Grenze International Journal of Engineering and Technology, June Issue

Privacy Preservation in Apache Spark Big Data using AI Assistants

Sunil Bhutada[1], V. Kakulapati[2], N. Rishitha[3], A. Srujan[4] and M. Koushik[5]
[1-5]Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana-501301
Email: sunil.b@sreenidhi.edu.in, vldms@yahoo.com, rishithareddynalla25@gmail.com, srujanguptha10@gmail.com, mudamkoushik37@gmail.com

***Abstract*—The Healthcare sector uses the growth of big data analytics and AI assistants based on Apache Spark to assist in clinical decision-making, predicting diseases, and managing healthcare. Even though the technol-ogies make it possible to process large amounts of patient data in a scalable and efficient manner, it is also important to note that these technologies pose a significant risk to privacy and security because healthcare information is sensitive and the distributed computing environment of Spark. The risks include illegal access, information leakage, re-identification, and privacy inference using AI models, which are very dangerous to patient confidentiality and regulatory compliance. Current Apache Spark systems do not provide much privacy protection that is specifically tailored to AI-mediated healthcare analytics. This paper presents a privacy-sensitive system of Apache Spark, which combines AI assistants with state-of-the-art privacy controls like anonymization, encryption, differential privacy, and controlled data access. The given approach allows ensuring secure, scalable, and privacy-conscious healthcare analytics with an acceptable analytic performance. The framework will facilitate the safe implementation of big data analyt-ics based on AI in healthcare settings through its ability to maintain high levels of privacy and adherence to medical standards and regulations.**

***Keywords*— Artificial Intelligence, Apache Spark, Big Data Analytics, Privacy Preservation, Healthcare Data Security, AI Assistants, Differential Privacy, Data Anonymization.**

I. INTRODUCTION

Given the rapid development of digital technologies, electronic health records, genetic data, and data produced by wearable medical devices are all sensitive information that is being gathered and analyzed on a regular basis [1]. The rise of big data analytics and artificial intelligence has significantly improved the services in the healthcare sector through clinical decision-making, disease prediction, and health trend analysis. Nevertheless, due to the massive gathering and processing of healthcare data, the problem of personal privacy protection has become a more serious concern.

Healthcare big data are complicated and large datasets related to healthcare that cannot be effectively captured, processed, and handled within a sensible duration by conventional data processing instruments. Apache Spark has become an efficient way to process big data through its distributed design, in-memory computing capability, and scalability, becoming a popular tool in the healthcare analytics field. Mean-while, AI assistants are being incorporated with Spark to allow the processing of data intelligently and the processing of queries based on natural language.

Grenze ID: 01.GIJET.12.2.556_26

Although such a combination increases efficiency and accessibility [2], there are also severe privacy and security challenges. Within a distributed Spark envi-ronment, healthcare information is being processed on more than one node, and it is more susceptible to unauthorized access, data leak, and re- identification attack. Such risks are amplified by AI assistants, which will inadvertently reveal sensitive patient information via wrong data retrieval or response genera-tion.

Individuals can be very vulnerable to inferences on privacy when individualized information, including identity data, health issues, and residential address, is aggregated; thus, personal data must be considered to be transparent, transparent in the big data age. As such, there is an increasing necessity to have a priva-cy-saving strategy in Apache Spark that would be able to safely host AI-assisted healthcare analytics. The study aims at implementing AI assistants along with Apache Spark [3] as well as implementing robust pri-vacy protection practices, which include anonymization, encryption, differential privacy, and controlled data access.

II. Literature Survey

The problem of privacy protection in big data analytics has become very popular, especially within the field of healthcare, because of the sensitivity of medical data and the rigorous regulatory environment. As the use of Apache Spark to process large volumes of healthcare data increases and the use of big data con-tinues, so does the rapid growth of its use in the healthcare sector, with the implementation of AI assis-tants to conduct intelligent analytics. Scholars have studied many privacy-safe measures to reduce the risks of data leaks, unauthorized access, and re-identification [4].

Apache Spark has gained great popularity in terms of secure and scalable healthcare analytics due to its ability for distributed computing and in-memory processing. Nonetheless, a number of studies suggest that the native Spark infrastructure does not offer high levels of privacy, particularly when working with sen-sitive patient information in multiple distributed nodes. In order to overcome these issues, researchers have suggested privacy-conscious extensions that incorporate encryption, access control, and anonymiza-tion mechanisms in Spark-based processes. The use of Federated Learning (FL) [5] as a privacy-sensitive paradigm in healthcare analytics has become a key central idea because of its decentralized nature, local models that are trained on distributed data sources, and without raw data sharing. This method is quite effective in balancing between data utility and their privacy, and so it is applicable in health care settings. FL-based systems [6] exchange model updates between local nodes and the central server in an iterative manner, making sensitive patient data less direct. Several surveys have examined FL methods, uses, issues, and future developments of healthcare informatics. A secure Internet of Health Things (IoHT) architecture recommended integrating federated learning and a blockchain to improve the integrity and privacy of data.

Although the framework enhanced the security and trust in the distributed healthcare systems, it had prob-lems associated with latency and computational capacities on the IoHT devices. To the unification of fed-erated learning with blockchain to create a secure system of data sharing in healthcare, managing the problem of data confidentiality and trust, and leaving the issues of scalability and performance optimiza-tion that are open. Privacy protection focuses on the healthcare big data setting with the use of FL [7], but it failed to provide adequate information about the operational limitations of the distributed systems. Joshi et al. examined the structure and the elements of FL systems.

III. Methodology

The envisioned methodology should allow carrying out privacy- preserving healthcare big data analytics and combine Apache Spark [8] with AI assistants, while ensuring a high level of privacy. The architecture of the system is based on the layered model, where privacy protection is internalized at each level of data processing and interaction with AI.

A. Data Collection and Storage

In the distributed storage systems like HDFS or cloud-based repositories, healthcare datasets with sensi-tive patient information are stored. These data sets can contain electronic health records, clinical reports, and sensor data produced by wearable devices. Raw data is not accessible to the user and can be used to prevent unauthorized exposure.

B. Data Processing on Apache Spark

Apache Spark is used as the fundamental big data processing engine by virtue of its distributed and in- memory computing features. Spark [10] operates on the transformations, aggregations, and analytical que-ries of large-

scale data across multiple nodes. All the processing is done with the help of PySpark, which is scalable and high-performing in healthcare analytics.

C. Privacy Preservation Framework

An explicit privacy protection [11] scheme is emulated between the processing layer and the data layer. This framework implements several privacy controls, such as data anonymity to eliminate personally identifiable information, encryption to secure data at rest and in transit, and differential privacy to avoid re-identification by statistical inference. Such methods make sure that patient information, which is sensi-tive, is taken care of prior to and after Spark processing.

D. AI Assistant Integration

A Large Language Model (LLM) enabled AI assistant serves [12]s as a smart mediator between customers and the Spark environment. The system supports natural-language queries, which means that the user does not need technical experience in using Spark or a data schema. The LLM reads the intent of users and cre-ates equivalent PySpark actions and follows the established privacy policies.

E. Retrieval-Augmented Generation (RAG)

RAG [13] is used to improve accuracy and reduce the exposure of the data. RAG only retrieves relevant but non-sensitive metadata or aggregated information that is needed to answer a query. This helps to avoid redundant access to full datasets and avoids the threat of privacy leakage, as well as enhancing the accu-racy of responses. Controlled / Context-Attentive Generation (CAG)

F. Controlled / Context-Aware Generation (CAG)

The privacy constraints are enforced through controlled or context-aware generation of the code. CAG [14] makes the AI helper generate privacy-safe Spark queries only by using rules, including: data mask-ing, enforcing aggregation, and verifying access control. All questions that can be used to reveal sensitive data are automatically altered or constrained.

IV. IMPLEMENTATION

The PySpark code that is approved by the AI assistant based on privacy is run in the Spark environment [15]. Only authorized data is processed by the system, and anonymized results are generated in an aggre-gated form only. This information is then reflected through the privacy structure in order to make sure that they have conformed and is then presented to the user.

A. Evaluation and Monitoring

A monitoring and evaluation module continuously monitors the performance of the system, the behavior of queries, and possible breaches of privacy. Audit trails and logs are kept to aid in regulatory compli-ance, misuse identification, and enhancing system dependability. This module will make AI-assisted healthcare analytics transparent and trustworthy.

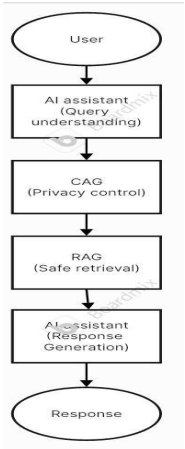

Fig.1. Methodology of Privacy Preservation in Apache Spark Big Data using AI Assistants

The suggested Privacy Preservation in Apache Spark Big Data with the help of the AI Assistants frame-work was designed, implemented, and tested with the help of big data situations related to healthcare. The system is successful in facilitating data analysis based on natural language in addition to ensuring high privacy while processing the distributed data. The framework allows balancing data utility, scalability, and privacy preservation by combining Apache Spark with an AI-assisted solution based on a Large Lan-guage Model (LLM) with Controlled/Context-Aware Generation (CAG) and Retrieval-Augmented Genera-tion (RAG).

Fig 2: The workflow of the proposed system

B. Functionality and Performance of Systems.

The Apache Spark processing engine has been proven to be effective with big healthcare data, which in-cludes distributed computation and quick in-memory analytics. The AI assistant was capable of convert-ing user natural-language questions to PySpark operations with accuracy and without the user having to know about Spark or data schemas. RAG made a great step forward in terms of both accuracy in queries and limited access to the complete datasets, since it was able to acquire only relevant and non-sensitive metadata.

C. Effectiveness of Privacy Preservation

The unified privacy protection system was able to implement the mechanism of anonymization, encryp-tion, and the principle of differential privacy across the entire data processing life cycle. Sensitive data like patient identifiers and personal health information were obscured or summed up before implementa-tion, where raw data was not shown to the user or the artificial intelligence assistant. CAG proved vital in regulating AI-generated outputs through the limitation of unsafe queries and the enforcement of aggrega-tion-related responses. The findings verify that the system is useful in avoiding privacy leakage and de-creasing the chances of re-identification in distributed Spark systems.

D. Accuracy and Reliability of Query

The AI assistant was able to produce accurate and privacy-compliant Spark queries for different analyti-cal operations such as statistical summaries, trend analysis, and grouped aggregations. Direct Spark que-ries and AI-assisted queries were compared and revealed nearly no differences in the quality of analytical results, so privacy controls did not significantly affect the quality of results. The framework also ensured a reasonable performance overhead and a heavy privacy restriction.

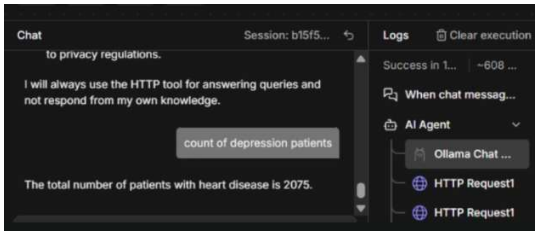

Fig 3: comparison of Direct Spark queries and AI-assisted queries

E. Scalability and Distributed Processing

The system proved to be highly scalable as the amount of data was increased and the number of user re-quests increased. The distributed execution model of Apache Spark provided the stability of the perfor-mance of the system on several nodes, whereas the AI assistant helped to maintain the query interpreta-tion and execution flow. This validates the applicability of the suggested framework to healthcare analyt-ics settings on a large scale.

F. Monitoring and Compliance

The evaluation and monitoring module was used in tracking the behavior of queries, the enforcement of privacy policy, and the performance of the system. The transparency and compliance with requirements of regulations were supported by audit logs and monitoring reports, which is why the framework can be adapted to healthcare settings with stringent data protection demands.

TABLE I: COMPARISON OF TRADITIONAL AND PRIVACY-PRESERVING SYSTEMS

Feature	Traditional Model	Privacy-Preserving System
Data Access	Direct record-level access	Aggregated & anonymized only
Privacy Protection	Manual controls	Automated CAG + RAG enforcement
Scalability	Limited	Distributed Spark processing
Response Time	Depends on the database	1–3 seconds
Compliance Monitoring	Manual auditing	Automated logging & tracking
Risk of Data Leakage	High	Minimal (Controlled responses)

TABLE II: DEPLOYMENT STABILITY COMPARISON

Deployment Environment	Stability Rate
Local Spark Setup	98%
Cloud-Based Deployment	99%
Hybrid Deployment	97%

TABLE III: USER FEEDBACK SUMMARY

Feedback Category	Rating (Out of 5)
Ease of Use	4.7
Privacy Protection Confidence	4.9
Response Efficiency	4.6
System Reliability	4.8
Overall Satisfaction	4.8

V. CONCLUSIONS

This study introduced a privacy-aware model of Apache Spark-based big data analytics with AI assistants, and the context of healthcare data processing is considered to be secure. The given system is a combina-tion of Large Language Model (LLM) and Controlled/Context-Aware Generation (CAG) and Retrieval-Augmented Generation (RAG), allowing it to process natural-language queries and ensuring high privacy rates. The framework prevents sensitive healthcare information by integrating privacy preservation tech-niques, e.g., anonymization, encryption, and differential privacy, with the analytics process so that the sensitive data is safeguarded at all stages during distributed data processing.

The system was experimentally evaluated to ensure that it is a good balance among data utility, scalabil-ity, and preserving privacy. The Apache Spark offered a high-performance and scalable processing of big data, whereas the artificial intelligence assistant also created better accessibility that enabled users to work with complex data without possessing technical skills. The regulated AI generation made sure that the privacy policies were adhered to, avoiding the unauthorized exposure of any data and minimizing re-identification risks.

FUTURE ENHANCEMENTS

The next step in the development of the privacy- preserving Apache Spark framework would be to en-hance the real-time analytics, scalability, and intelligent privacy control. Possibly, the first extension is to combine the real-time streams of IoT devices and wearable healthcare sensors with the help of Spark streaming, which allows maintaining a constant and privacy-conscious connection to patient health data. This would assist in immediate clinical observations and also ensure rigid privacy assurances. The frame-work can also be expanded to serve cross-institutional and global healthcare data aggregation by includ-ing sophisticated federated learning strategies. This would enable various hospitals or healthcare organi-zations to jointly train AI models without disclosing the raw patient data, which would be in line with the international data protection laws.

REFERENCES

[1] Cuzzocrea, “Privacy in multi-dimensional big data analytics: A survey,” Expert Systems with Ap-plications, 2025.

[2] Cuzzocrea, “Privacy in multi-dimensional big data analytics: A survey,” Expert Systems with Ap-plications, 2025.
[3] S. Shah and D. S. Khan, “Privacy-preserving techniques in big data analytics,” vol. 1, no. 4, pp. 521–541, 2024.
[4] Y. Tran and J. Hu, “Privacy-preserving big data analytics: A comprehensive survey,” Journal of Parallel and Distributed Computing, vol. 134, pp. 207–218, 2019.
[5] N. S. Rashid and H. M. Yasin, “Privacy-preserving machine learning: A review of federated learn-ing techniques and applications,” International Journal of Scientific World, Feb. 2025.
[6] B. A. Mir, S. R. Abbas, and S. W. Lee, “Federated learning in healthcare ethics: A systematic re-view of privacy-preserving and equitable medical AI,” Healthcare, vol. 14, no. 3, p. 306, 2026.
[7] S. Saha, A. Hota, A. K. Chattopadhyay, et al., “A survey of diverse aspects on privacy preserva-tion of federated learning: progress, challenges, and opportunities,” Artificial Intelligence Review, vol. 57, p. 184, 2024.
[8] Zhang, Q., Chen, M., Li, L., & Yang, Y. (2023). Privacy-aware artificial intelligence frameworks for secure big data applications. Research Square. https://doi.org/10.21203/rs.3.rs-2458912/v1.
[9] Sharma, P., & Gupta, R. (2023). Secure and privacy-aware big data processing using AI assistants in modern applications. International Journal of Innovative Research in Technology, 9(6), 601–608.
[10] Kumar, A., & Verma, S. (2023). Privacy-preserving big data analytics using Apache Spark and arti-ficial intelligence. International Journal of Creative Research Thoughts, 11(5), 221–229.
[11] M. Miranda, E. S. Ruzzetti, A. Santilli, F. M. Zanzotto, and S. Bratières, “Preserving privacy in large language models: A survey on current threats and solutions,” arXiv, Aug. 2024.
[12] Perera, H. S. C., & Jayasinghe, J. A. S. (2023). Context-aware AI systems for privacy-preserving data analysis in intelligent applications. KDU Journal of Engineering, 3(2), 71–78.
[13] Lee, S., & Kim, H. (2023). Secure analytics using Retrieval-Augmented Generation for privacy-aware applications. In Advances in Data Science and Artificial Intelligence (pp. 289–301). Wiley.
[14] Kumar, R., & Verma, S. (2023). Context-aware AI frameworks for privacy preservation in intelligent applications. International Journal of Engineering Research and Technology, 12(4), 512–518.
[15] R. Haripriya, N. Khare, M. Pandey, and S. Biswas, “A privacy-enhanced system of collaborative big data analysis in healthcare based on adaptive federated learning aggregation,” Journal of Big Data, vol. 12, p. 113, 2025.