Empowering Educational Researchers with a Privacy-Centric Data Platform: Design, Implementation, and Implications

Isanka WIJERATHNE[a*], Brendan FLANAGAN[b] & Hiroaki OGATA[a]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Center for Innovative Research and Education in Data Science, Kyoto University, Japan*
*Wijerathne.isanka.6z@kyoto-u.ac.jp

Abstract: The Educational Research Data Platform (EREDA) developed at Kyoto University addresses the challenges of managing and analyzing educational data while prioritizing privacy and security. EREDA seamlessly integrates with the existing Learning and Evidence Analytics Framework (LEAF) and offers advanced data management and analysis capabilities, including a robust Data Integrity Checker (DIC) tool and backfilling. By employing real-time data streaming, robust anonymization techniques, and a multi-tenancy design, EREDA empowers researchers to conduct secure and efficient educational research. The platform also features collaborative tools that enhance knowledge sharing and accelerate insight generation. As EREDA continues to evolve, its potential to drive data-driven decision-making in educational institutions and influence educational policies is significant. Future developments will focus on scaling the platform to reach a broader user base and enhancing its functionality to better support researchers.

Keywords: Learning Analytics, Educational Data Management, Collaborative Research, Big Data, Data Privacy

1. Introduction

Educational institutions increasingly rely on data-driven insights to improve teaching and learning experiences. However, managing and analyzing large datasets poses significant challenges, particularly in protecting personal information and ensuring data security (Pardo & Siemens, 2014; Siemens & Long, 2011). The Educational Research Data Platform (EREDA) was developed at Kyoto University to address these concerns. EREDA complements the existing Learning and Evidence Analytics Framework (LEAF) (Flanagan & Ogata, 2018) by providing a privacy-centric environment for educational research. The platform features a data streaming module that efficiently transfers data from MongoDB to ClickHouse for real-time analysis while anonymizing identifiable student information and excluding data from students who opt out of research use. Researchers can access EREDA securely, with VPN access required for off-campus usage, enabling comprehensive data analysis and visualization through its integration with Apache Superset. This paper discusses the design, implementation, and implications of EREDA, highlighting its role in facilitating secure, data-driven educational research, fostering collaboration, and adhering to ethical standards.

2. System Architecture

EREDA's architecture is designed to handle educational data securely and efficiently while prioritizing privacy. The platform (Figure 1) incorporates several key components that ensure robust data management:

Design Principles: EREDA is built with privacy, security, scalability, and legal compliance as core principles. Anonymization techniques protect personal data, and robust security measures, including encryption and role-based access control (RBAC), safeguard data throughout its lifecycle (Iwase, 2019).

Data Pipeline: EREDA employs a custom-built data streaming module that transfers data from MongoDB to ClickHouse, a high-performance column-oriented database. During this process, student data is anonymized, and those who opt out are excluded from the dataset, ensuring compliance with privacy regulations. Real-time data streaming allows for continuous updates, while backfilling capabilities ensure that historical data is accurately captured.

Data Storage and Analysis: ClickHouse serves as the central repository for processed data, supporting real-time analysis and aggregation. The integration of Apache Superset provides researchers with powerful tools for data exploration and visualization. Superset's user-friendly interface allows for effective data analysis, facilitating the derivation of actionable insights from educational data.

Data Privacy Measures: EREDA includes a Data Integrity Checker (DIC) tool that regularly validates data consistency and accuracy. The DIC performs checks against the source databases, ensuring that the data remains reliable and trustworthy. Additionally, access to the platform is restricted through VPNs, and RBAC ensures that only authorized users can access sensitive data. These measures collectively create a secure environment for educational research, protecting both data privacy and integrity (Domingus, 2017).

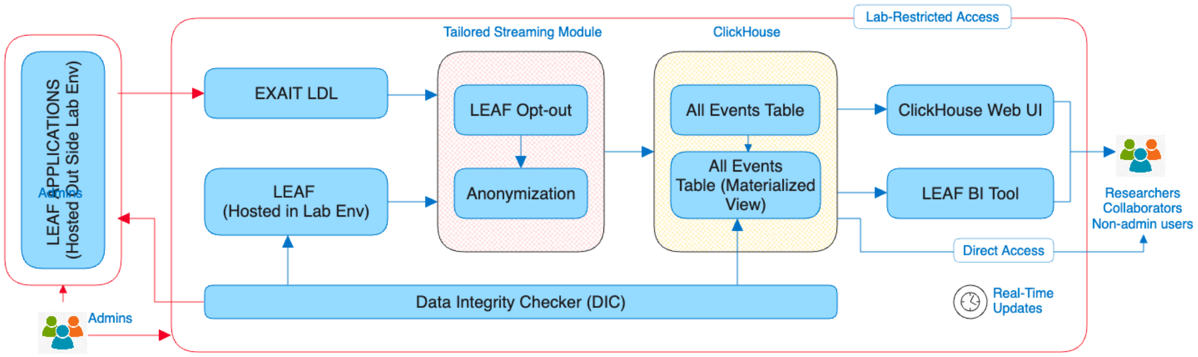

Figure 1: System Overview of EREDA.

3. Implementation and Collaborative Features

The implementation of EREDA is centered on creating a robust and secure data pipeline, supported by a carefully selected technology stack.

Technology Stack: MongoDB, known for its flexibility in managing unstructured data, serves as the Learning Record Store (LRS) for LEAF. Data retrieval and processing are handled by a custom Rust application, chosen for its performance and memory safety. Processed data is stored in ClickHouse, optimized for real-time analysis. Apache Superset is integrated for data visualization, providing researchers with an intuitive interface for exploring and analyzing educational data (Matsakis & Klock, 2014).

Collaborative Research Features: EREDA supports collaborative research by enabling secure access to datasets and visualizations among team members. Researchers can access the platform through the university network or via VPN from remote locations. ClickHouse allows for the creation and sharing of tables and views, promoting data organization and collaborative analysis. Apache Superset further enhances collaboration by allowing researchers to create and share interactive charts and dashboards, fostering a collaborative research environment that accelerates insight generation (Lepouras et al., 2014).

4. Implications and Future Directions

EREDA’s privacy-centric design and advanced data management capabilities have significant implications for educational research. The platform enables researchers to conduct their studies more efficiently by providing secure, real-time access to educational data. This focus on privacy and security allows researchers to dedicate their time to analysis rather than data management concerns (Martinovic & Ralevich, 2007).

Research Impact: The insights generated through EREDA have the potential to inform data-driven decision-making in educational institutions. By providing researchers with up-to-date and comprehensive datasets, EREDA facilitates the development of evidence-based recommendations for improving educational outcomes. This has far-reaching implications for policy-making and resource allocation in educational settings (Siemens, 2012) intuitive interface for exploring and analyzing educational data.

Scalability and Adaptability: EREDA's architecture is designed to scale, accommodating increasing data volumes as the platform gains wider adoption. The flexibility of the data pipeline and storage solutions ensures that EREDA can adapt to various educational contexts, making it a versatile tool for researchers across different institutions (Machado et al., 2019).

Ethical Considerations: Ongoing monitoring of privacy and security measures is essential to maintaining the platform's ethical standards. EREDA's commitment to ethical data management ensures that sensitive information remains protected, supporting the platform's long-term sustainability (Ocheja et al., 2023).

5. Conclusion

The Educational Research Data Platform (EREDA) represents a significant advancement in educational research by providing a privacy-centric environment for data management and analysis. Through its integration with LEAF, EREDA offers a robust and scalable solution that empowers researchers to conduct secure and efficient studies. As EREDA continues to evolve, its impact on educational research is expected to grow, enabling data-driven decision-making and contributing to the development of strategies that improve educational outcomes. Future developments will focus on enhancing EREDA's capabilities and expanding its reach to support a broader community of researchers.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

References

Domingus, M. (2017). Capability Maturity Model for Safeguarding Privacy in Academic Research.

Iwase, H. (2019). Overview of the Act on the Protection of Personal Information. *Eur. Data Prot. L. Rev.*, *5*, 92.

Lepouras, G., Katifori, A., Vassilakis, C., Antoniou, A., & Platis, N. (2014). Towards a learning analytics platform for supporting the educational process. In *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications* (pp. 246-251).

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning*, *10*(4), 469-484.

Machado, J. S., Farah, J. C., Gillet, D., & Rodríguez-Triana, M. J. (2019). Towards Open Data in Digital Education Platforms. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, *2161-377X*, 209-211.

Martinovic, D., & Ralevich, V. (2007). Privacy issues in educational systems. *International Journal of Internet Technology and Secured Transactions*, *1*(1-2), 132-150.

Matsakis, N. D., & Klock, F. S. (2014). The rust language. *ACM SIGAda Ada Letters*, *34*(3), 103-104.

Ocheja, P., Flanagan, B., Ogata, H., & Oyelere, S. S. (2023). Visualization of education blockchain data: trends and challenges. *Interactive Learning Environments*, *31*(9), 5970-5994.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British journal of educational technology*, *45*(3), 438-450.

Siemens, G. (2012, April). Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 4-8).

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*(5), 30.