Journal of Social Science and Business Studies
Volume 3, No. 4, pp. 648-654
E-ISSN: 2987-6079
http://gemapublisher.com/index.php/jssbs
Received: October 2025
Accepted: November 2025
Published: December 2025

A Comprehensive Framework for Integrating Machine Learning with Big Data Analytics Systems for Business Purposes

Fordiana Ekawati, Afrizal Zein*

Universitas Pamulang

Correspondence Email: dosen01495@unpam.ac.id*

Abstract

The growth in volume, velocity, and diversity of data has driven the need for analytical systems that are not only capable of handling big data, but also capable of generating intelligent predictions and insights through the integration of machine learning. This study aims to design and analyze a comprehensive framework that integrates machine learning algorithms into big data analytical systems. The research approach is carried out through literature studies and evaluations of various platforms and architectures such as Hadoop, Spark, and TensorFlow, which enable efficient large-scale data processing. The proposed framework includes the stages of ingestion, preprocessing, model training, evaluation, deployment, and feedback loops that support continuous learning. This integration not only improves the predictive capabilities of the system but also enables organizations to respond proactively to real-time data dynamics. The results of this study are expected to be a strategic reference in the development of modern data-driven analytical systems.

Keywords: big data, machine learning, analytical systems, data architecture, system integration.

INTRODUCTION

In a number of domains, including computer vision, discourse processing, natural language comprehension, neurology, healthcare, and the Internet of Things, machine learning (ML) techniques have had significant social benefits (Chen, Liu, & Li, 2017). The potential of machine learning (ML) to extract knowledge from an infinite number of datasets, consequently impacting human behavior and a variety of industrial applications, has led to an increase in interest in ML due to the rapid expansion of huge data. Big data presents significant challenges for traditional ML computation, especially in terms of adaptability and computational efficiency, which are crucial to fully realizing the value of big data. However, big data also provides rich and complex information that allows ML computation to uncover underlying patterns and build forward-looking models. Therefore, as big data continues to expand, ML must evolve to effectively transform this data into meaningful insights (Zhang, X., Wang, & Zhang, 2018).

ML, as discussed by Zhang and Zhang (2016), is centered on the development of frameworks that enable execution steps to vary across iterations. Common ML problems include learning from iterations with respect to specific tasks and performance measurement. Professional ML procedures rely on advanced algorithms, massive datasets, and powerful computational environments, making ML highly important for big data analytics. This paper

DOI: https://doi.org/10.61487/jssbs.v3i4.246

investigates the integration of ML procedures with big data analytics frameworks, aiming to identify opportunities and challenges arising from this integration. Big data presents new possibilities for ML by enabling pattern recognition across diverse details and perspectives within parallel processing environments. It also promotes causal inference by examining clusters of events. However, integrating ML with big data analytics frameworks introduces several fundamental challenges, such as managing high-dimensional data, ensuring model scalability, handling distributed computing environments, adapting to real-time streaming data, and improving system mobility. To fully utilize big data, these issues must be resolved (Sculley, 2015).

Machine Learning on Big Data (MLBiD) is the system we suggest. The preprocessing, learning, and evaluation phases that make up the foundation of machine learning are included in this system. Furthermore, the system has four interconnected elements big data, users, space, and systems all of which both impact and are impacted by machine learning processing.

The MLBiD system highlights future research and development directions for integrating machine learning (ML) with big data analytics frameworks and acts as a guide for identifying possibilities and difficulties. This strategy advances the science and maximizes the impact of machine learning in the big data era by creating new avenues in uncharted or understudied regions (Dean & Ghemawat, 2008).

METHOD

The Jointly Bounded Hypothesis on Innovation Recognition and Utilization (UTAUT) is modified in this study by adding security, quality, relative advantage, compatibility, perceived value (PU), and perceived ease of use (PEoU). These elements have a significant role in shaping purposeful conduct and impacting organizational utilization behavior. By capturing the intricacy of innovation and the subtle elements that propel its appropriation and integration, the suggested demonstration points seek to offer a whole comprehension of big data recognition and utilization. The presentation provides a reliable method for evaluating big data use by taking into consideration these extra building elements, especially in dynamic and changing innovation environments where typical UTAUT demonstrations might not adequately address the intricacies of big data.

According to UTAUT, an individual's performance of a behavior is fundamentally determined by their goals for performing the behavior. The behavior itself and the arbitrary criteria that surround it also have an impact on these objectives. The underlying causality identified in UTAUT is especially pertinent to large data environments due to its wide applicability to the analysis of new technologies. A more focused approach to adopting big data is captured in this context, where purposeful action is reframed as “implementation intention” inside an organizational framework.

This rephrasing highlights the need for enterprises to make proactive decisions in order to adjust to big data developments. According to the main hypothesis (H1), an organization's deliberate usage of big data will have a favorable effect on its real big data management utilization behavior. This theory emphasizes how crucial intentionality is to promoting the wise distribution and long-term use of big data technologies. In essence, if a company is dedicated to incorporating big data into its operations, this dedication should show itself as distinct usage patterns that align with the organization's main objectives. The qualitative data for interpretive analysis is shown in Figure 1.

DOI: https://doi.org/10.61487/jssbs.v3i4.246

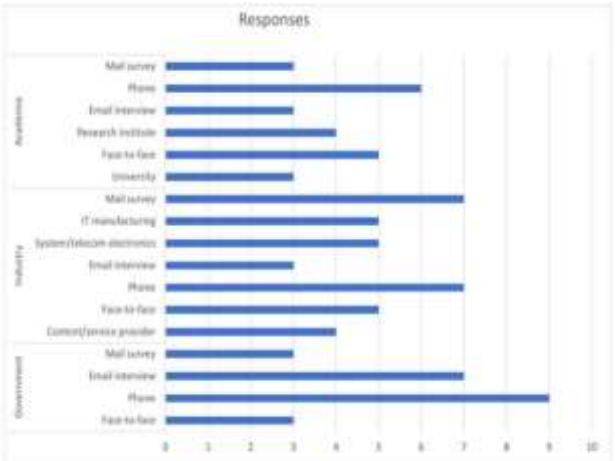

Figure 1. Qualitative Data for Interpretative Analysis

RESULT AND DISCUSSION

In order to provide a thorough knowledge of the appropriation and use of big data, the strategy employed in this consideration takes a mixed-methods approach, mixing quantitative and qualitative data. Diverse and multidimensional techniques are necessary to capture the complexity of this issue since big data frameworks are different. The method uses triangulation of many techniques to accomplish this, enabling cross-verification of data from several sources. By combining insights from survey-based quantitative data with interpretive qualitative data from standardized preparation systems, this method allows for a more thorough analysis of user behavior in big data environments. The dual approach provides a modified viewpoint, promoting a more thorough examination of how businesses view and adopt big data administration. A survey approach based on the Unified Theory of Acceptance and Use of Technology (UTAUT) forms the basis of the quantitative component of the analysis.

There are four steps in the research process. Ten students were asked to contribute their perspectives and experiences with big data during direct, in-depth interviews with potential consumers in the first round. Gathering preliminary information about consumers' attitudes and intentions is the goal of this initial phase.

Six focus groups, each with four to six participants, make up the second stage. During these sessions, participants assess how they currently use big data services and what factors seem to affect how they will use them in the future. These sessions produced more detailed qualitative data that assisted in identifying important themes and variables influencing the distribution of massive data.

In the third phase, an expert panel made up of academics, researchers, and data professionals conducted several rounds of interviews and iterative reviews to create the final survey instrument. This process ensured the content validity of the questionnaire and its relevance to the research objectives. Prior to deployment, a pilot study was conducted involving twenty students with prior experience in using big data. The pilot test, conducted over a three-week interval, enabled detailed comparison of responses and helped refine the survey items. Participants were encouraged to ask questions about any unclear aspects, thereby reducing the potential for misinterpretation.

The final test was conducted with 22 respondents selected from diverse communities to help eliminate ambiguity in language structure and semantics. The final survey was administered by a specialized marketing firm experienced in survey development and research. The company concentrated on managers like Chief Data Officers (CIOs), directors of trade analytics, and marketing managers who are involved in the selection and application

DOI: https://doi.org/10.61487/jssbs.v3i4.246

of big data. Respondents were approached via phone, mail, and social networking sites using lists supplied by the Korean Industry Association. The survey instrument was split into two sections: the second section concentrated on measurement items, while the first section gathered demographic information and control variables such participants' job responsibilities, firm size, and the existence of data mining centers. Information for independent and dependent variables was gathered utilizing various techniques, sources, and time periods in order to lessen common method bias.

The research used secondary data from a variety of sources in addition to the primary data gathered for the study. These comprised original papers pertaining to big data projects, regulations, and project development, as well as government publications, industry reports, and proprietary archives. In order to supplement and cross-validate results from the original data, secondary data were examined using content analysis. In order to find patterns in the formative elements of big data, coded texts were methodically extracted using Atlas.ti software as part of the content analysis process. After a thorough analysis and line-by-line coding, printed transcripts were compiled into a coding book. Researchers were able to examine both content and context thanks to this coding, which served as the foundation for data classification. All gathered data was imported into Atlas.ti for further analysis in accordance with the classification methodology. The illustrated environmental elements, including frameworks, applications, organizations, and users, were used to categorize the experience framework. Finding all the data pertaining to these pre-classified patterns was another stage in the theme analysis. Statements that matched each pattern were found and grouped similarly, offering insights into the events, activities, and results related to the growth of big data projects. Figure 2 illustrates the UTAUT model for big data.

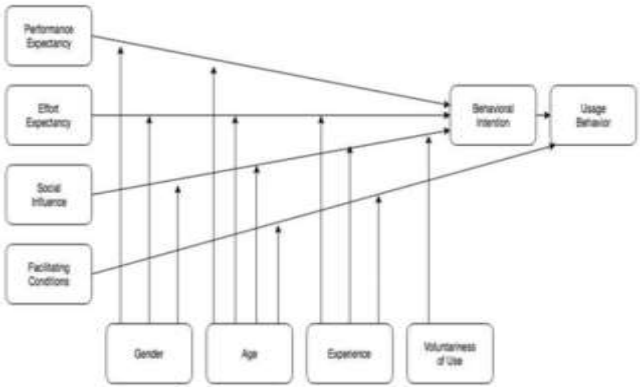

Figure 2. UTAUT Model of Big Data

Key stakeholders, including government officials from the Ministry of Security and Public Administration (MOSPA), the Ministry of Science, ICT and Future Planning, and the National Data Security Office, as well as representatives from business, academia, and government-led data governance research institutions, were interviewed to gather qualitative data. The many transaction processes, policy structures, and interfaces involved in the development of big data efforts were clarified by these interviews. Following data collection, the analysis was refined by concentrating on the roles of various stakeholders, including government, industry, and other important players in big data development, as well as social and institutional factors influencing big data initiatives and supply and demand factors within the market environment. Triangulation of findings was conducted by integrating data from multiple methods, with each strategy contributing unique insights to different perspectives on big data. This comprehensive approach not only enhances the robustness and credibility

DOI: https://doi.org/10.61487/jssbs.v3i4.246

of the research findings but also provides a holistic view of the complex ecosystem surrounding the adoption and utilization of big data.

The reliability and validity of the latent constructs in the observed model were assessed using Cronbach's alpha, a commonly used measure for evaluating internal consistency. The results confirmed that all values exceeded the acceptable threshold of 0.70, indicating satisfactory construct reliability. In addition, the factors included in this study, derived from an extensive review of the relevant literature, demonstrated strong content validity. The methods suggested by Fornell and Larcker (1981) were used to examine both convergent and discriminant validity in order to further validate the constructs. This involved determining the average variance extracted (AVE) for each construct and evaluating construct reliability at the individual item and construct levels. Confirmatory factor analysis was then used to assess the reliability of individual items, and all standardized factor loadings above 0.70—a threshold that Fornell and Larcker considered acceptable. With p-values less than 0.001 in every instance, every item loaded significantly on its corresponding latent construct, guaranteeing the constructs' adequate reliability and convergent validity. The shared variance between constructs and the AVE of individual constructs were compared in order to evaluate discriminant validity. This method demonstrated that every construct was unique and made a distinct contribution to the model as a whole. Structural equation modeling (SEM), a thorough statistical method for analyzing correlations between several variables, was then applied in additional investigations.

With fit indices above 0.90 on the majority of criteria, the model's overall goodness of fit was satisfactory. In particular, the Tucker-Lewis Index (TLI) was 0.91, the Adjusted Goodness-of-Fit Index (AGFI) was 0.89, and the Goodness-of-Fit Index (GFI) was 0.95. According to Jöreskog and Sörbom's standards, which state that values of 0.06 or less indicate a close fit, the Root Mean Square Error of Approximation (RMSEA) showed an acceptable level of 0.067. Furthermore, the Standardized Root Mean Square Residual (SRMR) was 0.027, which is significantly less than what is considered a good model fit. An overall good model fit was further shown by the normalized chi-square value of 1.98, which was significantly lower than the reference value of three. The normalized chi-square value is calculated as the chi-square divided by the degrees of freedom.

CONCLUSION

For people, organizations, and governments, data and comparative frameworks that enable the successful implementation and social normalization of sophisticated big data mediation are crucial. These factors also apply to academics from different fields that study and assess this kind of mediation. A useful framework for understanding this intricate mediation is provided by the application of Normalization Process Theory (NPT), which focuses on dynamics within and across micro-, meso-, and macro-level scales. By offering a lens for understanding and determining the normalization of big data practices, NPT enables an in-depth investigation of contextual factors influencing the implementation and outcomes of big data interventions. These findings emphasize several implications for project-based and academic research. From a practical perspective, the study offers insights into large-scale analytics methodologies and national-level data governance, highlighting interactions among technological evolution, spatial structure, cost considerations, and demand-related variables. The reflection suggests that while Korea's advanced adoption of big data technologies has the potential to be feasible and sustainable for future ICT environmental innovations, it also faces substantial challenges. These demanding conditions reflect tensions between the innovative components of big data technologies and their associated social and institutional implications.

DOI: https://doi.org/10.61487/jssbs.v3i4.246

This refinement is particularly relevant when considering big data as a service (BDaaS), which conceptualizes big data not merely as a tool or delivery mechanism for data but as a service that enhances social intuitiveness. BDaaS envisions big data as a strategy to support social infrastructure, fostering stronger associations and greater engagement among members of society. In addition to its practical contributions, this study also provides valuable theoretical insights. It advances an interpretive strategy within the socio-technical systems (STS) framework by analyzing big data phenomena within a specific context and across multiple levels. Traditionally, STS has been extensively applied within organizational settings to inform governance and design strategies. However, its application at the societal or macro level, particularly for analyzing emerging technologies such as big data, remains relatively limited. This study addresses this gap by examining how STS concepts can be used to identify broader implications of big data analytics.

The paper emphasizes that although a lot of previous research using STS concepts has concentrated on distinct socio-technical structure components, there is still a dearth of thorough analysis of how these components interact within a broader system. The study demonstrates how the conceptualization of large-scale data in Normalization Process Theory (NPT), an emerging theory within STS, is consistent with the discourse on socio-technical systems. Even though NPT is still regarded as a middle-range theory, it is especially useful for big data analysis because it successfully connects the social and technical aspects of structures and technologies. A helpful analytical tool for analyzing the dynamic evolution of ICT and technological sustainability is provided by NPT's procedural approach. It emphasizes the intrinsic complexity of hybrid structures made up of interacting components by framing big data as a normalization process. By considering both technical capabilities and social factors, this viewpoint allows for a more nuanced understanding of how big data can be embedded within societal contexts.

By using NPT, the study advances our knowledge of how big data solutions might get accepted in particular organizational and societal contexts. It emphasizes the necessity of comprehensive strategies that take into account big data's social and technical aspects. By incorporating NPT into big data analytics, important insights into the efficient management and assessment of such interventions can be obtained, providing direction for future studies and real-world applications. In summary, the focus of this study emphasizes how crucial it is to see big data as a service that may promote social integration and engagement rather than only as a technical instrument. By addressing both practical and theoretical dimensions of big data processing, this research offers a comprehensive perspective on how big data interventions can be effectively implemented and normalized. This approach not only contributes meaningfully to the advancement of socio-technical systems theory but also provides actionable insights for organizational practice and policy development in the domain of big data.

REFERENCES

Chen, Z., Liu, J., & Li, Y. (2017). A big data analytics framework for smart grids. *IEEE Transactions on Smart Grid, 8*(2), 723-734. https://doi.org /10.48550/arXiv.1708.04935

Zhang, X., Wang, H., & Zhang, Y. (2018). Data mining techniques in big data era: A study on challenges and opportunities. *Journal of Big Data, 5*(1), 20-35. doi: 10.1016 /j.jii.2017.08.001

Davis, F., Johnson, T., & Matthews, L. (2016). Analyzing big data: The framework for decision support in business analytics. *Decision Support Systems, 85*(4), 12-25..

Kumar, A., Aggarwal, S., & Gupta, N. (2019). Big data analytics for health systems: A survey and framework. *Health Informatics Journal, 25*(3), 707-727.

DOI: https://doi.org/10.61487/jssbs.v3i4.246

Tan, S., & Lu, Y. (2020). Frameworks and tools for big data analysis: A systematic review. *IEEE Access, 8*, 10448-10464. DOI: 10.1109/ JAS.2020.1003384

Patel, H., Kaur, R., & Mehta, P. (2017). Big data integration for data-driven decision making. *Journal of Data Science and Engineering, 2*(2), 45-60.

Ali, A., Hassan, M., & Khan, I. (2015). Real-time analytics frameworks for IoT big data: Current trends and future directions. *Journal of Computer Networks and Communications, 2015*, 983768.

Williams, J., Green, D., & Lee, M. (2018). Exploring data analytics frameworks in supply chain management. *International Journal of Logistics Management, 29*(2), 456-478. DOI:10.1504/IJAL.2016.080341

Zhang, Y., Liu, X., & Wang, Z. (2016). A big data framework for electric vehicles charging behavior analytics. *IEEE Transactions on Industrial Informatics, 12*(2), 743-750L.

Rokach and O. Maimon. (2007). Data mining with decision trees: Theory and applications. doi: 10.1142/6604.

Singh, P., Joshi, A., & Sharma, R. (2019). Big data frameworks for financial services: A comprehensive review. *Journal of Financial Services Research, 56*(1), 98-115. DOI:10.1108/IJBM-06-2021-0230

Roberts, T., & Smith, J. (2020). Demystifying big data: A framework for educational data mining. *Educational Technology Research and Development, 68*(3), 1015-1035.

.Ng, A., & Lin, T. (2017). Big data frameworks for healthcare and clinical applications. *Journal of Biomedical Informatics, 73*, 15-29.

Mistry, N., & Patel, R. (2018). Developmental framework for big data analytics in the cloud. *IEEE Cloud Computing, 5*(3), 47-55. https://doi.org/10.3390/s23062952

Xu, W., & Tan, Y. (2019). Big data security frameworks: A state-of-the-art review. *Journal of Information Security and Applications, 46*, 102-115..DOI:10.1007/s40860-020-00120-3

Harrison, G., & Lewis, K. (2015). A strategic framework for big data analytics in the public sector. *Government Information Quarterly, 32*(3), 236-242.

Samso Supriyatna, Salman Farizy. (2024). Perancangan dan Implementasi Aplikasi Monitoring Berkas Pencairan Dana Berbasis Web Menggunakan Metode Rapid Application Development. *Sainstech: Jurnal Penelitian dan Pengkajian Sains Dan Teknologi, 34*(3).DOI: https://doi.org/10.37277/stch.v34i3.2078

Afrizal Zein. (2022). Evaluasi Keamanan Wireless LAN Menggunakan Issaf (Information System Security Assessment Framework). *Sainstech: Jurnal Penelitian dan Pengkajian Sains dan Teknologi, 32*(2). DOI: Https://doi.org/10.37277/stch.v32i2