Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 8, 2025

A SECURITY-CENTRIC ARCHITECTURE FOR BIG DATA

Shaukat Ali*[1], Muhammad Zubair[2], Zulfiqar Ali[3]

*[1,2]Department of Computer Science, Islamia College University, Peshawar, Pakistan,
[3]City University of Science and Information Technology, Peshawar, Pakistan,

*[1]shaukat@icp.edu.pk, [2]zubair@icp.edu.pk, [3]zulfiqarali@cusit.edu.pk

DOI: https://doi.org/10.5281/zenodo.16919143

Keywords
Big Data, Big Data Security, Big Data Architecture, Granular Access Control

Article History
Received: 21 May, 2025
Accepted: 26 July, 2025
Published: 21 August, 2025

Copyright @Author
Corresponding Author: *
Shaukat Ali

Abstract
Big Data is a hot area of research and a buzzword in the research community and industry. The term Big Data refers to the huge volume of data, having properties of high production of even more data (velocity) with different forms of structured and unstructured data (Variety). Big data also has big problems associated with it; one of these problems is the issue related to security and privacy. In this paper, a security-centric architecture is proposed for big data. The proposed architecture protects data at the application layer using a granular access control mechanism and secures the data under management at the storage layer using a data aggregation method. The proposed architecture has a layered approach. This architecture provides a base for the community of big data security. The paper also explains different layers of architecture, especially those related to data security and privacy issues.

INTRODUCTION

Big data is a buzzword in the research community and industry [1]. The term big data refers to a huge volume of data generated in the digital universe from governments, information companies, and individuals. Big Data is not just a database or Hadoop system problem, but it has complex components of storage, processing, and visualization to deliver results to the required target applications. Although Hadoop and database systems constitute the core components and technologies for the processing of large-scale data and its analysis [2-4], Big data, however, is involved in almost all fields of life. Human beings contribute to big data in all its roles. The contribution of users to big data in the form of content generation raises some new problems of security and privacy. Big data has some issues that need to be addressed. For example, the huge volume of data, velocity of data or speed of data production, different types or variety of data (structured, Unstructured, Semi-Structured data) [5]. These properties of big data create new issues related to big data security and privacy [6-8]. The traditional process for big data collection and analytics is shown in Figure 1. In the traditional life cycle of big data, the focus of the organizations and users is the process of analytics and extracting more and more information only from the data, which can lead to security and privacy breaches. In our work, the focus is on the storage of big data in such a way that the privacy of the data and the user remain preserved.

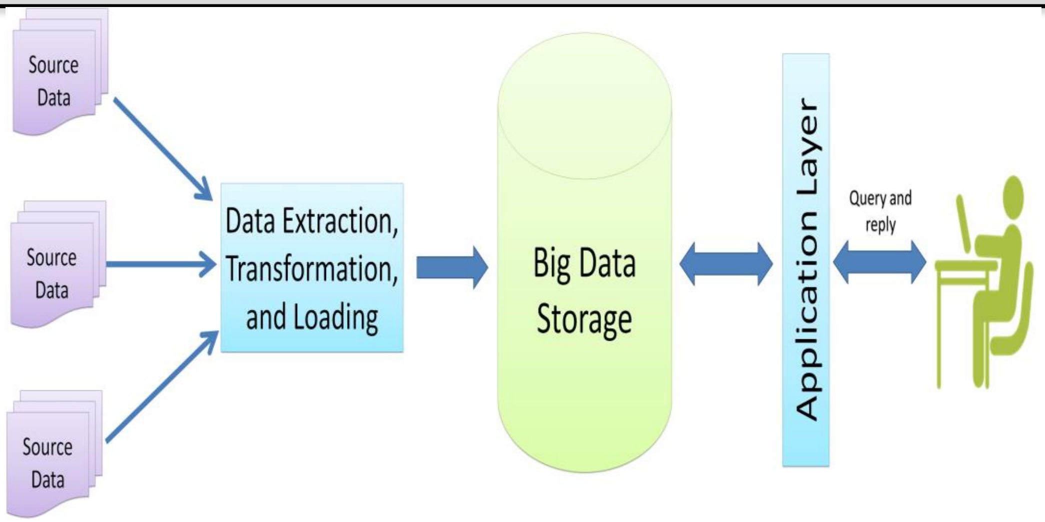

Fig. 1. Life Cycle model for Big Data

In this paper, an architecture for big data security is proposed. The focus of this work is to protect data at different layers during its life cycle, for example, to protect data at the application layer, storage layer, etc. The overall architecture is explained with a focus on data security and privacy related to each layer. A granular access control is proposed for securing data at the application layer. The privacy of the data is maintained under management at the storage layer using data aggregation.

The rest of the paper is organized as Section II discusses the related work. Security-centric big data architecture is presented in Section III. Finally, the paper is concluded with future work in Section IV.

II. RELATED WORK

In the literature, most of the work related to big data has been done on dealing with the huge amount of data, but security and privacy have not been explored that much. Big data security and privacy issues are still open challenges and have already started grabbing the attention of the research community [9]. To start with, the background of big data, some definitions from big data experts and consulting companies are presented first. The IDC (International Data Corporation) defines Big Data as:" A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture, discovery, and/or analysis". The Cloud Security Alliance (CSA) highlighted ten issues in big data security [10]. The CSA established a Big Data Working Group (BDWG) in 2012, which is working on big data security and privacy issues [11]. The NIST Big Data Public Working Group is working on the reference model of big data. The CSA (Cloud Security Alliance) Big Data Working Group (BDWG) [11] and NIST Big Data Public Working Group [12] are working on the big data security issues and big data reference architecture, respectively. The works of both groups are in early stages, and a paper has been published on the big data architecture ecosystem [13]. The Center for Big Data (called The Science and Technology Center for Big Data) was established in the Laboratory of Computer Science and Artificial Intelligence at the Massachusetts Institute of Technology. The group published its initial Execution Plan in March 2013 [14].

Demchenko et al [1] define Big Data using 5Vs: Volume, Variety, Velocity, Value, and Veracity. While dealing with Big Data, it is very important to address the 5Vs. The term Veracity in this research work is used to discuss security issues related to big data. Similarly, a simple definition by Jason Bloomberg [15]: Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. Gartner defines Big Data as follows, with three parts [16] Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

The term big data was used for the first time by Michael et al. [17] in their paper published in the proceedings of the Conference on Visualization in 1997. Before this, Gray et al. [18] used the term Information Explosion to represent the huge volume of data. Currently, the term big data is used very commonly.

III. SECURITY-CENTRIC BIG DATA ARCHITECTURE

Although a little work has been done on big data security architecture, it is still a demanding area of research [13]. All the previous big data architectures have focused on the storage of data, and security is not the main concern. The proposed architecture focuses on the security of data and has a layered approach to secure the data. The layered approach is used because of the variable nature of security needs during storage, transfer, and use. A layer for validating both the source and data is provided, which ensures data validity, thus supporting the integrity of the managerial-level decision-making. After the validation process, the data is aggregated to anonymize the data and thus protect user privacy. The subject-oriented nature of the data has been exploited to allow easy implementation of granular access control mechanisms during the ETL (Extraction, Transformation, and Loading). The term Subject-Oriented means storing similar types of data having the same security label. For example, The Patient data is stored in a subject Patient, and only doctor-type users will have access to the Patient's subject data.

Similarly, all data are stored with a Subject as a security label. It is also easy to check the authorization of the query that accesses cross-subject data. Only those users are allowed to cross query different Subjects' data who have clearance to access all those subjects' data. The consumer of data passes through authentication, and then a granular access control is enforced before data access. It makes it easy to keep the grain level of access as the subject. The sensitive and Personally Identifiable Information are encrypted before storage. Data will also be encrypted while transferring through the network in a distributed environment. Strong and granular access control security is implemented at the application layer. Metadata is stored along with the data across all layers for granular audit. Access to data is ensured to be available to all authorized users. The following are the components/layers of the proposed security architecture (Shown in Figure 2).

Application Layer

The application layer is the data consumers' layer. The user's security is implemented using this layer through an authentication mechanism. A granular access control is implemented in this layer. The application layer also has some other responsibilities to present data to the user for better and easier understanding and analysis.

Storage layer

In this section, the storage layer is discussed concerning Security and Privacy. Storage is one of the basic components of big data. The storage of big data is normally distributed across the network. The following are some mechanisms proposed for a security-centric Big Data architecture to store data securely and protect its privacy concerning individuals' data.

1) **Subject-oriented data:**

Big data has some privacy issues. In big data, there is a lot of information collected and stored about individuals. The privacy protection and security of individuals' information are a concern in the era of big data. It contains some Personally Identifiable Information (PII), and the security of such data is important from the data provider's point of view. If the data is stored according to Subject, the privacy of individuals can be protected from unauthorized users. If a user cannot cross-query the data having different Subjects, he/she cannot connect different components of data to find the person's private data. If the data in big data infrastructure is stored according to subject (e.g., sales, accounts, etc.), the privacy risk can be minimized.

2) **Aggregated Data:**

The data aggregation means the summarization of similar types of data to a single record instead of a few records. This aggregation will ensure the anonymity of private data and will ensure privacy. The aggregation of data not only ensures the privacy of data, but also helps the data analytics. The analysis of aggregated data is easier compared to granular data in terms of time and processing.

Network Infrastructure Layer

One of the basic components of big data is the volume of data. Big data has a huge volume of data; therefore, it can be stored in a distributed environment. Data may be stored in a cloud-type environment and will be connected through a high-speed network. The main aim of this research work is data security in all its stages. The data needs to be secured in place (storage) or in transit (network). The data through the network can be easily secured using encryption.

Source Systems

Many sources contribute to big data. The data may be structured, unstructured, or semi-structured data. Some of the data comes from organizations' data, but much of the data comes from individuals through social media and blogs [19]. Both the organization's data and individuals' data need security and have some privacy issues that need to be addressed.

Source System and Data Validation Layer

The developer of the big data environment will validate both the source and the data itself. This ensures data integrity, which means the correctness of data. Data integrity is crucial for data security because decisions are made through data analytics techniques. Therefore, decisions will be accurate if they are based on correct data. Similarly, validating the source is important because data from reputable sources has greater value. We must protect data from illegal use by consumers and establish trust in the data environment developer.

Data Aggregation and Classification

When data is coming from a source system, its source and the data itself are verified first. After source and data validation, the data is aggregated to hide the individual's private data. While aggregating the data, it will be classified according to subject area. There may be more than one subject area in a single storage. First, it can be used for fast data analytics. Second, it can be used for security. For example, if the data is stored according to subject area, the authorization and granular access control mechanism can easily be implemented for such data using the subject of the data as a grain.

Data Availability and Transparency Layer

The size of big data itself is a problem in big data [20]. Big data needs distributed storage to fulfill the requirement of storage. Transparency is needed in this distributed environment, which is used for the storage of big data. There is location transparency, which ensures the security of data location from the attacks of unauthorized users. Similarly, the transparency property hides the complexity of the system from the end users.

Meta Data/Audit Log layer

Metadata is an important component for data management. Metadata and its management become more important when dealing with big data and often multi-sourced complex data. Metadata can be used across for easy and clear interpretation of the enterprise. Metadata or Audit log can also be used for the security of data. Whenever there is a security breach of data, it can easily be checked through the audit log.

Data Ownership

According to Loshin [21], data has fundamental value, having added value as a byproduct of data processing. At the core, the degree of ownership is driven by the value that each interested party derives from the use of that information. There are different types of data owners. For example, the creator of data, the person who modifies data, the person who compiles data, Purchaser/Licenser as Owner, etc [22]. These different users are called the Subject of data, and the data item/record he/she can access is called the Object. It is required in a big data environment to ensure that the ownership of data means that only the authorized users can access the required information. A granular access control is needed to address such types of problems.

A labeled type security is needed in such situations. The data is stored according to Subject in the proposed method. Subject means a similar type of data is stored collectively using the same security label. Similarly same type of security label is also assigned to each user group. Each user belongs to a user group. The user has access rights of a particular user group in which the user is a member. Whenever a user wants to access a data object from a big data environment, the user's access rights label is compared to the Subject of the data. If the security label of data and the user's access labels match, the user will be given access to that particular group of data; otherwise, the access will be denied.

J. Encryption

Encryption is the conversion of data to some other format to make it secure. Encryption is one of the best options for data or information security, and it is considered to be the final layer of security [23]. Although encryption is a good option for security, it is very difficult to encrypt all data in a big data environment. Therefore, only sensitive data will be encrypted at rest (at the storage layer), and all data will be encrypted at transmission through the network. Since big data is distributed across a distributed system and will be connected through a network, therefore, encryption is the best option when data is traveling through a network.

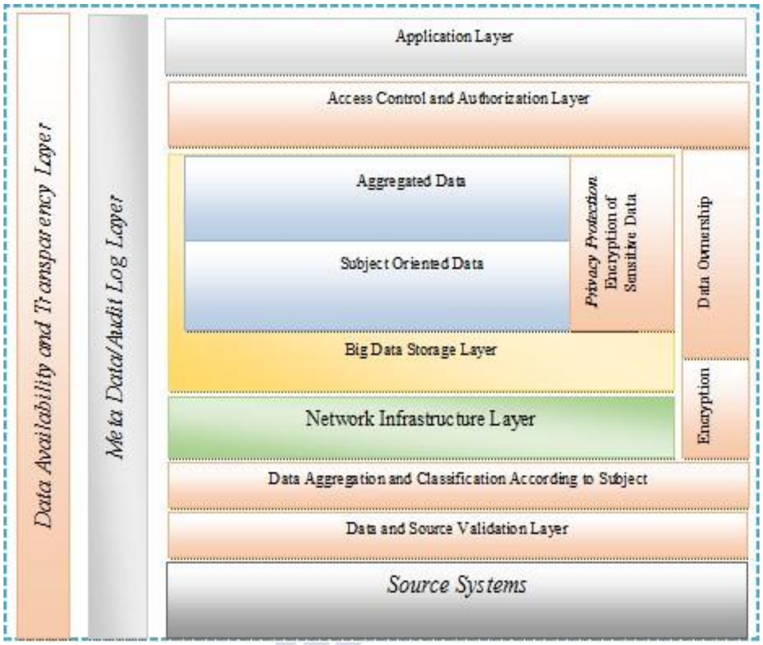
Data Availability and Transparency Layer
Meta Data/Audit Log Layer
Application Layer
Access Control and Authorization Layer
Aggregated Data
Subject Oriented Data
Privacy Protection
Encryption of Sensitive Data
Data Ownership
Big Data Storage Layer
Network Infrastructure Layer
Encryption
Data Aggregation and Classification According to Subject
Data and Source Validation Layer
Source Systems

Fig. 2. Big Data security-centric architecture

IV. CONCLUSION AND FUTURE WORK

In this paper, an architecture for big data security is proposed that consists of different layers. A layered approach is used for the big data life cycle in the proposed architecture, and each layer is explained in a security-centric. The data is secured at the storage layer through the subject-oriented property of big data that provides granular access control, and is helpful for the security and privacy of individuals' data. In the future, the proposed architecture will be implemented and tested for different privacy and security issues arising in big data analytics.

REFERENCES

[1] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *2013 International conference on collaboration technologies and systems (CTS)*, 2013: IEEE, pp. 48-55.

[2] P. Alvaro, T. Condie, N. Conway, K. Elmeleegy, J. M. Hellerstein, and R. Sears, "Boom analytics: exploring data-centric, declarative programming for the cloud," in *Proceedings of the 5th European conference on Computer systems*, 2010, pp. 223-236.

[3] Z. Liu, H. Zhang, and L. Wang, "Hierarchical spark: A multi-cluster big data computing framework," in *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, 2017: IEEE, pp. 90-97.

[4] I. Alam, A. Hameed, and R. A. Ziar, "Exploring sign language detection on smartphones: A systematic review of machine and deep learning approaches," *Advances in Human-Computer Interaction*, vol. 2024, no. 1, p. 1487500, 2024.

[5] I. Khan, S. Khusro, and I. Alam, "Smartphone distractions and its effect on driving performance using vehicular lifelog dataset," in *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2019: IEEE, pp. 1-6.

[6] O. Tene and J. Polonetsky, "Privacy in the age of big data: a time for big decisions," *Stan. L. Rev. Online*, vol. 64, p. 63, 2011.

[7] Y. Yao, L. Zhang, J. Yi, Y. Peng, W. Hu, and L. Shi, "A framework for big data security analysis and the semantic technology," in *2016 6th International Conference on IT Convergence and Security (ICITCS)*, 2016: IEEE, pp. 1-4.

[8] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1891-1899, 2017.

[9] I. Alam, A. Basit, and R. A. Ziar, "Utilizing Age-Adaptive Deep Learning Approaches for Detecting Inappropriate Video Content," *Human Behavior and Emerging Technologies*, vol. 2024, no. 1, p. 7004031, 2024.

[10] P. K. Murthy, "Top ten challenges in Big Data security and privacy," in *2014 International test conference*, 2014: IEEE, pp. 1-1.

[11] R. Patgiri, "A taxonomy on big data: Survey," *arXiv preprint arXiv:1808.08474*, 2018.

[12] W. Chang, D. Boyd, and O. Levin, "NIST big data interoperability framework," *Architectures White Paper Survey*, 2019.

[13] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," in *2014 International conference on collaboration technologies and systems (CTS)*, 2014: IEEE, pp. 104-112.

[14] M. Stonebraker, S. Madden, and P. Dubey, "Intel" big data" science and technology center vision and execution plan," *ACM SIGMOD Record*, vol. 42, no. 1, pp. 44-49, 2013.

[15] C. L. Borgman, "Big data and the long tail: Use and reuse of little data," 2013.

[16] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *AIP conference proceedings*, 2015, vol. 1644, no. 1: American Institute of Physics, pp. 97-104.

[17] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization'97 (Cat. No. 97CB36155)*, 1997: IEEE, pp. 235-244.

[18] M. Camilli, "Coping with the State Explosion Problem in Formal Methods: Advanced Abstraction Techniques and Big Data Approaches," 2015.

[19] Z. Ali, I. Alam, F. Idrees, S. Muhammad, M. H. U. Qureshi, and A. Basit, "Predicting Optimal Links in Complex Human Networks Using Structural Pattern Analysis," *Spectrum of Engineering Sciences*, vol. 3, no. 8, pp. 252-267, 2025.

[20] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, no. 2012, pp. 1-16, 2012.

[21] R. E. Gliklich, M. B. Leavy, and N. A. Dreyer, "Principles of registry ethics, data ownership, and privacy," in *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 4th edition*: Agency for Healthcare Research and Quality (US), 2020.

[22] M. Jan, S. Khusro, I. Alam, I. Khan, and B. Niazi, "Interest-Based Content Clustering for Enhancing Searching and Recommendations on Smart TV," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 3896840, 2022.

[23] S. Ali, A. Rauf, and H. Javed, "A technique for handling range and fuzzy match queries on encrypted data," *Int. Arab J. Inf. Technol.*, vol. 10, no. 3, pp. 239-244, 2013.