International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Exploring Frontiers in Big Data: Privacy-Preserving Exchange and Data Lake Innovations

Priya Kumari[1], Dr. Gobi Natesan[2], Manish Kumar[2]

[1]Jain (Deemed to be) University, Bangalore, India
[2]Assistant Professor, Department of CS&IT, Jain University, Bangalore, India
[2]Assistant Professor Cum Jr. Scientist, Department of SS&AC, BAU Sabour, India
[1]Priyakumari9608067@gmail.com,[2]Gobi.n@jainuniversity.ac.in

DOI: https://doi.org/10.55248/gengpi.5.0324.0637

ABSTRACT

With an emphasis on two crucial areas—privacy-preserving data exchange and data lake innovations—"Exploring Frontiers in Big Data: Privacy-Preserving Exchange and Data Lake Innovations" explores the rapidly changing field of big data management. As enormous datasets are shared within data federation structures, big data interchange transcends standard paradigms and presents substantial issues for privacy preservation. This work offers a perceptive summary of concepts and problems associated with large data exchange that protects privacy, examining various strategies and tactics. In addition to future research directions focused at increasing privacy preservation in massive data sharing, application scenarios including social networks, bio-informatics tools, and smart city systems are taken into consideration. Meanwhile, massive data lakes have become essential parts of contemporary data architectures, requiring developments in models, frameworks, and technique.

Keywords: Big data, data management, analytics, research directions, innovations, privacy-preserving techniques, data exchange, data federation architectures, data lakes, and privacy.

INTRODUCTION

Large dataset interchange is essential in the continuously changing world of big data technology and applications. The exchange that takes place in federated environments such as healthcare systems has a number of issues, the most important of which being privacy preservation. Ensuring the confidentiality and security of sensitive information during exchange operations becomes increasingly important as big data continues to permeate numerous sectors.

Data integration techniques have historically been used to manage heterogeneous datasets. But in federated settings, when several entities work together, massive data sharing techniques take precedence over integration protocols. The main obstacle is to enable smooth data transmission while protecting privacy concerns. Imagine a situation in which medical facilities share clinical datasets to support joint research or patient care projects. Information sharing is necessary to improve healthcare outcomes, however there are worries that private patient information, including Social Security numbers (SSNs), may be disclosed to uninvited parties.

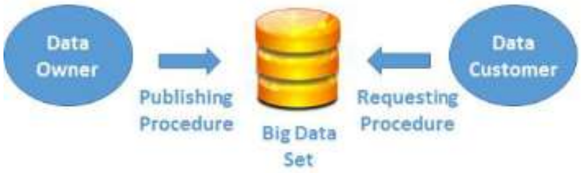

Fig 1: Big data exchange system

To reduce these concerns, it becomes clear that strong privacy-preserving algorithms and methods are required.

This study investigates potential solutions to these problems and examines the complications surrounding privacy in big data exchange networks. Motivated by practical uses, especially in the healthcare field, we investigate the subtleties of privacy maintenance in federated settings. By means of an extensive examination of current literature and developing patterns, our goal is to illuminate the changing terrain of privacy-improving technology in large. Within the framework of healthcare systems, and especially in the context of the COVID-19 pandemic, it is not uncommon for patients to be screened at one hospital (Hi) and then need to be screened at another (Hj). In this case, Hospital Hi is required to share with Hospital Hj the clinical dataset (Di) that was gathered during the first screening campaign. Hospital Hi is dedicated to working with Hospital Hj to improve the overall clinical/screening process, but it is worried that when Hospital Hj accesses Di for big data exchange, it might unintentionally obtain private patient data, like Social Security numbers (SSNs), that is protected by strict privacy laws. Notably, even while SSNs are usually relevant, they do not directly support the objectives of the targeted clinical data exchange procedure. Big data lake techniques: how can several heterogeneous big data repositories be successfully and efficiently integrated into a single big data lake? How can large data be queried within a big data lake in an efficient and effective manner? A common use case for the core privacy-preserving big data exchange system is shown in Figure 2, which includes three hospitals: Hospital A, Hospital B, and Hospital C. These healthcare facilities participate in privacy-preserving big data exchanges that facilitate the extraction of valuable insights from shared datasets while respecting privacy preservation guidelines, which are essential for making decisions in the healthcare setting. Driven by these reasons, this work provides a summary of models and problems in the field of privacy-preserving big data exchange research along with important future objectives for the advancement of next-generation research. The following sections outline the most recent approaches to privacy-preserving massive data exchange problems (Section II) and future directions for this field of study (Section III). In Section IV, final thoughts and directions for additional research are discussed.

This study has been prompted by these difficult questions. To address these latter, we present in this paper: (i) an overview of the cutting-edge methods at the core of big data lake research, particularly those that aim to address these difficult questions; and (ii) novel open problems and issues that inform future research directions on advancing the trend in big data lake research.

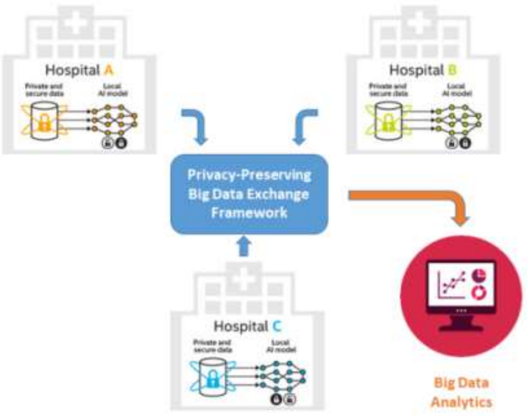

Fig.2: Application Scenario: Healthcare Analytic Big Data Exchange Framework with Privacy Preservation.

I. Up-To-Date Proposals In Big Data Exchange Research That Protects Privately:

Several strategies have been used in the literature to draw attention to massive data exchange research issues that protect privacy. We list the ones that we think are more important in the following summary.

Draw attention to the fact that data has emerged as the most valuable asset in recent years. The number of online marketplaces for exchanging data is increasing. These marketplaces assist both data consumers and data owners in locating suitable services and in publishing their datasets. However, data is a unique commodity as opposed to conventional items like food and clothing. Copyright and privacy protection are particularly difficult in today's data exchange markets. Furthermore, many organizations who possess data services find it challenging to maintain data services, which calls for specialized IT skills. Ensuring fairness and privacy in the rapidly developing field of big data exchange while enabling transactions without the need for middlemen is crucial. In order to establish an environment that supports peer-to-peer data exchange, this study suggests a novel solution that makes use of blockchain technology. Our method, in contrast to conventional marketplaces, does away with the need for middlemen, giving data owners an easy way to monitor data usage and safeguard privacy and copyright. Motivated by the increasing significance of reliable data in increasingly digitalized societies, we tackle the shortcomings of current trading procedures and offer an equitable framework for big data sharing. Our method protects anonymity and allows for autonomous and fair transactions through the use of smart contracts and unaware transfer protocols. The incorporation of an electronic check system improves the convenience and equity of transactions while enabling.

II. Big Data Lakes: Cutting Edge Research Ideas:

Significant progress has been made in the subject of big data lake research in recent years, specially in the three main research axes (models, frameworks, and methodologies) listed in Section I. An overview of some of the most noteworthy proposals in this field is given in this section.

In the current big data landscape, provides an extensive overview of the definition and rising prominence of big data lakes. The authors recognize the rise of data lakes as the go-to method for planning and building next-generation systems that can handle changing big data problems. Even with data lakes' allure, big businesses have a lot of worries and uncertainties about how to use them. In-depth discussion of data lake conception is provided, along with strategies and ideas to deal with related issues.

Examines relevant research challenges and dives deeper into the principles of huge data lakes. The authors draw attention to the rise of two other ideas that have emerged in addition to big data: data lakes and rapid data. People wonder if these are truly new concepts, or just new titles for the well-worn term "big data" in marketing. This study aims to clarify how these three ideas relate to one another and highlight any potential overlaps or differences between them. This paper describes a practical use of large data lakes in the context of architectures for personalized healthcare service recommendations. The authors draw attention to the difficulties in customizing medicine recommendations in customized healthcare services by leveraging big data analytics and relational patient data. They point out that converting the mostly unstructured healthcare data into a relational format takes a lot of time and work. In order to tackle these issues, the research suggests a unique design for data lakes that aims to improve analytics precision, streamline data input procedures, and promote interaction with other data sources including insurance companies, chemists, and clinical labs. In order to accommodate both structured and unstructured data, the suggested data lake architecture makes use of the Hadoop Distributed File System (HDFS). This eliminates data silos and facilitates the seamless integration of various data sources. To help with focused healthcare recommendations, the design also makes use of the K-means clustering method to discover patient groupings with comparable health issues. In addition, the Support Vector Machine (SVM) algorithm is used for every patient cluster to identify the best treatment suggestions.
The data lake architecture is effective in decreasing the time it takes to import data from different vendors, regardless of the type of data, as shown by the results of the experiments. The suggested architecture has the potential to transform personalized healthcare services and enhance patient outcomes by offering a single platform for data storage, analysis, and connectivity.

III. Research Challenges and Directions for Big Data Lakes:

The rapidly developing subject of big data lakes offers a wide range of research opportunities and problems that will influence the course of next-generation big data research. In this section, we discuss some of the major issues raised by massive data lakes and future directions that this study suggests exploring.

large Data Lake Architectures: Creating an efficient and successful large data lake architecture is a basic difficulty, much like creating an effective data warehousing system. Big data lakes' effectiveness and efficiency are greatly influenced by their architectural design, which calls for novel approaches based on early data warehousing experiences.

Conventional Data Population/Alimentation Frameworks: Given the 3V nature of big data and data lake architectures, data ingesting and population provide particular issues not found in typical data warehousing solutions. It is essential to find innovative solutions, such sampling-based methods, to successfully handle these problems.

Query-Rewriting Paradigms: Big data lakes use query rewriting and schema mapping to facilitate data access while adopting lazy integration tactics that preserve repositories in their original state. Creating effective query-rewriting strategies is a crucial task for facilitating smooth data access and analytics in large data lakes.

Contemporary Metadata Managers: To enable big data analysis in data lakes, efficient metadata management is essential. Modern big data applications require novel metadata manager design and implementation paradigms that are adapted to the intricacies of big data lakes, while traditional approaches are insufficient.

Big Data Governance Methodologies: Large data sets have the capacity to propel data-driven decision-making across a range of socioeconomic sectors. For data-driven policy creation and analysis to effectively handle next-generation societal concerns like the COVID-19 pandemic, strong governance procedures employing massive data lakes are required.

Privacy-Preserving Big Data Lakes: Because big data lakes contain sensitive data, privacy protection is crucial, in contrast to standard data warehousing designs. The key problem for future research efforts is designing fundamental procedures with privacy protection paradigms in order to achieve privacy-preserving big data lakes.

New Applications for Big Data Lakes: Big data lakes are the cornerstones of a wide range of new applications, including e-science solutions, smart city applications, healthcare analytics, and social evolution analysis.

Big Data Exchange with Privacy Preservation in IoT Environments: The amount and diversity of data created has grown dramatically with the spread of Internet of Things (IoT) devices. However, there are particular difficulties in preserving privacy in huge data interchange inside Internet of Things environments, including heterogeneous data, resource limitations, and decentralized data processing. It is recommended that future research concentrate on creating specific privacy-preserving methods for IoT environments in order to protect sensitive data and facilitate easy data transmission and analysis.

Scalability and Effectiveness in Big Data Exchange with Privacy Preserving. Ensuring scalability and efficiency in privacy-preserving data sharing is becoming more and more important as big data grows in size. To tackle scalability issues, future research should investigate cutting-edge strategies including parallel processing methods and distributed computing frameworks. The development of dynamic privacy-preserving mechanisms that can adapt to changing data sharing requirements while maintaining robust privacy protection is essential in dynamic environments. User acceptance and usability are crucial for the successful adoption of privacy-preserving big data exchange systems. Future research should prioritize user-centric design principles to ensure that privacy-preserving mechanisms are clear, transparent, and easy to use for both data owners and other stakeholders.

Privacy-Preserving Big Data Exchange: Scalability Challenges, Big data is known to provide tight scalability challenges, which makes privacy preserving big data sharing no different from other big data processing activities. The privacy-preservation constraint deteriorates in this case, and the problem scales towards increasingly complex configurations. Thus, integrating privacy-preservation with scalability will be essential to the development of large data sharing research in the future.

CONCLUSION AND FUTURE WORK

We have now explored massive data lakes in this paper, looking at models, frameworks, and approaches as well as the problems and potential paths for future research. It is clear from our investigation that big data lakes have the potential to completely transform data management and analytics, but further research is necessary to realize these potentials. Our inquiry has led us to conclude that there is an urgent need for more research and development in the big data lake space. There are still problems and questions that need to be investigated further, even with tremendous progress. One interesting approach to improving the efficacy and efficiency of these systems is the incorporation of data semantics into large data lake processing, which is motivated by related research projects. In order to support increasingly complicated data management and analytics settings, future development should concentrate on expanding the large data exchange framework's privacy-preserving capabilities. To guarantee the framework’s applicability and usefulness in real-life situations, this means incorporating it with new technologies and approaches. It is also important to work on improving the large data lake systems' usability and user-centric design in order to encourage broader acceptance and adoption among stakeholders.

This work essentially paves the stage for future research projects that seek to fully realize the promise of big data lakes to spur innovation and insights in the big data era by improving our understanding of and application for them.

ACKNOWLEDGMENTS

The "Lorraine Université d' Excellence" French PIA project (reference ANR-15-IDEX-04-LUE) has provided some funding for this study.

REFERENCES

[1] R. Venkatesan, B. Sikaria, S. Shastry, A. Sikaria, R. Draves, N. Sharman, Z. Xu, Y. Barakat, C. Douglas, R. Li, M. Manu, S. Michaylov, R. Ramos, B. Krishnamachari-Sampath, K. Krishnamoorthy, P. Li, R. Draves, S.S. Naidu, S. Shastry, A. Sikaria, S. Sun, and R. Venkatesan. (2017). A Hyperscale Distributed File Service for Big Data Analytics is Azure Data Lake Store. pp. 51–63 in SIGMOD Conference 2017.

[2] Wang, C.-L., Liu, H., and S. Rangarajan. (2017). Big Data Lake-Based Scalable Architecture for Personalized Healthcare Service Recommendations. Pages 65–79 in ASSRI 2017, pages.

[3] Y.A.-R.I. Mohamed and A.A. Munshi. 2018). Big Data Analytics with Data Lake Lambda Architecture for Smart Grids. Access IEEE, 6, 40463–40471.

2015; H. Fang [4]. In the Big Data Era, Managing Data Lakes: What's A Data AND WHY HAS It became popular in data management system.

[6] Dayal, U., and S. Chaudhuri (1997). An Overview of OLAP Technology and Data Warehousing. 26(1) SIGMOD Rec., pp. 65–74.

[7] J.D. Ullman, A. Cuzzocrea, and D. Saccà (2003). Big Data: An Agenda for Research. 198–203 in: IDEAS 2013, pp.

[8] Jagadish, H.V., and A. Labrinidis (2012). Big Data's Opportunities and Challenges. VLDB Endow. Proc., 5(12), 2032–2033. Chen, M., Mao, S., and Liu, Y. (2014). Big Data: An Overview. Mobile Networks Applications, 19(2), 171-209. D. Laney (2001) [10]. Managing 3D Data: Regulating Data Amount, Speed, and Variety. META Group Technical Report.

[11] Furfaro, F., Masciari, E., Saccà, D., and Sirangelo, C. (2004). Estimated Question Response on Sensor Network Data Streams. In "GeoSensor Networks," edited by A. Stefanidis and S. Nittel.

[15] Eaton, C., and P. Zikopoulos. (2011). Recognizing Big Data: Analytics for Streaming Data and Enterprise Class Hadoop, First Edition, McGraw-Hill Osborne Media, Inc.

[16] Cuzzocrea A, De Maio C, Fenza G, Loia V, and Parente M. 2016. OLAP Analysis of Multidimensional Tweet Streams to Provide Advanced Analytics Support. In: SAC 2016, pages 992–999.

(17) Papaemmanouil, G. Papastefanatos, and N. Bikakis. (2019). Analytics, Visualization, and Exploration of Big Data. Big Data Res., 18, art. 100123.

[18] Wang, X., Qi, D., Lin, W., Yu, M., Zheng, N., Zhou, and P. Chen (2018). A Broad Structure for Knowledge Integration and Discovery in Big Data. Concurr. Comput. Pract. Exp., 30(13), art. 100123.

[19] W. Lehner, M. Thiele, and J. Eberius. The year is 2017. Ad-hoc exploratory analytics for large-scale datasets. As in: A.Y. Zomaya.