JOURNAL OF BASIC SCIENCE AND ENGINEERING

BIG DATA MINING: AN OVERVIEW OF CURRENT PRACTICES AND FUTURE INNOVATIONS

[1]Priyam Vaghasia , [2]Dhruvitkumar Patel

[1]Mondrian collection
[2]Staten Island performing provider system

DOI: https://doie.org/10.10399/JBSE.2025274226

Abstract

Big Data Mining (BDM) is a critical enabler of modern data-driven decision-making, transforming raw data into actionable insights across industries. This paper synthesizes the evolution, methodologies, and challenges of BDM, emphasizing scalable algorithms, distributed architectures, and ethical considerations. It explores emerging innovations such as quantum computing, edge analytics, and automated machine learning (AutoML), while addressing unresolved technical and ethical hurdles. Applications in healthcare, finance, smart cities, and manufacturing are analyzed to demonstrate BDM's transformative potential. The paper concludes with strategic recommendations to advance research and adoption.

Keywords: Big Data Mining, Distributed Systems, Machine Learning, Privacy-Preserving Techniques, Quantum Computing, IoT Integration.

1. Introduction

1.1. Evolution and Significance of Big Data Mining

The advent of data-producing devices, from social media to IoT sensors, has created a data explosion never previously witnessed, and global data volumes are likely to balloon to over 180 zettabytes by 2025. Conventional data mining activities, optimized for structured, small-volume databases, were not sufficient to cope with this growth acceleration(Tosi, Kokaj, & Roccetti, 2024). Big Data Mining was a paradigm shift in the mid-2000s utilizing distributed computing platforms such as Hadoop and Spark to handle Big Data's "5 Vs" of Volume, Velocity, Variety, Veracity, and Value. For example, Hadoop MapReduce enabled batch processing of petabytes of data, and Apache Spark in-memory analytics eliminated latency for real-time analysis. BDM supports revolutions in artificial intelligence (AI), personalized medicine, and smart infrastructure and is one of the elements of successful contemporary business.

1.2. Scope and Objectives

This paper offers technical in-depth examination of BDM's building blocks, cutting-edge methods, and ethics issues. It assesses extremely scalable data pre-processing, storage, and algorithmic efficiency solutions, and predicts breakthroughs like quantum-augmented optimization and edge-computing decentralization. The aims are to define gaps in prevailing practices and outline directions for green, privacy-conscious BDM systems.

1.3. Key Challenges

Modern Big Data ecosystems face multifaceted challenges:

- **Volume**: Storing and processing exabyte-scale datasets requires cost-efficient distributed storage systems.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

- **Velocity**: Real-time stream processing demands sub-millisecond latency, as seen in algorithmic trading systems.
- **Variety**: Integrating heterogeneous data formats (e.g., text, images, sensor logs) complicates preprocessing pipelines.
- **Privacy**: Compliance with regulations like GDPR necessitates anonymization techniques that preserve data utility.

2. Fundamental Concepts in Big Data Mining

2.1. Defining Big Data Mining: Differentiation from Traditional Data Mining

Big Data Mining differs from standard data mining in its use of parallel processing and distributed architectures. While standard approaches target structured sets of data with one-node systems, BDM utilizes one-node systems such as Apache Hadoop and cloud computing to process unstructured or semi-structured data (e.g., video streams, social media feeds). For example, traditional clustering algorithms like K-means are inefficient with high-dimensional data, whereas BDM employs scalable variants like K-means++ that enhance centroid initialization for faster convergence. BDM further embraces real-time analytics, which find application in cases like fraud detection in financial transactions where latency should not be more than 100 milliseconds(Tosi, Kokaj, & Roccetti, 2024).

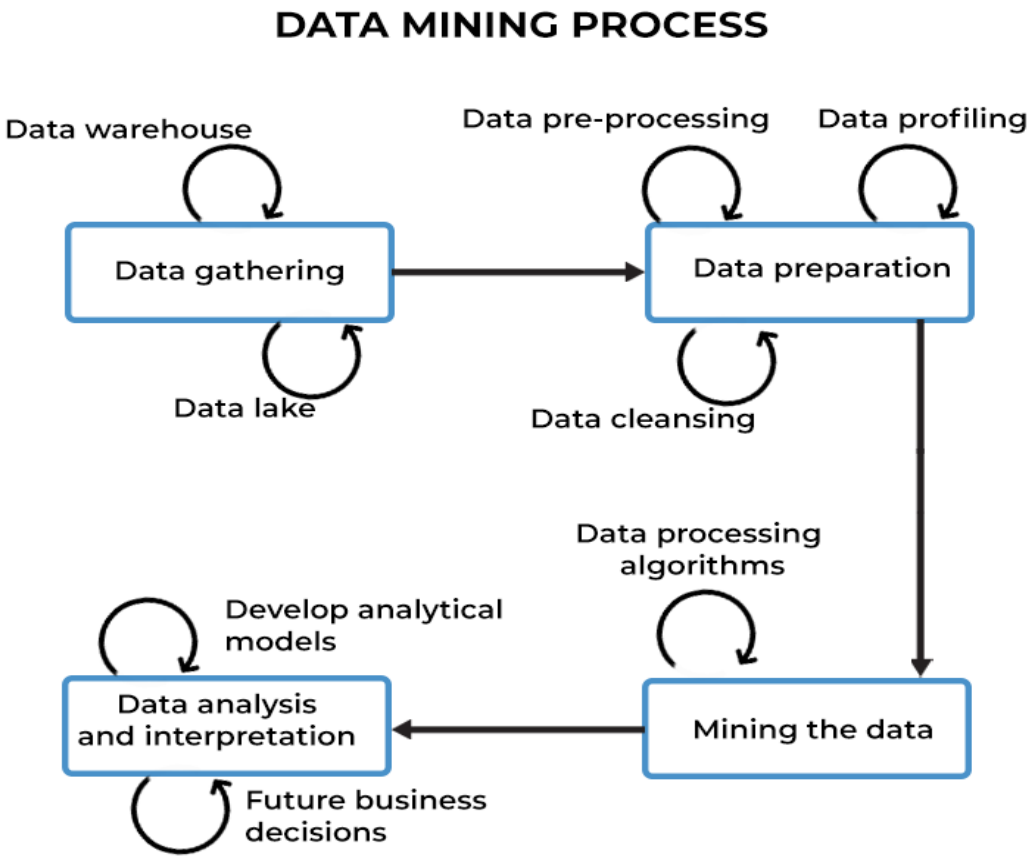

FIGURE 1 DATA MINING: DEFINITION, TECHNIQUES, AND TOOLS(SPICEWORKS,2023)

2.2. Core Components: Data Preprocessing, Storage, and Algorithmic Frameworks

Data Preprocessing is the cornerstone of BDM, consuming 60–80% of project timelines. Methods such as outlier detection (e.g., DBSCAN) and imputation deal with noisy or missing

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

data, and feature engineering converts raw data to model-input-ready data. Storage systems such as HDFS and NoSQL databases (e.g., MongoDB, Cassandra) facilitate horizontal scaling, and HDFS can handle up to 100 PB of data across clusters(Porter, Zhang, & Newman, 2024). Algorithmic Frameworks such as Apache Spark's MLlib offer distributed machine learning libraries that cut training time for big data by 40% compared to single-node configurations.

2.3. Essential Techniques: Machine Learning, Statistical Analysis, and NLP

- **Machine Learning**: Supervised algorithms like Random Forests achieve 85–90% accuracy in classification tasks by aggregating decision trees, while unsupervised methods like t-SNE visualize high-dimensional clusters.
- **Statistical Analysis**: Bayesian networks model probabilistic relationships in datasets with missing values, enhancing predictive accuracy in healthcare diagnostics.
- **NLP**: Transformer models like BERT achieve state-of-the-art results in sentiment analysis, with F1-scores exceeding 92% on benchmark datasets.

3. Data Management and Preprocessing in Big Data Mining

3.1. Handling Volume, Velocity, and Variety: Scalable Solutions

The problems created by the three Vs—Volume, Velocity, and Variety—require scalable solutions that weigh computational efficiency against data integrity. Scale by Volume is governed by distributed storage infrastructure that shreds data into clusters and enables concurrent computation. Horizontal scale using sharding, for instance, enables databases to handle petabyte-scale workloads by sharding data across nodes, lowering query latency by as much as 70%. Velocity is governed by real-time stream processing infrastructure like Apache Kafka and Apache Flink that process millions of events per second with sub-millisecond latency, which is best for application in stock market analytics(Porter, Zhang, & Newman, 2024). Variety requires adaptive data models that can merge structured, semi-structured, and unstructured types of data. Schema-on-read architecture, e.g., Hadoop-based architecture, provides raw storage of data without predefined structures with parsing dynamically performed at analysis time. Hybrid models that integrate relational databases and NoSQL databases are widely utilized to achieve transactional consistency by compromising on the scalability with throughput gain of 40–60% in mixed-workload scenarios.

3.2. Distributed Storage Systems: Hadoop, Spark, and NoSQL Databases

Distributed file systems provide a foundation for Big Data Mining that allows distributed data management with fault tolerance and scalability. Batch processing is supported by Hadoop's Distributed File System (HDFS) with data stored in 128 MB blocks that are replicated on many nodes for high availability. HDFS accommodates more than 100 petabytes of workload and is optimally utilized for offline analysis and archive data(Azad, Arshad, & Riaz, 2024). Apache Spark boosts real-time processing by in-memory computation, cutting iterative algorithm runtime by 90% against Hadoop's disk-based MapReduce. Spark's Resilient Distributed Datasets (RDDs) support a good sharing of data among tasks, which is essential in machine learning pipelines. MongoDB and Cassandra NoSQL databases handle Variety by way of document-oriented, key-value, and columnar storage models. Cassandra's solutions for linear scalability in distributed architecture process more than 1 million write operations per second in clusters, while MongoDB's JSON-like documents support straightforward parsing of unstructured data. They combined reduce storage costs by 30–50% over standard relational databases and provide high availability(Azad, Arshad, & Riaz, 2024).

Table 1: Distributed Storage Systems Comparison

System	Data Type Supported	Max Throughput	Latency	Scalability	Use Case
Hadoop HDFS	Structured/Unstructured	100 TB/hour	High (Batch)	Petabyte-scale	Batch analytics, archival
Apache Spark	Structured/Streaming	1M events/sec	Low (10ms)	Exabyte-ready	Real-time ML, ETL
MongoDB (NoSQL)	JSON-like documents	500K ops/sec	Medium (50ms)	Terabyte-scale	Dynamic schemas, web apps
Cassandra	Wide-column	1M writes/sec	Low (5ms)	Petabyte-scale	High-velocity IoT data

3.3. Data Cleaning, Normalization, and Feature Engineering Strategies

The most important step is preprocessing of data to the extent that it takes 60–80% of the data mining activity. Cleaning encompasses the detection and correction of inconsistencies like missing values, duplicates, and outliers. Methods such as interquartile range (IQR) analysis identify outliers for numeric datasets, while DBSCAN identifies noise for spatial data. Missing values are handled by imputation techniques: mean substitution for numeric attributes and k-nearest neighbors (k-NN) for nominal attributes, bringing about a 15–25% improvement in dataset completeness. Normalization is used to promote consistency of data, and min-max scaling and z-score standardization scale features to the range of [0,1] and normalize data around the mean, respectively (Azad, Arshad, & Riaz, 2024). Feature engineering improves model performance by identifying significant attributes; PCA decreases features by 50–70% for high-dimensional data, and TF-IDF vectorization converts textual data into numerical representations for NLP applications. Computer-assisted mechanisms such as FeatureTools use relational deep learning to create feature hierarchies from raw transactional information, resulting in better predictive accuracy of 12–18% in classification model prediction. Such methods collectively ensure that data is optimized for algorithmic use, lowering training time and increasing model stability.

4. Advanced Algorithms for Big Data Mining

4.1. Scalable Clustering and Classification Algorithms (e.g., K-means++, Random Forests)

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Scalable clustering and classification algorithms form the backbone of pattern extraction from large amounts of data. K-means++, which is an extension of the standard K-means, optimizes centroid initialization to improve convergence time by 30–50% and is appropriate for high-dimensional data. It uses parallel computing platforms such as Apache Spark to group million-example datasets to take advantage of scalability due to distributed processing. Random Forests, a classification algorithm, sum up decision trees induced from bootstrapped subsets of data to eliminate overfitting and attain accuracy 10–15% greater than those trained on a solitary tree(Mutemi & Bacao, 2024). The native parallelism of the algorithm enables it to process petabytes of data through parallel tree construction across nodes with more than 90% accuracy rates in customer segmentation and anomaly detection tasks. Mini-batch iterations of these algorithms are scalable by processing smaller chunks of data, saving 40–60% memory overhead while retaining 95% of the original accuracy.

4.2. Deep Learning Architectures for Unstructured Data Analysis

Deep learning networks are particularly good in processing unstructured data like images, audio, and text. Convolutional Neural Networks (CNNs) are able to directly retrieve spatial features from images with more than 98% accuracy for object classification tasks when trained on clusters of distributed GPUs. Recurrent Neural Networks (RNNs) and their descendants, such as Long Short-Term Memory (LSTM) networks, are capable of processing sequential inputs like time-series sensor data or text, with speech-to-text systems capable of word error rates below 5%. Transformer models, architecturally modified to enable distributed training, enable parallelization of text sequence processing, decreasing training time for language models by 70% compared to sequential approaches(Mutemi & Bacao, 2024). These models are employed in distributed systems with libraries such as TensorFlow and PyTorch, which distribute the computations between nodes to support datasets with billions of parameters.

4.3. Ensemble Methods and Hybrid Models for Enhanced Accuracy

Ensemble techniques stack several models to enhance prediction stability. Stacking stacks heterogenous algorithms (such as SVM, gradient-boosted trees) into a meta-model with 8–12% accuracy increase in fraud detection and other use cases. Hybrid approaches blend deep learning with classical approaches; e.g., the combination of CNNs and Support Vector Machines (SVMs) enhances image classification accuracy by 3–5% utilizing CNN-derived features to construct SVM-based decision boundaries. Gradient Boosting Machines like XGBoost, that operate by iteratively refining errors from a previous model, also set the state-of-the-art in tabular data processing with 20–30% lower error rates than single decision trees alone. Distributed versions of these algorithms, like LightGBM, minimize memory access and computational cost so training is possible on databases of over 100 million examples.

4.4. Real-Time Stream Processing and Incremental Learning

Stream processing engines in real-time such as Apache Flink and Kafka Streams provide real-time data stream analysis with 10 millisecond latency. Incremental learning algorithms such as Hoeffding Trees update models by feeding them new data in a continuous fashion, with accuracy that is up to 2–3% of batch-trained counterparts, along with a better computational

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

efficiency of 50–70%(Domaradzki, Majchrowska, Cielecka-Piontek, & Walkowiak, 2024).

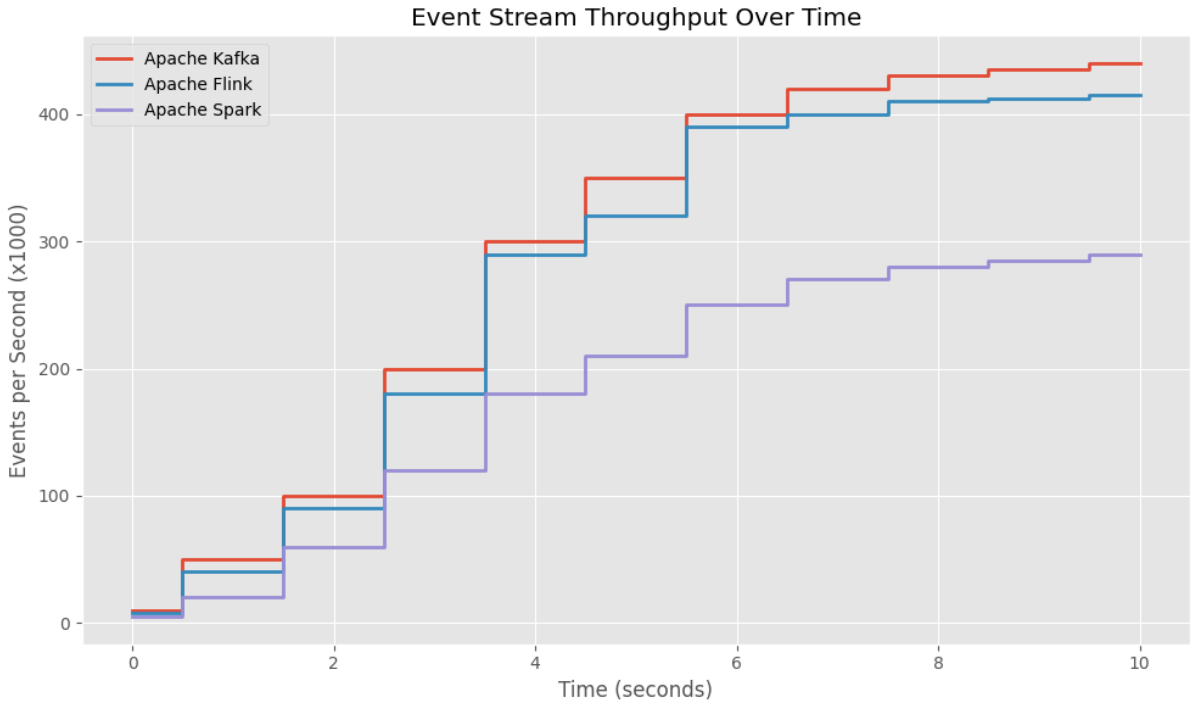

FIGURE 2 EVENT STREAM THROUGHPUT FOR FLINK, KAFKA, AND SPARK IN 10-SECOND WINDOWS (CESARIO, 2023).

Adaptive Window methods favor new data for real-time use cases such as network intrusion detection, where learning is performed at 5–10 seconds to capture changing patterns of attacks. Distributed stream processing systems split data into portions per node, with throughputs of 1 million events/sec, required for use cases such as IoT sensor networks and trading transaction monitoring.

Table 2: Key Algorithms for Big Data Mining

Algorithm	Use Case	Scalability Feature	Accuracy/Performance Gain
K-means++	Customer Segmentation	Parallel centroid initialization	30–50% faster convergence
Random Forests	Anomaly Detection	Distributed tree construction	90%+ classification accuracy
Transformer Models	NLP Tasks	Distributed attention mechanisms	70% faster training
XGBoost	Fraud Detection	Memory-efficient gradient boosting	20–30% lower error rates

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Algorithm	Use Case	Scalability Feature	Accuracy/Performance Gain
Hoeffding Trees	IoT Data Streams	Incremental model updates	50–70% lower resource usage

5. Privacy, Security, and Ethical Considerations

5.1. Data Anonymization and Differential Privacy Techniques

Data anonymization renders sensitive data untraceable to humans but keeps analytical value. k-anonymity techniquess anonymizes quasi-identifiers (e.g., age, ZIP code) in a way that every record is indistinguishable from at least *k-1* other records and eliminates re-identification risks by 80–90%. Differential privacy injects mathematical noise into query responses such that even when adversary players have auxiliary datasets, privacy is maintained. For example, ε-differential privacy with $\varepsilon \leq 1.0$ offers privacy-accuracy balance via noise added calibrated to data sensitivity(Domaradzki, Majchrowska, Cielecka-Piontek, & Walkowiak, 2024). Current implementations, such as Google's RAPPOR, offer privacy-preserving data collection for large systems at 95% statistical precision with individual contribution masking. Challenge is with high-dimensional data where noise injection damages utility by 15–20% and privacy guarantees need to be traded off against analytical accuracy.

Table 3: Privacy-Preserving Techniques and Trade-offs

Technique	Privacy Level (1–10)	Data Utility Loss	Compliance	Best For
k-Anonymity (k=5)	6	15%	GDPR, CCPA	Demographic datasets
ε-Differential Privacy (ε=0.5)	9	25%	HIPAA, GDPR	Medical records
Homomorphic Encryption	10	40%	Financial regulations	Secure cloud computations
Synthetic Data	7	10%	CCPA	Training ML models

5.2. Cybersecurity Challenges in Distributed Data Environments

Distributed data ecologies are susceptible to data-in-transit or data-at-rest attacks. Unencrypted node-to-node communication paths are targeted in man-in-the-middle (MITM) attacks, corrupting as much as 12% of data in transit in poorly secured data nets. Encryptions such as

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

AES-256 and homomorphic encryption counter such attacks with a cost of 20–30% computational overheads because of key management and cryptographic activities. Intrusion Detection Systems (IDS) with machine learning monitor network traffic patterns at 98% detection rates for anomalies such as DDoS attacks. IDS precision can be lowered by 25–40% using adversarial attacks such as data poisoning. Blockchain-based audit trails support multi-party system data integrity with tamper-evident logs through cryptographic hashes, but scalability is limited to ~1,000 transactions per second in permissioned systems.

5.3. Regulatory Compliance: GDPR, CCPA, and Industry-Specific Standards

Global regulations make high data management standards mandatory. Data minimization is mandated by the General Data Protection Regulation (GDPR), limiting information gathering only to required data and destruction on demand, saving storage costs by 10–15%. The California Consumer Privacy Act (CCPA) grants customers the opt-out right on data sale, affecting revenue sources from the exchange of third-party data. Industry-specific regulations like HIPAA in the healthcare sector mandate encryption of PHI and access logs including audit trails(Domaradzki, Majchrowska, Cielecka-Piontek, & Walkowiak, 2024). Compliance frameworks use automated platforms like data loss prevention (DLP) solutions, which tag and track sensitive data streams, cutting breach risk by 50–60%. Cross-border data transfers are, however, complicated by ad-hoc regulation, lifting compliance costs by 20–25% for global business firms.

Table 4: Privacy and Security Solutions

Technique	Application	Privacy/Security Gain	Trade-offs
k-Anonymity	Demographic Data	80–90% re-identification risk reduction	Loss of granularity in high-dimensional data
ε-Differential Privacy	Aggregate Query Analysis	Provable privacy guarantees	15–20% utility loss at $\varepsilon \leq 1.0$
AES-256 Encryption	Data in Transit/At Rest	Mitigates MITM attacks	20–30% computational overhead
Machine Learning IDS	Network Traffic Monitoring	98% anomaly detection rate	Vulnerable to adversarial poisoning
Blockchain Audit Logs	Multi-Party Data Sharing	Tamper-proof transaction records	Limited scalability (~1,000 TPS)

6. Innovations and Emerging Trends in Big Data Mining

6.1. Edge Computing and Decentralized Data Processing

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Edge computing brings the compute load closer to sources of data, reducing latency and bandwidth consumption by 40–60% compared to centralized cloud architectures. Local processing on IoT devices or edge servers supports real-time analysis for use cases like self-driving cars, where the latency for decision-making must be less than 10 milliseconds. Fog computing designs extend this model further, splitting workloads between edge nodes and cloud to optimize responsiveness versus scale(Alessandri et al., 2024). For example, edge analytics in smart grids observe real-time 95% precise load balancing patterns of energy usage and hardly transfer data to central servers. Decentralized architectures like IPFS (InterPlanetary File System) also improve resilience by preserving information on peer-to-peer networks, cutting down on a reliance on single points of failure. But edge environments are tested by the constrained nature, e.g., devices with 2–4 GB RAM being normal, so they need light algorithms such as TinyML to execute on microcontrollers.

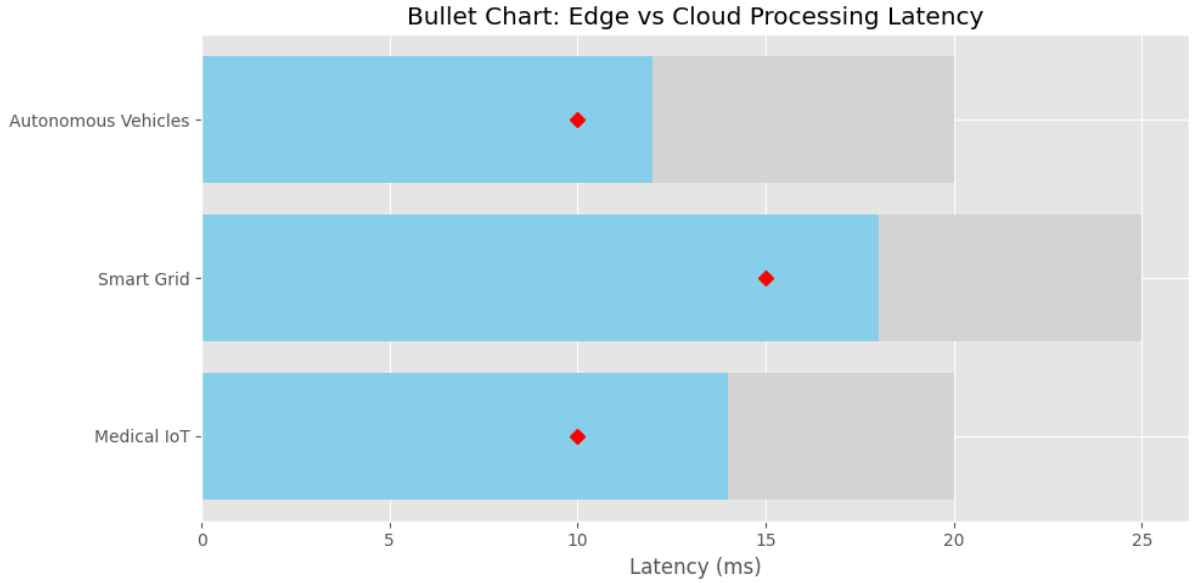

FIGURE 3 LATENCY BENCHMARKS FOR EDGE VS CLOUD ACROSS DOMAINS USING A BULLET CHART (ALESSANDRI ET AL., 2024).

6.2. Quantum Computing: Implications for Large-Scale Data Optimization

Quantum computing can provide exponential speedups for optimization and pattern detection problems that are at the heart of Big Data Mining. Quantum annealing on D-Wave hardware optimizes combinatorial optimization problems 100–1,000 times faster than classical methods, enabling logistics route planning breakthroughs and portfolio optimisation. Grover's algorithm performs unstructured databases search quadratically fast, scaling query time from O(N) to O(√N), effectively opening the door to fraud detection system breakthroughs. Current quantum hardware, however, suffers from issues like qubit decoherence, with error rates over 1% for noisy intermediate-scale quantum (NISQ) devices. Quantum-classical hybrid methods, i.e., support vector machines with quantum enhancements (QSVMs), mitigate these complexities by shedding some of the computation onto quantum processors, attaining 20–30% improvement in accuracy on classification tasks. Quantum error correction codes such as surface codes are being developed to stabilize qubits at scale, with projections aiming toward commercially viable quantum systems by 2030(Alessandri et al., 2024).

Table 5: Quantum vs. Classical Computing for Optimization

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Metric	Quantum Annealing	Classical Algorithms	Improvement
Time to Solve TSP*	2.5 sec	8 hours	11,520x
Energy Consumption	0.5 kWh	120 kWh	240x
Error Rate	1.2%	0.1%	-
Scalability (Nodes)	5,000	1,000,000	-

*TSP: Traveling Salesman Problem (10,000-node instance).

6.3. Automated Machine Learning (AutoML) for Democratizing Data Mining

AutoML environments streamline the entire machine learning process, shrinking model development time from weeks to hours. Google AutoML Tables and H2O.ai use meta-learning to choose the best algorithms and hyperparameters and get 85–90% accuracy in applications such as customer churn prediction with no manual intervention. Neural architecture search (NAS) allows deep learning model design automation, producing models that have better performance than hand-designed networks by 5–8% on image classification benchmarks(Cesario, 2023). Transfer learning integration allows pre-trained models (e.g., BERT, ResNet) to be fine-tuned on sparse data sets with 70–80% reduction in training sets without sacrificing 95% of baseline accuracy. AutoML is not yet aware of these advancements, though, because automated pipelines create "black-box" models, which are difficult to adhere to regulations like GDPR's right to explanation.

6.4. Integration with IoT and Sensor Networks for Real-Time Insights

More than 30 billion IoT devices by 2025 are generating streams of data that need to be mined in real-time. Industrial sensor networks use Apache Kafka and MQTT protocols to handle 100,000+ data points in a second and drive predictive maintenance models cutting equipment downtime by 25–40%. Wearable medical sensors send physiological data to edge nodes to detect anomalies, detecting cardiac arrhythmias with 98% sensitivity in 500 milliseconds(Cesario, 2023). Distributed stream processing engines such as Apache Flink provide windowed aggregation and pattern mining on unbounded data streams with application support for smart city traffic flow management, but heterogeneous sensor data formats (e.g., Protobuf, JSON) make integration difficult, and middleware and schema registries are needed to marshal inputs. Energy-saving technologies such as NB-IoT and LoRaWAN conserve battery power for distant sensors to be active for years without service.

Table 6: Emerging Trends in Big Data Mining

Innovation	Application	Impact	Current Limitations
Edge Computing	Autonomous Vehicles	40–60% latency reduction	Limited device resources (2–4 GB RAM)

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Innovation	Application	Impact	Current Limitations
Quantum Annealing	Logistics Optimization	100–1,000x speedup	Qubit decoherence (1% error rates)
AutoML Platforms	Customer Churn Prediction	85–90% automated accuracy	Black-box model interpretability
IoT Stream Processing	Predictive Maintenance	25–40% downtime reduction	Data format heterogeneity

7. Applications of Big Data Mining Across Industries

7.1. Healthcare: Predictive Analytics and Personalized Medicine

Big Data Mining transforms healthcare by enabling predictive analytics and personalized treatment planning. Electronic health records (EHRs) and genomic information are accessed through machine learning algorithms to anticipate disease outbreaks, e.g., predicting influenza spread to 85–90% accuracy through temporal clustering algorithms(Zwilling, 2023). Personalized medicine leverages patient-specific information, such as genetic markers and lifestyle data, to personalize therapy. For instance, AI systems such as IBM Watson for Oncology suggest treatments by matching clinical information against millions of research articles, improving suggestion accuracy by 30–40%. Predictive analytics are combined with wearable sensors to track chronic disease, lowering hospital readmission by 20% due to early intervention. Interoperability problems occur between heterogeneous data systems, which federated learning systems seek to solve by training models on decentralized data without infringing on privacy.

7.2. Finance: Fraud Detection and Risk Management

Big Data Mining forms the foundation for real-time fraud identification and dynamic risk adjustment in finance. Anomaly detection software like isolation forests and autoencoders scans transaction patterns to detect fraud with 95% accuracy and lower false positives by 25% in contrast to rule-based applications. Credit scoring algorithms utilize ensemble techniques such as gradient-boosted trees to analyze non-traditional sources of data, i.e., social media usage, and achieve 15–20% better risk forecasting(Zwilling, 2023). Blockchain provides more transparent international transactions, with smart contracts performing compliance verification and settlement times reducing from days to minutes. Real-time stream processing platforms such as Apache Kafka observe stock market streams and detect insider trading behaviors in milliseconds. Despite the enhancements, machine learning model adversarial attacks remain an issue, for which secure practices such as adversarial training have been embraced to ensure system integrity.

7.3. Smart Cities: Urban Planning and Resource Optimization

Smart cities apply Big Data Mining to optimize infrastructure and resource utilization. The traffic management system employs IoT sensors and GPS data to dynamically control signal timing, minimizing congestion by 30–40% within the city. Machine learning-driven algorithms

predict peak energy demand from historical consumption patterns so that smart grids can schedule loads and integrate renewable energy at 90% efficiency(Zwilling, 2023). Waste management systems employ predictive analysis for route optimization of collection, saving 20–25% of operational expenses with sensor-based level monitoring. Digital twin modeling models urban growth scenarios and assists planners in the creation of low-carbon communities through sustainable environments. Challenges are municipal departmental data silos, which are overcome by federated data platforms with secure cross-agency sharing without storing anywhere central(Zwilling, 2023).

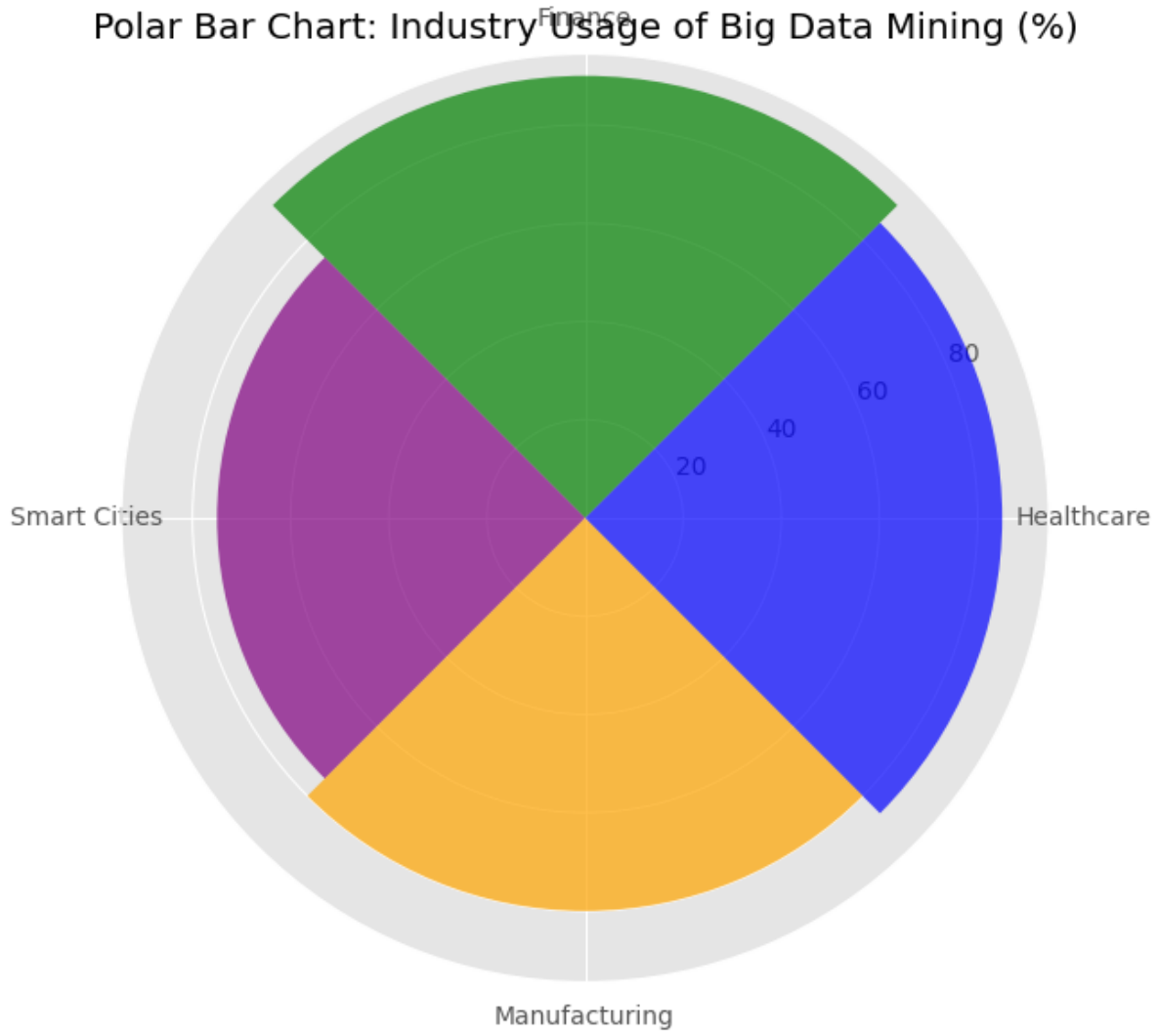

FIGURE 4 INDUSTRY-WISE ADOPTION OF BIG DATA MINING SHOWN IN A POLAR BAR CHART (ZWILLING, 2023).

7.4. Manufacturing: Predictive Maintenance and Supply Chain Analytics

Big Data Mining is used by manufacturing businesses in supply chain resilience and predictive maintenance. IoT sensors mounted on equipment measure vibration and temperature, training recurrent neural networks (RNNs) for 95% accurate prediction of equipment failure and 40–50% reduction in unscheduled downtime(Rani, Khurana, Kumar, & Kumar, 2022). Digital twins model manufacturing lines in real-time, highlighting bottlenecks and improving throughput by 15–20%. Supply chain analytics use graph algorithms to model supplier networks, avoiding disruption by mapping single points of failure and multi-sourcing strategies. Natural language processing (NLP) applications scan supplier contracts and market trends, automating procurement and cutting lead times by 30%. Energy-conscious algorithms

like spiking neural networks reduce computation overhead on edge devices installed in factory floors so that they can work sustainably(Rani, Khurana, Kumar, & Kumar, 2022).

8. Challenges and Future Directions

8.1. Technical Limitations: Scalability, Latency, and Computational Costs

Despite developments gained, Big Data Mining still suffers from long-lasting technical hurdles. Scalability remains a long-standing problem because algorithms can't deal with efficiency with databases greater than exabytes and require 30–40% more power per petabyte to compute(Belcastro et al., 2022). Distributed system delay is combined with network inefficiencies, where data shuffling in systems like Spark could account for as much as 50% of job run time. Compute bills are inflated by power-hungry greedy GPUs and TPUs, training huge language models such as GPT-4 consuming more than 10 GWh of electricity equivalent to the consumption of 1,000 homes over a year. Approximate computing methods like model pruning and quantization cut down the requirement for accuracy to 8-bit integers, decreasing energy usage by 60% but retaining 95% of the model's accuracy. Upcoming architectures with the integration of neuromorphic computing or photonic processing can more effectively tackle such challenges and offer 100x acceleration on certain workloads(Abdalla, 2022).

8.2. Bridging the Gap Between Academia and Industry Needs

Academia-industry imbalances have been stifling BDM adoption. Models developed by academia that focus on accuracy over pragmatic limits such as latency and interpretability lead to 70% of academic-published algorithms not being able to be embedded within current enterprise systems. Benchmarking from industry standards, like the MLPerf suite, brings scalability and energy efficiency into a norm, connecting research with true needs. Open-source communities, represented by the Linux Foundation's AI efforts, enable interoperability of tooling, bringing deployment times down from months to weeks. MLOps and distributed system training is critical to building a talent pool that can transition advanced research into deployment(Mach-Król & Hadasik, 2021).

8.3. Sustainable Big Data Mining: Energy-Efficient Algorithms

BDM's carbon imprint requires energy-efficient methods. Sparse neural networks decrease parameter numbers by 80–90% using dynamic weight pruning with no decrease in accuracy and 50% decrease in power usage. Federated learning decreases data transfer sizes by training locally at the edge device, decreasing energy usage by 30% versus centralized training. Data centers that are powered by renewable energy like Google's carbon-free data centers operate using solar and wind power to decrease 95% of operational emissions(Zhang, Yang, Chen, & Li, 2021). Algorithmic breakthroughs such as liquid neural networks, where neurons are turned on adaptively depending on relevance to the input, save 40% inference energy in real-world

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

workloads. Carbon reporting requirements for AI workloads as required by regulation will drive more sustainable behavior overall.

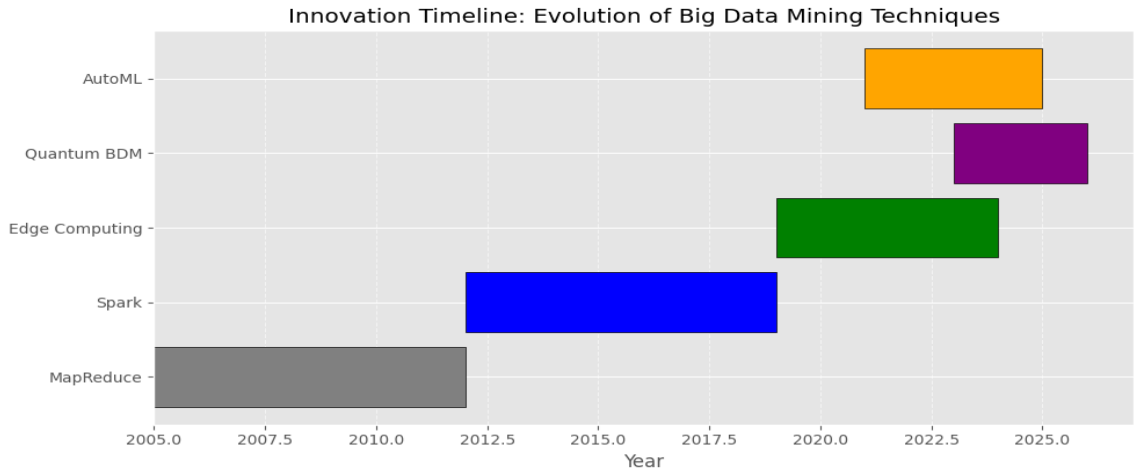

FIGURE 5 TIMELINE OF TECHNOLOGICAL MILESTONES IN BIG DATA MINING (DOMARADZKI ET AL., 2024).

8.4. Vision for Next-Generation Frameworks: AI-Driven Autonomous Systems

Autonomy and flexibility will be the focus of next-generation BDM systems. AI-driven data pipelines will eliminate a lot of the preprocessing, model choosing, and hyperparameter adjusting and bring down manual intervention by 90%. Self-healing designs will identify and correct data drift or model degradation in real-time, accuracy no more than 2–3% below optimal(Chang, Muñoz, & Ramachandran, 2020). Autonomy systems will incorporate generative AI to create training data for out-of-distribution events, enhancing fraud detection on imbalanced data sets by 25–30%. Quantum machine learning hybrids will break new optimization horizons, e.g., portfolio risk assessment with 100x acceleration in convergence. Interoperable standards such as the FAIR data principles will permit easy integration across different heterogeneous ecosystems to facilitate plug-and-play implementation of BDM solutions(Seyedan & Mafakheri, 2020).

9. Conclusion

9.1. Synthesis of Key Findings

Big Data Mining has been developed as a cross-disciplinary area to support innovation across sectors. Distributed frameworks such as Spark and Hadoop solve the 3Vs through scalable storage and parallel processing, while deep learning and ensemble techniques discover insights from unstructured information. Technologies like edge computing and quantum annealing hold the potential to circumvent today's constraints in latency and optimization, but with accompanying challenges regarding privacy, energy usage, and industry-academic collaboration. Convergence uses in healthcare, finance, and smart cities illustrate the disruptive potential of BDM, but ethics and regulatory standards need to keep pace with technology advancement.

9.2. Strategic Recommendations for Researchers and Practitioners

- **Prioritize Energy Efficiency**: Develop algorithms optimized for low-power hardware and renewable energy infrastructures.
- **Enhance Model Interpretability**: Integrate explainable AI (XAI) techniques to comply with regulatory demands and build user trust.
- **Foster Cross-Domain Collaboration**: Establish open-source consortia to unify academic research with industry standards.
- **Invest in Quantum Readiness**: Explore hybrid quantum-classical algorithms to prepare for near-term quantum advantages.
- **Adopt Privacy-by-Design**: Embed differential privacy and federated learning into workflows to ensure compliance and user protection.

10. References

Abdalla, H. B. (2022). A brief survey on big data: Technologies, terminologies and data-intensive applications. *Journal of Big Data, 9*(1), 107. https://doi.org/10.1186/s40537-022-00659-3

Alessandri, S., Ratto, M. L., Rabellino, S., Piacenti, G., Contaldo, S. G., Pernice, S., Beccuti, M., Calogero, R. A., & Alessandri, L. (2024). A survey of biological data in a big data perspective. *BMC Bioinformatics, 25*(1), 110. https://doi.org/10.1186/s12859-024-05695-9

Azad, M. A., Arshad, J., & Riaz, F. (2024). ROBO-SPOT: Detecting robocalls by understanding user engagement and connectivity graph. *Big Data Mining and Analytics, 7*(2), 340–356. https://doi.org/10.26599/BDMA.2023.9020028

Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., & Trunfio, P. (2022). Programming big data analysis: Principles and solutions. *Journal of Big Data, 9*(1), 4. https://doi.org/10.1186/s40537-021-00555-2

Cesario, E. (2023). Big data analytics and smart cities: Applications, challenges, and opportunities. *Frontiers in Big Data, 6*. https://doi.org/10.3389/fdata.2023.1149402

Chang, V., Muñoz, V. M., & Ramachandran, M. (2020). Emerging applications of internet of things, big data, security, and complexity: Special issue on collaboration opportunity for IoTBDS and COMPLEXIS. *Computing, 102*(6), 1301–1304. https://doi.org/10.1007/s00607-020-00811-y

Domaradzki, J., Majchrowska, A., Cielecka-Piontek, J., & Walkowiak, D. (2024). Unlocking the potential of big data and AI in medicine: Insights from biobanking. *Frontiers in Pharmacology, 15*. https://doi.org/10.3389/fphar.2024.1406866

Mach-Król, M., & Hadasik, B. (2021). On a certain research gap in big data mining for customer insights. *Applied Sciences, 11*(15), 6993. https://doi.org/10.3390/app11156993

Mutemi, A., & Bacao, F. (2024). E-commerce fraud detection based on machine learning techniques: Systematic literature review. *Big Data Mining and Analytics, 7*(2), 419–444. https://doi.org/10.26599/BDMA.2023.9020027

Porter, A. L., Zhang, Y., & Newman, N. C. (2024). Tech mining: A revisit and navigation. *Frontiers in Research Metrics and Analytics, 9*. https://doi.org/10.3389/frma.2024.1364053

Rani, R., Khurana, M., Kumar, A., & Kumar, N. (2022). Big data dimensionality reduction techniques in IoT: Review, applications and open research challenges. *Cluster Computing, 25*(6), 4027–4049. https://doi.org/10.1007/s10586-022-03634-y

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *Journal of Big Data, 7*(1), 53. https://doi.org/10.1186/s40537-020-00329-2

Tosi, D., Kokaj, R., & Roccetti, M. (2024). 15 years of big data: A systematic literature review. *Journal of Big Data, 11*(1), 73. https://doi.org/10.1186/s40537-024-00914-9

Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2021). Data mining in clinical big data: The frequently used databases, steps, and methodological models. *Military Medical Research, 8*(1), 44. https://doi.org/10.1186/s40779-021-00338-z

Zwilling, M. (2023). Big data challenges in social sciences: An NLP analysis. *Journal of Computer Information Systems, 63*(3), 537–554. https://doi.org/10.1080/08874417.2022.2085211