Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02589-x

REVIEW

A comprehensive survey of stream reasoning and its integration with knowledge graphs

Gözde Ayşe Tataroğlu Özbulak[1,2,4] · Gaetano Manzo[1,3] · Yash Raj Shrestha[2] · Jean-Paul Calbimonte[1,4]

Received: 16 May 2024 / Revised: 14 August 2025 / Accepted: 22 August 2025

Abstract

The rapid expansion of decentralized, complex streaming data across diverse domains such as the Internet of Things, healthcare, and smart cities presents significant technical challenges. These challenges–data heterogeneity (integration of diverse formats and sources), dynamicity (handling real-time data evolution), and high-volume throughput (efficient processing of large, rapidly arriving data)–are the central focus of this study and are examined in depth. To address these critical issues necessitates advanced methods capable of seamless integration, effective real-time reasoning, and continuous learning from heterogeneous streaming data, thus enhancing real-time decision-making capabilities. This study provides an extensive review of existing research at the intersection of streaming data, machine learning, and reasoning. The literature review categorizes Stream Reasoning approaches into three key groups: Streaming Machine Learning, Streaming Linked Data, and Streaming Knowledge Graphs. Each category is critically examined in terms of strengths, limitations, ongoing challenges, and future opportunities identified in recent studies. Additionally, potential integrative solutions that leverage Knowledge Graph structures and advanced Stream Reasoning techniques are highlighted, illustrating how state-of-the-art modeling methods can effectively address Stream Reasoning related challenges. The analysis concludes that combining Knowledge Graph and Machine Learning approaches significantly enhances the capability to manage and overcome complex Stream Reasoning challenges.

✉ Gözde Ayşe Tataroğlu Özbulak
gozdeayse.tatarogluozbulak@unil.ch

Gaetano Manzo
gaetano.manzo@hevs.ch

Yash Raj Shrestha
yashraj.shrestha@unil.ch

Jean-Paul Calbimonte
jean-paul.calbimonte@hevs.ch

1 University of Applied Sciences and Arts Western Switzerland HES-SO, TechnoPôle 3, 3960 Sierre, Switzerland

2 University of Lausanne, Unicentre, 1015 Lausanne, Switzerland

3 Computational Health Research Branch, National Library of Medicine, Bethesda, Maryland 20894, USA

4 The Sense Innovation and Research Center, Av de Provence 83, 1007 Lausanne, Switzerland

Published online: 28 October 2025

Keywords Stream reasoning · Streaming machine learning · Streaming linked data · Streaming knowledge graphs

1 Introduction

Stream Reasoning (SR) processes information in real-time by making logical deductions from a dynamic flow of data from various sources [1]. As stream processing technology advances, it becomes increasingly integral to decision-making across different fields. Commonly used in Internet of Things (IoT) applications, big data analytics, and monitoring complex systems like healthcare, SR finds diverse applications, from smart cities to manufacturing and financial analysis [2, 3]. This technology speeds up data analysis and allows systems to quickly adapt to changing conditions.

However, SR encounters several challenges inherent in the nature of streaming data, commonly characterized by heterogeneity, dynamicity, and high volume. *Heterogeneity* arises from the diverse sources and formats of the incoming data streams, leading to semantic conflicts and integration issues during modeling and reasoning [4–6]. *Dynamicity* manifests in the continuous evolution of data streams, requiring constant adaptation in modeling, updating, and reasoning over non-stationary data [7]. Moreover, the sheer scale and speed of data stream pose the challenge of *high volume*, necessitating efficient storage and processing methods for scalable operations [8, 9]. Although these main challenges —heterogeneity, dynamicity, and high volume— are addressed by various methods in literature, there is still no comprehensive method that targets all of them. Based on the characteristics of each challenge, the targeted application area and its scope, the chosen processing method or approach may vary.

Within SR, numerous approaches utilize Knowledge Graphs (KGs) and Machine Learning (ML) methods to address its challenges. KGs provide versatile semantic modeling and querying capabilities that interconnect heterogeneous data and, by structuring and correlating high-volume streams, simplify large-scale data management [10, 11]. In parallel, the integration of ML with SR has led to Streaming Machine Learning (SML), where models continuously learn and adapt to evolving data and concept drift in dynamic streams [12]. These advancements address specific challenges across various domains. For example, in healthcare, real-time SR applications are crucial for promptly responding to patient needs through monitoring systems and personalized healthcare programs [13]. Similarly, in industrial processes, real-time management of dynamic, large-scale stream data is essential. Stream based ML methods are employed in industrial engines, aiding in the identification of potential disruptions amidst dynamic and heterogeneous data structures [14]. These examples highlight the broad applicability of SR solutions across diverse sectors, each confronting its unique set of challenges.

To identify research opportunities for improved solutions, it is essential to comprehensively study, classify, and analyze various approaches and methods documented in the literature. This study aims to highlight different approaches grouped into three main categories: ML, Linked Data, and KGs. It provides an overview of SR-based methods and assesses their effectiveness on high-volume, heterogeneous, real-time streams. Moreover, this study identifies ongoing challenges in SR, to shape further research opportunities. Primary focus is placed on how the integration of SR and KGs helps coping with those challenges, as analyzed in Section 6. This work further maps how hybrid SR–KGs approaches mitigate main SR challenges defined in this study (heterogeneity, dynamicity, and high-volume). The contributions of this study can be listed as follows:

- Introducing primary challenges —heterogeneity, dynamicity, and high volume— faced in SR, while categorizing approaches based on their methodologies for addressing these challenges.
- Exploring studies on SR challenges, including current SML and Deep Learning strategies, and discussing the application of KGs to tackle heterogeneity, dynamicity, and high volume.
- Identifying open challenges in SR and investigating opportunities in integration of SR and KGs to alleviate challenges.
- Gathering the metrics and criteria defined in literature that can evaluate the performance of SR based approaches.

The remainder of this paper is structured as follows. Section 2 presents the fundamental notions and definitions that underpin the study, including both KGs and SR. Section 3 describes the methodology used for the survey, detailing the research questions, search strategy, and scope. Section 4 outlines the key challenges in Stream Reasoning and relates them to existing work. Section 5 reviews current approaches, classified into SML, Streaming Linked Data (SLD), and Streaming Knowledge Graphs (SKG), and concludes with open challenges to investigate further opportunities. Section 6 investigates how the integration of SR and KGs can address these challenges. Section 7 discusses evaluation approaches and metrics employed in the literature. Section 8 presents a broader discussion that synthesizes the findings. Section 9 provides the conclusions of the survey. Finally, Section 10 outlines a structured agenda for future research.

2 Background

In this section, we introduce the main characteristics and definitions regarding KGs and SR. Then the relationships between these paradigms are highlighted.

2.1 Knowledge graphs

The general definition of KGs in the literature is a data model that represents knowledge using contextual relationships and semantic structures [15]. The primary purpose of creating KGs is to construct a meaningful structure by linking data through semantic relationships, thereby facilitating efficient and intuitive access to information [16]. KGs can be configured manually or automatically. In manual structuring, domain experts label data elements and determine their relationships based on literature, common sense, and expert knowledge [15, 17]. In contrast, automatic structuring relies on Machine Learning (ML) techniques, such as Named Entity Recognition (NER), to identify entities and infer their relationships from textual data [18]. Additionally, semantic analysis methods and ontologies from the literature support the automation of this process, enabling the extraction of meaningful relationships between Knowledge Graph (KG) nodes and edges [17, 18].

For example, the Artificial Intelligence in Knowledge Graph (AI-KG) framework introduced in [19] applies an automated pipeline that combines various tools and evaluation methods to extract entities and their relationships. Owing to AI capabilities, KGs can integrate, monitor, and process dynamic data from heterogeneous sources through decentralized architectures. The approach in [10] demonstrates decentralized KG construction for a deep recommender system. To manage dynamic data, KG frameworks also incorporate semantic web technologies and ontologies by allowing real-time graph updates [20]. As a result, data

can be queried, transferred, and accessed via a semantically defined common language within the KG.

2.2 Stream reasoning

SR techniques interpret rapidly changing complex data streams, extract patterns, and enable real-time decisions [9].

Also, SR offers to interpret the complex structure of big data through stream-based modeling that adapts to rapidly changing conditions in real-time [21]. Similarly, SR has the potential to maintain integrity and reliability of large-scale stream data at high speed coming from various sources or devices. This issue causes challenges while creating streaming model and robustness of the model can decrease [1]. At this point, the use of new approaches with state-of-the-art ML methods to increase efficiency of SR has become widespread [22]. When the stream data is not stable, utilizing ML techniques to process large volumes of data and to extract patterns from data streams can increase streaming model accuracy [12]. As noted in Section 1, ML–SR integration gave rise to SML. The learning paradigms of SML are detailed in Section 5.1, and their relation to SR challenges is summarized in Table 3. In addition to ML, Streaming Linked Data, which is considered among SR approaches, also plays a role in improving SR processes.

The subsequent sections will elaborate on the specific features of SR that lead to various challenges, as well as the ways in which these challenges can be addressed through the integration of KGs.

3 Research methods

Within the scope of this study, we conducted a detailed review of the methods presented in the existing literature. This process took into account the research criteria commonly used in systematic reviews and was inspired by the methodology of [23]. The aim is to better understand the challenges associated with stream reasoning (SR), how these challenges have been addressed by different approaches, and the specific contexts in which they arise. In addition, this review seeks to identify the limitations of current methods and potential opportunities for future research in the field. Following the approach of [23], we define a set of methodological research questions in this section to guide our subsequent analysis. We also describe the search process, along with the inclusion and exclusion criteria, as illustrated in Figure 1.

3.1 Research questions

This review study seeks answers to the following main questions:

- RQ1: What are the primary challenges of SR and their causing factors, linked to the intrinsic nature and characteristics of streaming data?
- RQ2: What are the approaches in the literature that address one or more of these primary challenges in SR?
- RQ3: How does the integration of SR and KGs alleviate challenges in SR?

In the background section, we discussed the importance and relationship between SR and KGs. Then, in this section, we introduced the research questions that will guide our literature

review. The purpose of question RQ1 is to identify which are the primary challenges that arise because of the intrinsic characteristics of streaming data, and investigate their causes. In RQ2, the goal is to understand how these challenges are evaluated in literature approaches over various research fields. While identifying which method addresses which SR challenge, we also aim to reveal additional open issues that may require further investigation. In the last research question RQ3, we target to see whether integrated approaches that combine KGs and SR in the literature can effectively tackle stream processing issues. These research questions have been designed to guide our literature search and structure the review process, thus allowing us to identify directions for improving SR methods from different perspectives—*i.e.,*ML, Linked Data and KGs.

3.2 Search process

The selection of review papers was guided by the criteria illustrated in Figure 1. We focused on peer-reviewed conference and journal articles published in English within the last 15 years, and accessible through major databases such as DBLP, Web of Science, and Google Scholar. The studies that did not match the selected keywords, were duplicates or redundant, exceeded the 15-year threshold, or were not fully written in English were excluded from the review.

At the end of the evaluation steps by considering inclusion and exclusion criteria, 152 main articles were included in our review study. The papers in this review study were selected by considering methodological terms in SR, SML, SLD and SKG. The reason behind the evaluation of SR approaches over these different search domains is to present state-of-the-art methods and techniques which meet different streaming challenges. Therefore, it would be easy to see current issues in SR, and to propose innovative solutions by following developments in recent years. It is also necessary to clarify why we chose the research keywords related to these fields. In this context, ML methods bring the advantages of applying data-centric methods to incoming streams, potentially providing real-time inductive reasoning through different learning strategies. It also offers methods to reduce heterogeneity, high volume and especially dynamicity problems caused by the characteristics of data streams. On the other hand, Linked Data methods structure data streams according to ontology models, which are useful for better understanding relationships between stream data elements and their data sources. As a consequence they help to cope with SR challenges arising from data heterogeneity and dynamicity. Moreover, KGs—which have attracted attention in recent years—offer opportunities for SR enhancement. Semantic enrichment of large amounts of data structured as KGs can be utilized by streaming approaches for diminishing the challenges in reasoning and interpreting dynamic and heterogeneous stream data.

After the final selection of related papers by using the proposed search process, we analyzed them according to the research questions presented above. At the end of this research study, we expected to answer to the following sub-questions: *What are the main challenges in SR*; *What are the factors leading to these challenges*; *Which SR based approaches are addressing the challenges*; *Which types of methodology meets with which challenges*; *Which research areas are involving all the challenges*; *How open challenges arise in this scope*; *Which type of approaches tackle the open challenges*; *What are the advantages and disadvantages of the evaluated approaches*; *What are the main evaluation criteria of the approaches*; *What are the future directions in this area*.

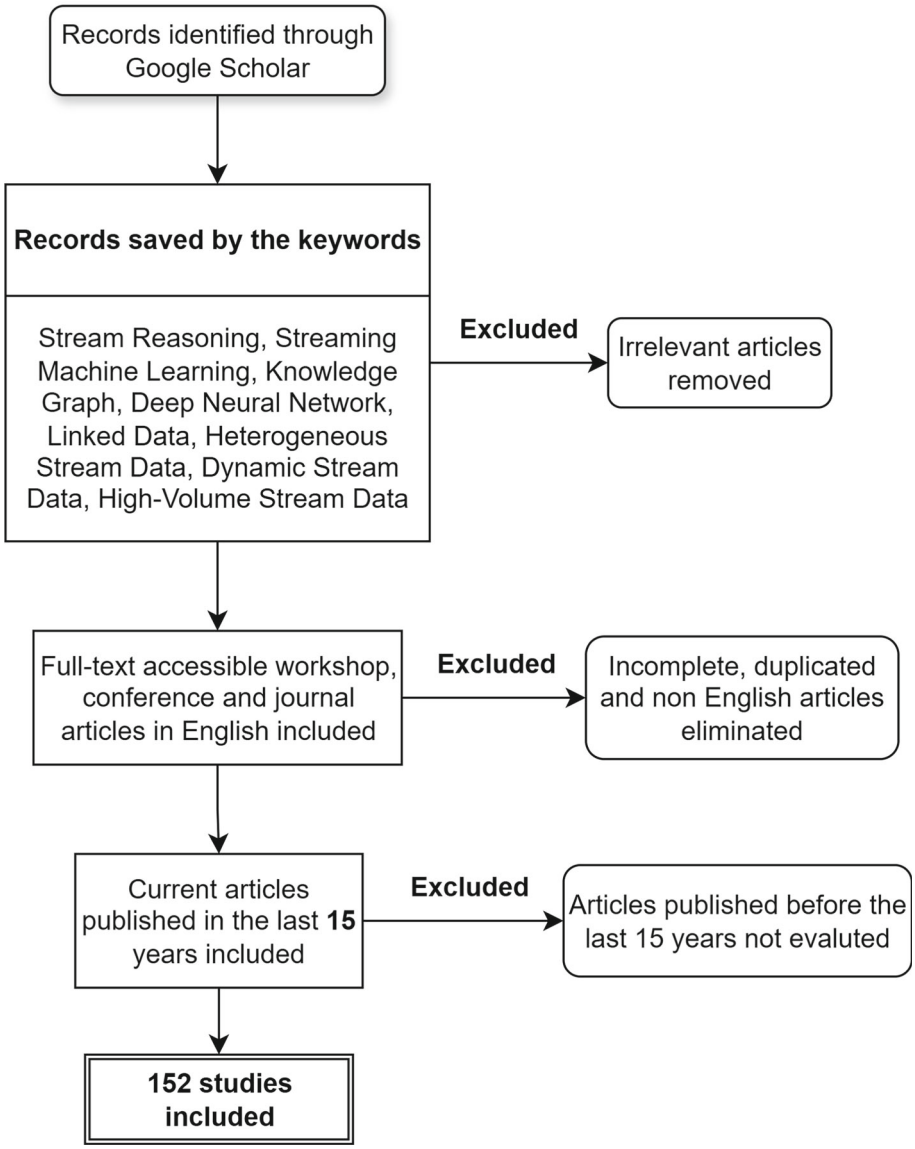

Fig. 1 Schematic view of the systematic literature search

3.3 Related survey comparison

Across recent literature, the fields of SR, dynamic KGs, and SML have each been extensively surveyed. However, existing reviews tend to treat these topics in isolation. There is still no clear approach that combines semantic reasoning with adaptive learning over heterogeneous and evolving, high volume data streams. Yet in real-world settings, such integration is increasingly needed to reason and learn from complex, streaming data. This motivates a clearer positioning of our work in relation to existing literature. To situate this survey within

the existing literature, Table 1 offers a structured comparison of nine recent and representative surveys, selected based on their thematic relevance and citation impact. Each survey is analyzed based on four key aspects: (i) surveys' primary research objective and technical scope, (ii) the extent to which it explicitly addresses three fundamental challenges–data heterogeneity (H), temporal dynamicity (D), and high-volume evaluation (V), (iii) the level of integration between semantic KG structures and SR or ML components, categorized as none, partial, or full, and (iv) notable limitations of the survey, especially in contrast to the aspects that are comprehensively addressed in our work.
As summarized in the survey matrix, the majority of prior surveys exhibit one or more of the following gaps:

- Most existing surveys focus exclusively on either SR or ML, with limited attention to their integration within KGs centric frameworks.
- Few studies adequately address the necessity of reasoning and learning over heterogeneous and temporally evolving semantic data, where schemas and formats vary across sources and time.
- Limited attention has been given to real-time, high-throughput streaming environments, where both reasoning and learning must function under stringent temporal and scalability constraints.

However, our work uniquely targets the intersection of SR, dynamic KGs, and ML approaches to provide a unified and scalable framework. The motivation for this integration is driven by the demands of real-world applications such as IoT, healthcare, and smart-city systems, where data are heterogeneous, evolves continuously, and must be interpreted semantically while adapting to concept drift. Unlike previous reviews, our work explicitly targets all three challenges (heterogeneity, dynamicity, high-volume) and promotes ML–SR integration as a central requirement for scalable, context-aware reasoning.

While Table 1 contrasts surveys thematically, a more technical evaluation is provided in Table 4, which synthesizes metrics and methodologies in across benchmark studies. This complementary view further highlights the gaps in current evaluation practices and motivates the integrated criteria proposed in our work.

4 Challenges in stream reasoning

In this section, we highlight the primary challenges associated with SR. Our focus particularly centers on the issues of *heterogeneity*, *dynamicity*, and *high-volume*, given their widespread occurrence in the literature [32]. Addressing these challenges enables the development and optimization of efficient SR algorithms and systems [32]. To this end, we begin by providing definitions of *heterogeneity*, *dynamicity*, and *high-volume*, explaining their characteristics and factors through the exploration of our research question RQ1. We then present prior literature and methodologies for identifying paths of improvement.

4.1 Heterogeneity

Data streams consumed by stream reasoners comprise various types of observations, events, and instances continuously generated by multiple sources [7, 8, 33]. These stream items may encompass textual and numerical data, JSON objects from web-based applications, or measurements from IoT devices [4, 34, 35]. This diversity in incoming data streams introduces aspects of *heterogeneity* [36].

Table 1 Comparative survey matrix positioning our work within related literature

Survey(Year–Title–Citation)	Focus & Scope	H	D	V	KG–SR Integration Level	Limitations
2017–*Stream Reasoning: A Survey and Outlook* [24]	Foundational overview of SR concepts, engines and query models	✓	✓		Partial-Ontology semantics; no dynamic KG updates	ML perspective; high-volume evaluation
2019–*Large-Scale Semantic Integration of Linked Data* [25]	Techniques for integrating heterogeneous Linked-Data sets into unified graphs	✓		✓	Partial-Static RDF graphs; no streaming updates	Streaming reasoning; dynamic KG updates; ML integration
2023–*Logical Rule Based Knowledge Graph Reasoning* [26]	Comprehensive analysis of rule-based reasoning techniques for KGs	✓	✓		Full-Dynamic KG rule reasoning; no stream engines	Streaming reasoning; high-volume evaluation; ML integration
2024–*A Survey on the Evolution of Stream Processing Systems* [27]	Architectural evolution of large-scale stream-processing engines		✓	✓	None-KG not addressed	KG semantics; data heterogeneity; ML integration
2024–*Grounding Stream Reasoning Research* [28]	Research agenda and systematic overview of open challenges in SR	✓	✓		Full-Conceptual SR with KG semantics; no ML layer	high-volume benchmarking; machine-learning integration
2025–*Supervised Learning from Data Streams* [29]	Overview of supervised learning and concept-drift techniques for data streams		✓	✓	None- exclusively ML; KG dimension absent	KG semantics; heterogeneous data formats
2025–*Knowledge Graph Evolution: Proliferation, Dynamic Embedding & Versioning* [30]	Methods for evolving, versioning and embedding dynamic KGs	✓	✓		Partial-Dynamic KG evolution; no SR	High-volume streaming; SR+ML integration
2025–*Neural Symbolic Methods for Knowledge Graph Reasoning* [31]	Survey of neural-symbolic techniques for KG reasoning	✓			Partial-Neural-symbolic KG reasoning; no SR layer	SR; high-volume stream data evaluation
2025–Our survey: *A Comprehensive Survey of SR and Its Integration with KGs*	**Integrated review combining SR, dynamic KGs and SML**	✓	✓	✓	**Full-KG schema, live updates, ML for concept drift**	—

Column keys: H=heterogeneity; D=dynamicity; V=high-volume; ✓=explicitly addressed
Limitations=aspects not covered in the respective survey but addressed in our work

Combining stream data from various sources on a single platform facilitates a more comprehensive and insightful analysis within the model [37]. This amplifies stream data correlation, aids in pattern identification, and improves information accessibility [38]. Nevertheless, integrating this diverse data and ensuring system compatibility can be complex due to variations in data formats, protocols, and interfaces [39]. Therefore, integrating heterogeneous data presents significant challenges, especially on large streaming data platforms, often leading to delays. Implementing a distributed architecture is crucial to streamline the merging of these diverse data formats across the system [38]. This integration problem is even more challenging while processing complex and heterogeneous multimedia streaming data due to the high computational cost incurred in the pre-processing steps [40].

Integrating continuously incoming diverse stream elements necessitates addressing heterogeneity at the stream sources initially. Beginning with syntactic heterogeneity, such as varying structures like JSON or relational-based streams, several approaches have focused on transforming different formats into structure of KGs like RDF [41]. Mapping-based techniques, including RML or R2RML languages, have been proposed over the years, enabling the creation of materialized [42–44] and virtual views [45, 46] of streams over semantic models. While materialization may streamline integration, its practicality hinges on the stream's velocity or consumption method. Virtualization holds promise in optimizing transformation or leveraging query rewriting to delegate most processing to the stream sources.

Modeling heterogeneous stream data poses significant challenges due to the complexity of expressing such data in a semantically rich and coherent format [47]. Accommodating these differences using a single model or ontology is difficult since data from various sources are articulated using different concepts, often with partial overlaps [37, 48]. This scenario presents obstacles in maintaining semantic integrity and establishing accurate relationships between data during modeling [49]. Even with the utilization of a common semantic schema to represent heterogeneous data stream elements, widely adopted KGs frameworks like RDF have demonstrated limitations in modeling capabilities. The inherent temporal nature of RDF prompted the development of extensions enabling the representation of time-annotated graphs, incorporating point-in-time and interval-based semantics [50–52]. These extensions serve as the foundation for continuous query processing for SKG, as elaborated in Section 5.2.

The characteristics of heterogeneous stream data can pose challenges in effectively processing data and deriving meaningful results during real-time reasoning [53]. Reasoning over heterogeneous stream data remains a challenging task due to the presence of semantic and syntactic differences [7, 54]. Esposito *et al.* [55] delve into heterogeneity and complex processing issues in big data analytics, emphasizing the importance of better semantic representation for data integrity. Furthermore, analyzing heterogeneous data at varying speeds and volumes may necessitate further optimization and incur high distributed computing costs in dedicated streaming infrastructures [55, 56]. Real-time SR complexity is significantly influenced by the heterogeneous nature of stream data, as adaptation to diverse expressiveness levels is often required [57]. The heterogeneous data at different speeds and volumes can make it hard to meet modeling need in terms of scalability and performance requirements [58, 59]. For dense and heterogeneous data streaming, some data points can be missed, thus affecting the accuracy and reliability of the model [53]. Overcoming these challenges involves representing data in a richer form by amalgamating heterogeneous information from different sources, while ensuring overall coherence [60]. Additionally, enhancing reasoning performance over complex data arising from heterogeneous streaming data is crucial [8].

In addressing the challenge of heterogeneity, various works in the literature have made significant contributions. For instance, Corral Plaza *et al.* [4], Su *et al.* [33], and Peng *et al.* [61] have targeted distinct objectives related to managing heterogeneous stream data. While the

general difficulty of handling heterogeneous stream data is discussed in [32, 62], Corral Plaza *et al.* [4] and Su *et al.* [33] specifically focus on heterogeneity within the IoT domain, particularly related to device diversity. Peng *et al.* [61] aim at integrating heterogeneous data in dynamic environments, while Brewka *et al.* [63] address the topic of multi-layer heterogeneous data integration in health and home settings.

4.2 Dynamicity

In addition to the challenges posed by the diverse structure of streaming data, we also explore the challenges arising from its constant changes. This continuous flow of changes, known as *dynamicity*, [64] is influenced by factors such as the source or environment of the streams. The primary challenge with dynamic streaming data is its rapid and high-volume nature in various formats and structures [65, 66]. Furthermore, the data structure may change during streaming, making it challenging to process quickly during SR. Data streams consumed by stream reasoners consist of a combination of different types of observations, events and instances continuously produced by various sources [7, 8]. This continuous stream requires instant updates, especially in high-speed scenarios where unexpected changes occur [56].

Because of dynamicity, maintaining consistent modeling and reasoning processes in real-time requires rapid adaptation, making standard approaches impractical. Extracting high-level knowledge from time-annotated dynamic data for decision-making also presents a significant challenge. Additionally, during the reasoning process, data quality directly impacts performance, making it crucial to manage input stream inconsistencies and noise [67]. Improving data quality helps to prevent decisions based on inaccurate or incomplete information [68]. To tackle these challenges, it's important to swiftly detect instant changes and anomalies to adapt to existing patterns effectively [69]. Also, dynamic systems must be able to recognize, merge, and understand diverse, dispersed data streams to automatically analyze incoming data [32]. Similarly, Anicic *et al.* [70] propose a new language model for SR and to bolster semantic web tools, which may struggle to process rapidly changing complex data provided by event streams.

The constantly changing nature of stream data, occurring at different speeds and volumes, poses challenges in analysis, especially with unreliable data sources [1]. This issue is particularly prominent in common use cases like IoT or the Web of Things, where noise, incomplete, and erroneous information from different sources may compromise the reliability of analyses [3, 53]. This unreliability of data during SR can hinder the detection of abnormal situations in industrial processes [71]. Moreover, effective resource management becomes imperative in the face of instantly changing data dynamics [72]. Dynamic SR demands a high level of computational power to process continuous and variable data streams in real-time [7]. Limited computing resources may slow down data processing and analysis, thus affecting efficiency and model performance [66, 73].

The challenges posed by the dynamism of stream data have been extensively discussed in various methodological and review studies, which examine SR from different perspectives. Notably, this dynamic nature is not solely defined by the intrinsic velocity of data sources, but also by continuous changes in data quality, update frequency, source provenance, and usage context [1].

4.3 High volume

Another major challenge in stream data processing is dealing with continuous and dense stream data at high speed and in large quantities [74]. This type of large-scale streaming data is often termed as *high-volume* stream data and is crucial in real-time applications [75]. There are several factors contributing to the challenges of processing high-volume stream data. The main issues include the need for real-time processing, high computation costs, and managing the rapid growth of data at high speed [21, 76].

Managing high-volume stream data presents several challenges stemming from both infrastructural and algorithmic limitations. A key issue is the difficulty of real-time processing, often due to inadequate system capacity to ensure accurate SR [77]. Additionally, existing algorithms may lack the efficiency to process continuous, high-speed data within limited time constraints [78]. The dynamic and complex nature of streaming data can further hinder model accuracy and reliability [32]. Finally, high computational costs represent a significant barrier to effective high-volume stream data management. This problem is often worsened by limited memory resources [21].

When the memory required to process high-volume data exceeds the limits of conventional data processing systems, computational constraints become evident [79]. In such cases, instant analysis and the extraction of meaningful insights become increasingly difficult. Effective management of these limitations is crucial for achieving accurate analysis and for minimizing redundant computations through efficient resource allocation [80]. Moreover, the development of optimized learning algorithms for processing large-scale streaming data remains a critical challenge [73]. The high speed and dimensionality of the data further complicate anomaly detection within the strict time constraints of real-time environments [81, 82]. Even big data tools often struggle to manage, monitor, and process these continuously growing data streams within acceptable latency thresholds [53]. Collectively, these factors underscore the complexity of handling high-volume streaming data.

To handle high-volume streaming data, various algorithms and approaches have been developed, supported by appropriate technological infrastructures and resources. In pursuit of this goal, researchers have investigated and addressed numerous challenges through reviews and diverse approaches. For example, studies such as [12, 22] focus on the challenges faced in managing high volumes of streaming data, identifying open challenges in specific areas such as SML. Similarly, in another study, [33] explores all the challenges highlighted in our study (including heterogeneous, dynamic, and high-volume stream data management), aiming to extract high-level information from various IoT devices. Furthermore, challenges arising during the reasoning and processing of high-volume stream data are tackled across different application domains, offering solutions or exploring these challenges further.

Table 2 outlines the three principal challenges in SR (heterogeneity, dynamicity, and high-volume) as identified through recurring patterns in the literature. These challenges reflect systemic obstacles in processing and reasoning over streaming data in real time, and are critical to the design of scalable and adaptive SR systems. For each challenge, the table identifies a set of contributing factors that underlie its emergence. These factors arise from the intrinsic properties of streaming data, such as diverse formats, temporal instability, and scale. They are also shaped by the limitations of current infrastructures and modeling frameworks. The "Causal Relationship" column provides a structured explanation of how each factor leads to the corresponding challenge. These links are not merely descriptive but reveal the underlying mechanisms through which stream data properties manifest as reasoning bottlenecks. For example, the lack of a unified data representation complicates semantic integration and this leads directly to the challenge of heterogeneity. Similarly, unstable or rapidly changing

Table 2 Core Stream Reasoning challenges, their causal factors, the relationship between each factor and challenge, and some representative studies that relate to these challenges

Challenges	Causal Factors	Causal Relationship	Reference Studies
Heterogeneity	Streaming data integration	higher latency	stream processing, schema mapping, semantic integration [37–41]
	Stream data modeling	schema alignment	ontology-based modeling, RDF transformation [38, 46–49]
	Real-time SR	costly mappings	query processing, low latency analytics, semantic querying [4, 33, 53, 57, 60]
	Diversity in data format	no unified representation	heterogeneous format handling, RDF mapping, semantic representation [8, 37, 39, 41, 47, 62]
	Data quality	reprocessing overhead	quality assessment, anomaly detection, error handling [53, 67–69]
Dynamicity	Adaptation of reasoning model	distribution shift	concept drift adaptation, incremental updating, continuous learning [32, 64, 69, 70]
	Instant updates	increased latency and retraining cost	real-time updates, dynamic query processing, incremental reasoning [54, 64–66, 69]
	Continuous process	continual adaptation need	stream continuity management, adaptive processing, real-time inference [7, 8, 64, 70]
	Data quality	noisy data	stream data quality, error reduction, robustness [3, 53, 67]
High-Volume	Real-time processing	latency/memory constraints	low latency processing, high-performance analytics, streaming scalability [21, 75–77, 81, 82]
	High-speed unbounded data	scalability issues	big data analytics, streaming frameworks, scalable processing [21, 74, 76, 78, 81, 82]
	High computational needs	high throughput	computational efficiency, optimization techniques, high stream management [21, 33, 73, 79, 80]

input distributions result in model degradation, which characterizes the challenge of dynamicity. High throughput requirements and memory limitations increase computational loads, leading to the high-volume problem. On the other hand, the "Reference Studies" column includes selected references that discuss these challenge-related factors in applied or theoretical contexts. These works do not necessarily offer complete solutions but are relevant in how they expose, examine, or mitigate aspects of the challenges. For example, studies on schema alignment and RDF transformation relate to the modeling of heterogeneous streams; others explore concept drift adaptation or propose scalable data processing techniques. While diverse in scope and methodology, these studies share a thematic connection to the structural origins of the SR challenges. To sum up, the table provides a comprehensive conceptual structure for understanding the foundational difficulties in SR and how they relate to various operational and representational constraints. It also serves as a bridge to the next section (Section 5), where we examine how different methodological approaches address these challenges through specific technologies and paradigms.

In brief, the reasoning of heterogeneous, dynamic, and high-volume streaming data presents various challenges, including different data formats, semantic incompatibilities, synchronization issues, and scalability concerns. To address these challenges effectively necessitates the application of advanced data processing algorithms and techniques. Given this context, it becomes crucial to thoroughly examine and analyze the approaches developed using diverse techniques and algorithms in the literature. This step clarifies key challenges in streaming data, while helping to design effective solutions. Each approach is reviewed to find strengths and weaknesses, thus allowing new methods to emerge and existing ones to improve. The next section addresses these goals as part of RQ2.

5 Existing approaches addressing challenges in stream reasoning

SR methods allow for real-time data processing, semantic inference, instant data integration, and adaptive modeling to tackle the challenges of diverse, constantly changing, and high-volume data streams. In this section, we highlight advanced SR approaches and identify the previously discussed challenges. We offer insights into addressing RQ2, which focuses on how these approaches manage high-volume, dynamic, and heterogeneous data streams. These methodologies are grouped into three categories based on their methodological or technological approach:

- *SML* is predominantly utilized to address the challenges associated with dynamic and high-volume streaming data, although it touches on the challenge of heterogeneous data to some extent [83].
- *SLD* approaches primarily focus on handling heterogeneity and data integration, ensuring continuous updates while preserving semantic relationships [84].
- *SKG* approaches offer effective solutions for enhancing real-time data comprehension in dynamic environments, while allowing for flexible management of large volume stream data and the heterogeneous structure and complex relationships of this data [85].

We show how these SR approaches relate to each other and explain how they handle the challenges above presented. We conclude the section by highlighting the ongoing challenges posed by the models, setting the stage for RQ3. To support this analysis with a conceptual overview, Figure 2 synthesizes the methodological landscape and challenge coverage across the main categories.

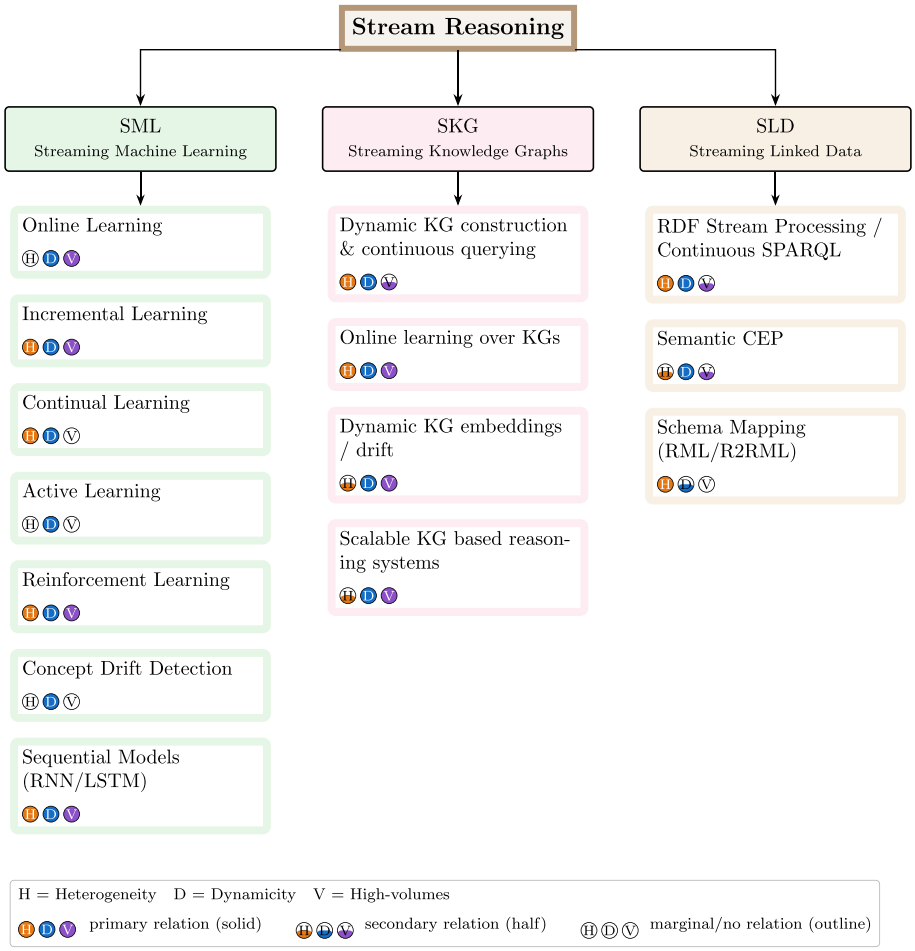

Fig. 2 Taxonomy of stream reasoning approaches, organized into three main categories (SML, SLD, and SKG), each comprising representative subtypes. H/D/V markers indicate the extent to which each sub-approach addresses the key challenges of heterogeneity, dynamicity, and data volume: a solid circle denotes strong coverage, a half-filled circle indicates partial relevance, and an empty circle signifies minimal or no direct focus. This structure facilitates a clearer understanding of the relationships between methodological classes and the challenges they aim to address

Figure 2 provides a taxonomy of stream reasoning approaches, structured along three major methodological paradigms: SML, SLD, and SKG. Each category includes representative sub-approaches that reflect the diversity of techniques proposed in the literature. The figure also integrates three key system-level challenges defined in this study (heterogeneity, dynamicity, and volume) by indicating, for each sub-approach, the degree to which it addresses these aspects. This classification is designed to serve as a conceptual reference framework that guides the discussion in the remainder of the survey to help organize the broad and complex solution space covered. Among the categories, SML approaches exhibit particularly high methodological diversity and structural depth by combining techniques such as online, continual, or drift-aware learning. While Figure 2 offers a high-level conceptual

view, Table 3 complements it by providing a more granular and literature-grounded breakdown of the various SML methods, explicitly mapping them to the corresponding challenges they address.

5.1 Streaming machine learning

SML is based on ML techniques that incorporate adaptive models and real-time updates, which adjust to dynamic structures for processing streaming data [22]. Several evolving algorithms and tools in ML, along with streaming algorithms and frameworks, are designed specifically to handle large data stream volumes [12]. Thanks to their flexibility and adaptability, they can interpret and process different data types and structures, effectively tackling challenges arising from data heterogeneity [86].

We categorize the current technologies or methods provided by ML based on their underlying learning approaches. Additionally, we incorporate existing works based on the modeling strategies, encompassing hybrid algorithmic dependency models and KGs.

Online learning enables the model to quickly adapt to dynamic and changing environmental conditions on constantly incoming data streams, so it can produce accurate and up-to-date results in real time [87]. Due to the high dynamicity the models have difficulty adapting to new incoming data, which may cause model performance to decrease, a phenomenon known as Concept Drift [88]. To deal with the dynamic structure of streaming data by addressing concept drift, [89] proposes an online semi-supervised method with a group of micro-clusters. Another study related with Online Learning [90] targets the challenge of high-volume with high speed in big data streams to address limitations of batch learning methods.

Incremental Learning allows new incoming data to be continuously integrated into the stream and reliably updates the model over time by adapting to changing data distributions [91]. This learning presents cumulative learning and decision-making capabilities with adaptation the speed, size, and variability of large scale data for in-depth understanding of stream sequences [92]. Owing to the opportunities brought by Incremental Learning, targeted SR challenges can be improved. In this line, the study in [93] offers a solution to cope with imbalanced data streams that occur because of non-stationary dynamic environment. Nikpour and Asadi [94] propose a dynamic hierarchical incremental learning-based supervised clustering method for data streams, addressing challenges related to concept drift, high speed and large volumes. Abdallah *et al.* [95] provide a comprehensive review of real-time activity recognition techniques over evolving data streams, emphasizing the role of SR in heterogeneous sensor environments. Le-Phuoc *et al.* [96] introduce a scalable neural-symbolic stream reasoning approach that combines learning and incremental reasoning to improve the performance of real-time monitoring systems under dynamic streaming conditions.

Continual Learning enables models to retain previously acquired knowledge while adapting to new, incoming data. It improves performance by handling continuous updates in data streams and adjusting to changing conditions over time [97]. In the context of SR, Criado *et al.*[98] apply Continual Learning strategies within a federated learning framework to classify heterogeneous and evolving data. To address the real-time and dynamic nature of streaming environments, Ashfahani *et al.*[99] propose Online Continual Learning algorithms that make efficient use of computational resources and mitigate catastrophic forgetting [100].

Active Learning facilitates the selection of highly representative data samples and plays a critical role in semi-supervised learning by reducing labeling costs in dynamic streaming environments [101]. To support this objective, Ienco *et al.* [102] propose an Active Clustering Learning approach for data streams. This method uses a pre-clustering mechanism

to identify the most informative instances, helping to mitigate the effects of concept drift caused by stream dynamics. Similarly, Žliobaitė *et al.* [103] introduce three Active Learning strategies aimed at selecting the most uncertain samples for real-time concept drift detection. Wassermann *et al.* [104] extend earlier work by introducing a stream-based Active Learning approach. It applies Reinforcement Learning to decide the most suitable time for querying the oracle, improving adaptation to concept drift while minimizing labeling cost.

Reinforcement Learning has been applied alongside Active Learning in supervised modeling to address the challenge of dynamicity in streaming data [104]. Reinforcement Learning enables agents to learn optimal actions by interacting with their environment and using reward-based feedback [105]. Dodaro *et al.* [72] improve SR performance under high-throughput conditions by using Reinforcement Learning to define learned constraints and enhance cache management. Russo *et al.* [106] employ Reinforcement Learning algorithms to manage high-volume streaming data from heterogeneous sources. In addition, attention-based deep Reinforcement Learning methods are used to enhance SR over large-scale, complex KGs by learning dynamic relationships between paths [107].

Concept Drift Detection Methods are specifically geared towards managing and understanding the dynamic nature of data streams rather than the adaptations of learning models [108]. Therefore, they solely address the challenge of dynamicity in stream data by detecting changes (drifts) over time that affect the performance of learning models. In contrast to learning strategies, concept drift detection methods focus more on monitoring and adapting to changes in data distribution [109]. There are three common approaches. Statistical methods track changes in data distribution [16]. Window-based methods compare recent data with earlier segments within defined time intervals. ML-based methods monitor model performance in real time to identify drift [110, 111].

Sequential Models are essential for handling data with time dependencies and tracking temporal changes in stream [112]. They offer flexibility in processing dynamic, large volume of stream data from different sources in dynamic environments. For example, [113] overcomes complex reasoning over time-annotated high-volume and speed stream data by sequential Recurrent Neural Network (RNN) with Answer Set Programming. [114] benefits from RNN and CNN based time series modeling capabilities for improved reasoning with C-SPARQL to address real-time high-frequency stream data. On the other hand, [115] develops Long-Short-Term memory (LSTM) based multi-task learning system to tackle with challenges in processing of heterogeneous, sparse, complex stream data coming from user activity sensors.

Table 3 presents ML-based approaches that address key challenges in SR, particularly those related to heterogeneity, dynamicity, and high data volume. Among these, Incremental Learning and Sequential Model approaches have shown to tackle all three core challenges. The table also maps each SML method to the corresponding SR challenge and its supporting literature. Dynamicity appears to be the most commonly addressed challenge across all SML methods. Overall, Incremental Learning, Reinforcement Learning, and Sequential Models demonstrate broader applicability by covering all identified challenges. However, the most appropriate method may vary depending on the data type, application context, and current technological capabilities.

5.2 Streaming linked data

SLD focuses on methods that improve the integration and utility of real-time data from diverse web sources by using semantic queries and continuously updated streams [84, 118].

Table 3 Group of SML Based Studies in Literature According to Their Relation with the Highlighted Main Challenges

Challenge	SML Approaches	Studies
Heterogeneity	Incremental learning	[116]
	Continual learning	[98]
	Reinforcement Learning	[106, 117]
	Sequential Models	[115]
Dynamicity	Online Learning	[87, 88]
	Incremental learning	[91, 93, 96]
	Continual learning	[98, 99]
	Active Learning	[102–104]
	Reinforcement Learning	[72, 104]
	Concept drift detection	[109–111]
	Sequential Models	[114]
High-Volume	Online Learning	[90]
	Incremental learning	[92, 94]
	Sequential Models	[113]
	Reinforcement Learning	[106, 107]

In particular, Linked Data approaches view the web as a global data space by using URIs for unique identification, HTTP for access and dereferencing, and RDF for data modeling and representation [118].

To overcome these limitations, the literature proposes streaming extensions to Linked Data, enabling the processing of stream data and logical inferences [119]. These technologies collectively address challenges posed by high-volume, dynamic, and heterogeneous stream data across various learning paradigms. For instance, [120] employs RDF and Semantic Web technologies for real-time representation, integration, and querying of heterogeneous stream data in the big data domain. Similarly, Calbimonte *et al.* [121] employ RDF Stream Processing (RSP) engines and propose extensions to SPARQL that support query rewriting and continuous querying over heterogeneous RDF data streams. These enhancements aim to improve the efficiency and scalability of stream data integration and reasoning.

Furthermore, semantic web technologies such as Complex Event Processing (CEP) and RSP are employed to detect and extract meaningful patterns from data streams [51]. CEP is used to identify semantically complex events in heterogeneous streaming data, while RSP integrates, analyzes, and reasons over RDF streams from diverse sources [9].

Within the framework of Semantic Web technologies and SR, heterogeneous data streams from IoT devices can be instantly harmonized with ontological models, enabling semantic integration and real-time decision-making [54]. In this context, semantic reasoning techniques are explored to support responsive IoT applications, with a focus on context interpretation and semantic information extraction [122].

Maarala *et al.* [122] compare the scalability of various distributed reasoners using semantic methods in real-time settings. Similarly, Abeykoon *et al.* [123] integrate static knowledge from Semantic Web ontologies with dynamic data from IoT devices to monitor patients' health conditions in real time, demonstrating the enhanced utility of semantic enrichment in SR.

To process high-volume semantic data during reasoning, Oren *et al.* [124] propose the use of RDF engines based on a divide-conquer-swap parallel computing strategy. Additionally,

Della Valle *et al.* [62] and Margara *et al.* [7] explore real-time data integration and dynamic semantic interpretation for continuous stream data using Semantic Web technologies.

5.3 Streaming knowledge graphs

KGs effectively organize and represent information, making it applicable across a wide range of SR applications [125]. Some KGs are specifically designed to manage continuously updated data streams, offering flexible structures and embeddings that support real-time integration, analysis, and decision-making [126, 127]. The ability of dynamic KGs to process real-time updates accelerates high-volume data processing by continuously reflecting the latest information, thereby supporting timely and informed decisions [128].

For example, Omran *et al.* [10] proposed an approach to learn temporal rules from dynamic KG streams using the StreamLearner system. Similarly, Barry *et al.* [128] examined graph-based online learning for high-velocity data streams in large-scale networks. To manage big data, Bellomarini *et al.* [129] introduced a KG management system with advanced reasoning capabilities, designed to handle the scalability of streaming data with efficient computational complexity. Furthermore, SR models can integrate with heterogeneous data sources, while KGs effectively assimilate information from diverse formats and origins [130].

Due to the dynamic nature of stream data, SR models must rapidly adapt to evolving inputs, which often leads to concept drift [88]. To address this, Chen *et al.* [131] propose an enhanced drift detection method that leverages KGs embeddings and semantic components. Among the related studies, Barry *et al.* [128] is particularly notable, as it addresses all three core challenges examined in this paper–namely, heterogeneity, dynamicity, and high-volume stream data. Their graph-based online learning approach enables real-time reasoning over fast and large-scale data while effectively handling structural and temporal variability.

5.4 Open challenges

We investigate SR approaches developed to address the challenges posed by heterogeneous, dynamic, and high-volume data streams. Despite significant progress, persistent problems remain. In particular, efforts to tackle these fundamental issues have inadvertently led to the emergence of additional open challenges, which remain unresolved in the literature. While this study does not explore these in detail, we introduce them by highlighting their origins and related approaches discussed in previous work.

Dynamic data stream management is a key driver behind many of these open challenges. As models continuously process real-time dynamic data, they may overfit to recent inputs while forgetting earlier information, i.e., exhibiting *Catastrophic Forgetting* [100]. Techniques such as Elastic Weight Consolidation (EWC) and Gradient Episodic Memory (GEM) have been proposed to address this issue, but their applicability to real-world scenarios remains limited [132]. Another major challenge is *Concept Drift*, where statistical properties of streaming data evolve over time, degrading the performance of reasoning models [13, 88]. Despite the development of numerous detection techniques, this issue persists due to complexities such as drift heterogeneity, high false detection rates, class imbalance, and increased computational burden [133].

Scalability also remains a critical open problem, especially when processing high-volume streaming data. This is often caused by system-level limitations and the complexity of heterogeneous semantic representations [134, 135]. Scalability refers to a system's ability to manage growing volumes of complex data without compromising performance [136]. Closely tied to

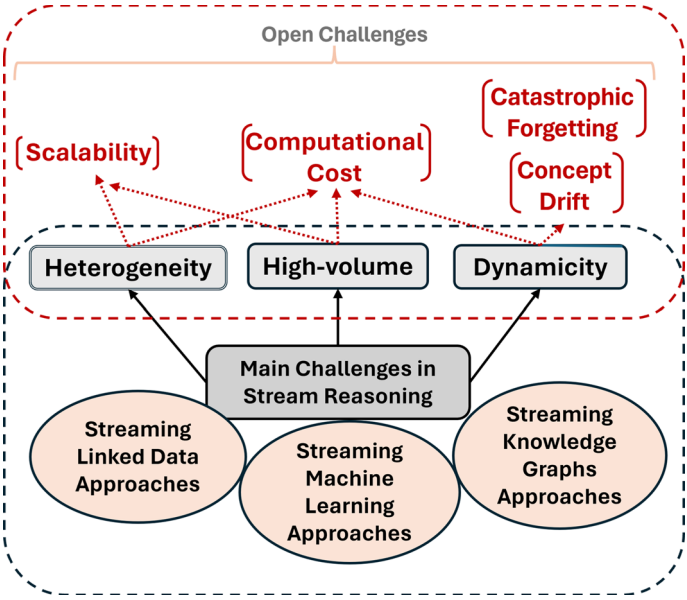

Fig. 3 Hierarchical view of some open challenges arising from main challenges (heterogeneity, high-volume, dynamicity) which are addressed by various approaches

scalability is the challenge of *high computational cost*, which varies depending on the complexity of the models and algorithms used to process evolving data streams [137]. Even with distributed processing architectures, unpredictable data patterns can still lead to increased resource consumption [38].

Figure 3 illustrates the hierarchical relationship between the main SR challenges discussed in this study and the open challenges that emerge as a consequence. The figure is constructed in a bottom-up manner to reflect the structure of the analysis. It first shows how approaches within categories such as *SML*, *SLD*, and *SKG* address the primary SR challenges, namely heterogeneity, dynamicity, and high volume. It then illustrates how these challenges give rise to open issues, including scalability, computational cost, concept drift, and catastrophic forgetting. The arrows depict the links between specific main challenges and the open issues they contribute to.

In summary, while existing SR approaches provide useful solutions to core challenges, they also uncover new problems that require further research. As real-time systems continue to deal with fast, heterogeneous, and large-scale data, developing more adaptive, scalable, and efficient methods remains essential for advancing the field. The next section (Section 6) analyzes how integrating SR with KGs can enhance stream-data management and answers RQ3.

6 Stream reasoning and knowledge graphs integration

In this section, we discuss SR–KGs integration for improving streaming data management. First, we outline the benefits of KGs for SR; then, we examine how SR exploits KG-based

reasoning. We analyze how the integration mitigates heterogeneity, dynamicity, and high-volume constraints, and illustrate this with representative use cases from the literature.

Current streaming models have issues while reasoning over rapidly changing knowledge because of the arising challenges in a dynamic environment. For instance, Dynamic KGs, which have a structure that can adapt to constantly developing and changing data, can be used to improve reasoning processes [126]. We introduced this improvement in the previous section by sharing how KGs based approaches address the basic difficulties encountered during SR through a few approaches from the literature.

KGs offer a flexible and scalable solution for managing constantly evolving stream data in dynamic environments [7]. By enforcing logical relationships and semantic context, KGs enable SR, which facilitates data integration and supports real-time decision-making across diverse applications [125, 138]. SR contributes to the evolution of KGs by enabling real-time processing and offering key functionalities such as inference over data streams, pattern extraction, and adaptive modeling [85, 138]. These capabilities allow KGs to handle large-scale stream data, support continual learning, recognize complex events, and integrate heterogeneous data sources [139]. Barry *et al.* [128] propose a KG-based approach with online learning that supports real-time processing and pattern recognition. Their method dynamically updates streaming models, handles heterogeneous data, and scales effectively for large networks under high-volume data streams. Similarly, Le *et al.*[140] introduce a live KGs based software engine that processes fast, high-volume streaming data from heterogeneous IoT sources. The system integrates Linked Stream Data [141] to ensure real-time and scalable data management.

The integration of KGs and SR offers a promising direction for addressing unresolved challenges in stream data management. One of the most prominent of these challenges is the *concept drift* problem [13], as discussed in the previous section. A key difficulty in managing concept drift arises from the inability to semantically interpret inconsistent or evolving information in the data stream [108, 131]. To address this issue, some approaches incorporate KG embeddings to extract semantic features from streaming data, improving the handling of concept drift. This integration demonstrates that challenges traditionally associated with SR can be re-examined and improved by using KGs. In line with this idea, Trivedi *et al.* [85] propose a deep learning framework that models non-linear temporal dynamics and updates embeddings over time, offering a solution for temporal reasoning in dynamic KGs. Another notable example is the self-developing reasoning method introduced by Wang *et al.* [142], which adapts autonomously to environmental changes in real time. Their autonomic KG system combines semantic connections between nodes to continuously update and suggest task-user relationships. These examples illustrate how the integration of KGs and SR can enable more effective management of dynamic by evolving information through tailored reasoning strategies.

The approaches emerging from KGs and SR target another ongoing challenge, *Scalability*, and offer suggestions for its solution. For instance, Zhu *et al.* [143] presents a shortest path algorithm which uses only highly required nodes and edges to handle scalability problem in large-scale KGs reasoning. Similarly, Oliveira *et al.* [144] also offers a KGs management system that supports continuous querying to process high-frequency data stream efficiently and quickly on the Edge device. This system is based on RDF graphics that support different SR models. It offers a scalable query processing structure that can detect anomalies. In another example, Ren *et al.* [145] recommends a scalable framework that enables a multi-hop reasoning in KGs, which provide efficient reasoning even in large-scale complex KGs. Their approach improves reasoning efficiency by operating directly on the graph and opti-

mizing resource usage by storing embeddings on the CPU while minimizing GPU memory consumption.

Beyond scalability, another persistent challenge is *computational resource management*, which often varies depending on the complexity of the proposed methods. Yang *et al.* [146] tackle the challenge of computational efficiency in complex factoid question answering by designing a two-step reasoning framework. First, they construct an evidence graph that filters and retains only the most relevant entities and relations from the knowledge base, based on the input question. Then, they combine this graph with the syntactic and semantic features of the question and feed them into a graph neural network for reasoning. This targeted representation significantly reduces computational overhead while enhancing both accuracy and response time. Similarly, Zhu *et al.*[147] address high resource consumption through a model distillation framework that lowers embedding dimensionality and enhances reasoning efficiency in KGs. This builds upon prior findings by Ruffinelli *et al.*[148], which emphasized the computational burden of managing high-dimensional embeddings. Finally, Bellomarini *et al.* [149] introduce the Vadalog system–based on a datalog fragment as a scalable solution for knowledge representation and reasoning. By incorporating a termination strategy, the system avoids unnecessary reasoning steps, balancing computational complexity with expressive power, even in large-scale KGs [150].

When the capabilities of KGs and SR are combined, their complementary benefits in processing streaming data and enhancing reasoning processes as in following:

- KGs represent relationships and connections between data in semantically rich contexts, enabling SR methods to perform real-time inference and support decision-making in dynamic environments.
- While KGs semantically integrate heterogeneous data from diverse sources with varying structures, SR methods enable continuous processing and updating of this integrated information.
- The deep, structured knowledge stored in KGs can serve as a comprehensive foundation when enriched with real-time data processed by SR methods. This integration enhances scalability and supports efficient handling of high-volume data streams.
- By combining historical knowledge in KGs with current observations from SR, systems can achieve more generalizable and robust models.

In light of these insights, it can be concluded that the integration of SR and KGs offers significant advantages in addressing the challenges posed by heterogeneous, dynamic, and high-volume streaming data.

7 Evaluation of streaming approaches

The performance evaluation criteria and metrics of SR models vary across application domains, despite the existence of commonly accepted standards in the literature. This section discusses the prominent evaluation criteria used in streaming approaches, organized under the main categories of *SML*, *SLD*, and *SKG*. The aim is to provide an innovative contribution by jointly and comparatively analyzing these criteria across different approaches.

To the best of our knowledge, no previous SR-related survey has presented the evaluation methods of all relevant studies in a consolidated manner. In this regard, compiling these criteria offers a novel perspective for the literature. Accordingly, the evaluation criteria identified in the approaches discussed in Section 5 are summarized in Table 4. The table presents the evaluation metrics under 3 main headings and 7 sub-headings of SML, based

Table 4 Evaluation criteria of streaming approaches are discussed in Section 5. SML:Streaming Machine Learning; SLD:Streaming Linked Data; SKG:Streaming Knowledge Graphs; OL:Online Learning; IL:Incremental Learning; CL:Continuous Learning; AL:Active Learning; RL:Reinforcement Learning; CDDM:Concept Drift Detection Methods; SM:Sequential Models. TPR/FPR:True Positive Rate/False Positive Rate; Receiver Operating Characteristics/Area Under Curve; Hit@K:K means number of test instance

Criteria	**Main Streaming Approaches**								
	SML							**SLD**	**SKG**
	OL	IL	CL	AL	RL	CDDM	SM		
Accuracy	[87, 89]	[92, 94–96]	[97–99]	[102–104]	[104]	[16, 108, 109]	[113, 114]	[122, 124]	[10, 126, 128]
Average Number of Error	[87]	X	X	X	X	[109, 111]	X	X	X
Error Rate	[90]	X	X	X	X	[16, 110]	X	X	X
(Execution)Time Cost/Complexity	[90]	[93–96]	[99]	X	[72, 106]	[108]	[113, 114]	[9, 51, 122, 123]	[10, 127]
Page-Hinkley	X	X	X	[104]	[104]	[109]	X	X	X
ROC/AUC	X	[92, 93]	X	X	[107]	[109]	[115]	X	[126, 128]
Memory Usage/Payload	X	[93]	X	X	[72, 106]	X	X	[9]	X
Cluster Purity	X	[94]	X	X	X	X	X	X	X
V-Measure	X	[94]	X	X	X	X	X	X	X
Jaccard Coefficient	X	[94]	X	X	X	X	X	X	X
Number of Parameters	X	X	[99]	X	[106]	X	X	X	X
Precision/Recall	X	X	X	[103]	X	[109]	X	[124]	[126]
Computation Cost	X	X	X	X	[106]	[108]	X	[122]	X
Hit@K	X	X	X	X	[107]	X	X	X	[10, 126–128]
Mean Reciprocal Rank (MRR)	X	X	X	X	[107]	X	X	X	[10, 126, 127]
Mean Square Error(MSE)	X	X	X	X	X	[109, 111]	[115]	X	X
Loss Error Rate	X	X	X	X	X	[109]	X	[124]	[127]
TPR/FPR	X	X	X	X	X	[16, 110]	X	X	X
(Detection) Delay	X	X	X	X	X	[110]	X	[9, 122]	X
F-Measure	X	X	X	X	X	[16, 110]	X	X	[126]
Throughput	X	X	X	X	X	X	X	[9, 119, 124]	[10, 128]

on the approaches in the literature, as categorized in Section 5. Thus, common metrics used by the methods shared in these approaches were revealed. Accordingly, we observe that the most common evaluation criteria are *Accuracy* and *Time Cost* which name varies based on the method as Execution Time or Time Cost. However the *Cluster Purity*, *V-Measure* and *Jaccard Coefficient* are specified for only Incremental Learning related approaches. On the other hand, the *Throughput* criterion, which refers to the number of samples processed per unit of time, is commonly employed in SLD and SKG approaches but is generally absent in SML studies. This criterion, along with other performance metrics, is used to evaluate the effectiveness of the proposed methods. The comparative performance results are reported by using datasets with varying distributions across multiple application domains.

In SML based approaches, model performance is closely tied to timely adaptation and accurate prediction due to the continuous learning strategies offered by methods such as Online Learning, Incremental Learning, and Continual Learning. Conversely, Active Learning and Reinforcement Learning employ semi-supervised strategies on representative sub-datasets and evaluate model performance under evolving stream structures by using metrics such as Page-Hinkley [104], which support real-time drift detection. Concept Drift Detection Methods (CDDM) are particularly suited to the dynamic nature of streaming data. These approaches offer the broadest range of evaluation metrics for drift detection, which vary depending on dataset characteristics and algorithm design. Time-series-based Sequential Models commonly rely on metrics related to processing time and error rate over fixed intervals. Among the primary streaming reasoning approaches, SLD evaluates performance based on whether Web-based linked data can be processed within acceptable delay thresholds using accurate, optimized memory management.

On the other hand, KG based approaches stand out for their broad evaluation criteria, ranking second only to CDDM methods. These criteria are used for tasks like missing entity detection, link prediction, and assessing how models handle data drift. The diversity of evaluation criteria observed in these approaches within the context of SR highlights their methodological richness. In particular, CDDM and SKG approaches appear to provide more suitable foundations for developing adaptive methods tailored to stream data.

8 Discussion

This survey presents a comprehensive review of current research in SR, focusing on its core challenges in the context of real-time decision-making. By utilizing state-of-the-art techniques from ML, Neural Networks, and KGs, the study addresses key challenges associated with dynamic, heterogeneous, and high-volume data streams.

The research begins by outlining the methodological criteria used in the literature, followed by brief explanations of SR and KG concepts. It then categorizes major challenges in stream processing, examining their origins and how they are defined in existing works. Prominent solutions proposed in the literature are discussed under three main categories: SML, SLD and SKG. In addition, unresolved secondary challenges that emerge from these core issues are identified, and the potential of SR–KG integration in addressing them is explored.

While streaming approaches offer several benefits, they also present notable limitations in real-time reasoning over streaming data. These issues can be observed across the main categories of SML, SLD, and SKG. Within SML, learning strategies such as Online, Incremental, and Continual Learning are generally considered suitable for continuously streaming data. However, their performance may degrade in the presence of concept drift [88], where

sudden changes in data format or type hinder adaptation. Although these models can process dynamic data without retraining for each new dataset and can reduce certain concept biases through adaptive learning and memory mechanisms [151], they remain vulnerable to catastrophic forgetting–over-adapting to new inputs while losing previously acquired knowledge. Continuous learning methods propose mitigation strategies, but no complete solution exists yet [152]. Active Learning can reduce labeling costs, but selecting an optimal representative dataset becomes complex in heterogeneous and rapidly evolving streams. Similarly, Reinforcement Learning applied to high-volume, high-velocity streams, especially in feedback-driven environments [72], can be computationally expensive. Moreover, constantly changing stream characteristics may impair RL models' generalization ability. Concept drift detection methods, though designed for real-time adaptability, often fail to scale effectively with high-speed and large-volume streams. Accurate detection requires retaining historical data and addressing catastrophic forgetting, which in turn demands significant computational resources. This requirement is also a limitation for other learning strategies. Sequential models such as LSTM and RNN, which rely on storing and analyzing past data, can also struggle with massive, dynamic streams. Insufficient storage and processing capacity may prevent them from capturing evolving patterns by limiting their suitability for SR compared to other strategies.

Semantic models used in SLD for integrating and interpreting large volumes of heterogeneous data from multiple sources may lack the flexibility and speed needed to capture and process inter-data complexity in real time. While this integration is essential for managing data heterogeneity, it can hinder accurate and efficient real-time SR. Similarly, SKG provide an effective infrastructure for storing and processing heterogeneous and dynamic data streams. However, the need for continuous knowledge base updates increases computational demands during reasoning and can reduce overall efficiency. Conversely, KGs enhance SR by supplying meaningful, high-volume data representations to ML models, enabling them to capture relationships more effectively and improve performance.

As discussed in Section 6, integrating SR with KGs, and when appropriate with ML, can help address unresolved challenges. KGs semantically organize large-scale, continuously streaming data, assisting ML models in knowledge retention and mitigating issues such as catastrophic forgetting and scalability constraints. This integration creates a more adaptive system for handling dynamic streams. Additionally, the constantly updated nature of KGs allows ML models to adapt during the SR process, improving responsiveness to concept drift.

Despite these benefits, such integrations require complex and computationally expensive modeling of high-volume, dynamic data streams, as well as substantial processing power and memory resources. Incorporating noisy and heterogeneous data from multiple sources can also create compatibility issues, increasing error rates and reducing system performance.

The approaches discussed in the categories of SKG, Linked Data, and ML in relation to SR share common limitations when processing heterogeneous, dynamic, high-volume streams. Real-time reasoning can be constrained by time requirements, and the variable quality of incoming data can negatively affect model accuracy. Furthermore, storing and processing large-scale streams imposes both computational and memory costs. Nonetheless, integrating KGs, SR, and ML offers promising state-of-the-art solutions to these challenges. This study emphasizes the importance of such integrations and aims to guide the development of SR-based solutions aligned with current technologies.

9 Conclusion

This study identified the main challenges in SR and their underlying causes by reviewing the literature. We evaluated various algorithmic and technological approaches proposed to address these streaming challenges and grouped them into categories (summarized in Table 4). Thereby, we also uncovered several open challenges that emerge indirectly from these issues and remain unsolved. A key finding of our analysis is the importance of integrating SR with KGs structures, since KGs provide a strong framework for storing and representing large-scale streaming data. Through examples from the literature, we illustrated how this integration can effectively address the identified challenges. Moreover, incorporating state-of-the-art techniques into these approaches yields more effective solutions for managing streaming challenges. The basic challenges in SR discussed in this study are heterogeneity, dynamicity, and high volume, which stem from the nature of streaming data. We categorized existing solutions for these challenges into three main paradigms: SML, SLD, and SKG, and analyzed the suitability of each for tackling different issues. Within SML, we explored the advantages and limitations of state-of-the-art online, incremental, continuous, active, and reinforcement learning strategies, as well as concept drift detection and time-series methods. These approaches vary in their application and methodology: some address multiple challenges, while others focus on a specific one. Based on our analysis, SLD (using ontologies to connect diverse data formats and semantics) is well-suited for addressing heterogeneity. SKG can be more effective for heterogeneity, owing to its graph structure that accurately represents relationships among complex data. For dynamicity, SML approaches leverage adaptive, continuous learning algorithms to capture evolving patterns, and SKG approaches provide flexible data infrastructures to manage continuous dynamic data streams. This makes both categories effective for this challenge. To handle high-volume streams, techniques such as parallel processing for fast queries, SML with incremental/online learning strategies, and distributed processing capabilities proved most efficient. In addition, we compared the effectiveness of specific SML methods for each challenge and observed that reinforcement learning strategies can address all SR challenges most comprehensively.

In brief, our findings reveal that integrating SR with KGs can help address open challenges such as scalability, concept drift, and computational cost. Including machine learning and neural network methods in this integration further improves the management of these challenges. For example, GNN based approaches have been shown to enhance SR by leveraging the structured information in KGs. Therefore, we suggest that future work focus on developing a state-of-the-art GNN architecture in conjunction with KGs to minimize the remaining challenges in SR.

10 Future work

The prospective research agenda for SR–KG integration should be systematically structured to address both technical approaches and application domains, with explicit connections to the main challenges and open challenges identified in this survey. Following subsections present targeted directions for advancing SR–KG methods. The first subsection focuses on methodological developments, while the second explores how these solutions can be applied across diverse real-world scenarios.

10.1 Technical approaches

Future research on SR over KGs should focus on incremental and drift-aware inference strategies. These strategies should address the main challenges in Section 4 and the open challenges in Section 5.4, including concept drift, catastrophic forgetting, scala- bility, and computational cost. In this context, each proposed technical direction is explicitly designed to tackle one or more of these challenges: incremental and drift-aware inference directly targets concept drift and catastrophic forgetting; principled window semantics aim to ensure temporal consistency while supporting scalability; change-propagation mechanisms reduce computational cost by limiting redundant reasoning; and approximate or anytime reasoning addresses the trade-off between latency constraints and resource limitations. For learning models, graph neural networks on evolving KGs are promising when paired with update-efficient embeddings, enabling adaptation to shifting data without frequent retraining–thus mitigating both concept drift and scalability issues. For evaluation frameworks, SR-specific protocols should measure throughput, window latency, end-to-end latency, incremental update cost, memory footprint, and drift-handling performance (see Section 7). These metrics map directly to the identified challenges by quantifying system responsiveness, efficiency, and robustness under dynamic conditions. When learning is involved, ranking metrics such as Hit@K and Mean Reciprocal Rank (MRR) are suitable, providing standardized measures that support reproducible and comparable benchmarks.

10.2 Application domains

In IoT and broader cyber–physical systems, SR should place lightweight KGs and temporal operators close to data sources. This reduces delays and helps manage high data rates–addressing the main challenges of scalability and latency. Periodic consolidation of local results into a global reasoning layer can improve scalability and maintain semantic consistency, directly mitigating related open challenges such as distributed data integration. In multi-site deployments with data residency constraints, federated continual learning allows local model training and incremental updates without moving raw data, thus addressing privacy and compliance challenges while supporting scalable knowledge sharing. Parameter sharing preserves global model coherence while reducing communication costs. In healthcare monitoring, SR over clinical event streams should combine uncertainty-aware inference with drift or change-point detection, directly targeting open challenges related to noisy, incomplete, and evolving data. This enables timely responses to patient state changes and supports clinical decision-making with interpretable explanations–an essential aspect of trust and accountability. Beyond these examples, SR–KG integration can be extended to other relevant application domains, such as finance, transportation, and environmental monitoring, where addressing the identified main and open challenges is crucial for reliable and efficient real-time reasoning.

Acknowledgements This work was supported by the Swiss National Science Foundation through the StreamKG project with grant number 213369.

Author Contributions As the first author, Gözde Ayşe Tataroğlu Özbulak took part in every stage of the study (determining the research topics and questions, collecting and evaluating relevant academic articles, designing and writing...). Jean-Paul Calbimonte and Gaetano Manzo actively contributed to the direction of the study, determination of research questions, and design. Jean-Paul Calbimonte supervised the study as the main supervisor. Yash Rash Shrestha contributed to revising and improving previous versions of the article. The first version of the draft was written by Gözde Ayşe Tataroğlu Özbulak and updated according to the

comments of Jean-Paul Calbimonte, Gaetano Manzo and Yash Rash Shrestha. All authors read and approved the final draft.

Funding Open access funding provided by University of Lausanne.

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Dell'Aglio D, Della Valle E, Harmelen F, Bernstein A (2017) Stream reasoning. A survey and outlook Data Sci 1(1–2):59–83
2. Bourgais M, Giustozzi F, Vercouter L (2021) Detecting situations with stream reasoning on health data obtained with iot. Proc Comput Sci 192:507–516
3. Mileo A (2015) Web stream reasoning: from data streams to actionable knowledge, reasoning web. web logic rules: 11th international summer school 2015, berlin, germany, july 31-august 4, 2015. Tutorial Lectures 11:75–87
4. Corral-Plaza D, Medina-Bulo I, Ortiz G, Boubeta-Puig J (2020) A stream processing architecture for heterogeneous data sources in the internet of things. Comput Stand Interfac 70:103426
5. Jung HS, Yoon CS, Lee YW, Park JW, Yun CH (2017) Cloud computing platform based real-time processing for stream reasoning. In: 2017 Sixth International Conference on Future Generation Communication Technologies (FGCT). IEEE
6. Dessì D, Osborne F, Reforgiato Recupero D, Buscaldi D, Motta E (2021) Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Futur Gener Comput Syst 116:253–264
7. Margara A, Urbani J, Harmelen F, Bal H (2014) Streaming the web: reasoning over dynamic data. J Web Semant 25:24–44
8. Stuckenschmidt H, Ceri S, Della Valle E, Van Harmelen F (2010) Towards expressive stream reasoning. In: Dagstuhl seminar proceedings. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
9. Anicic D, Rudolph S, Fodor P, Stojanovic N (2012) Stream reasoning and complex event processing in etalis. Semant Web 3(4):397–407
10. Omran PG, Wang K, Wang Z (2019) Learning temporal rules from knowledge graph streams. In: Proceedings of the AAAI 2019 spring symposium on combining machine learning with knowledge engineering (AAAI-MAKE 2019). CEUR-WS.org
11. Lu R, Cai Z, Zhao S (2019) A survey of knowledge reasoning based on kg. In: IOP conference series: materials science and engineering, 569(5). IOP Publishing, 052058
12. Gomes HM, Read J, Bifet A, Barddal JP, Gama J (2019) Machine learning for streaming data: state of the art, challenges, and opportunities. ACM SIGKDD Explor Newsl 21(2):6–22
13. Hoens TR, Polikar R, Chawla NV (2012) Learning from streaming data with concept drift and imbalance: an overview. Progress Artif Intell 1(1):89–101
14. Zhao Y, Wang X, Chen J, Wang Y, Tang W, He X, Xie H (2022) Time-aware path reasoning on knowledge graph for recommendation. ACM Trans Inf Syst 41(2):1–26
15. Zhang Z, Zhuang F, Zhu H, Shi Z, Xiong H, He Q (2020) Relational graph neural network with hierarchical attention for knowledge graph completion. In: Proceedings of the AAAI conference on artificial intelligence, 34(05). Association for the Advancement of Artificial Intelligence (AAAI), pp. 9612–9619

16. Wang H, Abraham Z (2015) Concept drift detection for streaming data. In: 2015 International Joint Conference on Neural Networks (IJCNN). IEEE
17. Guo L, Yan F, Li T, Yang T, Lu Y (2022) An automatic method for constructing machining process knowledge base from knowledge graph. Robot Comput -Integr Manuf 73:102222
18. Al-Moslmi T, Gallofre Ocana M, Opdahl AL, Veres C (2020) Named entity extraction for knowledge graphs: a literature overview. IEEE Access 8:32862–32881
19. Dessì D, Osborne F, Recupero DR, Buscaldi D, Motta E, Sack H (2020) Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: The semantic web–ISWC 2020: 19th international semantic web conference. Greece, November, 2020, Proceedings, Part II 19. Springer International Publishing: Springer International Publishing, pp. 127–143
20. Yahya M, Breslin JG, Ali MI (2021) Semantic web and knowledge graphs for industry 4.0. Appl Sci 11(11):5110
21. Safaei AA (2017) Real-time processing of streaming big data. Real-Time Syst 53:1–44
22. Bajwa WU, Cevher V, Papailiopoulos D, Scaglione A (2020) Machine learning from distributed, streaming data [from the guest editors]. IEEE Signal Process Mag 37(3):11–13
23. Fotia L, Delicato F, Fortino G (2023) Trust in edge-based internet of things architectures: state of the art and research challenges. ACM Comput Surv 55(9):1–34
24. Dell'Aglio D, Della Valle E, Harmelen F, Bernstein A (2017) Stream reasoning: a survey and outlook: a summary of ten years of research and a vision for the next decade. Data Sci 1(1–2):59–83
25. Mountantonakis M, Tzitzikas Y (2019) Large-scale semantic integration of linked data: a survey. ACM Comput Surv 52(5):103:1-103:40
26. Zeng Z, Cheng Q, Si Y (2023) Logical rule-based knowledge graph reasoning: a comprehensive survey. Mathematics 11(21):4486
27. Fragkoulis M, Carbone P, Kalavri V, Katsifodimos A (2024) A survey on the evolution of stream processing systems. VLDB J 33(2):507–541
28. Bonte P, Calbimonte J-P, Leng D, Dell'Aglio D, Valle ED, Eiter T, Giannini F, Heintz F, Schekotihin K, Le-Phuoc D, Mileo A, Schneider P, Tommasini R, Urbani J, Ziffer G (2024) Grounding stream reasoning research. Trans Graph Data Knowl 2(1):2:1-2:47
29. Read J, Zliobaite I (2025) Supervised learning from data streams: an overview and update, ACM Computing Surveys, 57(12) to appear, article in press
30. Jin X, Wang Z, Duan M, Shao Y, Hong X, Wang Y, Oh B (2025) A survey on knowledge graph evolution: proliferation, dynamic embedding, and versioning. Int J Web Grid Serv 21(1):88–111
31. Cheng K, Ahmed NK, Rossi RA, Willke T, Sun Y (2025) Neural-symbolic methods for knowledge graph reasoning: a survey. ACM Trans Knowl Discov Data 18(9):1–44
32. Della Valle E, Ceri S, Harmelen F, Fensel D (2009) It's a streaming world! reasoning upon rapidly changing information. IEEE Intell Syst 24(6):83–89
33. Su X, Gilman E, Wetz P, Riekki J, Zuo Y, Leppänen T (2016) Stream reasoning for the internet of things: challenges and gap analysis. In: Proceedings of the 6th international conference on web intelligence, mining and semantics, ser. WIMS '16. ACM
34. Bansal SK (2014) Towards a semantic extract-transform-load (etl) framework for big data integration. In: 2014 IEEE International Congress on Big Data. IEEE, pp. 522–529
35. Razzaque MA, Milojevic-Jevric M, Palade A, Clarke S (2015) Middleware for internet of things: a survey. IEEE Internet Things J 3(1):70–95
36. Vongsingthong S, Smanchat S (2015) A review of data management in internet of things. Asia-Pac J Sci Technol 20(2):215–240
37. Corral-Plaza D, Medina-Bulo I, Ortiz G, Boubeta-Puig J (2020) A stream processing architecture for heterogeneous data sources in the internet of things. Comput Stand Interfaces 70:103426
38. Akanbi A, Masinde M (2020) A distributed stream processing middleware framework for real-time analysis of heterogeneous data on big data platform: case of environmental monitoring. Sensors 20(11):3166
39. Hendler J (2014) Data integration for heterogenous datasets. Big data 2(4):205–215
40. Bartolini I, Patella M (2019) Real-time stream processing in social networks with ram3s. Future Internet 11(12):249
41. Dimou A, Vander Sande M, Colpaert P, Verborgh R, Mannens E, Van de Walle R (2014) Rml: a generic language for integrated RDF mappings of heterogeneous data. Ldow, 1184
42. Oo SM, Haesendonck G, De Meester B, Dimou A (2022) RMLStreamer-siso: an rdf stream generator from streaming heterogeneous data. In: International semantic web conference. Springer, pp. 697–713
43. Santipantakis GM, Kotis KI, Vouros GA, Doulkeridis C (2018) Rdf-gen: Generating RDF from streaming and archival data. In: Proceedings of the 8th international conference on web intelligence, mining and semantics, pp. 1–10

44. Gerber D, Hellmann S, Bühmann L, Soru T, Usbeck R, Ngonga Ngomo A-C (2013) Real-time RDF extraction from unstructured data streams. In: The semantic web–ISWC 2013: 12th international semantic web conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12. Springer, pp. 135–150
45. Mauri A, Calbimonte J-P, Dell'Aglio D, Balduini M, Brambilla M, Della Valle E, Aberer K (2016) Triplewave: spreading RDF streams on the web. In: the semantic web–ISWC 2016: 15th international semantic web conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15. Springer, pp. 140–149
46. Calbimonte J-P, Corcho O, Gray AJ (2010) Enabling ontology-based access to streaming data sources. In: The semantic web–ISWC 2010: 9th international semantic web conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I 9. Springer, pp. 96–111
47. Taheriyan M, Knoblock CA, Szekely P, Ambite JL (2016) Learning the semantics of structured data sources. J Web Semant 37:152–169
48. Tan R, Chirkova R, Gadepally V, Mattson TG (2017) Enabling query processing across heterogeneous data models: a survey. In: 2017 IEEE international conference on big data (Big Data). IEEE, pp. 3211–3220
49. Bansal SK, Kagemann S (2015) Integrating big data: a semantic extract-transform-load framework. Computer 48(3):42–50
50. Dell'Aglio D, Della Valle E, Calbimonte J-P, Corcho O (2014) RSP-QL semantics: a unifying query model to explain heterogeneity of RDF stream processing systems. Int J Semant Web Inf Syst(IJSWIS) 10(4):17–44
51. Dell'Aglio D, Dao-Tran M, Calbimonte J-P, Le Phuoc D, Della Valle E (2016) A query model to capture event pattern matching in RDF stream processing query languages. Springer International Publishing, pp. 145–162
52. Keskisärkkä R, Blomqvist E, Lind L, Hartig O (2019) RSP-QL: enabling statement-level annotations in RDF streams. In: International conference on semantic systems. Springer, pp. 140–155
53. Wu X, Zhu X, Wu G-Q, Ding W (2013) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
54. Shi F, Li Q, Zhu T, Ning H (2018) A survey of data semantization in internet of things. Sensors 18(1):313
55. Esposito C, Ficco M, Palmieri F, Castiglione A (2015) A knowledge-based platform for big data analytics based on publish/subscribe services and stream processing. Knowl-Based Syst 79:3–17
56. Della Valle E, Dell'Aglio D, Margara A (2016) Taming velocity and variety simultaneously in big data with stream reasoning: tutorial. In: Proceedings of the 10th ACM international conference on distributed and event-based systems, pp. 394–401
57. Ali MI, Ono N, Kaysar M, Shamszaman ZU, Pham T-L, Gao F, Griffin K, Mileo A (2017) Real-time data analytics and event detection for iot-enabled communication systems. J Web Semant 42:19–37
58. Hu H, Wen Y, Chua T-S, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687
59. Read J, Bifet A, Holmes G, Pfahringer B (2012) Scalable and efficient multi-label classification for evolving data streams. Mach Learn 88:243–272
60. Kenda K, Kažič B, Novak E, Mladenić D (2019) Streaming data fusion for the internet of things. Sensors 19(8):1955
61. Peng C, Goswami P (2019) Meaningful integration of data from heterogeneous health services and home environment based on ontology. Sensors 19(8):1747
62. Della Valle E, Ceri S, Braga D, Celino I, Frensel D, van Harmelen F, Unel G (2009) Research chapters in the area of stream reasoning, SR2009, 466
63. Brewka G, Ellmauthaler S, Gonçalves R, Knorr M, Leite J, Pührer J (2018) Reactive multi-context systems: heterogeneous reasoning in dynamic environments. Artif Intell 256:68–104
64. Lécué F, Kotoulas S, Mac Aonghusa P (2012) Capturing the pulse of cities: opportunity and research challenges for robust stream data reasoning. In: Workshops at the twenty-sixth AAAI conference on artificial intelligence
65. Yang Y, Wu X, Zhu X (2005) Combining proactive and reactive predictions for data streams. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, pp. 710–715
66. Barbieri D, Braga D, Ceri S, Della Valle E, Grossniklaus M (2010) Stream reasoning: where we got so far. In: Proceedings of the 4th workshop on new forms of reasoning for the semantic web : scalable & dynamic
67. Teh HY, Kempa-Liehr AW, Wang KI-K (2020) Sensor data quality: a systematic review. J Big Data 7(1):1–49

68. Carlo B, Daniele B, Federico C, Simone G (2011) A data quality methodology for heterogeneous data. Int J Database Manag Syst 3(1):60–79
69. Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M (2019) Real-time big data processing for anomaly detection: a survey. Int J Inf Manag 45:289–307
70. Anicic D, Fodor P, Rudolph S, Stojanovic N (2011) Ep-sparql: a unified language for event processing and stream reasoning. In: Proceedings of the 20th international conference on World wide web, pp. 635–644
71. Giustozzi F, Saunier J, Zanni-Merk C (2019) Abnormal situations interpretation in industry 4.0 using stream reasoning. Proc Comput Sci 159:620–629
72. Dodaro C, Eiter T, Ogris P, Schekotihin K (2020) Managing caching strategies for stream reasoning with reinforcement learning. Theory Pract Logic Program 20(5):625–640
73. Gama J (2012) A survey on learning from data streams: current and future trends. Progress Artif Intell 1(1):45–55
74. Kolajo T, Daramola O, Adebiyi A (2019) Big data stream analysis: a systematic literature review. J Big Data 6(1):47
75. Thudumu S, Branch P, Jin J, Singh J (2020) A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7:1–30
76. Maarala AI, Rautiainen M, Salmi M, Pirttikangas S, Riekki J (2015) Low latency analytics for streaming traffic data with apache spark. In: 2015 IEEE international conference on big data (Big Data). IEEE, pp. 2855–2858
77. Liu X, Iftikhar N, Xie X (2014) Survey of real-time processing systems for big data. In: Proceedings of the 18th international database engineering & applications symposium, pp. 356–361
78. Krempl G, Žliobaite I, Brzeziński D, Hüllermeier E, Last M, Lemaire V, Noack T, Shaker A, Sievi S, Spiliopoulou M et al (2014) Open challenges for data stream mining research. ACM SIGKDD Explor Newsl 16(1):1–10
79. Acharjya DP, Ahmed K (2016) A survey on big data analytics: challenges, open research issues and tools. Int J Adv Comput Sci Appl 7(2):511–518
80. Wang C, Chen M-H, Schifano E, Wu J, Yan J (2016) Statistical methods and computing for big data. Stat Interface 9(4):399
81. Souiden I, Omri MN, Brahmi Z (2022) A survey of outlier detection in high dimensional data streams. Comput Sci Rev 44:100463
82. Cao L, Wang J, Rundensteiner EA (2016) Sharing-aware outlier analytics over high-volume data streams. In: Proceedings of the 2016 international conference on management of data, pp. 527–540
83. L'heureux A, Grolinger K, Elyamany HF, Capretz MA, (2017) Machine learning with big data: challenges and approaches. IEEE Access 5:7776–7797
84. Le-Phuoc D, Dao-Tran M, Xavier Parreira J, Hauswirth M (2011) A native and adaptive approach for unified processing of linked streams and linked data. In: international semantic web conference. Springer, pp. 370–388
85. Trivedi R, Dai H, Wang Y, Song L (2017) Know-evolve: deep temporal reasoning for dynamic knowledge graphs. In: Proceedings of the 34th international conference on machine learning - Volume 70, ser. ICML'17. JMLR.org, pp. 3462–3471
86. Wang L (2017) Heterogeneous data and big data analytics. Autom Control Inf Sci 3(1):8–15
87. Beyazit E, Alagurajah J, Wu X (2019) Online learning from data streams with varying feature spaces. Proc AAAI Conf Artif Intell 33(01):3232–3239
88. Wang S, Schlobach S, Klein M (2011) Concept drift and how to identify it. J Web Semant 9(3):247–265
89. Din SU, Shao J, Kumar J, Ali W, Liu J, Ye Y (2020) Online reliable semi-supervised learning on evolving data streams. Inf Sci 525:153–171
90. Wang D, Wu P, Zhao P, Wu Y, Miao C, Hoi SC (2014) High-dimensional data stream classification via sparse online learning. In: 2014 IEEE international conference on data mining. IEEE
91. Ikonomovska E, Gama J, Džeroski S (2011) Learning model trees from evolving data streams. Data Min Knowl Disc 23:128–168
92. He H, Chen S, Li K, Xu X (2011) Incremental learning from stream data. IEEE Trans Neural Netw 22(12):1901–1914
93. Chen S, He H (2011) Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach. Evol Syst 2(1):35–50
94. Nikpour S, Asadi S (2022) A dynamic hierarchical incremental learning-based supervised clustering for data stream with considering concept drift. J Ambient Intell Humaniz Comput 13(6):2983–3003
95. Abdallah ZS, Gaber MM, Srinivasan B, Krishnaswamy S (2018) Activity recognition with evolving data streams: a review. ACM Comput Surv (CSUR) 51(4):1–36

96. Le-Phuoc D, Eiter T, Le-Tuan A (2021) A scalable reasoning and learning approach for neural-symbolic stream fusion. Proc AAAI Conf Artif Intell 35(6):4996–5005
97. Aljundi R, Kelchtermans K, Tuytelaars T (2019) Task-free continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11254–11263,
98. Criado MF, Casado FE, Iglesias R, Regueiro CV, Barro S (2022) Non-iid data and continual learning processes in federated learning: a long road ahead. Inf Fusion 88:263–280
99. Ashfahani A, Pratama M (2019) Autonomous deep learning: continual learning approach for dynamic environments. In: Proceedings of the 2019 SIAM international conference on data mining. SIAM, pp. 666–674
100. Castro FM, Marín-Jiménez MJ, Guil N, Schmid C, Alahari K (2018) End-to-end incremental learning. In: Proceedings of the European conference on computer vision (ECCV). Springer International Publishing, pp. 241–257
101. Fu Y, Zhu X, Li B (2013) A survey on instance selection for active learning. Knowl Inf Syst 35:249–283
102. Ienco D, Bifet A, Žliobaitė I, Pfahringer B (2013) Clustering based active learning for evolving data streams. In: International conference on discovery science. Springer, pp. 79–93
103. Žliobaitė I, Bifet A, Pfahringer B, Holmes G (2013) Active learning with drifting streaming data. IEEE Trans Neural Netw Learn Syst 25(1):27–39
104. Wassermann S, Cuvelier T, Casas P (2019) Ral - improving stream-based active learning by reinforcement learning. In: Proceedings of the workshop on interactive adaptive learningco-located with European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD 2019), 2444. CEUR-WS.org, 32–47. [Online]. Available: https://ceur-ws.org/Vol-2444/ialatecml_paper3.pdf
105. Dezfouli A, Balleine BW (2012) Habits, action sequences and reinforcement learning. Eur J Neurosci 35(7):1036–1051
106. Russo GR, Cardellini V, Presti FL (2019) Reinforcement learning based policies for elastic stream processing on heterogeneous resources. In: Proceedings of the 13th ACM international conference on distributed and event-based systems, pp. 31–42
107. Wang Q, Hao Y, Cao J (2020) Adrl: an attention-based deep reinforcement learning framework for knowledge graph reasoning. Knowl-Based Syst 197:105910
108. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Comput Surv (CSUR) 46(4):1–37
109. Bayram F, Ahmed BS, Kassler A (2022) From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl-Based Syst 245:108632
110. Agrahari S, Singh AK (2022) Concept drift detection in data stream mining: a literature review. J King Saud Univ -Comput Inf Sci 34(10):9523–9540
111. Zenisek J, Holzinger F, Affenzeller M (2019) Machine learning based concept drift detection for predictive maintenance. Comput Ind Eng 137:106031
112. Sovilj D, Budnarain P, Sanner S, Salmon G, Rao M (2020) A comparative evaluation of unsupervised deep architectures for intrusion detection in sequential data streams. Expert Syst Appl 159:113577
113. Ferreira J, Lavado D, Gonçalves R, Knorr M, Krippahl L, Leite J (2021) Faster than laser-towards stream reasoning with deep neural networks. In: Progress in artificial intelligence: 20th EPIA conference on artificial intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20. Springer, pp. 363–375
114. Ferreira R, Lopes C, Gonçalves R, Knorr M, Krippahl L, Leite J (2021) Deep neural networks for approximating stream reasoning with c-sparql. In: EPIA conference on artificial intelligence, pp. 338–350
115. Qin Z, Cheng Y, Zhao Z, Chen Z, Metzler D, Qin J (2020) Multitask mixture of sequential experts for user activity streams. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3083–3091
116. Qi P, Zhou X, Ding Y, Zheng S, Jiang T, Li Z (2022) Collaborative and incremental learning for modulation classification with heterogeneous local dataset in cognitive IoT. In: IEEE transactions on green communications and networking
117. Fu X, Yu FR, Wang J, Qi Q, Liao J (2019) Dynamic service function chain embedding for nfv-enabled iot: a deep reinforcement learning approach. IEEE Trans Wireless Commun 19(1):507–519
118. Bizer C (2009) The emerging web of linked data. IEEE Intell Syst 24(5):87–92
119. Wang S, Wan J, Li D, Liu C (2018) Knowledge reasoning with semantic data for real-time data processing in smart factory. Sensors 18(2):471
120. Saber A, Al-Zoghby AM, Elmougy S (2018) Big-data aggregating, linking, integrating and representing using semantic web technologies. In: advances in intelligent systems and computing, S. I. Publishing, Ed. Springer International Publishing, pp. 331–342

121. Calbimonte J-P, Mora J, Corcho O (2016) Query rewriting in RDF stream processing. In: The semantic web. Latest advances and new domains: 13th international conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13. International Publishing. Springer International Publishing, pp. 486–502
122. Maarala AI, Su X, Riekki J (2017) Semantic reasoning for context-aware internet of things applications. IEEE Internet Things J 4(2):461–473
123. Abeykoon V, Kamburugamuve S, Govindrarajan K, Wickramasinghe P, Widanage C, Perera N, Uyar A, Gunduz G, Akkas S, Laszewski GV (2019) Streaming machine learning algorithms with big data systems. In: 2019 IEEE international conference on big Data (Big Data). IEEE
124. Oren E, Kotoulas S, Anadiotis G, Siebes R, Teije A, Harmelen F (2009) Marvin: distributed reasoning over large-scale semantic web data. J Web Semant 7(4):305–316
125. Ji S, Pan S, Cambria E, Marttinen P, Philip SY (2021) A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 33(2):494–514
126. Kazemi SM, Goel R, Jain K, Kobyzev I, Sethi A, Forsyth P, Poupart P (2020) Representation learning for dynamic graphs: a survey. J Mach Learn Res 21(70):1–73
127. Wu T, Khan A, Yong M, Qi G, Wang M (2022) Efficiently embedding dynamic knowledge graphs. Knowl-Based Syst 250:109124
128. Barry M, Bifet A, Chiky R, El Jaouhari S, Montiel J, El Ouafi A, Guerizec E (2022) Stream2graph: dynamic knowledge graph for online learning applied in large-scale network. In: 2022 IEEE international conference on big data (Big Data). IEEE
129. Bellomarini L, Gottlob G, Pieris A, Sallinger E (2018) Swift logic for big data and knowledge graphs: overview of requirements, language, and system. In: SOFSEM 2018: theory and practice of computer science: 44th international conference on current trends in theory and practice of computer science, Krems, Austria, January 29-February 2, 2018, Proceedings 44. Springer, pp. 3–16
130. Narayanan SN, Ganesan A, Joshi K, Oates T, Joshi A, Finin T (2018) Early detection of cybersecurity threats using collaborative cognition. In: 2018 IEEE 4th international conference on collaboration and internet computing (CIC). IEEE, pp. 354–363
131. Chen J, Lecue F, Pan JZ, Deng S, Chen H (2021) Knowledge graph embeddings for dealing with concept drift in machine learning. J Web Semant 67:100625
132. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S (2019) Continual lifelong learning with neural networks: a review. Neural Netw 113:54–71
133. Agrahari S, Singh AK (2022) Concept drift detection in data stream mining: a literature review. J King Saud Univ Comput Inf Sci 34(10):9523–9540
134. Gulisano V, Jimenez-Peris R, Patino-Martinez M, Valduriez P (2010) Streamcloud: A large scale data streaming system. In: 2010 IEEE 30th international conference on distributed computing systems. IEEE, 126–137
135. Maarala AI, Su X, Riekki J (2016) Semantic reasoning for context-aware internet of things applications. IEEE Internet Things J 4(2):461–473
136. Henning S, Hasselbring W (2021) How to measure scalability of distributed stream processing engines? In: Companion of the ACM/SPEC international conference on performance engineering, pp. 85–88
137. Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M (2017) Ensemble learning for data stream analysis: a survey. Inf Fusion 37:132–156
138. Chen X, Jia S, Xiang Y (2020) A review: knowledge reasoning over knowledge graph. Expert Syst Appl 141:112948
139. Chen Y, Li H, Li H, Liu W, Wu Y, Huang Q, Wan S (2022) An overview of knowledge graph reasoning: key technologies and applications. J Sens Actuator Netw 11(4):78
140. Le-Phuoc D, Quoc HNM, Quoc HN, Nhat TT, Hauswirth M (2016) The graph of things: a step towards the live knowledge graph of connected things. J Web Semant 37:25–35
141. Le-Phuoc D, Dao-Tran M, Pham M-D, Boncz P, Eiter T, Fink M (2012) Linked stream data processing engines: facts and figures., In: International semantic web conference. Springer, pp. 300–312
142. Wang J, Yan Y, Zhao G (2023) Self-evolving reasoning for task-user relationships in mobile crowdsensing via the autonomic knowledge graph. Artif Intell Rev 56(Suppl 3):3789–3819
143. Zhu Z, Yuan X, Galkin M, Xhonneux L-P, Zhang M, Gazeau M, Tang J (2024) A* net: a scalable path-based reasoning approach for knowledge graphs. Advances in neural information processing systems, 36
144. de Oliveira J, Callé C, Xu W, Calvez P, Curé O (2022) Knowledge graph stream processing at the edge. In: Proceedings of the 16th ACM international conference on distributed and event-based systems, pp. 115–125

145. Ren H, Dai H, Dai B, Chen X, Zhou D, Leskovec J, Schuurmans D (2022) Smore: knowledge graph completion and multi-hop reasoning in massive knowledge graphs. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pp. 1472–1482
146. Yang X, Chiang M-F, Lee W-C, Chang Y (2021) Cost-effective knowledge graph reasoning for complex factoid questions. In: 2021 international joint conference on neural networks (IJCNN). IEEE, pp. 1–8
147. Zhu Y, Zhang W, Chen M, Chen H, Cheng X, Zhang W, Chen H (2022) Dualde: dually distilling knowledge graph embedding for faster and cheaper reasoning. In: Proceedings of the fifteenth ACM international conference on web search and data mining, pp. 1516–1524
148. Ruffinelli D, Broscheit S, Gemulla R (2019) You can teach an old dog new tricks! on training knowledge graph embeddings. In: international conference on learning representations
149. Bellomarini L, Benedetto D, Gottlob G, Sallinger E (2022) Vadalog: a modern architecture for automated reasoning with large knowledge graphs. Inf Syst 105:101528
150. Calì A, Gottlob G, Lukasiewicz T, Marnette B, Pieris A (2010) Datalog+/-: a family of logical knowledge representation and query languages for new applications. In: 2010 25th annual IEEE symposium on logic in computer science. IEEE, pp. 228–242
151. Mulinka P, Wassermann S, Marín G, Casas P (2018) Remember the good, forget the bad, do it fast-continuous learning over streaming data. In: Continual learning workshop at NeurIPS 2018
152. Ebrahimi S, Petryk S, Gokul A, Gan W, Gonzalez JE, Rohrbach M, Darrell T (2021) Remembering for the right reasons: explanations reduce catastrophic forgetting. Appl AI Lett 2(4):e44