SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 8 | August 2025 - Peer Reviewed Journal

A REVIEW ON REAL-TIME DATA PIPELINES FOR E-COMMERCE TRANSACTIONAL DATA ANALYTICS

Mahesha K[1], Sagar BR[2], Tejonidhi M[3], Yashwanth N[4] , Ambika V[5]

[1]*Dept of CSE-Data Science, ATME College of Engineering, Mysuru*
[2]*Dept of CSE-Data Science, ATME College of Engineering, Mysuru*
[3]*Dept of CSE-Data Science, ATME College of Engineering, Mysuru*
[4]*Dept of CSE-Data Science, ATME College of Engineering, Mysuru*
[5]*Assistant Professor, Dept of CSE-Data Science, ATME College of Engineering, Mysuru*

Article DOI: https://doi.org/10.36713/epra23838
DOI No: 10.36713/epra23838

ABSTRACT

The exponential growth of e-commerce has necessitated the development of robust real-time data pipelines capable of processing high-velocity transactional data for instant decision-making. This review paper systematically examines existing methodologies, tools, and challenges in designing and implementing such pipelines, focusing on ingestion, processing, and analytics layers. By analyzing recent literature, we evaluate architectural frameworks (e.g., Kafka, Spark, Flink), performance trade-offs (latency vs. throughput), and emerging trends (AutoML, event-driven architectures). The paper highlights gaps in scalability, privacy, and interpretability while proposing future directions for intelligent, self-optimizing pipelines.

KEYWORDS: *Real-Time Data Processing, E-Commerce, Stream Processing, Apache Kafka, Fraud Detection, Automl,Genai, E-Commerce Data*

I. INTRODUCTION

E-commerce platforms generate vast streams of transactional data, including purchases, refunds, and user interactions, which require immediate processing to enable dynamic pricing, fraud detection, and personalized recommendations. Traditional batch processing systems fail to meet these demands due to inherent latency. This paper reviews state-of-the-art solutions for real-time data pipelines that ingest, process, and analyze transactional data continuously, aligning with the problem statement: "Design and implement a data pipeline that ingests, processes, and analyzes transactional data from an e-commerce platform in real-time."

We synthesize findings from 9 key studies (2019–2025) to evaluate technologies, architectures, and challenges in building scalable, low-latency pipelines.

II . LITERATURE REVIEW

II.a Data Ingestion Layer

- High-throughput tools: Apache Kafka and Amazon Kinesis dominate ingestion layers due to their scalability and fault tolerance (George, 2022; Saaed et al., 2024).
- Micro-batch trade-offs: Kushal Shah (2023) advocates for micro-batching in high-volume scenarios where sub-second latency is acceptable.
- Event-Driven Architectures (EDA): Pala (2024) demonstrates EDA's superiority in decoupling systems for resilience, e.g., PayPal's payment log processing (Vignesh et al., 2023).

II.b Stream Processing Layer

- Low-latency frameworks: Apache Flink and Spark Streaming enable in-memory processing for fraud detection and dynamic pricing (Bagam, 2022).
- Lambda Architecture: Matcha & Siddharth (2025) highlight its hybrid (batch + stream) approach for balancing accuracy and speed.
- AutoML integration: Thirunagalingam (2024) shows how AutoML automates model tuning for real-time recommendations, reducing manual effort.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 8 | August 2025 - Peer Reviewed Journal

II.c Analytics & Storage Layer

- Time-series databases: InfluxDB and Cassandra optimize query performance for real-time dashboards (Saaed et al., 2024).
- Explainable AI (XAI): Kehinde (2023) emphasizes the need for interpretable models in fraud detection to maintain regulatory compliance.

The paper by Bamigboy Kehinde highlights how real-time data pipelines enhance e-commerce personalization by enabling instant analysis of user behavior for dynamic recommendations, pricing, and engagement. It reviews technologies like stream processing, event-driven architectures, and ML techniques such as collaborative filtering and deep learning. Case studies from Amazon, Netflix, and Alibaba show business benefits, while challenges like latency, scalability, and privacy are noted. Future directions include explainable AI, federated learning, quantum computing, and edge analytics to improve trust and personalization accuracy.[1]

The paper by Mr. Jobin George emphasizes the need for real-time data pipelines to replace traditional batch processing for faster, smarter decision-making. It proposes a scalable AWS-based architecture using services like Kinesis, Lambda, DynamoDB, S3, Redshift, and QuickSight for seamless ingestion, processing, storage, and analytics. The framework supports diverse data sources, including IoT, web apps, and social media, while addressing scalability, cost, security, and latency. Real-world applications in healthcare, finance, and logistics showcase its adaptability. Ultimately, the study positions AWS as a robust platform for building end-to-end real-time analytics systems that transform raw data into actionable intelligence.[2]

The paper by Kushal Shah (2025) demonstrates how e-commerce companies are shifting from traditional batch processing to real-time ETL pipelines that achieve sub-second latencies and 99.999% data consistency. The research shows that modern architectures using micro-batch approaches and optimized state management can process over 500GB/day while reducing memory usage by 80% and query times by up to 76%. Key business applications include dynamic pricing (adjusting prices within 43 seconds of competitor changes), inventory optimization (reducing stockouts by 42%), and real-time personalization that significantly improves conversion rates. The paper concludes that real-time ETL has become a competitive necessity for e-commerce organizations seeking faster decision-making and measurable ROI improvements.[3]

The paper by Matcha & Siddharth (February 2025) presents a comprehensive study on building robust real-time data pipelines by integrating Apache Kafka, Apache Spark, and StreamSets technologies. The research demonstrates how Kafka provides scalable data ingestion, Spark delivers low-latency in-memory analytics, and StreamSets orchestrates the entire pipeline with advanced monitoring and automation capabilities. Through experimental evaluation across various workloads, the study reveals that throughput scales effectively with Kafka partitioning, though latency increases proportionally with data volume. The paper validates this architecture through real-world applications in e-commerce, healthcare, and finance industries. The authors conclude that this integrated approach creates a powerful, reliable, and efficient solution for enterprise-scale real-time analytics systems.[4]

The paper by Vignesh S, Srinath N K, and Sandeep R V presents a comprehensive data-driven framework for detecting payment fraud in e-commerce by analyzing transactional log data and user behavior patterns. The researchers employ statistical analysis and machine learning techniques, including Decision Trees, Random Forest, and Logistic Regression, to identify anomalies in transaction features such as user ID, amount, payment type, timing, and location. Their comparative analysis demonstrates that combining multiple features with appropriate algorithms significantly enhances fraud detection accuracy and precision, while visualizations provide clear representations of fraud probability and transaction patterns. The study addresses key challenges including false positives, imbalanced datasets, and evolving fraud patterns, emphasizing the critical need for real-time detection systems in fast-paced online commerce environments. The authors advocate for continuous learning systems and propose future enhancements including streaming data integration and deep learning implementation to create scalable, adaptive fraud identification models that can evolve with emerging fraud strategies.[5]

The paper by Naresh Pala presents a comprehensive event-driven architecture framework designed to build scalable and resilient real-time data processing systems for high-volume domains like e-commerce and IoT. The framework emphasizes modular design with decoupled components (ingestion, streaming processing, storage, and analytics) that enable independent scaling and fault tolerance, integrating technologies like Kafka, Spark Streaming, and NoSQL databases. Key technical features include message replication, checkpointing, replay mechanisms for high availability, cloud-native design with Kubernetes compatibility, and advanced strategies for backpressure handling and dynamic resource allocation. The authors validate their approach through real-world case studies across finance, healthcare, and retail sectors, while addressing production challenges such as event ordering, schema evolution, and monitoring. The paper concludes by advocating for future systems to incorporate AI-driven optimization and self-healing capabilities, providing a solid foundation for next-generation real-time analytics infrastructure.[6]

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 8 | August 2025 - Peer Reviewed Journal

The paper by Naveen Bagam (2022) explores the critical importance of real-time data analytics in e-commerce and retail, demonstrating how processing data as it's generated enables businesses to respond rapidly through dynamic pricing, instant fraud detection, personalized recommendations, and efficient inventory management. The research outlines comprehensive real-time system architectures encompassing data ingestion, processing, storage, and visualization components, while examining how industry leaders like Amazon, Netflix, and Alibaba leverage these technologies to enhance customer experience and operational agility. Key challenges addressed include latency optimization, scalability issues, privacy concerns, and complex data source integration, while emerging trends like edge computing, federated learning, and AI-driven personalization are identified as future directions. The study emphasizes that businesses investing in real-time analytics gain significant competitive advantages in customer satisfaction and market adaptability, while advocating for more intelligent and explainable systems to ensure user trust and regulatory compliance. The paper provides a foundational framework for implementing next-generation analytics systems capable of scaling with evolving data volumes and business requirements.[7]

The paper by Arunkumar Thirunagalingam (2024) provides a comprehensive review of real-time big data processing architectures, focusing on the industry shift from traditional batch processing to low-latency, high-throughput systems across e-commerce, finance, and healthcare sectors. The research examines core system layers including data ingestion, processing, storage, and analytics, evaluating key technologies like Apache Kafka, Spark, Flink, and Storm alongside NoSQL databases and distributed file systems for their streaming data capabilities. The authors analyze different processing models (micro-batch vs. continuous stream) while emphasizing critical requirements for fault tolerance, scalability, and data consistency in modern pipelines, supported by real-world use cases in retail and sensor data processing. Key challenges identified include data heterogeneity, integration complexity, and infrastructure costs, while future directions point toward integrating real-time analytics with machine learning, edge computing, and predictive modeling for intelligent automated responses. The paper establishes a foundational framework for designing scalable, efficient, and responsive big data systems that can meet evolving industry demands for real-time insights and decision-making.[8]

“Real-Time Analytics with Apache Cassandra and Apache spark by Sultan saaed, John Olusegun, Edwin Frank” paper discusses a framework for implementing real-time analytics by integrating Apache Kafka for high-throughput data ingestion and Apache Cassandra for scalable, low-latency data storage. It explains the advantages of using Kafka for processing live data streams and Cassandra for real-time querying and analytics across distributed systems. The authors highlight how this combination supports event-driven architecture and enables businesses to analyze incoming data instantly. The architecture is suitable for applications like fraud detection, e commerce transactions, and IoT systems where fast response to data is critical. Key components include Kafka producers, brokers, stream processors, and Cassandra clusters. The paper also covers data modeling strategies in Cassandra to optimize write/read performance. Real-world use cases and performance benchmarks show that the system achieves high availability, fault tolerance, and scalability. Challenges such as data duplication, message loss, and query complexity are acknowledged, with mitigation techniques suggested. The authors conclude by emphasizing the role of modern NoSQL and streaming tools in enabling truly real-time data systems and suggest areas for future research including AI integration and adaptive analytics pipelines.[9]

III. OUTCOME OF LITERATURE SURVEY

The literature reviewed reveals that real-time data processing plays a crucial role in enabling responsive, personalized, and intelligent decision-making in e-commerce platforms. Various architectures and tools, including Apache Kafka, Spark, StreamSets, and Cassandra, are commonly used to achieve high throughput, low latency, and scalability. Integration of automated machine learning (AutoML) has emerged as a powerful enhancement, enabling adaptive and efficient model management within pipelines. Event-driven architectures and micro-batch processing techniques further contribute to system resilience and improved performance. Real world implementations demonstrate success in areas such as fraud detection, dynamic pricing, and real-time recommendations. Despite advancements, challenges remain around data quality, model interpretability, privacy, and resource management. These findings highlight the need for further innovation in building scalable, explainable, and intelligent real-time data pipelines for e commerce.

IV METHEDOLOGY

This project is implemented in multiple interconnected stages to achieve real-time ingestion, processing, and analytics of e-commerce transactional data

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 8 | August 2025 - Peer Reviewed Journal

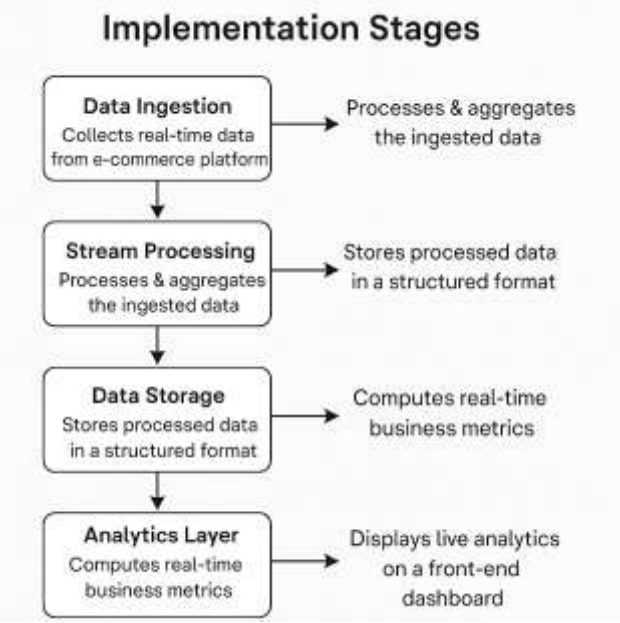


Fig 1 Data pipeline Implementation layers

Data Ingestion Real-time data is collected from simulated e-commerce activities such as customer purchases, cart updates, payments, and inventory operations. Tools like Apache Kafka or REST APIs are used to stream these events into the pipeline with low latency.

Stream Processing The ingested data is processed using a real-time computation engine (such as Apache Spark Streaming). Key operations include filtering, validation, transformation, and aggregation to ensure data cleanliness and usability for downstream analytics.

Data Storage Processed data is stored in structured formats using scalable storage systems (like Delta Lake or PostgreSQL). This enables fast retrieval and historical tracking of transactional activities.

Analytics Layer Real-time business metrics (e.g., total sales, user activity, inventory changes) are computed using pre-defined logic. This layer ensures that insights are derived continuously as new data flows in.

Dashboard & Visualization A front-end dashboard (built using tools like React.js or any web framework) visualizes these analytics live. It dynamically updates KPIs such as sales trends, user activity spikes, low-stock items, etc., enabling data-driven decisions.

V. CHALLENGES

- Scalability: Managing spikes during sales events (e.g., Black Friday).
- Privacy: GDPR-compliant anonymization of transactional data.
- Resource Intensity: AutoML models require significant computational power (Thirunagalingam, 2024).

Emerging Solutions

- Edge computing: Reduces latency by processing data closer to source (Bagam, 2022).
- Federated learning: Enhances privacy in personalized analytics (Kehinde, 2023).

VI. FUTURE SCOPE & RESEARCH GAPS

real-time data pipelines are pivotal for e-commerce agility, but gaps remain in interpretability, energy efficiency, and cross-platform standardization. future work should explore:

1. Unified frameworks combining Kafka, Flink, and AutoML.
2. Quantum computing for ultra-low-latency processing.
3. Ethical AI guidelines for transparent real-time decision-making. This review underscores the need for adaptive pipelines that balance speed, accuracy, and ethics in the evolving digital econom

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 8 | August 2025 - Peer Reviewed Journal

VII .CONCLUSION

This project highlights the growing importance of real-time data pipeline es in transforming e-commerce platforms into intelligent, responsive systems. By continuously ingesting, processing, and analyzing transactional data, businesses can enhance personalization, detect fraud, and optimize inventory management in near real-time. The literature shows that integrating tools like Kafka, Spark, and AutoML significantly boosts system scalability, adaptability, and analytical power. Event-driven architectures and micro-batch processing approaches further improve throughput and fault tolerance. Real-world case studies affirm the effectiveness of these technologies in delivering actionable insights across domains. However, challenges such as data privacy, model interpretability, and computational overhead still persist. Incorporating explainable AI and edge computing may address these concerns while pushing the boundaries of real-time analytics. Ultimately, this project serves as a foundational step toward building smarter, self-optimizing, and customer-centric e-commerce systems.

Acknowledgment: We are deeply thankful to Prof.Ambika V, CSE–Data Science, at ATME College of Engineering, Mysuru, for her guidance and continuous encouragement throughout our research.
We would also like to express our gratitude to all faculty members for their support and helpful feedback. A special thanks goes to the authors of the studies we reviewed your contributions greatly shaped our understanding of data pipelines and its role in analysis of real time data of e-commerce platforms

REFERENCES

1. *Real-Time Data Processing Techniques for E-Commerce Personalization by Bamigboy kehinde (2023).*
2. *Build A Realtime Data Pipeline: Scalable Application Data Analytics On Amazon Web Services by MR. JOBIN GEORGE PartnerEngineering Lead, Data Analytics Science & Technology CA, USA*
3. *Real-Time Analytics in E-commerce: Strategies for Implementing Near Real-Time ETL Pipelines by Kushal Shah Fairleigh Dickinson University, USA*
4. *Building Robust Data Pipelines: Real-Time Data Processing with Spark, Kafka, and StreamSets Sasibhushana Matcha & Er. Siddharth, Visvesvaraya Technological UniversityMachhe, Belagavi, Karnataka 590018, &2Bennett University Greater Noida, Uttar Pradesh 201310, India*
5. *Data Analytics on E-Commerce Transaction Logs for Payment Management by Vignesh S &*
6. *Sandeep R V Payments Engineering, PayPal Inc. Bangalore, India, Srinath N K Dean, Computer Science & Engineering, R V College of Engineering Bangalore, India.*
7. *Understanding Event-Driven Architecture: A Framework for Scalable and Resilient Systems Naresh Pala The Kroger Co, USA*
8. *Real-Time Data Analytics in E-Commerce and Retail Naveen Bagam Independent Researcher, USA*
9. *Transforming Real-Time Data Processing: The Impact of AutoML on Dynamic Data Pipelines Arunkumar Thirunagalingam1, Department of Business Intelligence and Reporting, Santander Consumer, Texas, United States of America.*
10. *Real-Time Analytics with Apache Cassandra and Apache spark by Sultan saaed, John Olusegun, Edwin Frank (2024)*
11. *Advancements in Data Ingestion: Building High-Throughput Pipelines with Kafka and Spark Streaming by Chandrakanth Lekkala(July 2020)*
12. *Apache Kafka and real-time data streaming by Chandrakanth Lekkala Florida Institute of Technology (July 2021)*
13. *Harnessing real-time data analytics for strategic customer insights in e-commerce and retail by Olamide Raimat Amosu 1, Praveen Kumar 2, Adenike Fadina 3, Yewande Mariam Ogunsuji 4, Segun Oni 5 and Kikelomo Adetula 6 (aug 2024)*
14. *E–Commerce Trends and Future Analytics Tools by Premkumar Balaraman and Sabarinathan Chandrasekar (August 2016)*