Machine Learning Applications in Databases and Big Data Analytics

No Author Given

No Institute Given

Abstract. The convergence of machine learning (ML) and big data technologies redefines the landscape of data-driven systems, enabling adaptive, scalable, and intelligent solutions across diverse domains. This survey systematically explores the integration of ML techniques in database systems and big data frameworks, highlighting advances in query optimization, data quality assurance, feature engineering, and real-time stream processing. Furthermore, it analyzes real-world applications, including predictive maintenance, recommendation systems, fraud detection, and healthcare analytics, demonstrating the operational value of ML in data-intensive environments. Finally, the survey concludes by identifying key challenges in scalability, interpretability, and privacy-preserving learning, outlining promising research directions to foster the next generation of robust and trustworthy big data analytics.

Keywords: Batabases · Machine Learning · Big Data · Data Processing

1 Introduction

The rapid growth of data generated by digital systems, sensors, and online platforms has driven the evolution of traditional data management systems into large-scale, complex ecosystems. In parallel, ML has emerged as a transformative paradigm for extracting value from this data, offering intelligent capabilities for decision support, pattern recognition, and predictive modelling. The intersection of ML with big data infrastructures has become central to modern analytics, enabling scalable, adaptive, and context-aware systems [31].

As big data platforms evolve, they increasingly integrate ML models not only for analytics but also as operational components of database management, stream processing, and application-level intelligence. However, this integration introduces a new set of technical, architectural, and ethical challenges, ranging from real-time inference at scale to the need for explainable and privacy-respecting solutions [12].

Despite the substantial advances in both ML and big data systems, their convergence is not straightforward. ML models must be tailored to function effectively in environments characterised by distributed architectures, heterogeneous data formats, and continuous data flows. Furthermore, production-level applications demand not only scalability and accuracy but also transparency, fairness,

and strict data privacy compliance. Understanding how ML is operationalised within big data ecosystems is crucial for designing intelligent, high-impact, and trustworthy data-driven systems [14].

To address this gap, the present survey offers a structured and in-depth analysis of the current landscape. Its key contributions are as follows:

- Comprehensive overview of ML integration in both database systems and large-scale big data frameworks, including techniques for query optimisation, anomaly detection, and streaming analytics.
- Detailed use case analysis covering four high-impact domains: predictive maintenance, personalised recommendations, financial fraud detection, and healthcare analytics.
- Structured taxonomy of challenges, highlighting technical barriers in scalability, interpretability, and privacy, alongside current solutions and future research directions.

The rest of this paper is organised as follows. ML techniques in databases are noted in Section 2. Moreover, in Section 3, ML in big data frameworks is outlined. Section 4 discusses applications and use cases. Next, Section 5 provides challenges and future directions. Finally, Section 6 concludes the present survey.

2 Machine Learning Techniques in Databases

The integration of ML into core database components marks a shift from static, rule-based mechanisms to adaptive, data-driven optimisation. This section examines how ML models improve query planning, indexing strategies, and data quality assurance in modern database systems.

2.1 Query Optimization and Learned Indexes

Query optimization in relational databases has traditionally relied on heuristics and static cost models, which often struggle with dynamic workloads and data skew. Recent developments use reinforcement learning (RL) to model query planning as a sequential decision process, enabling adaptive execution strategies based on observed query patterns and latency feedback [32,35].

Simultaneously, learned index structures have emerged as a data-driven alternative to traditional indexing. Neural models, such as piecewise linear regression and recursive indexes, approximate the cumulative distribution function of keys, enabling faster lookups and improved performance in skewed or clustered datasets without requiring manual tuning [34,40].

2.2 Data Quality and Anomaly Detection

Data anomalies, such as out-of-distribution values, incomplete records, and semantic inconsistencies, undermine the reliability of downstream analytics. ML-based methods like autoencoders enable proactive detection by reconstructing

input tuples and identifying corrupted entries, particularly in high-dimensional or partially labelled datasets [23].

Probabilistic and generative models, including variational autoencoders (VAEs) and Gaussian mixture models (GMMs), capture attribute dependencies and reveal inconsistencies often missed by rule-based systems. In temporal databases, recurrent models (e.g., Long short-term memory (LSTM)) are used to detect delayed updates or irregular patterns, crucial in domains like finance and the Internet of Things [24,10].

These techniques can be embedded within the database engine to operate during data ingestion, supporting real-time anomaly-aware processing and scalable, automated data quality assurance [39].

Table 1 summarizes the key differences between ML-based query optimization and data quality enhancement techniques in databases. It highlights their objectives, underlying ML models, integration points, and deployment modes. The comparison highlights how each approach makes a distinct contribution to performance and reliability within data management systems.

Table 1. ML Techniques in Database Optimization and Data Quality.

Aspect	**Query Optimization & Learned Indexes**	**Data Quality & Anomaly Detection**
Objective	Speed up query execution and data access	Detect and correct data anomalies
ML Techniques	RL, Regression Models	Autoencoders, VAEs, GMMs, LSTMs
Target Component	Query planner, index structures	Data validation, ingestion pipeline
Adaptability	Adapts to workload shifts and data distributions	Adapts to structural, semantic, and temporal anomalies
Deployment Mode	Embedded in execution engine	Inline or batch validation during ingestion
Key Benefit	Reduced latency and smarter indexing	Improved data integrity and reliability

3 Machine Learning in Big Data Frameworks

The application of ML in big data ecosystems requires careful integration with distributed computing infrastructures and data-intensive processing workflows. This section examines how ML models are scaled, engineered, and adapted within large-scale batch and streaming environments.

3.1 Integration with Distributed Architectures

Modern big data frameworks like Apache Spark, Hadoop YARN, and Flink enable distributed computation essential for large-scale ML training. Built-in libraries

(e.g., Spark MLlib) and scalable wrappers (e.g., Horovod, TensorFlow on Spark) support co-located data processing and model computation, thereby reducing I/O overhead and duplication. These platforms use in-memory data sharing and resilient distributed datasets to parallelise preprocessing and iterative training [25,30,13].

Efficient integration depends on optimizing task scheduling, fault tolerance, and model checkpointing across nodes. Graphics processing unit (GPU) acceleration and container orchestration (e.g., Kubernetes) further enhance training efficiency in resource-intensive scenarios [51].

3.2 Feature Engineering at Scale

The effectiveness of ML models depends heavily on the quality and discriminative power of their input features. In large-scale environments, feature engineering is challenged by data volume, dimensionality, and heterogeneity. To address this, distributed transformations, such as joins, normalization, and encoding, are applied using functional programming and lazy evaluation [22].

Techniques like distributed principal component analysis (PCA), feature hashing, and statistical aggregations support dimensionality reduction and contextual enrichment. Automated platforms further leverage meta-learning and Bayesian optimization to generate feature sets efficiently, enabling scalable and reproducible pipelines suited for real-time processing [18,42].

3.3 Online Learning for Streaming Data

Big data systems increasingly process continuous data streams from sensors, applications, and user interactions, rendering static batch learning ineffective. Online learning algorithms, such as Hoeffding trees, online passive-aggressive models, and incremental stochastic gradient descent, enable real-time model updates without retraining from scratch [27,9].

Frameworks like Apache Flink, Kafka Streams, and Spark Structured Streaming provide essential support for temporal consistency through windowing, watermarking, and event-time processing. These systems enable stateful learning with low latency and memory efficiency. The focus lies on deploying stable, stream-compatible models capable of handling concept drift and integrating seamlessly with real-time ingestion pipelines [29,4].

Table 2 provides a comparative summary of ML integration across three key areas in big data frameworks. It highlights their distinct goals, supporting platforms, core methods, and performance priorities. The table illustrates how each approach contributes to scalability, adaptability, and efficiency in large-scale ML workflows.

4 Applications and Use Cases

The integration of ML with big data platforms has given rise to intelligent applications across diverse domains. By exploiting large-scale, heterogeneous data

Table 2. ML Techniques in Big Data Frameworks.

Aspect	**Distributed Integration**	**Feature Engineering**	**Online Learning**
Goal	Parallel training and data locality	Scalable feature transformation	Real-time model adaptation
Frameworks	Spark, Flink, TensorFlowOnSpark	Spark MLlib, Featuretools	Flink, Kafka Streams, Spark Streaming
Key Methods	Task scheduling, GPU use, checkpointing	PCA, feature hashing, meta-learning	Incremental models, concept drift detection
Output	Accelerated training/inference	Compact, informative features	Continuously updated models
Optimization Focus	Resource efficiency, fault tolerance	Dimensionality and execution time	Latency, memory usage, adaptation speed

and real-time analytics, these systems achieve impactful outcomes. This section highlights representative use cases that demonstrate their operational value and architectural significance.

4.1 Predictive Maintenance

Predictive maintenance utilises time-series data and event logs from industrial systems to anticipate equipment failures before they occur. Traditional threshold-based approaches often yield high false positives and lack adaptability. In contrast, supervised ML models, such as random forests (RFs), support vector machines (SVMs), and recurrent neural networks (RNNs), learn complex degradation patterns by combining historical records with real-time sensor data [49,6].

Big data platforms ingest multi-source telemetry (e.g., vibration, temperature, cycles) and process it using distributed systems like Apache Kafka and Flink to enable real-time anomaly detection and health scoring. Graph-based models further enhance diagnostics through root-cause analysis across interconnected components [37,47].

Deployment challenges include ensuring low-latency inference at the edge, particularly in environments with constrained connectivity. Federated learning (FL) and compact deep learning models offer practical solutions by enabling on-device training with periodic global updates [50].

4.2 Personalized Recommendations

Personalized recommendation systems are essential to e-commerce, media streaming, and online education platforms. They leverage user behavior data—such as browsing history and interaction logs—to suggest relevant items, using models like collaborative filtering, matrix factorization, and deep neural networks (DNNs) [36,21].

To handle the scale and sparsity of interaction data, these systems rely on NoSQL (Structured Query Language) databases (e.g., Cassandra, MongoDB)

and real-time inference pipelines built with Spark Structured Streaming or Flink, which enable fast updates to feature stores [43,46].

RL methods, including contextual bandits and deep Q-networks, are increasingly employed to optimize recommendation strategies based on user feedback, effectively balancing exploration and exploitation to enhance engagement [17].

4.3 Financial Risk and Fraud Detection

The financial sector is increasingly leveraging ML for risk assessment, credit scoring, and fraud detection, thereby surpassing the limitations of traditional rule-based systems. Advanced models like gradient boosting, anomaly ensembles, and graph neural networks detect complex, evolving fraud patterns by learning subtle behavioral correlations across accounts and transactions [44,33].

Big data platforms process high-velocity streams from automated teller machines, point-of-sale devices, and online systems using tools like Apache Kafka and Samza, enabling real-time pattern analysis with sliding-window features and temporal embeddings. Hybrid pipelines combining supervised and unsupervised models enhance the detection of both known and novel threats [20].

Graph analytics is central to uncovering fraud rings and synthetic identities, with graph convolutional networks (GCNs) identifying suspicious links in heterogeneous graphs. Privacy and compliance concerns further drive the use of FL and encryption-based training methods [8].

4.4 Healthcare Data Analytics

Healthcare analytics is a highly sensitive and impactful domain for ML and big data. Patient records, diagnostic images, lab results, and clinical notes, often stored in heterogeneous systems, are analyzed using models such as decision trees, support vector machines, convolutional neural networks (CNNs), and LSTMs to predict disease onset, assess patient risk, and guide personalized treatments [3].

Big data platforms integrate Electronic Health Records (EHRs), imaging data, and genomics for multi-modal analysis, supported by scalable storage systems like Hadoop distributed file system (HSFS) and cloud services. Real-time frameworks enable early warning systems for critical conditions, while neuro-linguistic programming (NLP) techniques extract structured insights from clinical narratives to enhance the usability of data [11,16].

Ensuring model interpretability and fairness remains a major challenge. Tools like Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) facilitate the explanation of predictions to clinicians, thereby supporting trust and regulatory compliance. Furthermore, privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) drive the adoption of FL, enabling the development of collaborative models that do not require the sharing of sensitive patient data [2].

Table 3 summarizes ML applications across key sectors, highlighting the interplay between data modalities, learning methods, infrastructure demands,

and domain-specific constraints. This multidimensional view clarifies how ML models are tailored to meet distinct operational goals under the architectural and regulatory pressures of each application context.

Table 3. Overview of ML-Driven Applications in Big Data Environments.

Domain	**Primary Data Sources**	**Core ML Techniques**	**Big Data Infrastructure Needs**	**Unique Challenges**	**ML Objectives**
Predictive Maintenance	Sensor telemetry, equipment logs	Time-series models, RNNs, RFs, SVMs	Kafka, Flink, edge computing support	Latency at the edge, root-cause tracing	Failure forecasting, degradation scoring
Personalized Recommendations	User behavior logs, interaction histories, metadata	Matrix factorization, DNNs, RL	NoSQL stores, Spark Streaming, real-time feature stores	Sparse data, cold-start problem, feedback loops	Real-time content/product personalization
Financial Fraud Detection	Transactional streams, user profiles, entity graphs	Boosting models, GNNs, anomaly ensembles	Kafka Streams, graph engines, sliding-window analytics	Detection latency, class imbalance, adversarial behavior	Fraud identification, risk scoring, behavior profiling
Healthcare Analytics	EHRs, medical imaging, clinical text, genomics	CNNs, LSTMs, Decision Trees, NLP models	HDFS, cloud-native analytics, FL	Privacy compliance, data heterogeneity, interpretability	Disease prediction, risk stratification, treatment support

5 Challenges and Future Directions

While ML continues to advance big data systems, challenges in scalability, interpretability, and privacy remain significant. Addressing these issues is essential to ensure robust, adaptive, and trustworthy ML solutions. This section highlights key obstacles and outlines future research directions for real-world, data-intensive integration.

5.1 Scalability vs. Model Complexity Trade-offs

As data volumes grow and application domains diversify, the need for more expressive ML models increases. However, greater model complexity, through deeper architectures or rich feature interactions, often hinders computational scalability. High-capacity models like DNNs and ensembles demand significant memory, processing power, and training time, making them challenging to deploy in large-scale, real-time settings [48,19].

To maintain low-latency responses in big data environments, solutions such as distributed training, approximate inference, and model compression techniques like quantization and distillation are employed. Still, the dynamic adjustment of model granularity based on workload and resource constraints remains a largely unexplored yet critical research direction [26].

Future directions include the development of self-adaptive model architectures, where computational depth or resolution is adjusted on-the-fly, and model-parallel learning frameworks that can distribute components of a single large model across multiple compute nodes in an efficient, synchronized manner [38].

5.2 Interpretability and Trust in ML-Integrated Systems

While ML models achieve high predictive accuracy, their integration into decision-support systems presents challenges in terms of transparency and user trust, particularly in critical domains such as healthcare, finance, and law. Stakeholders require not just accurate outputs but also understandable justifications, which black-box models often fail to provide [45].

Interpretability tools such as SHAP, LIME, and counterfactual reasoning offer post-hoc explanations but may be fragile under model drift or adversarial inputs. Beyond individual predictions, ensuring system-wide auditability and traceability remains a major concern [5].

Future trustworthy ML will depend on intrinsically interpretable models, formal verification methods, and adaptive explanation interfaces tailored to different user roles. Bridging these capabilities with human-centered design and ethical principles is essential for widespread and responsible adoption [41].

5.3 Privacy-Preserving Learning and Edge Intelligence

The convergence of distributed data sources, edge computing, and privacy regulations presents significant challenges for training and deploying ML models. In many cases, especially involving sensitive data, centralized training is not viable due to legal or ethical constraints. Consequently, privacy-preserving paradigms, such as FL, have emerged as critical solutions [15,7].

While FL allows decentralized model training without sharing raw data, it introduces challenges in communication efficiency, convergence, and handling non-identically distributed data distributions. Techniques such as differential privacy, secure multiparty computation, and homomorphic encryption offer stronger guarantees but add computational and latency overheads [1].

Future research should develop hybrid frameworks that strike a balance between accuracy, privacy, and efficiency, while incorporating adaptive strategies for participant selection and communication. Embedding edge intelligence into model pipelines will be key to achieving scalable, privacy-compliant learning in modern distributed environments [28].

Table 4 outlines the primary challenges facing ML integration in big data environments. It contrasts current mitigation strategies with emerging research

directions, highlighting their systemic impact on scalability, trustworthiness, and privacy-aware intelligence.

Table 4. Key Challenges and Future Directions in ML for Big Data.

Challenge Theme	**Core Technical Barriers**	**Current Approaches**	**Open Research Directions**	**Impact on ML Systems**
Scalability vs. Model Complexity	Resource-intensive training, latency bottlenecks, deployment constraints	Model pruning, distillation, distributed training	Self-adaptive architectures, model-parallel learning	Balances performance with resource constraints at scale
Interpretability and Trust	Opaque model decisions, limited explanation robustness	SHAP, LIME, counterfactual reasoning	Intrinsically interpretable models, role-based explanation interfaces	Enhances transparency, user trust, and regulatory compliance
Privacy-Preserving Learning and Edge Intelligence	Data decentralization, heterogeneity, communication overhead	FL, differential privacy, encryption-based learning	Adaptive federated schemes, hybrid privacy frameworks	Enables secure, scalable learning across distributed and sensitive environments

6 Conclusion

The integration of ML with big data technologies marks a significant step toward building intelligent, adaptive, and large-scale data systems. This survey examined how ML techniques are applied across various layers of the data stack—from query optimization and feature engineering to real-time analytics and anomaly detection—highlighting both architectural advances and domain-specific implementations.

Through the analysis of representative use cases, the survey demonstrated the transformative impact of ML in key sectors such as manufacturing, e-commerce, finance, and healthcare. At the same time, it identified pressing challenges related to model scalability, interpretability, and privacy—factors that increasingly define the feasibility and societal acceptance of data-driven solutions.

As big data ecosystems continue to evolve, future research must focus on developing more efficient, transparent, and privacy-preserving ML models that can operate across decentralized, heterogeneous environments. By bridging the gap between algorithmic innovation and real-world deployment, these efforts will play a central role in shaping the next generation of robust and responsible intelligent systems.

References

1. Ardiç, E., Genç, Y.: Enhanced privacy and communication efficiency in non-iid federated learning with adaptive quantization and differential privacy. IEEE Access (2025)
2. Ashwani, S.: Advancing explainable ai in healthcare methods, applications, and ethical implications. In: Federated Learning for Neural Disorders in Healthcare 6.0, pp. 61–95. CRC Press (2025)
3. Beam, A.L., Kohane, I.S.: Big data and machine learning in health care. Jama **319**(13), 1317–1318 (2018)
4. Bhatt, N., Thakkar, A.: An efficient approach for low latency processing in stream data. PeerJ Computer Science **7**, e426 (2021)
5. Bhattacharya, A.: Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more. Packt Publishing Ltd (2022)
6. Bonnevay, S., Cugliari, J., Granger, V.: Predictive maintenance from event logs using wavelet-based features: an industrial application. In: 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019) Seville, Spain, May 13–15, 2019, Proceedings 14. pp. 132–141. Springer (2020)
7. Boussis, D., Dritsas, E., Kanavos, A., Sioutas, S., Tzimas, G., Verykios, V.S.: Mapreduce implementations for privacy preserving record linkage. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. pp. 1–4 (2018)
8. Chen, J., Huang, G., Zheng, H., Yu, S., Jiang, W., Cui, C.: Graph-fraudster: Adversarial attacks on graph neural network-based vertical federated learning. IEEE Transactions on Computational Social Systems **10**(2), 492–506 (2022)
9. Chen, R., Dai, T., Zhang, Y., Zhu, Y., Liu, X., Zhao, E.: Gbdt-il: Incremental learning of gradient boosting decision trees to detect botnets in internet of things. Sensors **24**(7), 2083 (2024)
10. Cook, A.A., Mısırlı, G., Fan, Z.: Anomaly detection for iot time-series data: A survey. IEEE Internet of Things Journal **7**(7), 6481–6494 (2019)
11. Dhayne, H., Haque, R., Kilany, R., Taher, Y.: In search of big medical data integration solutions-a comprehensive survey. IEEE Access **7**, 91265–91290 (2019)
12. Dritsas, E.: Efficient algorithms for big data management. Ph.D. thesis, University of Patras, Greece (2020)
13. Dritsas, E., Trigka, M.: Applying machine learning on big data with apache spark. IEEE Access (2025)
14. Dritsas, E., Trigka, M.: Exploring the intersection of machine learning and big data: A survey. Machine Learning and Knowledge Extraction **7**(1), 13 (2025)
15. Dritsas, E., Trigka, M.: Federated learning for iot: A survey of techniques, challenges, and applications. Journal of Sensor and Actuator Networks **14**(1), 9 (2025)
16. Dritsas, E., Trigka, M.: A survey on the applications of cloud computing in the industrial internet of things. Big Data and Cognitive Computing **9**(2), 44 (2025)
17. El Mimouni, I., Avrachenkov, K.: Deep q-learning with whittle index for contextual restless bandits: Application to email recommender systems. In: Northern Lights Deep Learning Conference 2025 (2025)
18. Garouani, M., Ahmad, A., Bouneffa, M., Hamlich, M., Bourguin, G., Lewandowski, A.: Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data. Journal of Big Data **9**(1), 57 (2022)
19. Gzar, D.A., Mahmood, A.M., Abbas, M.K.: A comparative study of regression machine learning algorithms: Tradeoff between accuracy and computational complexity. Mathematical Modelling of Engineering Problems **9**(5) (2022)

20. Habeeb, R.A.A.: Real-Time Anomaly Detection Using Clustering in Big Data Technologies. Ph.D. thesis, University of Malaya (Malaysia) (2019)
21. Lara-Cabrera, R., González-Prieto, Á., Ortega, F.: Deep matrix factorization approach for collaborative filtering recommender systems. Applied Sciences **10**(14), 4926 (2020)
22. Las-Casas, P., Papakerashvili, G., Anand, V., Mace, J.: Sifter: Scalable sampling for distributed traces, without feature engineering. In: Proceedings of the ACM Symposium on Cloud Computing. pp. 312–324 (2019)
23. Li, Y., Wang, Y., Ma, X.: Variational autoencoder-based outlier detection for high-dimensional data. Intelligent Data Analysis **23**(5), 991–1002 (2019)
24. Liao, W., Guo, Y., Chen, X., Li, P.: A unified unsupervised gaussian mixture variational autoencoder for high dimensional outlier detection. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 1208–1217. IEEE (2018)
25. Long, A., Han, W., Huang, X., Li, J., Wang, Y., Chen, J.: Distributed deep learning for big remote sensing data processing on apache spark: geological remote sensing interpretation as a case study. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. pp. 96–110. Springer (2023)
26. Luo, P., Yu, F.R., Chen, J., Li, J., Leung, V.C.: A novel adaptive gradient compression scheme: Reducing the communication overhead for distributed deep learning in the internet of things. IEEE Internet of Things Journal **8**(14), 11476–11486 (2021)
27. Marpu, R., Manjula, B.: Streaming machine learning algorithms with streaming big data systems. Brazilian Journal of Development **10**(1), 322–339 (2024)
28. Mughal, F.R., He, J., Das, B., Dharejo, F.A., Zhu, N., Khan, S.B., Alzahrani, S.: Adaptive federated learning for resource-constrained iot devices through edge intelligence and multi-edge clustering. Scientific Reports **14**(1), 28746 (2024)
29. Onishi, T., Michaelis, J., Kanemasa, Y.: Recovery-conscious adaptive watermark generation for time-order event stream processing. In: 2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI). pp. 66–78. IEEE (2020)
30. Polak, A.: Scaling machine learning with Spark: distributed ML with MLlib, TensorFlow, and PyTorch. " O'Reilly Media, Inc." (2023)
31. Potla, R.T.: Scalable machine learning algorithms for big data analytics: Challenges and opportunities. Journal of Artificial Intelligence Research **2**(2), 124–141 (2022)
32. Ramadan, M., El-Kilany, A., Mokhtar, H.M., Sobh, I.: Rl_qoptimizer: a reinforcement learning based query optimizer. IEEE Access **10**, 70502–70515 (2022)
33. Renganathan, K.K., Karuppiah, J., Pathinathan, M., Raghuraman, S.: Credit card fraud detection with advanced graph based machine learning techniques. Indonesian Journal of Electrical Engineering and Computer Science **35**(3), 1963–1963 (2024)
34. Setiawan, N.F., Rubinstein, B.I., Borovica-Gajic, R.: Function interpolation for learned index structures. In: Databases Theory and Applications: 31st Australasian Database Conference, ADC 2020, Melbourne, VIC, Australia, February 3–7, 2020, Proceedings 31. pp. 68–80. Springer (2020)
35. Shahrivari, H., Papapetrou, O., Fletcher, G.: Workload prediction for adaptive approximate query processing. In: 2022 IEEE international conference on big data (big data). pp. 217–222. IEEE (2022)
36. Sharma, S., Rana, V., Kumar, V.: Deep learning based semantic personalized recommendation system. International Journal of Information Management Data Insights **1**(2), 100028 (2021)
37. Shetty, S.: Improving processing of real-time Big Data in Smart Grids using Apache Flink and Kafka. Ph.D. thesis, Dublin, National College of Ireland (2019)

38. Shukla, K., Xu, M., Trask, N., Karniadakis, G.E.: Scalable algorithms for physics-informed neural and graph networks. Data-Centric Engineering **3**, e24 (2022)
39. Singh, K., Kushwaha, A.S.: Advanced techniques in real-time data ingestion using snowpipe. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN pp. 2960–2068 (2024)
40. Slanináková, T., Antol, M., OĬha, J., Kaňa, V., Dohnal, V.: Data-driven learned metric index: an unsupervised approach. In: Similarity Search and Applications: 14th International Conference, SISAP 2021, Dortmund, Germany, September 29–October 1, 2021, Proceedings 14. pp. 81–94. Springer (2021)
41. Sperrle, F., El-Assady, M., Guo, G., Borgo, R., Chau, D.H., Endert, A., Keim, D.: A survey of human-centered evaluations in human-centered machine learning. In: Computer Graphics Forum. vol. 40, pp. 543–568. Wiley Online Library (2021)
42. Sresth, V., Nagavalli, S.P., Tiwari, S.: Optimizing data pipelines in advanced cloud computing: Innovative approaches to large-scale data processing, analytics, and real-time optimization. INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS **10**, 478–496 (2023)
43. Sunny, B.K., Janardhanan, P., Francis, A.B., Murali, R.: Implementation of a self-adaptive real time recommendation system using spark machine learning libraries. In: 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES). pp. 1–7. IEEE (2017)
44. Thilagavathi, M., Saranyadevi, R., Vijayakumar, N., Selvi, K., Anitha, L., Sudharson, K.: Ai-driven fraud detection in financial transactions with graph neural networks and anomaly detection. In: 2024 International Conference on Science Technology Engineering and Management (ICSTEM). pp. 1–6. IEEE (2024)
45. Valente, F., Paredes, S., Henriques, J., Rocha, T., de Carvalho, P., Morais, J.: Interpretability, personalization and reliability of a machine learning based clinical decision support system. Data Mining and Knowledge Discovery **36**(3), 1140–1173 (2022)
46. Vonitsanos, G., Dritsas, E., Kanavos, A., Mylonas, P., Sioutas, S.: Security and privacy solutions associated with nosql data stores. In: 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA. pp. 1–5. IEEE (2020)
47. Wang, H., Wu, Z., Jiang, H., Huang, Y., Wang, J., Kopru, S., Xie, T.: Groot: An event-graph-based approach for root cause analysis in industrial settings. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 419–429. IEEE (2021)
48. Yu, F., Wang, D., Shangguan, L., Zhang, M., Tang, X., Liu, C., Chen, X.: A survey of large-scale deep learning serving system optimization: Challenges and opportunities. arXiv preprint arXiv:2111.14247 (2021)
49. Zainuddin, Z., EA, P.A., Hasan, M.: Predicting machine failure using recurrent neural network-gated recurrent unit (rnn-gru) through time series data. Bulletin of Electrical Engineering and Informatics **10**(2), 870–878 (2021)
50. Zhang, Z., Gao, Z., Guo, Y., Gong, Y.: Scalable and low-latency federated learning with cooperative mobile edge networking. IEEE Transactions on Mobile Computing **23**(1), 812–822 (2022)
51. Zhou, J., Zhang, K., Zhu, F., Shi, Q., Fang, W., Wang, L., Wang, Y.: Elasticdl: A kubernetes-native deep learning framework with fault-tolerance and elastic scheduling. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. pp. 1148–1151 (2023)