides the list of best practices to be followed and the reason for the implementation of these best practices.

TABLE I. BEST PRACTICES FOR INFRASTRUCTURE SECURITY

A. Infrastructure Security		
• **Secure Distributed Processing of Data**		
	Best Practice	*Reason to Implement*
i)	Establish initial trust	To ensure trustworthiness of mappers
ii)	Ensure conformance with predefined security policies	To achieve a high level of security in computations.
iii)	De-identify data	To prevent the identity of the data subject from being linked with external data. Such linking may compromise the subjects' privacy.
iv)	Authorize access to files with predefined security policy	To ensure integrity of inputs to the mapper.
v)	Ensure that untrusted code does not leak information via system resources	To ensure privacy
vi)	Prevent information leakage through output	To ensure security and privacy.
vii)	Maintain worker nodes	To ensure proper functionality of worker nodes.
viii)	Detect fake nodes	To avoid attacks in cloud and virtual environments.
ix)	Protect mappers	To avoid generating incorrect aggregate outputs.
x)	Check for altered copies of data	To avoid attacks in cloud and virtual environments
• **Security Best Practices for Non-Relational Data Stores**		
	Best Practice	*Reason to Implement*
i)	Protect Passwords	To ensure privacy
ii)	Safeguard data by data encryption while at rest	To reliably protect data in spite of weak authentication and authorization techniques applied.
iii)	Use transport layer security (TLS) to establish connections and communication	To maintain confidentiality while in transit; to establish trusted connections between the user and server; and to securely establish communication across participating cluster nodes.
iv)	Provide support for pluggable authentication modules	To certify users are able to program to pluggable authentication module (PAM) interface by using PAM library API for authentication-related services
v)	Implement appropriate logging mechanisms	To expose possible attacks.
vi)	Apply fuzzing methods for security testing	To expose possible vulnerabilities caused by insufficient input validation in NoSQL that engages hypertext transfer protocol (HTTP) to establish communication with users

		(e.g., cross-site scripting and injection).
vii)	Ensure appropriate data-tagging techniques	To avoid unauthorized modification of data while piping data from its source.
viii)	Control communication across cluster	To ensure a secure channel.
ix)	Ensure data replication consistency	To handle node failures correctly.
x)	Utilize middleware layer for security to encapsulate underlying NoSQL stratum	To have a virtual secure layer

TABLE II. BEST PRACTICES FOR DATA PRIVACY

B.	**Data Privacy**	
•	**Scalable and Composable Privacy-Preserving Analytics**	
	Best Practice	*Reason to Implement*
i)	Implement differential privacy	To protect privacy even when data is linked with external data sources.
ii)	Implement Utilize homomorphic encryption	To enable encrypted data to be stored and processed on the cloud.
iii)	Maintain software infrastructure	To avoid exploitation of improperly maintained software, a major vulnerability.
iv)	Use separation of duty principle	To provide robust internal control as well as information security.
v)	Be aware of re-identification techniques	To protect the privacy interests of consumers.
vi)	Incorporate awareness training with focus on privacy regulations	To avoid potential litigation issues into the future.
vii)	Use authorization mechanisms	To secure applications in the infrastructure stack.
viii)	Encrypt data at rest	To prevent access to sensitive information.
ix)	Implement privacy-preserving data composition	To address privacy concerns preemptively.
x)	Design and implement linking anonymized data stores	To ensure privacy.
•	**Cryptographic Technologies for Big Data**	
	Best Practice	*Reason to Implement*
i)	Construct system to search, filter for encrypted data	To balance data confidentiality and data utility.
ii)	Secure outsourcing of computation using fully homomorphic encryption	To enable outsourcing of computation while addressing security and privacy concerns.
iii)	Limit features of homomorphic encryption for practical implementation	To balance computational cost and versatility when handling encrypted data.
iv)	Apply relational encryption to enable comparison of encrypted data	To enable efficient comparison of encrypted data without sharing encryption keys.
v)	Reconcile authentication and anonymity	To balance security and privacy.
vi)	Implement identity-based encryption	To overcome difficulties associated with key management of a public-key cryptosystem.
vii)	Utilize attribute-based encryption and access control	To integrate access control and encryption.
viii)	Use oblivious RAM for privacy preservation	To prevent information leakage that may occur through access pattern analysis implemented by cloud providers.
ix)	Incorporate privacy-preserving public auditing	To enable public auditing without causing privacy concerns.
x)	Consider convergent encryption for deduplication	To improve efficiency of storage usage.
•	**Granular Access Control**	
	Best Practice	*Reason to Implement*
i)	Choose appropriate level of granularity required	To balance complexity and granularity of access control.
ii)	Normalize mutable elements, denormalize immutable elements	To design suitable access control mechanisms.
iii)	Track secrecy requirements	To implement a scalable access control system.
iv)	Maintain access labels	To make policy decisions on data with complex provenance.
v)	Track admin data	To re-key the system when necessary.
vi)	Use standard single sign-on (SSO) mechanisms	To reduce the administrative burden of supporting a large

		user base.
vii)	Employ proper federation of authorization space	To allow data providers to maintain control of access to their data when data analysis spans over multiple providers.
viii)	Incorporate proper implementation of secrecy requirements	To ensure secrecy requirements about data identity, user identity, user purpose and many other requirements.
ix)	Implement logical filter in application space	To prevent data abuse and leakage by application.
x)	Develop protocols for tracking access restrictions	To operate access control system as expected.

TABLE III. BEST PRACTICES FOR DATA MANAGEMENT

C.	**Data Management**	
•	**Secure Data Storage and Transactions Logs**	
	Best Practice	*Reason to Implement*
i)	Implement exchange of signed message digests	To address potential disputes.
ii)	Ensure periodic audit of chain hash or persistent authenticated dictionary (PAD)	To solve user freshness and update serializability issues.
iii)	Employ SUNDR (secure untrusted data repository)	To store data securely on untrusted servers
iv)	Use broadcast encryption	To improve scalability
v)	Apply lazy revocation and key rotation	To improve scalability
vi)	Implement proof of retrievability (POR) or provable data possession (PDP) methods with high probability	To enable a user to reliably verify that data uploaded to the cloud is actually available and intact, without requiring expensive communication overhead
vii)	Utilize policy-based encryption system	To avoid collusion attacks.
viii)	Implement mediated decryption system	To avoid collusion attacks (assuming users are willing to exchange private keys without exchanging decrypted content).
ix)	Use digital rights management	To counter collusion attacks where users are willing to exchange decrypted contents when access control is implemented by means of encryption.
x)	Build secure cloud storage on top of untrusted infrastructure	To store information in a confidential, integrity-protected way—even with untrusted cloud service providers—while retaining service availability, reliability and the ability for efficient data retrieval and flexible data sharing.
•	**Granular Audits**	
	Best Practice	*Reason to Implement*
i)	Create a cohesive audit view of an attack	To answer essential questions following an attack.
ii)	Evaluate completeness of information	To provide a full audit trail.
iii)	Ensure timely access to audit information	To accelerate incident response.
iv)	Maintain integrity of information	To ensure trust in audit data.
v)	Safeguard confidentiality of information	To prevent audit data from reaching the wrong hands.
vi)	Implement access control and monitoring for audit information	To safeguard audit information.
vii)	Enable all required logging	To build up an audit view.
viii)	Use tools for data collection and processing	To find actionable information without being overwhelmed by big data.
ix)	Separate big data and audit data	To enforce separation of duties.
x)	Create audit layer/orchestrator	To facilitate audit data analysis.
•	**Data Provenance**	
	Best Practice	*Reason to Implement*
i)	Develop infrastructure authentication protocol	To prevent malicious parties from accessing data.
ii)	Ensure accurate, periodic status updates	To collect data correctly.
iii)	Verify data integrity	To ensure trust in data.
iv)	Ensure consistency between provenance and data	To ensure provenance information is trustworthy.
v)	Implement effective encryption methods	To maintain security of provenance data.
vi)	Use access control	To prevent abuse and unauthorized disclosure of

		provenance records and data by malicious parties.
vii)	Satisfy data independent persistence	To preserve indistinguishability of provenance data.
viii)	Utilize dynamic fine-grained access control	To allow only authorized users to obtain certain data.
ix)	Implement scalable fine-grained access control	To protect large-scale provenance data.
x)	Establish flexible revocation mechanisms	To prevent access by unauthorized entities.

TABLE IV. BEST PRACTICES FOR INTEGRITY AND REACTIVE SECURITY

D.	**Integrity and Reactive Security**	
•	**Endpoint Input Validation/Filtering**	
	Best Practice	*Reason to Implement*
i)	Use trusted certificates	To ensure trust in communication and prevent attacks
ii)	Do resource testing	To avoid the drawback of managing certificates in a large enterprise
iii)	Use statistical similarity detection techniques and outlier detection techniques	To detect and filter out malicious input
iv)	Detect and filter malicious inputs at central collection system	To block malicious input data without requiring extra computation in resource constrained endpoint devices.
v)	Secure the system against Sybil attacks	To detect and prevent one entity from masquerading as multiple identities in a system.
vi)	Identify plausible ID spoofing attacks on the system.	To detect and prevent an attacker from assuming legitimate identities.
vii)	Employ trusted devices	To detect and prevent Sybil attacks and to prevent the compromise of endpoint devices and applications running on them.
viii)	Design parameter inspectors to examine incoming parameters	To detect and filter out malicious inputs.
ix)	Incorporate tools to manage endpoint devices	To prevent an attacker from compromising endpoint devices and applications running on the devices.
x)	Use antivirus and malware protection systems at endpoints	To prevent an attacker from compromising endpoint devices and applications running on the devices.
•	**Real-Time Security/Compliance Monitoring**	
	Best Practice	*Reason to Implement*
i)	Apply big data analytics to detect anomalous connections to cluster	To ensure only authorized connections are allowed on a cluster, as this makes up part of the trusted big data environment.
ii)	Mine logging events	To ensure that the big data infrastructure remains compliant with the assigned risk acceptance profile of the infrastructure.
iii)	Implement front-end systems	To parse requests and stop bad requests.
iv)	Consider cloud-level security	as data is moved to cloud, deployment of security on cloud infrastructure is necessary
v)	Utilize cluster-level security	To ensure that security methodology for big data infrastructure is approached on all components of infrastructure.
vi)	Apply application-level security	To secure applications in the infrastructure stack.
vii)	Adhere to laws and regulations	To avoid legal issues when collecting and managing data.
viii)	Reflect on ethical considerations	One cannot simply use big data, technical and ethical regulations should be abided.
ix)	Monitor evasion attacks	To avoid potential system attacks and/or unauthorized access.
x)	Track data-poisoning attacks	To prevent monitoring systems from being misled, crashing, misbehaving or providing misinterpreted data due to malformed data.

3. Conclusion

Big Data provides immense opportunity to organizations to process data, explore insight and patterns to help in better decision making. The collection, processing and storing process could lead to many security and privacy risk as identified. In this paper, the security and privacy challenges are discussed.

References

[1] Y. Gahi, M. Guennoun and H. T. Mouftah, "Big Data Analytics: Security and privacy challenges," *2016 IEEE Symposium on Computers and Communication (ISCC)*, Messina, 2016, pp. 952-957.

[2] M. Imran, M. V. Ahamad, M. Haque, and M. Shoaib, “Big data analytics tools and platform in big data landscape”. In *Handbook of Research on Pattern Engineering System Development for Big Data Analytics*, pp. 80-89, (2018). 10.4018/978-1-5225-3870-7.ch006.

[3] T. Choudhury, A. S. Chhabra, P. Kumar and S. Sharma, "A recent trends on Big Data analytics," *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, Moradabad, 2016, pp. 225-231.

[4] B. B. Jayasingh, M. R. Patra and D. B. Mahesh, "Security issues and challenges of big data analytics and visualization," *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Noida, 2016, pp. 204-208.

[5] H. Shuijing, "Big Data Analytics: Key Technologies and Challenges," *2016 International Conference on Robots & Intelligent System (ICRIS)*, Zhangjiajie, 2016, pp. 141-145.

[6] Cloud Security Alliance. (2013). *Expanded Top Ten Security and Privacy Challenges*. Retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

[7] Cloud Security Alliance. (2016). Big Data Security and Privacy Handbook: 100 Best Practices in Big Data Security and Privacy. Retrieved from https://downloads.cloudsecurityalliance.org/assets/research/big-data/BigData_Security_and_Privacy_Handbook.pdf