computer sciences and mathematics forum

MDPI

Proceeding Paper

The Evolution and Challenges of Real-Time Big Data: A Review †

Ikram Lefhal Lalaoui *, Essaid El Haji and Mohamed Kounaidi

Intelligent Automation and BioMed Genomics Laboratory, FST of Tangier, Abdelmalek Essaadi University, Tetouan 93000, Morocco; e.elhaji@uae.ac.ma (E.E.H.); mkounaidi@uae.ac.ma (M.K.)

* Correspondence: ikram.lefhallalaoui@etu.uae.ac.ma

† Presented at the International Conference on Sustainable Computing and Green Technologies (SCGT'2025), Larache, Morocco, 14–15 May 2025.

Abstract

The importance of real-time big data has become crucial in the digital revolution of modern society, in the context of increasing data flows from multiple sources, including social media, internet connected devices (IOT) and financial systems, real-time analysis and processing is becoming a strategic tool for fast and accurate decision making, we find applications in different domains such as healthcare, finance, and digital marketing, which is revolutionizing traditional business models. In this article, we explore the recent advances and future prospects of real-time big data. Our research is based on recent work published between 2020 and 2025, examining the technological advances, the difficulties encountered and suggesting ways of optimizing the efficiency of these technologies.

Keywords: Big Data; data analytics; data management; stream processing; NoSQL; spark streaming; Flink

1. Introduction

The digital transformation brought the concept of Big Data, which has become one of the most important elements of current information systems, enabling the collection, storage and analysis of large volumes of highly heterogeneous data [1]. With that, traditional systems based on batch processing and static architectures would fail to keep up with the trends in terms of real-time insight generation, low latency, and high scalability [2].

The ability to analyse real-time data in an efficient and reliable manner provides a decisive competitive advantage for organizations, especially in industrial and commercial environments [3]. On the other hand, the rapid growth of interconnected platforms, including devices in the Internet of Things (IoT), social media and online services, creates huge, high-velocity streams of data [4].

Technologies like Apache Kafka as well as other mechanisms have undergone a number of performance optimizations to address these challenges, including improvement in topic relevance and establishing automatic latency control [5]. More generally, a second ecosystem comprising stream processing engines, distributed storage systems, and artificial intelligence (AI) [6] has emerged that provides intelligent, real-time decision making.

This article provides a critical review of advances in the management of massive real-time data, focusing specifically on technologies, practical applications, and future opportunities in this field. We also explore ways of resolving the difficulties associated with the processing of massive real-time data, paving the way for new research to optimize these processes.

check for updates

Academic Editors: Hicham Gibet Tani, Mohamed Kouissi, Mohamed Ben Ahmed, Anouar Abdelhakim Boudhir and Lotfi Elaachak

Published: 1 July 2025

Citation: Lalaoui, I.L.; El Haji, E.; Kounaidi, M. The Evolution and Challenges of Real-Time Big Data: A Review. *Comput. Sci. Math. Forum* **2025**, *10*, 11. https://doi.org/10.3390/cmsf2025010011

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

2. Methodology

In this review, a qualitative and comparative approach is taken to understand the evolution and challenges of real-time Big Data technologies. It is learned from the data obtained from all academic publications in between 2020 to 2025.

Searches were made in major scientific databases, such as Institute of Electrical and Electronics Engineers (IEEE) Xplore, Science Direct, SpringerLink, and Elsevier, for relevant articles. We performed the literature search using a list of keywords including, but not limited to, "real-time Big Data", "stream processing", "Apache Kafka", "Spark Streaming", "Flink", "NoSQL", "incremental learning", "data architecture" and "edge computing".

The criteria for inclusion were peer-reviewed journal articles or conference proceedings related to real-time ingestion, processing, architecture, and intelligent systems based on novel tools as well as challenges during the 2020–2025 era. On the contrary, articles focusing exclusively on conventional batch processing or were limited in technical or methodological depth were excluded from the analysis.

About 70 different articles were analysed. Most importantly, the latest and influential works were selected to summarize the critical technological trends, principal frameworks, and future directions in research of real-time Big Data processing.

3. Background

Real-time big data management is based on a set of technologies that enable the collection, processing, storage and continuous analysis of data.

Recent advances in these domains have come from innovative studies and research, and this review highlights the main tools and frameworks used.

3.1. Data Flow Management

3.1.1. Apache Kafka

Apache Kafka is a distributed message management platform created for the continuous processing of large amounts of data, it is often used to guarantee reliable and scalable transmission of circumstances in real time [4]. Kafka operates on a publish-subscribe scheme that ensures data retention, therefore offering high fault tolerance, in their study [4] on optimising topic relevance in Kafka, BroMin and broMax improving latency and system resilience.

3.1.2. ETL Stream

The Stream ETL architecture, consisting of Kafka, spark and Cassandra, is designed to optimise the analysis of massive volumes of data. Ref. [7] explains how this method reduces latency and optimises the scalability of systems.

3.2. Real-Time Flow Processing

3.2.1. Apache Flink

Apache Flink is a stream processing engine designed to perform real-time data analysis tasks in situations of reliable latency and high availability, widely used in the field of analysis [8], using a "shared-nothing" architecture to improve scalability and reduce memory consumption by more than 50%.

3.2.2. Apache Storm

Apache Storm is a framework used for real-time distributed processing of targeted data, typically used for anomaly detection and analysis of social network data, ref. [9] applied this technology to the detection of anomalies in tweets related to COVID-19. In addition, ref. [2] highlights the importance of Apache storm in the industrial B2B market-

ing sector, emphasising its importance in the real-time processing of massive data from connected objects and social platforms, enabling companies to adjust their live advertising strategies and react more effectively to consumer behaviour.

3.3. Real-Time Analysis with Apache Spark Streaming

Spark Streaming is an extension of Apache Spark used to process continuous data streams, it offers the performance of a distributed computing engine for this type of processing, this technology is often used in areas such as the analysis of massive data streams on social networks, the analysis of sentiments in real time. Ref. [7] have developed a real-time data stream management framework, which combines technologies such as spark streaming, Kafka and Cassandra, with the aim to enhance sentiment analysis on twitter, their stream ETL framework facilitates real-time processing and visualisation of tweets, using spark streaming to extract, transform and load large volumes of data [7], Spark Streaming is also used in other sectors that require real-time processing, such as monitoring price variations, and optimising systems and setting tariffs according to market fluctuations [10].

3.4. NO_SQL Database for Real-Time

3.4.1. Cassandra

Cassandra is a NO_SQL database often used in big data contexts, and designed for distributed storage and efficient integration of data flows. It is designed to guarantee horizontal scalability and fault tolerance, making it an appropriate choice for real-time systems that require rapid management of large quantities of data [8].

3.4.2. Hbase

Hbase is a distributed column-based database developed for big data contexts that require instant access to information in real time. it is often used in parallel with hadoop for the storage and analysis of massive unstructured data. Ref. [9] have proven the effective incorporation of Hbase into their live processing framework for analysing sentiment on twitter. they use the Stream ETL system with Hbase to retain intermediate processing results ensuring robust data storage in a real time big data environment.

3.5. Machine Learning and AI for Big Data

Artificial intelligence (AI) and machine learning are instrumental in improving real-time data management. In response to the challenges posed by massive data flows, refs. [11,12] introduced the concept of a dynamic balanced quadtree (DB-quadtree) to solve the balancing problems encountered with traditional quadtree structures [12]. This method optimises the indexing of data flows, facilitating fast and efficient queries in real time, particularly in scenarios where data density is unpredictable. Alongside these innovations, incremental learning is emerging as an appropriate response to constantly changing environments. This method adjusts to the data in real time, without requiring a complete re-training of the models, ref. [13] have applied an incremental ensemble type framework to predict electricity prices in Spain in real time, using their [13] explainable method, they have strengthened the reliability of the forecasts while ensuring the interpretability of the results, which is crucial for decision-makers in the energy sector.

4. Literature Review

Real-time big data processing is a category based on architectures and technologies that process massive data with very reliable latency and high availability. This real-time data comes from a variety of sources, including IOT sensors, social networking platforms,

financial transaction systems, mobile devices, etc. These data flows require infrastructures for ingestion, processing and storage:

First generation (2000–2010): The rise of distributed database systems and data processing technologies has marked a significant advance in the management of massive data, and frameworks such as Apache Hadoop have enabled large quantities of data to be processed efficiently using batch processing techniques, becoming essential for coping with the sheer volume of massive data [4]. It was against this backdrop that Apache Kafka emerged around 2010, offering not only efficient management of large-scale data, but also extremely high processing speed. This capability has proved crucial for applications requiring real-time data flows [4,9]. This ability to distribute data and process flows instantaneously represents a relevant response to the challenges posed by the variability and speed of data [4].

Second generation (2010–2015) marked a transition towards architectures adapted to continuous data processing, with the growing adoption of dedicated tools such as Apache Storm. The latter has facilitated real-time event processing, particularly for monitoring social networks and detecting anomalies [14]. At the same time, HBase has become a popular solution for real-time storage, particularly in large enterprises requiring distributed databases [7].

With the growth of social networks, connected objects and online services, streaming processing systems have met the demand for near real-time analysis [3]. Technologies such as Kafka Streams have been adopted for their low latency, facilitating the analysis of real-time events and their application in various domains, such as infrastructure monitoring and anomaly detection [3].

In the financial sector, the adoption of Big Data technologies has encouraged the development of predictive models based on the analysis of transaction flows and markets [15].

Third generation (2015–2020), the focus was on integrating more powerful streaming technologies, such as Apache Kafka and Spark Streaming, to manage massive data streams. These tools have facilitated real-time analysis on platforms such as Twitter, particularly for sentiment analysis and critical event detection [7]. The need to visualise data fast and to optimise NoSQL databases as Cassandra has become obvious. For example, ref. [11] demonstrated how real-time Big Data technologies, used for food price analysis in Poland, support economic decision-making through the analysis of massive price streams collected daily. In addition, the development of Deep Learning techniques applied to real-time data has improved anomaly detection and hidden pattern recognition [16]. In the context of sentiment analysis on social platforms as Twitter, the massive data flow management with low latency has become crucial. The use of Kafka and Spark, coupled with sophisticated analytical tools, has made it possible to integrate essential features such as event-driven data processing, robust state management and horizontal scalability. These systems also facilitate interaction with real-time machine learning models. Nevertheless, these systems have presented challenges, including high resource requirements and complex management of system resilience [7].

Fourth generation (2020–2025): Real-time Big Data has reached a level of hyper-responsiveness, marked by improved ETL architectures, the integration of artificial intelligence and the optimisation of distributed systems. The automation of data flow processing has improved instantaneous decision-making, as exemplified by Complex Event Processing (CEP) and Stream Processing (SP) systems, which facilitate the analysis of heterogeneous flows in critical environments such as the monitoring of infrastructures and the management of digital services [3]. In tandem, advances in cybersecurity have enhanced real-time stream pre-processing, enabling improved anomaly detection through machine learning mechanisms [17].

There have been further important advances in distributed streaming technologies, with improvements in data stream management platforms. The optimisation of processing engines such as Flink has made it possible to efficiently analyse continuous streams, especially in recommender systems where latency and memory management are essential [8]. The application of real-time Big Data to economic forecasting has also progressed, with the exploitation of intelligent sensors to refine the analysis of price variations and improve strategic decisions in real time, as suggested by the article's discussion on the use of online prices [11]. In the e-commerce sector, advanced analysis of data flows has made it possible to optimise online sales strategies using streaming processing architectures that dynamically adapt recommendations to market trends [18]. Improved management of cloud infrastructures has increased the reliability and reduced the latency of data flows, in particular through optimisations of Apache Kafka which ensure better scalability of distributed infrastructures [4].

Explainable artificial intelligence (XAI) has emerged as an essential lever in real-time system optimisation, making predictive models more transparent and interpretable. The application of these techniques to energy forecasting has led to improved management of the electricity grid and increased reliability of energy demand estimates [19]. The exploitation of Big Data technologies for the analysis of online trends has also become more effective, particularly with the integration of Kafka, Spark Streaming and HBase, which allow massive and instantaneous processing of data from social networks [7].

The evolution of real-time Big Data therefore relies on more flexible, high-performance architectures capable of managing growing data flows in key sectors including economics, cybersecurity, energy and digital commerce. The convergence of streaming, AI and distributed systems is opening up new prospects for more intelligent, proactive management of real-time information.

5. Discussion

Real-time Big Data systems have come a long way in the past few years and seen improvements in data ingestion, processing, and decision making. Yet, even with these advances, major challenges remain. Comparing the latest technologies, as well as analysing their drawbacks and recommending strategies towards improving future systems, forms this section.

5.1. Recent Technological Advances

Modern stream processing frameworks like Kafka Streams, Apache Flink, and Spark Streaming have made great strides in reducing latency and scaling up the throughput in various high-throughput scenarios [4,8]. When embedded in optimized ETL pipelines, these tools have transformed industries including finance, energy, and digital commerce [17].

Explainable AI (XAI) and incremental learning algorithms have been added in these models to make it more adaptable and to add transparency into a domain like financial forecasting and cyber security where volatility is a key factor [11,17].

Moreover, partitioning strategies for Apache Kafka [4], distributed processing frameworks like TensorFlow and PyTorch, and hybrid architectures for cloud and edge computing have also been in the limelight and devised to augment resilience and responsiveness [19]. These advances help speed processing of large-scale, unstructured data across a variety of applications, including IoT monitoring and public health surveillance.

Multimodal methods employed for real-time sentiment analysis, real-time dynamic pricing, and real-time sales prediction are likewise seeing more adoption, with utilizes of DAG modeling for anomaly detection, strategic decision making, etc. [4,9,18]. Additionally,

these techniques enhance the resource management efficiency for Big Data workloads in the cloud with continuous improvements [5].

5.2. Remaining Challenges and Limitations

There are a number of obstacles preventing the ideal use of real-time Big Data systems. The sheer volume of data has been growing exponentially, putting tremendous pressure on both storage Infrastructures and Network Bandwith, usually, the volume has never been observed before and sometimes can exceed the limits of reliable Process and Storage [18]. Moreover, the absence of standardized data formats and the low interoperability of processing platforms limit the scalability and integration of these solutions [11].

Latency is still a major issue. Although distributed workloads are possible on platforms such as Kafka and Flink, these systems generally fail to fulfill performance consistency when multiple data sources are aggregated [4]. Data quality Another grand challenge is data quality heterogeneous unstructured high-velocity data causes various models to be inaccurate. In addition, this extraction has to be validated using complex preprocessing methods, including anomaly removal and real-time detection algorithms [11].

More severe technical constraints like quadtree management in high-speed data streams [12] and Skyline recomputation in large-scale databases [20], considerably restrict the room for optimization of the systems and call for new algorithmic patterns.

5.3. Recommendations and Future Directions

Looking forward, the integration of AI and ML in real-time pipelines will play a central role in automating decision-making and detecting anomalies proactively [19]. Decentralized, edge-based systems are expected to reduce latency even further while improving responsiveness at the data source.

Future work should focus on improving interoperability between cloud-native and edge-native systems. The adoption of open standards and dynamic orchestration of data flows will enable more efficient use of resources and support mission-critical analytics [7]. In the e-commerce space, streaming learning models and multimodal analytics are expected to drive marketing personalization and predictive insights [18].

In the energy and supply chain sectors, the application of incremental prediction models for electricity pricing [13] and logistics optimization [21] presents significant potential. Finally, emerging advances in cybersecurity, distributed analysis [22], and graph-based processing optimization [23] are expected to bolster the resilience and scalability of real-time Big Data architectures in the coming years.

6. Conclusions

Real-time Big Data has transformed many sectors by improving the analysis and processing of massive flows of information. The growth of distributed architectures and streaming learning models has enhanced the ability of systems to adapt to rapid changes in data [5,16].

Notwithstanding these advances, challenges persist, including effective resource management and data standardisation for improved interoperability [24]. Security and confidentiality also continue to be major challenges, requiring advanced solutions to protect real-time data flows [25].

In the future, the integration of improved predictive models and the optimisation of cloud infrastructures will increase the efficiency of Big Data systems. The move towards hybrid and distributed solutions will ensure better management of continuous flows and increased resilience of architectures [23].

Therefore, the Big Data real-time analysis provides a significant opportunity for the data analysis and exploitation. But development in this field needs constant working on technical difficulties, so that it is effective and secure to use. What we expect is that a combination of innovation and optimisation will drive a new solution that can realise the full potential of real-time Big Data in the next few years.

Author Contributions: Conceptualization, I.L.L.; methodology, I.L.L.; software, I.L.L. and E.E.H.; validation, I.L.L., E.E.H. and M.K.; formal analysis, I.L.L.; investigation, I.L.L.; resources, I.L.L.; data curation, I.L.L.; writing—original draft preparation, I.L.L.; writing—review and editing, I.L.L., E.E.H. and M.K.; visualization, I.L.L.; supervision, E.E.H. and M.K.; project administration, E.E.H. and M.K.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IOT	Internet of Things
IT	Information Technology
ETL	Extract, Transform, Load
AI	Artificial Intelligence

References

1. Yim, S.T.; Son, J.C.; Lee, J. Spread of E-commerce, prices and inflation dynamics: Evidence from online price big data in Korea. *J. Asian Econ.* **2022**, *80*, 101475. [CrossRef]
2. Jabbar, A.; Akhtar, P.; Dani, S. Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Ind. Mark. Manag.* **2020**, *90*, 558–569. [CrossRef]
3. Corral-Plaza, D.; Ortiz, G.; Medina-Bulo, I.; Boubeta-Puig, J. MEdit4CEP-SP: A model-driven solution to improve decision-making through user-friendly management and real-time processing of heterogeneous data streams. *Knowl. Based Syst.* **2021**, *213*, 106682. [CrossRef]
4. Raptis, T.P.; Cicconetti, C.; Passarella, A. Efficient topic partitioning of Apache Kafka for high-reliability real-time data streaming applications. *Future Gener. Comput. Syst.* **2024**, *154*, 173–188. [CrossRef]
5. Fu, X.; Pan, L.; Liu, S. To store or not: Online cost optimization for running big data jobs on the cloud. *Future Gener. Comput. Syst.* **2024**, *156*, 42–52. [CrossRef]
6. Tian, G.; Wu, W. Big data pricing in marketplace lending and price discrimination against repeat borrowers: Evidence from China. *China Econ. Rev.* **2023**, *78*, 101944. [CrossRef]
7. Ismail, A.; Sazali, F.H.; Jawaddi, S.N.A.; Mutalib, S. Stream ETL framework for twitter-based sentiment analysis: Leveraging big data technologies. *Expert. Syst. Appl.* **2025**, *261*, 125523. [CrossRef]
8. Hazem, H.; Awad, A.; Yousef, A.H. A distributed real-time recommender system for big data streams. *Ain Shams Eng. J.* **2023**, *14*, 102026. [CrossRef]
9. Amen, B.; Faiz, S.; Do, T.T. Big data directed acyclic graph model for real-time COVID-19 twitter stream detection. *Pattern Recognit.* **2022**, *123*, 108404. [CrossRef]
10. Pébereau, C.; Remmy, K. Barriers to real-time electricity pricing: Evidence from New Zealand. *Int. J. Ind. Organ.* **2023**, *89*, 102979. [CrossRef]
11. Macias, P.; Stelmasiak, D.; Szafranek, K. Nowcasting food inflation with a massive amount of online prices. *Int. J. Forecast.* **2023**, *39*, 809–826. [CrossRef]

12. Yang, G.; Wu, X.; Zhang, J. A dynamic balanced quadtree for real-time streaming data. *Knowl. Based Syst.* **2023**, *263*, 110291. [CrossRef]
13. Melgar-García, L.; Troncoso, A. A novel incremental ensemble learning for real-time explainable forecasting of electricity price. *Knowl. Based Syst.* **2024**, *305*, 112574. [CrossRef]
14. Pauwels, K.; Aksehirli, Z. Big data analytics democratized with clean collaboration and customer privacy choice. *J. Bus. Res.* **2025**, *188*, 115112. [CrossRef]
15. Bricongne, J.C.; Meunier, B.; Pouget, S. Web-scraping housing prices in real-time: The Covid-19 crisis in the UK. *J. Hous. Econ.* **2023**, *59*, 101906. [CrossRef]
16. Selmy, H.A.; Mohamed, H.K.; Medhat, W. Big data analytics deep learning techniques and applications: A survey. *Inf. Syst.* **2024**, *120*, 102318. [CrossRef]
17. Dębski, R.; Dreżewski, R. Real-time surrogate-assisted preprocessing of streaming sensor data. *Comput. Netw.* **2022**, *219*, 109422. [CrossRef]
18. Xu, W.; Cao, Y.; Chen, R. A multimodal analytics framework for product sales prediction with the reputation of anchors in live streaming e-commerce. *Decis. Support. Syst.* **2024**, *177*, 114104. [CrossRef]
19. Mari, A.; Remlinger, C.; Castello, R.; Obozinski, G.; Quarteroni, S.; Heymann, F.; Galus, M. Real-time estimates of Swiss electricity savings using streamed smart meter data. *Appl. Energy* **2025**, *377*, 124537. [CrossRef]
20. Bourahla, C.; Maamri, R.; Brahimi, S. Skyline recomputation in Big Data. *Inf. Syst.* **2023**, *114*, 102164. [CrossRef]
21. Esmaeeli, Z.; Mollaverdi, N.; Safarzadeh, S. A game theoretic approach for green supply chain management in a big data environment considering cost-sharing models. *Expert Syst. Appl.* **2024**, *257*, 124989. [CrossRef]
22. Berloco, F.; Bevilacqua, V.; Colucci, S. Distributed Analytics For Big Data: A Survey. *Neurocomputing* **2024**, *574*, 127258. [CrossRef]
23. Li, Z.; Liu, S.; Liu, J.; Zhang, Y.; Liang, T.; Liu, K. SIM: A fast real-time graph stream summarization with improved memory efficiency and accuracy. *Comput. Netw.* **2024**, *248*, 110502. [CrossRef]
24. Dwivedi, A.; Pant, R.P. An algorithmic implementation of entropic ternary reduct soft sentiment set (ETRSSS) using soft computing technique on big data sentiment analysis (BDSA) for optimal selection of a decision based on real-time update in online reviews. *J. King Saud. Univ.-Comput. Inf. Sci.* **2022**, *34*, 2118–2130. [CrossRef]
25. Kalra, R.; Singh, T.; Mishra, S.; Satakshi; Kumar, N.; Kim, T.; Kumar, M. An efficient hybrid approach for forecasting real-time stock market indices. *J. King Saud. Univ.-Comput. Inf. Sci.* **2024**, *36*, 102180. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.