International Journal of Information Management Data Insights 5 (2025) 100321

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Enhancing DataOps practices through innovative collaborative models: A systematic review

Aymen Fannouch *, Jihane Gharib, Youssef Gahi

Laboratory of Engineering Science, Ibn Tofail University, Kenitra, Morocco

ARTICLE INFO

Keywords:
DataOps
Agile methodologies
DevOps
Data Integration
Data Governance
Data Quality
Collaborative models

ABSTRACT

The rapidly evolving field of Data Operations (DataOps) is essential for enhancing data management within large-scale enterprises. However, persistent challenges, such as inefficiencies in data integration, delivery, and governance, limit its potential impact. These obstacles hamper the seamless implementation of DataOps strategies, slowing down operational processes and affecting organizational performance in data-driven environments. To address these issues, this research employs a systematic literature review, analyzing contributions from 2004 to 2024, to identify relevant solutions and innovations. The study highlights the value of frameworks, methodologies, and advanced technologies—such as automation, cloud platforms, and continuous delivery pipelines—that have reshaped the DataOps landscape. These contributions guide enterprises toward best practices in data strategy and foster improved collaboration across business and IT teams. Building on this analysis, our research also proposes a personal framework designed to offer a comprehensive approach to DataOps strategy. This framework integrates key insights from existing research and provides practical recommendations and best practices to streamline workflows, enhance data governance, and align IT operations with business goals. The enhanced DataOps practices derived from our framework demonstrate significant potential to boost operational efficiency, accelerate decision-making processes, and unlock new growth opportunities. Furthermore, the implementation of such practices sets the foundation for future innovations in data management and offers a path forward for organizations seeking sustainable, long-term value.

1. Introduction

The exponential surge in data generation and sharing has redefined the digital landscape, marking the emergence of the Big Data (BD) era. This era profoundly impacts every sector, transforming how information is produced, managed, and analyzed. Across businesses, governments, and scientific communities, data has grown exponentially, encompassing structured, semi-structured, and unstructured formats, alongside multimedia content like videos, images, and audio. These originate from diverse sources such as social media, IoT devices, and sensor networks (Agarwal & Dhar, 2014, Gandomi & Haider, 2015, Chen et al., 2014, Hashem et al., 2015). The daily global production of data, primarily unstructured, amounts to around 2.5 quintillion bytes, with over 40 zettabytes (Dobre & Xhafa, 2014, Gantz & Reinsel, 2012).

This complex and diverse data landscape, often termed the Data Deluge, underscores the importance of BD, which has been recognized as one of Gartner's Top 10 Strategic Technology Trends for 2013 and a critical trend for the subsequent five years (Savitz, 2012). BD, akin to the 'Digital Oil' of the 21st century, is considered a vital resource fueling modern economies and industries (Yi et al., 2014; Berners-Lee & Shadbolt, 2011). Its foundation lies in datafication, where human intelligence is digitized and disseminated (Mayer-Schönberger & Cukier, 2013). While this drives exponential data growth, it also introduces challenges in managing, processing, and deriving actionable insights from vast datasets. Anticipated advancements suggest that future data will predominantly be generated through automated machine interactions across interconnected networks (Van Dijck, 2014). Unlocking BD's potential requires effective management and advanced analytical tools capable of transforming raw data into valuable insights, enabling timely responses to opportunities and challenges (Chen et al., 2013; Fannouch et al., 2024; Yu et al., 2022). However, traditional computing infrastructures, such as relational databases and data warehouses, struggle to cope with the unprecedented velocity and volume of modern data streams, rendering them increasingly obsolete (Jukić et al., 2015; Bosch & Olsson, 2023; Avram, 2014). In response, advanced Big Data analytics technologies have emerged, bridging these gaps. These tools empower

* Corresponding author.
E-mail address: aymen.fannouch@uit.ac.ma (A. Fannouch).

https://doi.org/10.1016/j.jjimei.2025.100321

Available online 2 February 2025

organizations to process massive datasets, derive deep insights, foster innovation, and achieve competitive advantages in an increasingly data-driven world.

Building on this foundation, advanced Big Data (BD) analysis technologies, such as NoSQL Databases, BigQuery, MapReduce, Hadoop, Kafka, WibiData, and Skytree, enhance insights and improve decision-making across sectors like healthcare (Raghupathi & Raghupathi, 2014; Al-Quraishi et al., 2024; Arigbabu et al., 2024), economic productivity (Manyika et al., 2011; Liu et al., 2024), natural disaster prediction (Yi et al., 2014), and resource management in industries like offshore gas fields (Nemoto et al., 2023). In healthcare, the integration of predictive analytics with BD technologies, as highlighted in (Al-Quraishi et al., 2024), has opened new frontiers in personalized medicine by leveraging tools like HDFS, Spark, and Hive to analyze diverse datasets, including Electronic Health Records (EHRs), genomic data, and real-time data from wearable devices. Academic research highlights various analytical methods, including experiments, simulations, algorithms, and mathematical modeling, to address the challenges posed by Big Data (BD). These studies show that when managed effectively, BD can provide groundbreaking insights for businesses (Jukić et al., 2015). However, realizing these benefits requires a shift from traditional data management to more agile and collaborative frameworks like DataOps. While BD offers immense opportunities such as value creation, enhanced business intelligence, and improved resource management (Brown et al., 2011; Chen & Zhang, 2014; Kumar et al., 2013), it also presents challenges, including data integration complexities, skill shortages, security and privacy concerns, and infrastructure limitations (Gandomi & Haider, 2015; Kim et al., 2014; Barbierato et al., 2014; Jiang et al., 2015). Experts argue that unlocking BD's full potential demands scalable technologies capable of managing vast datasets beyond simple statistical analysis (Zhang et al., 2015). To address these complexities, methodologies from software development, particularly DevOps, have been adapted to data analytics. DevOps accelerates software delivery through cultural ideologies, methodologies, and tools (Fischer et al., 2020; Artac et al., 2017; Lwakatare et al., 2019), and its principles have been applied to data analytics to reduce time and effort in data acquisition (Kontostathis, 2017), streamline transformation phases (Jabbari et al., 2016), and support script maintenance via Continuous Integration (CI) servers (Bass et al., 2015) and centralized repositories (Kim et al., 2021). However, data analytics differs significantly from software development, especially regarding stateful data stores and strict process controls, which align more closely with data integration and business analysis (Ward-Riggs). This need for tailored solutions led to the emergence of DataOps (Yin et al., 2023).

DataOps, introduced in 2015, integrates people, processes, and technology to enable agile and automated data management. Drawing from DevOps, Agile Methodology, and Lean Manufacturing, it fosters collaboration, reduces silos, and enhances decision-making by leveraging continuous delivery, iterative workflows, and waste-minimization strategies (Fig. 1). DataOps has been formalized through its manifesto, which outlines 18 foundational principles emphasizing teamwork, experimentation, and interdisciplinary cooperation (The DataOps Manifesto; Dingsøyr et al., 2012; Hoda et al., 2012).

By automating the entire data lifecycle—from data cleaning and integration to analysis and reporting (Bergh et al., 2019)—DataOps maximizes business value while transforming data science into a streamlined operation. Despite being an emerging field compared to DevOps, DataOps has proven transformative, improving data quality (Elouataoui et al., 2024, 2023, 2023), streamlining data flow (Shojaee Rad & Ghobaei-Arani, 2024; Heck, 2024), and accelerating analytics processes (Bergh et al., 2019; Pelluru, 2022). These advancements foster collaboration, remove barriers, and support the efficient development of production-ready data products (Atwal, 2019; Xu et al., 2023). Ultimately, businesses adopting DataOps benefit from enhanced operational efficiency, innovation, and agility, enabling them to remain competitive in the evolving data-driven economy (Thusoo & Sarma, 2017; Xu et al., 2022).

1.1. Research questions

Research Question (RQ)	Description
RQ1: What are the key technological innovations and methodological approaches that advance DataOps practices?	To explore the latest technological innovations and methodologies that are transforming DataOps practices, focusing on how they contribute to more efficient data management and analysis in large enterprises.
RQ2: How can DataOps practices enhance data integration, delivery, and governance in large-scale enterprises?	To investigate the ways DataOps optimizes data integration, delivery, and governance, improving the overall operational efficiency and data quality within large organizations.
RQ3: What empirical evidence exists to support the effectiveness of DataOps in improving operational efficiency and decision-making?	To examine case studies and empirical research that provide evidence for the impact of DataOps on streamlining operations, enhancing decision-making, and improving data-driven workflows.

Section 2 provides an overview of DataOps, including its principles and objectives (Section 2.1), its evolution and core components (Section 2.2), and the challenges in its implementation (Section 2.3). Additionally, Section 3 outlines the methodology used, detailing the processes described in its subsections. Furthermore, Section 4 focuses on analyzing existing contributions and frameworks, highlighting their strengths and limitations. Section 5 consolidates best practices from existing frameworks and addresses identified gaps. Finally, the last section presents the conclusion.

2. Background

This section examines the foundational principles and objectives of DataOps, along with its evolution and implementation challenges, providing a comprehensive understanding of its role in optimizing data management practices and addressing the complexities of modern data operations.

2.1. Principles and objectives of DataOps

DataOps combines principles from DevOps and Agile methodologies, tailored to optimize the data lifecycle and foster collaboration across teams. Its foundational practices—illustrated in Fig. 2, such as continuous integration (CI), continuous delivery (CD), data quality governance, and scalability—streamline data operations from acquisition to reporting, enabling quicker and more reliable insights for decision-making (Thusoo & Sarma, 2017). By integrating CI/CD, DataOps ensures seamless data pipeline updates and real-time integration (Bergh

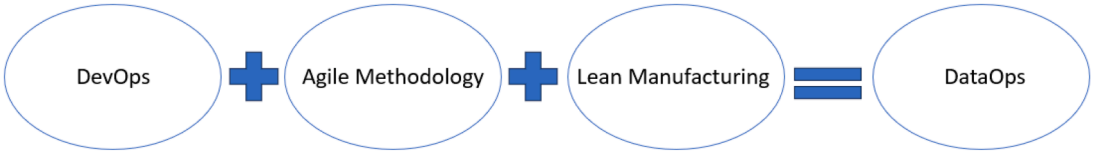

Fig. 1. The formation of DataOps from DevOps, Agile, and lean manufacturing.

et al., 2019).

Agile principles in DataOps provide organizations with the flexibility to adapt data processes to changing business needs, enhancing responsiveness and operational efficiency (Ereth, 2018). Automation of routine tasks, such as data validation and monitoring, reduces human error and promotes scalability (DeBellis et al., 2023). Cross-functional collaboration ensures consistent data flow and reliable model deployment, further optimizing organizational data management (Bergh et al., 2019). A key focus of DataOps is maintaining high data quality through continuous monitoring and alignment with governance frameworks, which are particularly vital in regulated industries (Elouataoui et al., 2024, 2023, 2023). By leveraging feedback loops, DataOps ensures ongoing improvement in data pipeline performance, aligning operational objectives with business goals (DeBellis et al., 2023). Its scalability empowers organizations to manage diverse data volumes and navigate complex technological environments (Tamburri et al., 2022). Beyond operational principles, DataOps emphasizes strategic objectives, transforming decision-making through enhanced transparency and accountability. It provides end-to-end visibility into data pipelines, fostering trust in data integrity and enabling strategic business decisions (Tamburri et al., 2022). By integrating accurate and high-quality data into business intelligence (BI) systems, DataOps maximizes data's business value, delivering actionable insights (Soares, 2010; Thusoo & Sarma, 2017). Through its agile approach, DataOps accelerates time-to-insight, enabling organizations to swiftly respond to market changes and seize emerging opportunities (Ereth, 2018). This adaptability not only optimizes data flow but also positions DataOps as a driver of sustained innovation and strategic agility.

As DataOps principles create the foundation for agile and efficient data practices, their evolution into a more structured discipline will be discussed in next section.

Fig. 2. Key DataOps principles for optimized data management.

2.2. Evolution and Core Components

The evolution of DataOps reflects a progressive shift in how data management practices have adapted to meet the challenges of Big Data. As shown in Fig. 3, key milestones highlight advancements across distinct periods:

Building on this comprehensive approach, (Munappy et al., 2020) provides a framework for the evolution of DataOps shows a progression from ad-hoc manual data analysis to fully automated data lifecycle management. Fig. 4 illustrates this progression, beginning with on-demand reports, advancing to semi-automated pipelines, followed by agile data science for iterative insights, and culminating in automated testing and monitoring for data quality. The final stage integrates DevOps, CI/CD, and agile practices, establishing a collaborative, automated data management system that eliminates data silos.

The evolving definition of DataOps highlights its pivotal role in aligning data operations with evolving business needs, fostering continuous improvement, and seamlessly integrating technological, organizational, and environmental factors. As depicted in Fig. 5, the framework identifies three essential components—Technology, Organization, and Environment—which collectively drive dynamic capabilities and deliver business value.

Technology, Organization, and Environment form the core components of DataOps, providing the foundation for efficient data

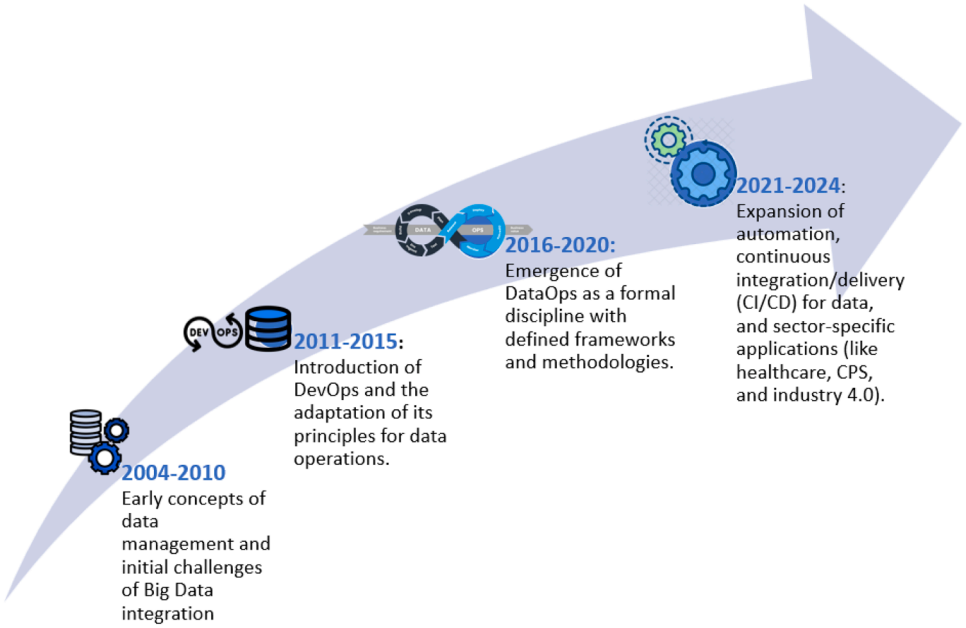

Fig. 3. Timeline of DataOps evolution and key advancements.

Fig. 4. Evolution of DataOps.

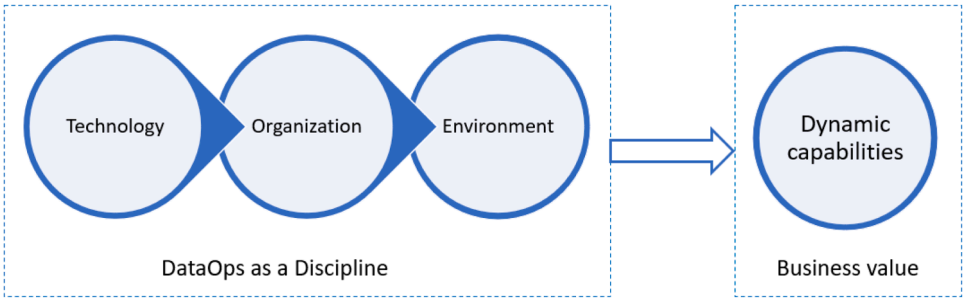

Fig. 5. Core components of DataOps discipline.

management. Technology supplies the tools and infrastructure for automation, while organizational alignment fosters collaboration and agility across departments. The environmental component addresses external factors like market trends, regulations, and industry influences.

When integrated effectively, these elements enhance dynamic capabilities, enabling organizations to adapt swiftly to changing conditions and achieve sustained business value. This holistic approach ensures DataOps remains adaptable, scalable, and aligned with organizational goals while addressing broader environmental challenges. Despite its advantages, implementing DataOps successfully still faces significant obstacles.

2.3. *Challenges in implementing DataOps*

The implementation of DataOps faces numerous challenges, including the need to align with evolving business goals and integrate siloed data across multiple platforms. These difficulties are further compounded by the complexities of managing unstructured data, which add significant layers of complexity to data operations (Rodriguez et al., 2020). As systems scale, ensuring the quality of data pipelines becomes increasingly challenging, with inefficiencies often stemming from manual processes, as illustrated in Fig. 6. Despite its transformative potential, automation remains underutilized, with many organizations dedicating insufficient resources to its development (DataKitchen). In sectors like exploration and production (E&P), these challenges are even more pronounced, given the need for specialized domain expertise and industry-specific knowledge (Jakobsen, 2022). Additional obstacles include fragmented data silos, pipeline scalability issues, heavy reliance on manual monitoring, and the absence of automated testing frameworks for ensuring data quality. Moreover, the integration of DataOps with DevOps often necessitates substantial organizational restructuring, adding further complexity to the implementation process (Munappy et al., 2020).

Another significant challenge to DataOps adoption is organizational resistance, often stemming from cultural inertia. This resistance manifests as hesitation among employees and stakeholders to embrace new methodologies or automation, driven by fears of job displacement, unfamiliarity with tools, or concerns over potential disruptions to established workflows. These cultural barriers are closely tied to resource constraints, including limited budgets, shortages of skilled personnel,

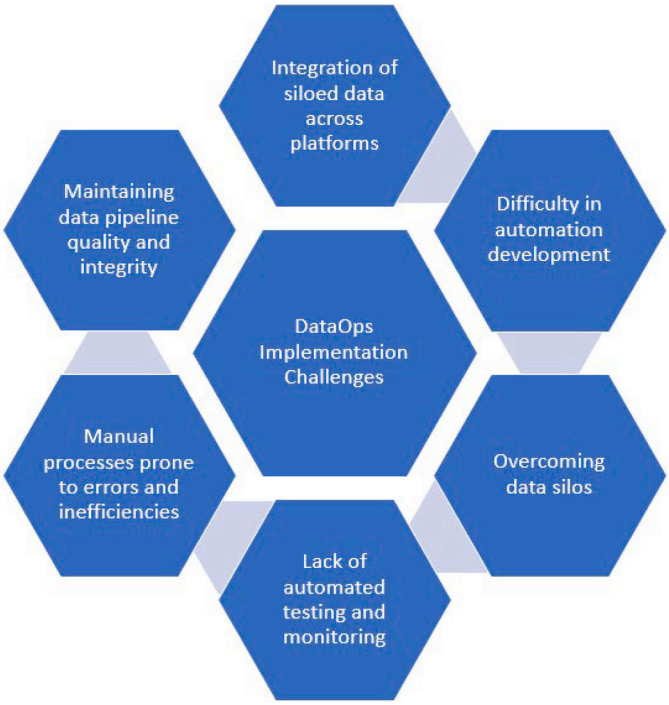

Fig. 6. DataOps implementation challenges.

and outdated technological infrastructure, which further exacerbate the difficulties of implementation. Together, these organizational and resource-related challenges create a significant roadblock to the widespread adoption of DataOps.

3. Methodology

The methodology section outlines the systematic approach used in this research to ensure rigor and consistency in reviewing relevant literature

3.1. Method

We conducted a systematic literature review to identify relevant research papers from databases and search engines, utilizing a carefully curated set of predefined search terms. This process was followed by a rigorous filtering phase, guided by established inclusion and exclusion criteria, to ensure the relevance and quality of the selected studies. Subsequently, we established a comprehensive set of parameters to compare the selected papers systematically. Finally, the filtered papers underwent a critical analysis to extract meaningful insights and identify key contributions to the field.

3.2. Search process

The search terms were carefully selected to reflect key concepts related to alignment and measurement. Using these predefined terms (Table 1), we conducted searches across specific research databases and search engines, focusing on Scopus, Web of Science, and Springer. The detailed search queries for these platforms are provided in Table 1.

Inclusion criteria for each subject area are outlined in Table 2, with relevance determined if a text satisfies any of the listed criteria (OR logic applied, not AND). Common exclusion criteria applied across all subjects include the following:

- The text is not fully accessible.
- The text is not written in English.

3.3. Data collection extraction and synthesis strategy

The systematic review process began with an initial search yielding a total of 770 records from three major databases: Scopus (300 records), Web of Science (120 records), and Springer (350 records) as of August 2024. After removing 250 duplicate records, 520 records remained for title and abstract screening. Through a thorough evaluation based on the inclusion and exclusion criteria, 420 records were excluded for not meeting the relevancy thresholds.

This left 100 papers for full-text assessment, during which 40 additional records were excluded due to reasons such as irrelevance to the research scope or failure to meet inclusion criteria. Ultimately, 60 studies were selected for detailed analysis Fig. 7. To ensure a comprehensive review, a snowballing approach was employed to identify additional relevant studies not captured during the initial search.

3.4. Analysis result

The presented figures provide a clear overview of the extracted articles in this study. Fig. 8 illustrates the types of publications contributing to the research, showing a mix of journal articles, conference papers, Book and Book Chapter. Fig. 9 reflects the thematic distribution of the articles, covering a wide range of topics, including foundational theories, framework development, challenges, technological innovations, and empirical case studies. Lastly, Fig. 10 highlights the evolving interest in DataOps over time, demonstrating an increasing trend in research output in recent years.

Building on these insights, Fig. 11 offers a quantitative analysis of the prominence of specific DataOps practices. This analysis provides valuable insights into how frequently key practices such as CI/CD pipelines, data version control, automated testing, metadata management, data governance, and monitoring & observability have been adopted or discussed over time. By quantitatively assessing these trends, the chart underscores

Table 1
Optimized search queries for each subject.

Search Subject	Search Engine	Query
DataOps Practices and Frameworks	Scopus, Web of Science, Springer	("DataOps" OR "data operations" OR "data lifecycle management" OR "collaborative data workflows") AND ("framework" OR "model")
Technological Innovations in DataOps	Scopus, Web of Science	("DataOps" OR "data operations") AND ("automation" OR "CI/CD pipelines" OR "metadata management" OR "data integration")
Data Governance and Policy in DataOps	Scopus, Springer	("Data governance" OR "data quality" OR "policy") AND ("DataOps" OR "data-driven workflows" OR "metadata standards")
Applications of DataOps in Enterprises	Scopus, Web of Science	("DataOps" OR "data operations") AND ("case studies" OR "enterprise data management" OR "real-world implementation")
Challenges and Limitations in DataOps Adoption	Scopus, Web of Science, Springer	("DataOps" OR "data operations") AND ("challenges" OR "limitations" OR "barriers to adoption" OR "organizational restructuring")

Table 2
Criteria for including search subjects.

Search Subject	Inclusion Criteria
DataOps Practices and Frameworks	- Proposes new frameworks or methodologies to enhance DataOps practices. - Explains the implementation of DataOps principles (e.g., CI/CD, Agile, Lean). - Focuses on integrating people, processes, and technologies. - Highlights strengths, drawbacks, and practical implications of frameworks.
Technological Innovations in DataOps	- Develops or evaluates technological tools specific to DataOps (e.g., workflow orchestration, automation). - Provides case studies demonstrating the application of CI/CD pipelines or automated data lifecycle management. - Discusses advancements in metadata management, data versioning, or real-time analytics. - Includes comparative analysis of tools for DataOps implementation.
Data Governance and Policy in DataOps	- Proposes governance models focused on data privacy, quality, and security within DataOps frameworks. - Explains integration of governance frameworks in data-driven workflows. - Focuses on regulatory compliance and organizational policy alignment. - Includes specific use cases of governance challenges in DataOps environments.
Applications of DataOps in Enterprises	- Describes real-world applications of DataOps in industries like healthcare, finance, and aviation. - Demonstrates measurable improvements in operational efficiency, data quality, or collaboration through DataOps practices. - Presents empirical evidence or outcomes of DataOps adoption. - Discusses challenges faced during DataOps implementation in enterprises.
Challenges and Limitations in DataOps	- Identifies organizational and technological barriers to DataOps adoption. - Explains the complexities of scaling DataOps practices across diverse environments. - Focuses on limitations in current research and gaps in knowledge. - Highlights recommendations for overcoming barriers in DataOps adoption.

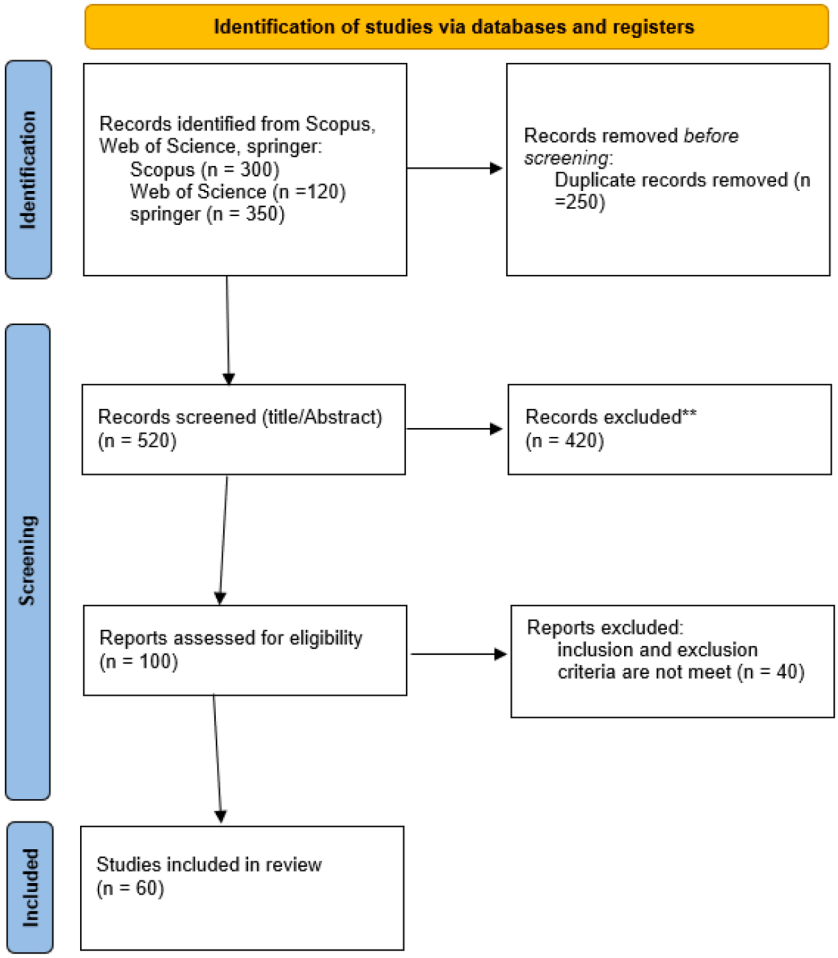

Fig. 7. The flowchart of selecting the research articles.

the evolving priorities and focus areas in DataOps research, offering a data-driven perspective that complements the broader discussion.

3.5. Data analysis techniques

Qualitative and quantitative analysis: After completing the selection of primary studies, we employed a combination of qualitative and quantitative analysis techniques to systematically synthesize and interpret the findings. The content of the selected studies was further classified according to predefined axes that align with the core aspects of our research focus:

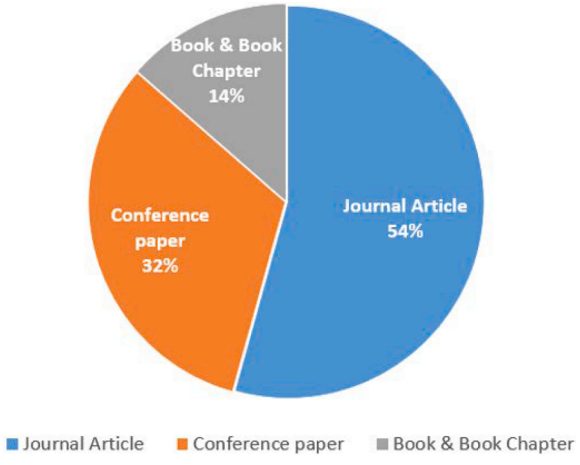

Fig. 8. Distribution of extracted articles by type.

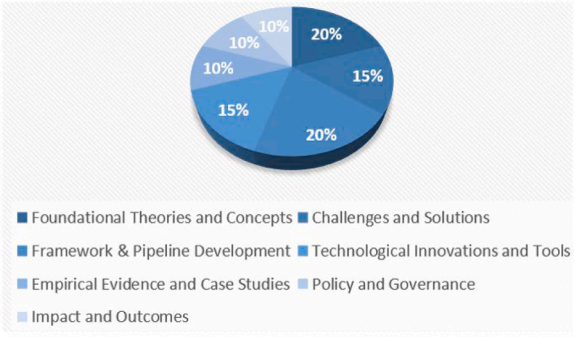

Fig. 9. Distribution of extracted articles by axes.

Thematic analysis: we utilized thematic analysis to identify recurring themes and patterns across the selected studies. This involved coding the qualitative data to categorize findings into key topics related to DataOps practices, such as data integration, quality, security, governance, and collaborative models. This approach ensured a structured and comprehensive organization of the data, allowing for the emergence of significant themes that aligned with the research objectives.

Classification by axes: The studies were organized into the following classification axes to provide a structured and focused analysis, as illustrated in Fig. 12:

Narrative synthesis: Given the diverse nature of the included studies, we employed narrative synthesis to integrate and summarize

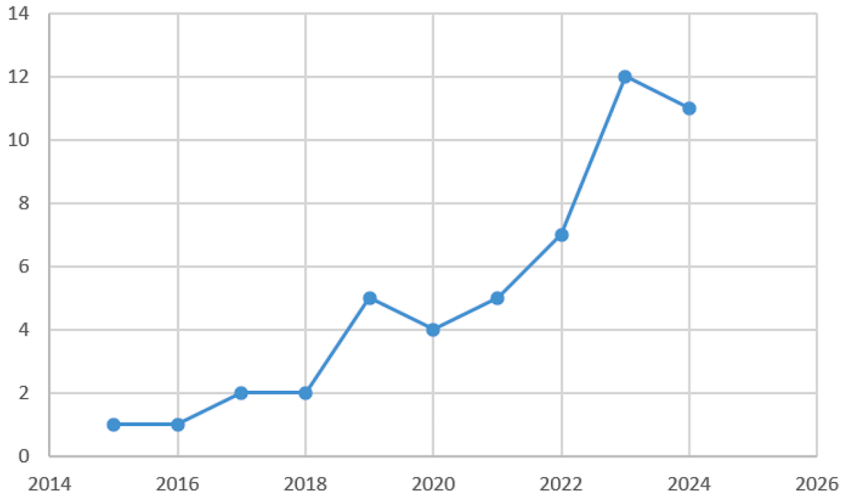

Fig. 10. Distribution of extracted articles by year.

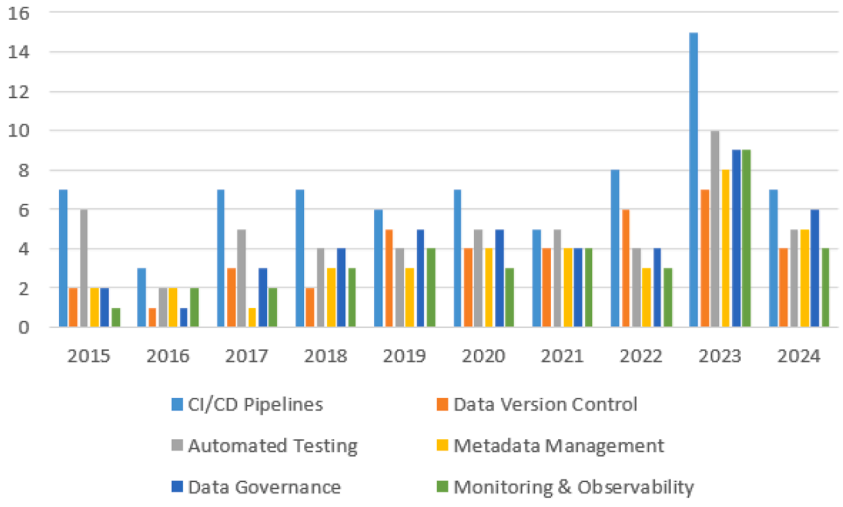

Fig. 11. Distribution of DataOps practices discussed over time.

findings from both qualitative and quantitative research. This method allowed us to construct a coherent narrative that linked various insights, providing a comprehensive understanding of how DataOps practices influence large-scale enterprises. In the following section, we will explore the frameworks extracted from the selected studies and provide detailed insights into their components and implications for DataOps practices.

4. Analysis framework

This section is structured into five key areas as illustrated in Fig. 13, each exploring critical dimensions of DataOps implementation and optimization across organizational contexts. It begins with a review of frameworks and methodologies that address data lifecycle challenges, followed by an analysis of technological innovations and tools enabling DataOps practices. The discussion then transitions to policy and governance frameworks essential for ensuring data quality, security, and compliance. Empirical evidence and case studies are then examined to showcase real-world applications and insights. Finally, a comparative analysis of existing frameworks highlights their strengths and limitations, offering a clearer understanding of areas for improvement and innovation. Together, these areas provide a structured understanding of best practices and innovative approaches that underpin successful DataOps initiatives.

4.1. Frameworks and methodologies

After exploring the methodology of our research, we turn our

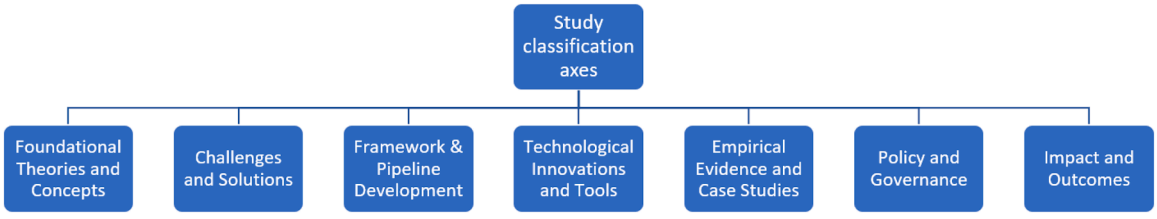

Fig. 12. Study classification axes for focused DataOps analysis.

Fig. 13. Analysis framework.

attention to the numerous contributions and frameworks that have been developed to address these issues comprehensively. These frameworks aim to streamline data operations, enhance collaboration, and improve the overall efficiency and effectiveness of data management processes. The DataOps lifecycle is a structured framework that enhances data management through collaboration, automation, and continuous improvement. Both (Bahaa et al., 2023) and (Rodriguez et al., 2020) provide valuable contributions to this lifecycle as illustrated in Figs. 14 and 15, respectively, tailored to different sectors but sharing common objectives of efficiency and quality.

In (Bahaa et al., 2023), the DataOps lifecycle is specifically adapted for the healthcare sector, addressing the complexities of managing sensitive and voluminous healthcare data. This framework outlines detailed stages such as defining the data domain, assembling a skilled team, data gathering, feature engineering, validation, modeling, testing, documentation, delivery, consumption, monitoring, and archiving or deletion. Each phase is meticulously designed to ensure data quality and improve collaboration, making it particularly well-suited to the healthcare environment. Conversely, the contribution from (Rodriguez et al., 2020) presents a DataOps lifecycle adapted to the software industry. While it includes many of the same stages—such as data gathering, modeling, and monitoring—the emphasis here is on integrating Agile, DevOps, and Lean Manufacturing principles. The stages in this lifecycle include planning, acquiring, transforming, storing, modeling, publishing, consuming, and monitoring, highlighting adaptability and rapid iteration as essential in the fast-paced software sector.

Despite their sector-specific adaptations, both frameworks demonstrate the versatility of DataOps in optimizing data operations. The healthcare-focused lifecycle (Bahaa et al., 2023) is more methodical, reflecting the regulated and detailed nature of healthcare data, while the software-focused framework (Rodriguez et al., 2020) stresses agility and iterative processes, suited to dynamic development environments. One significant contribution to the DataOps framework comes from (Mainali et al., 2021), where Fig. 16 illustrates a broader framework that aligns tools, people, and processes to convert data into insights. This framework integrates key components such as workflow orchestration, collaboration, deployment automation, and data governance, supporting the entire data lifecycle from collection to publishing. It is designed to streamline processes and ensure efficiency in data operations.

Building on this, (Garriga et al., 2021) presents a DataOps framework for cyber-physical systems (CPS), optimizing airport passenger flow predictions (Fig. 17). It integrates data sources, transformation processes, and machine learning for real-time insights, enhancing efficiency and passenger satisfaction. The use of machine learning techniques in this DataOps framework allows for continuous optimization, ensuring that the airport can respond swiftly to operational changes and enhance passenger satisfaction. This seamless integration of data sources, analytics, and visualization demonstrates the potential of DataOps in complex environments like CPS, where real-time decision-making is crucial.

In the enterprise context, (Zahid et al., 2018) presents a dependable architecture for Big Data Analytics pipelines, structured in seven layers, each utilizing specific tools to address key challenges. As illustrated in Fig. 18, The connection layer integrates data from various sources using APIs like PyMongo for MongoDB. The integration layer gathers and merges data into a master database, supported by Redis for metadata management. In the static layer, batch processing is handled by Hadoop and Spark, while the dynamic layer processes real-time data streams with Apache Kafka and Spark Streaming. The serving layer employs NoSQL databases such as MongoDB, Redis, HDFS, or HBase for storing and querying results. The interface layer uses Flask for implementing RESTful APIs, and the dashboard layer provides visualization through BI tools like Tableau or open-source libraries such as Plotly, ensuring seamless data flow, reliability, and scalability.

For **semantic data integration**, (Pinkel et al., 2015) proposes a **DataOps process**, illustrated in Fig. 19, consisting of three key stages—**Data Source, Mapper, and Post Processor**—designed to access, map, and transform data into RDF. This process offers a flexible solution for managing both **semantic and non-semantic data** from diverse sources.

Fig. 14. DataOps lifecycle framework.

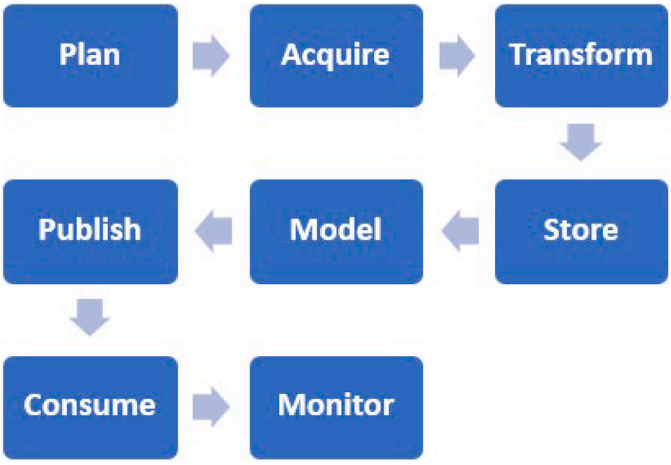

Fig. 15. DataOps lifecycle model.

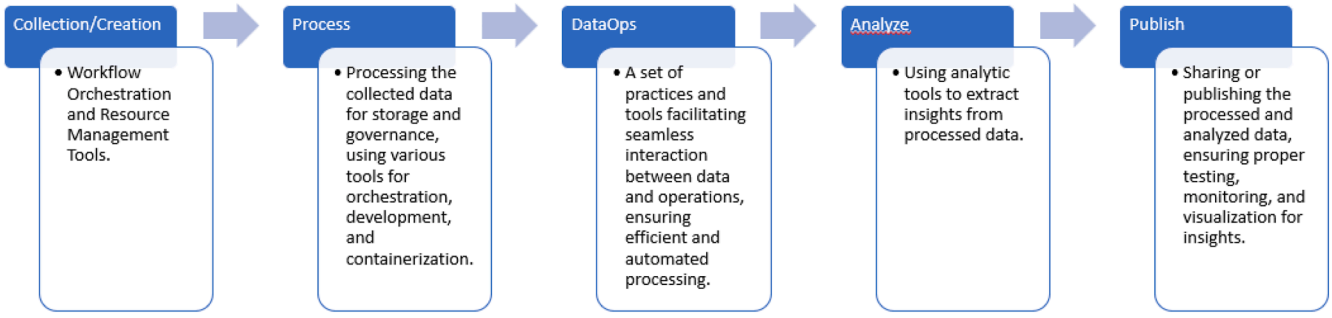

Fig. 16. End-to-end DataOps framework: from collection to publishing.

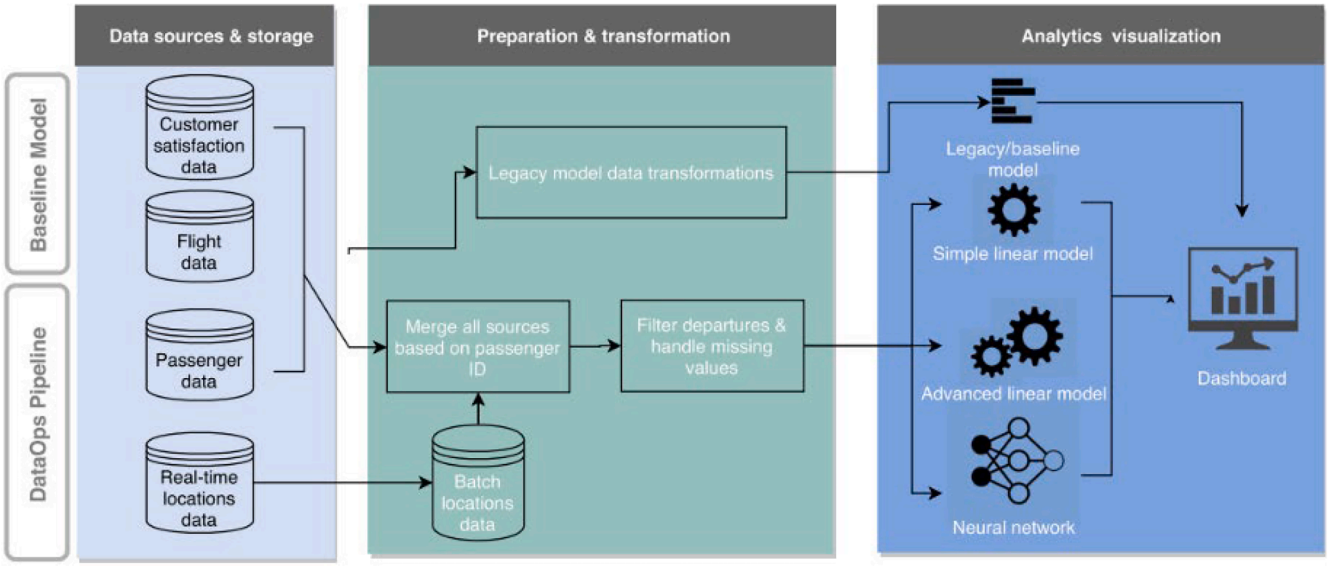

Fig. 17. Data pipeline for passenger flow prediction by (Garriga et al., 2021).

Data Source: Facilitates low-level access to raw data.

Mapper: Translates data into RDF triples, extracting and manipulating it for further use.

Post Processor: Merges data (reconciliation) and improves overall data quality.

This streamlined process ensures efficient data handling and transformation, supporting seamless integration across varied datasets.

Continuing the application of DataOps, (Tamburri et al., 2020) introduces two Big Data pipelines for labor market intelligence, focusing on skill extraction from job vacancies and resumes using machine learning models. As illustrated in Fig. 20, these pipelines automate processes like data selection, transformation, and interpretation to provide actionable insights and support decision-making in the labor market.

Furthermore, both (Tu et al., 2023) and (Bayram et al., 2023)

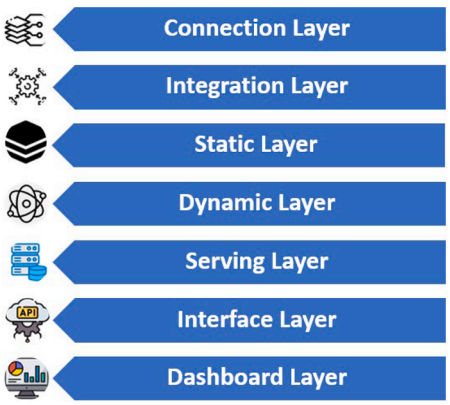

Fig. 18. Proposed dependable architecture for enterprise BDA pipelines.

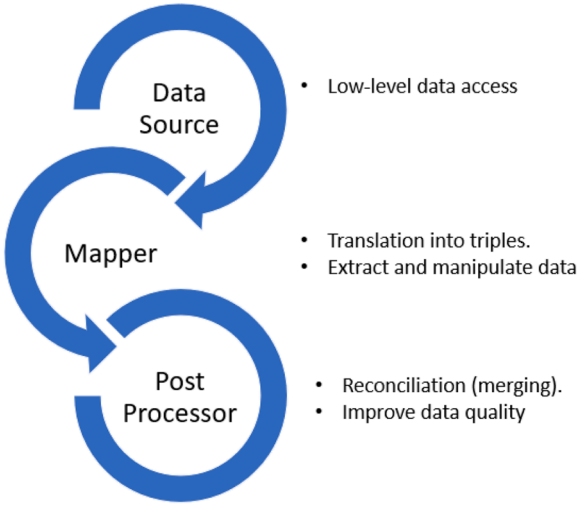

Fig. 19. DataOps process.

provide key contributions to advancing data quality management within DataOps, but they target distinct aspects of the data lifecycle. (Tu et al., 2023) introduces the Auto-Validate by-History (AVH) algorithm, a groundbreaking method for automating data validation in recurring pipelines. This contrasts sharply with existing tools such as TensorFlow Data Validation (TFDV) and Amazon Deequ, offer manual ways for engineers to define data quality constraints using declarative domain-specific languages (DSLs). While these tools improve data validation, they require manual programming of constraints for each column, making them difficult to scale across thousands of columns and pipelines in enterprise settings. The limitations of these existing methods, which rely on static, predefined configurations, become apparent when dealing with dynamic, recurring data pipelines. TFDV and Deequ require significant manual intervention, which limits their scalability and flexibility. In response, (Tu et al., 2023) presents the AVH algorithm as illustrated in Fig. 21 as a superior alternative that automates the entire data validation process by leveraging historical pipeline data. AVH dynamically programs data quality constraints based on past executions, significantly reducing the manual effort required and enhancing scalability.

On the other hand, (Bayram et al., 2023), As illustrated in Fig. 22, (Bayram et al., 2023) introduces the DQSOps framework, a powerful tool designed for scoring and monitoring data quality (DQ) in real-time data streams. This framework integrates machine learning models with traditional validation methods to provide continuous evaluation of data quality, ensuring that anomalies are detected and addressed promptly.

The DQSOps process begins by ingesting data from various sources into data windows, which are processed through a mutant simulator to simulate potential issues and prepare the data for further evaluation. A method activator triggers the necessary algorithms, applying multiple data quality dimensions (DQ_1, DQ_2, DQ_n) that measure key aspects such as completeness, accuracy, and consistency. The ML model predicts data quality scores in real-time, comparing them against a ground-truth DQ score repository for validation. If discrepancies are detected, retraining signals are initiated to fine-tune the model. A test oracle verifies these predictions, feeding them into performance evaluations that measure the system's accuracy and effectiveness. While DQSOps excels at stream-based assessments, offering rapid and adaptive responses to changing data conditions, the framework also integrates with AVH, which ensures long-term data consistency and scalability for recurring batch processes. Together, these contributions form a comprehensive toolkit for managing data quality across diverse operational contexts, balancing real-time precision with long-term reliability.

Finally, (Bergh et al., 2019) presents a comprehensive DataOps architecture, illustrated in Fig. 23, which integrates multiple tools such as Jenkins, Kubeflow, and Terraform for orchestrating, monitoring, and deploying data workflows across both on-premise and cloud environments. This architecture enables seamless collaboration by incorporating tools like Git for version control, Vault for managing secrets, and Okta for authentication. It supports both production and testing environments, ensuring consistent data processing and governance. The deployment pipeline leverages tools such as Tapestry and ArgoCD to automate environment creation and management. Additionally, the platform integrates metrics monitoring via tools like Tableau, which allows for real-time tracking of performance and data quality. Overall, this architecture enhances collaboration, accelerates cycle times, and ensures the robustness of data governance, ultimately benefiting organizational data operations by improving scalability, reliability, and security.

In summary, these contributions illustrate how DataOps frameworks can be adapted to various sectors, such as CPS, labor market intelligence, and enterprise environments, while addressing key challenges like data quality, integration, and workflow automation.

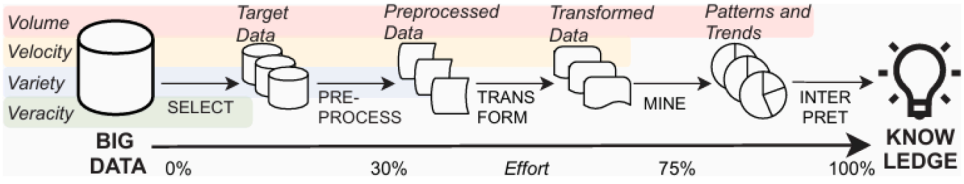

Fig. 20. Big data enables societal and labor market intelligence by (Tamburri et al., 2020).

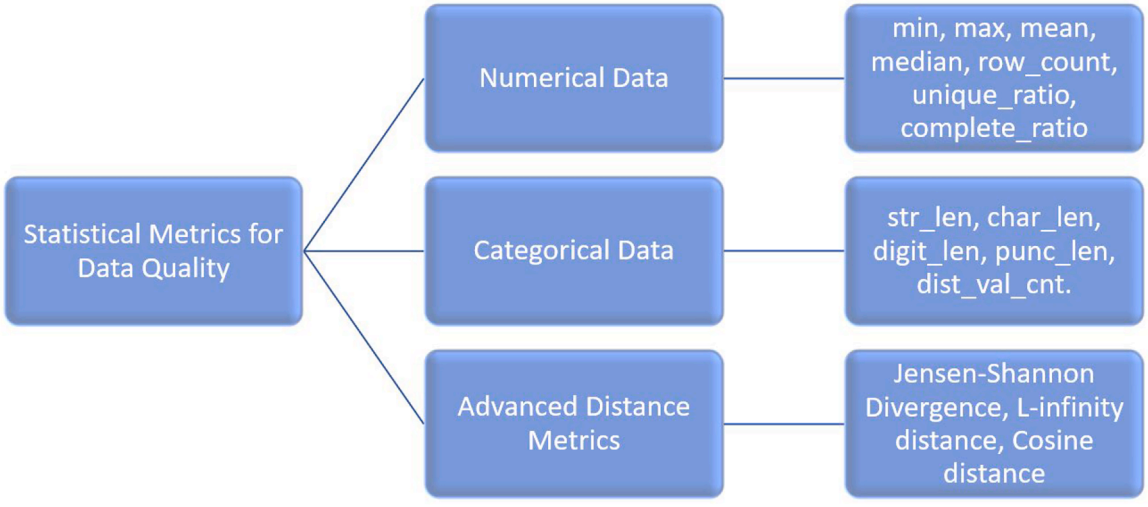

Fig. 21. Statistical metrics for data quality in AVH framwework.

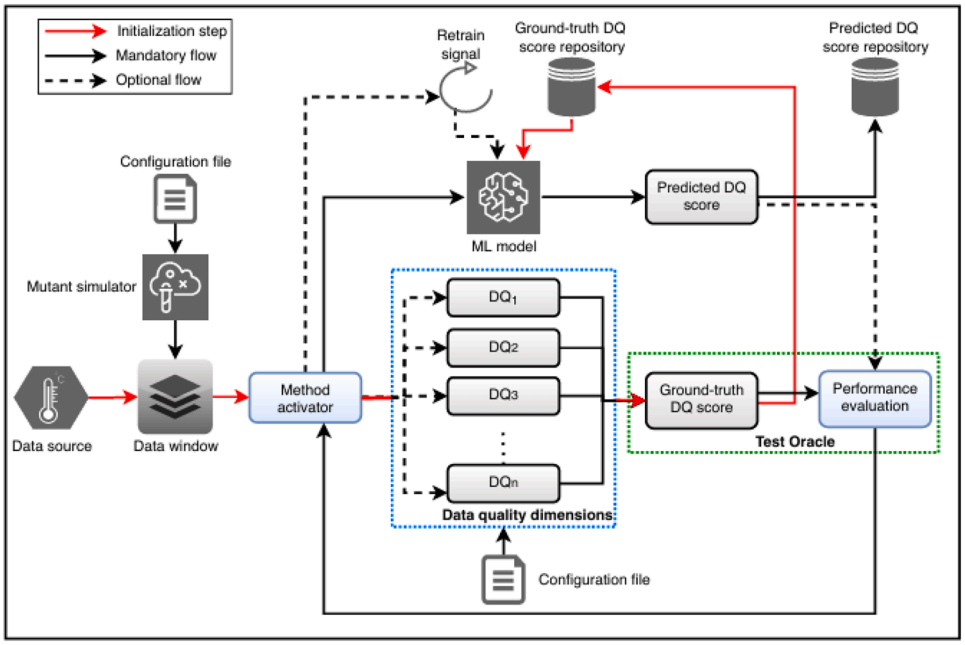

Fig. 22. DQSOps framework by (Bayram et al., 2023).

4.2. Technological innovations and tools

The rapid evolution of technology has driven the advancement and adoption of DataOps practices (Kannan & Jain, 2023). Cutting-edge innovations and specialized tools form the backbone of DataOps, enabling organizations to efficiently manage, integrate, analyze, and visualize data—the core pillars of this methodology.

Effective DataOps implementation relies on strategically utilizing a diverse array of technologies optimized for areas such as workflow orchestration, data governance, and real-time analytics. These tools enhance operational efficiency and promote seamless collaboration across teams and systems. Table 3 highlights essential tools used in DataOps environments, detailing their roles in the data lifecycle and key attributes like complexity and usability, providing a comprehensive guide for selecting tools that effectively support DataOps workflows.

In summary, the tools discussed are essential for the successful implementation of DataOps, contributing to the efficiency, scalability, and reliability of data operations. These tools enhance collaboration, streamline processes, and ensure high-quality data products. Their impact extends to faster decision-making, adaptability to changing

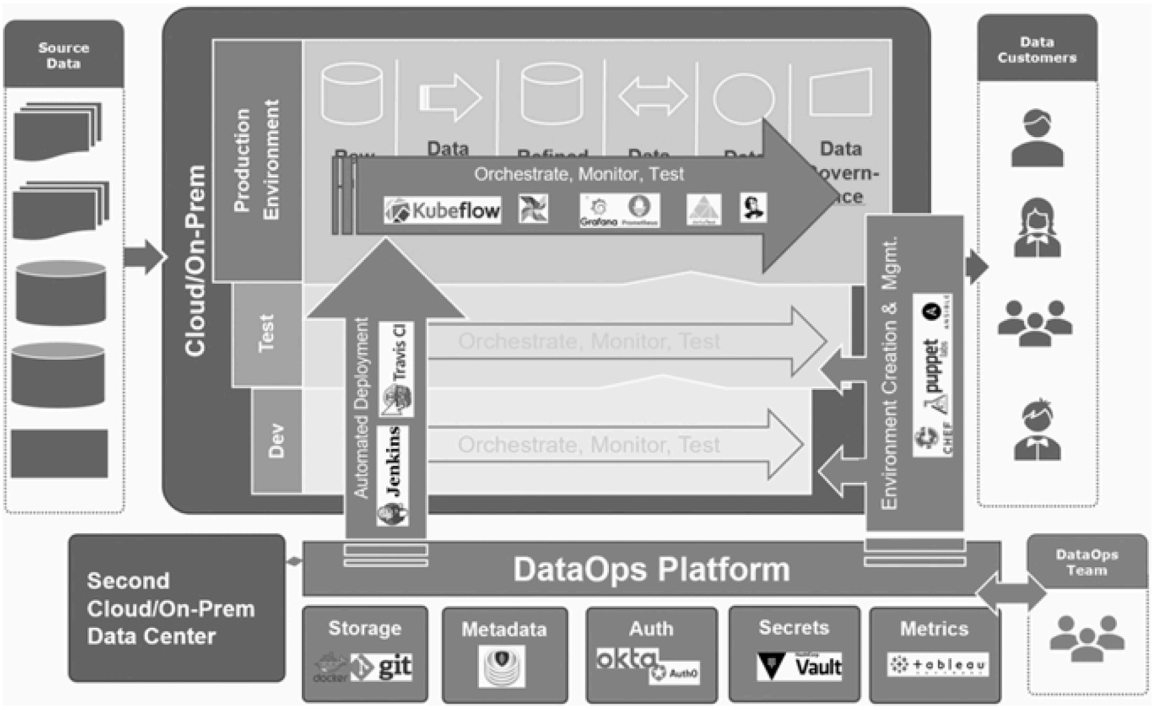

Fig. 23. DataOps data architecture with tools by (Bergh et al.,2019).

Table 3
Key DataOps tools by category.

Category	Tool	Complexity	Usability
Workflow Orchestration	Airflow	High	Medium
	Apache Oozie	High	Medium
	Reflow	High	Low
	Data Kitchen	Low	High
	BMC Control-M	Medium	Medium
	Argo Workflows	High	Low
	Apache NIFI	Medium	Medium
Testing & Monitoring	iCEDQ	Low	High
	Data Band	High	Low
	RightData	Medium	Medium
	Naveego	High	High
	DataKitchen	High	Medium
	Enterprise Data Foundation	High	Low
Deployment Automation	Jenkins	Medium	High
	Circle CI	Medium	Medium
	GitLab	Medium	Medium
	Travis CI	Medium	High
	Atlassian Bamboo	Low	High
Data Governance	Apache Atlas	High	Medium
	Talend	Medium	Medium
	Collibra	Low	Low
	IBM	Medium	High
	OvalEdge	Low	High
Code & Data Versioning	GitLab	Medium	High
	Github	Medium	High
	DVC	Medium	High
	DockerHub	Medium	High
Analytics & Visualization	Tableau	Low	Medium
	Power BI	Low	Medium
	QlikView	Low	Medium
Collaboration & Communication	Slack	Low	High
	Jira	Low	High
	Trello	Low	High

business needs, and a stronger data management framework, empowering organizations to fully leverage their data assets for innovation and competitive advantage. Next, we will explore data governance within DataOps, focusing on best practices that ensure data quality, security, and compliance.

4.3. Policy and governance

Data governance is crucial for managing and protecting data, covering aspects like privacy, security, and regulatory compliance. At different levels, it focuses on internal practices (micro), shared organizational principles (meso), and national or international regulations (macro) (Torre-Bastida et al., 2022). Organizations implement governance using established frameworks such as DAMA-DMBOK (International, 2017), TOGAF (Josey, 2016), COBIT (De Haes et al., 2013), and the DGI Data Governance Framework (Thomas, 2006). These frameworks ensure data quality, security, and compliance. However, as data ecosystems grow, traditional models are insufficient, requiring dynamic approaches like continuous metadata practices for real-time data flows (Underwood, 2023). In advanced applications such as AI-native 6G networks, DataOps and MLOps play a key role in privacy-aware data collection and governance (Saimler et al., 2024).

One significant advancement in this area the European Industrial Data Space (EIDS) framework was developed to ensure data sovereignty in the European industrial sector (Mertens et al., 2022). It introduces key governance policies, including federated data governance, allowing companies to retain data control while participating in shared ecosystems. EIDS also implements standardized certification processes based on the International Data Spaces (IDS) Reference Architecture, ensuring security and compliance, and enforces data usage control via smart contracts and blockchain. The trust framework ensures only certified entities participate in data exchanges. Companies like Volkswagen and Gestamp have successfully implemented EIDS in the BOOST 4.0 project.

While EIDS addresses industrial governance challenges, healthcare faces unique issues, such as data privacy, security, regulatory compliance (e.g., HIPAA), and ethical concerns (Tse et al., 2018). A framework for Big Data Algorithmic Systems (BDAS) (Janssen et al., 2020) focuses on ensuring the integrity of dynamic data streams used in critical decision-making systems. This governance model integrates regulations, cultural norms, and organizational policies. Key components include compliance with data protection laws, promoting professional norms, and developing policies for risk assessment, audits, and data quality management. The framework includes practical mechanisms for continuous data quality assessment, bias detection, and pattern recognition. It emphasizes the need for regular sampling and validation to prevent large-scale failures in algorithmic processing. Moreover, the framework supports processes for handling complaints, learning from feedback, and communicating decisions effectively to stakeholders, ensuring accountability and transparency. This comprehensive approach to data governance not only addresses technical and regulatory requirements but also fosters a culture of responsibility and ethical behavior, which is crucial for the sustainable and trustworthy development of AI technologies.

The transition from traditional to cloud-based systems introduces complexities in data governance, particularly in managing data quality, security, and compliance. The study (Al-Ruithe et al., 2019) provides a comprehensive review of both non-cloud and cloud data governance, identifying key challenges and critical success factors. It emphasizes the need for more rigorous strategies to address issues like data security, privacy, and regulatory compliance in cloud environments, as cloud data governance remains under-researched. Similarly, securing multi-tenant cloud environments, as highlighted in (Hashim & Hussein, 2024), presents unique challenges due to resource sharing among tenants, which exposes systems to vulnerabilities like data breaches, side-channel attacks, and insecure APIs. This underscores the importance of adopting advanced governance practices tailored to multi-tenant architectures, integrating countermeasures such as enhanced isolation, encryption, and access control mechanisms. These measures not only mitigate risks but also align with broader governance goals, ensuring compliance with frameworks like GDPR and HIPAA. Together, these insights offer valuable perspectives for developing robust governance frameworks capable of addressing the growing demands of modern cloud-based ecosystems. Further advancing the discussion on governance, (Soares, 2010) introduces the IBM Data Governance Unified Process, a comprehensive framework designed to enhance business value through structured data governance practices using IBM software. As illustrated in Fig. 24, the framework consists of several key steps and optional tracks, each focusing on different aspects of data governance. These steps include defining the business problem, obtaining executive sponsorship, conducting a maturity assessment, building a roadmap, and establishing an organizational blueprint. Additionally, the framework emphasizes the importance of building a data dictionary, understanding data, creating a metadata repository, setting metrics, implementing master data governance, managing analytics, ensuring security and privacy, overseeing the data lifecycle, and measuring results. This structured approach aims to ensure comprehensive governance across all stages of data management.

(Torre-Bastida et al., 2022) outlines a data governance framework for Big Data ecosystems as depicted in Fig. 25, addressing intra- and inter-organizational needs. It focuses on five key areas: data ownership, trust and security, data as a revenue-generating asset, data quality and traceability, and ethical management of personal data. The framework ensures comprehensive governance across all levels of data management.

In conclusion, strong policy and governance frameworks are crucial for ensuring DataOps practices meet regulatory and organizational standards. By focusing on data privacy, security, quality, and ethics, these frameworks protect sensitive information, maintain data integrity,

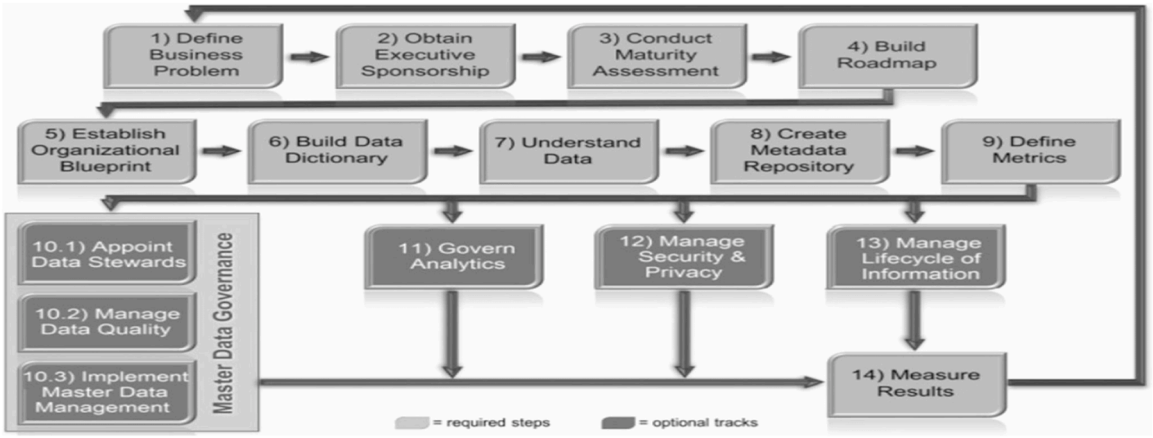

Fig. 24. IBM data governance framework by (Soares, 2010).

and promote accountability and transparency. Various models and strategies contribute to achieving effective data governance.

4.4. Empirical evidence and case studies

To validate the theoretical frameworks, methodologies, and tools discussed, exploring empirical evidence and real-world case studies is crucial. These examples provide insights into how DataOps practices are applied across industries, revealing practical challenges, successes, and lessons learned.

A notable case study by (Garriga et al., 2021) proposed a DataOps-driven methodology for predicting passenger flow, integrating data from sources like Wi-Fi and sensors into a real-time pipeline. This improved prediction accuracy by 60 %, enhancing resource planning and passenger satisfaction. Another study by (Zahid et al., 2018) focused on a dependable architecture for enterprise BDA pipelines, using a lambda architecture to handle both batch and real-time data. Implementations across five projects—including deep-learning trade prediction and data encryption platforms—demonstrated improvements in scalability and reliability, especially with cloud services like AWS for auto-scaling.

Further, (Tamburri et al., 2020) introduced a pipeline for analyzing job vacancies and resumes using Big Data and machine learning. Validated with Dutch-Flemish labor market data, models like BERT and a Simple Neural Network were tested, with BERT achieving the highest accuracy (0.816) and a high recall of 0.94, indicating precise skill prediction. The AVH framework (Tu et al., 2023), applied to 2000 Microsoft pipelines, showed superior performance compared to commercial solutions, automating data quality management and enhancing reliability. The DQSOps framework (Bayram et al., 2023), implemented in a steel

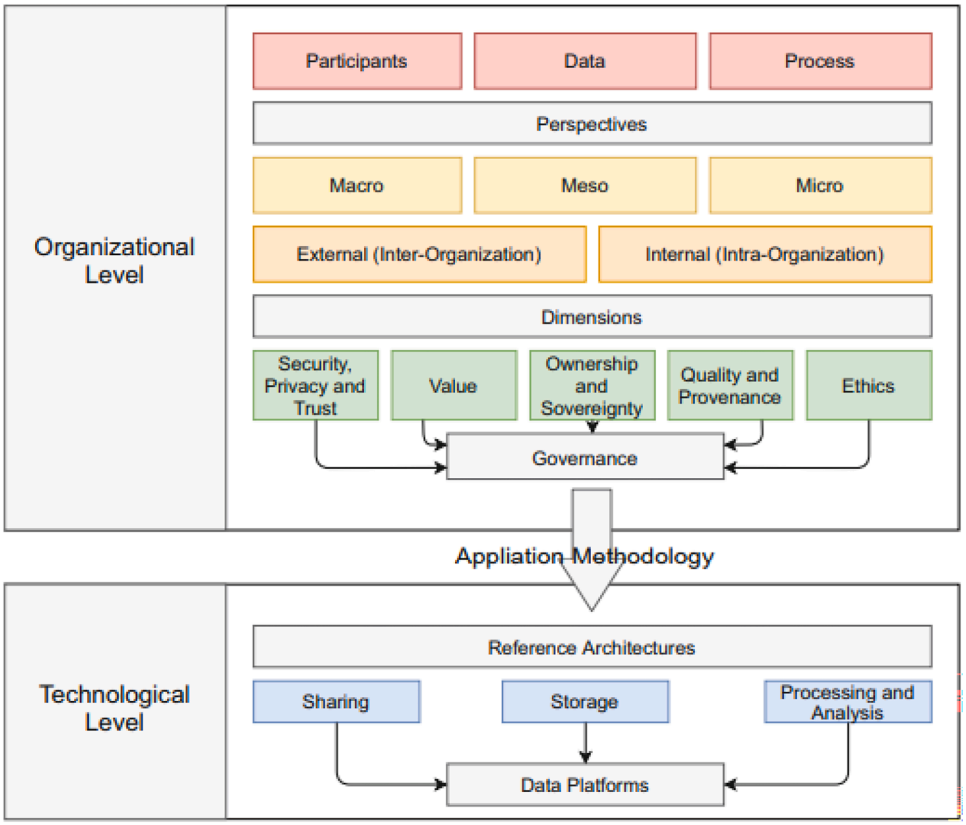

Fig. 25. Big data ecosystem governance components by (Torre-Bastida et al., 2022).

production process, provided computational speedup and high predictive accuracy for real-time data quality scoring using Apache Kafka. Lastly, (Bahaa et al., 2023) presented a DataOps lifecycle in a healthcare case study using the UCI Heart Disease dataset. The model with reduced features achieved better accuracy (87.32 %) and sensitivity (90.32 %), showing that reduced data can still perform effectively. These empirical cases affirm the benefits of DataOps, paving the way for further discussions on its impact and outcomes in data-driven environments.

4.5. Comparative analysis of existing frameworks

Table 4 presents a comparative overview of the strengths and limitations of existing DataOps frameworks, highlighting their capabilities and areas for improvement.

After analyzing the strengths and limitations of existing frameworks, the necessity for a new framework that can provide innovative solutions becomes evident.

5. Key Findings and impact of DataOps adoption

This section provides a better understanding of the impact of DataOps by exploring the impact of DataOps adoption. It introduces an innovative unified DataOps strategy framework, followed by a discussion on effective strategies for its implementation. Additionally, the section addresses the limitations of current practices and suggests future research directions.

5.1. Impact of DataOps adoption

The implementation of DataOps has a significant impact on organizations, driving improvements in data management and business operations. Key benefits include:

- **Agility and Adaptability:** DataOps enables businesses to develop dynamic capabilities, allowing them to respond swiftly to challenges and opportunities in a rapidly changing digital environment (Gür et al., 2022).
- **Operational Efficiency**: Through real-world examples, DataOps is shown to enhance operational efficiency, improve data quality, and promote better collaboration across teams by aligning production and development settings (Bergh et al., 2019).
- **Automated Orchestration**: Automation reduces errors and variability, ensuring data quality through continuous testing and monitoring, while providing decision-makers with timely, reliable insights (Pelluru, 2022).
- **Self-Service Infrastructure**: DataOps fosters the creation of scalable infrastructures, empowering data teams—such as engineers and administrators—to drive innovation and business success.

  Examples from companies like Facebook, Uber, LinkedIn, and eBay show how DataOps has:

  - Democratized data access.
  - Enhanced collaboration.
  - Increased agility in data operations.
  - Boosted productivity and profitability (Thusoo & Sarma, 2017).
- **Cloud Integration**: Leveraging DataOps within cloud environments amplifies benefits, including cost optimization and automation of data workflows, with the aid of DevOps methodologies, CI/CD pipelines, and Infrastructure as Code (IaC) (Voruganti, 2023).

Examples from companies like Facebook, Uber, LinkedIn, and eBay highlight how DataOps democratizes data access, enhances collaboration, and boosts agility and productivity ((Thusoo & Sarma, 2017)). By integrating data and operational practices, DataOps fosters adaptability and dynamic capabilities, driving improvements across management and business operations essential for competitiveness in a data-driven world. To quantify the impact of DataOps adoption, a survey conducted by **451 Research** analyzed **300 companies** across North America (451 Research 2020). This comprehensive study covered a wide range of industries, including **financial services, healthcare, technology, manufacturing, retail, government, and hospitality**. The respondents represented diverse leadership roles, with **30 %** being **CIOs, 24 %** as IT operations directors, and **9 %** holding positions in business intelligence and analytics. The survey findings revealed that **49 %** of the companies had between **1500 and 4999 employees**, while **51 %** had **5000 or more employees**. In terms of data volume, **38 %** of companies managed **2–5 PB (petabytes)** of data, **37 %** handled **5–10 PB**, and **25 %** managed **more than 10 PB**. These results highlight the widespread adoption and scaling of DataOps practices across industries, where organizations are achieving tangible benefits, including improved collaboration, increased operational efficiency, and enhanced cloud

Table 4
Strengths and limitations of DataOps frameworks.

Framework	Strengths	Limitations
DataOps Lifecycle Framework	Emphasizes principles like governance, agility, and scalability	No clear operational guidelines or real-world case integration
DataOps Lifecycle Model	Structured lifecycle for DataOps processes	Limited focus on governance and real-time analytics
Big Data Ecosystem Governance Components	Combines organizational and technological governance perspectives for comprehensive data ecosystem management	Focuses heavily on governance without addressing specific DataOps lifecycle automation
End-to-End DataOps Framework	Comprehensive lifecycle from collection to publishing	Lacks a focus on advanced data quality metrics, sector-specific adaptations, and pipeline monitoring.
Statistical Metrics for Data Quality in AVH Framework	Automates data validation using historical data and advanced statistical metrics	Limited to metrics-based validation, lacking broader governance integration
IBM Data Governance Framework	A comprehensive and flexible framework that integrates key aspects of data governance, including analytics, privacy, and result measurement	Lacks sector-specific adaptability, advanced metrics, and support for dynamic, real-time processes.
Data Pipeline for Passenger Flow Prediction	Provides an end-to-end pipeline integrating diverse data sources, advanced transformations, and multiple predictive models, ensuring flexibility and scalability for passenger flow prediction.	Lacks real-time adaptive feedback mechanisms and detailed handling of data quality issues during preparation and transformation stages.
Telecom DataOps Framework	Highly detailed with integration of tools like Kafka, Spark, and Jenkins for dynamic data processing and deployment	Telecom-specific focus and complex architecture may limit broader applicability
DataOps Data Architecture with Tools	Comprehensive integration of orchestration, monitoring, CI/CD, and data governance tools like Jenkins, Kubeflow, and Tableau	Complex setup may be resource-intensive for smaller organizations
Dependable Architecture	Focus on modularity and flexibility	Limited integration with emerging technologies
DataOps Process for Anything-to-RDF Integration	Streamlines RDF data integration with reconciliation and quality improvement	Focuses narrowly on RDF and lacks broader DataOps lifecycle considerations

integration. As summarized in the following Table 5, DataOps adoption has proven to be a game-changer for organizations aiming to unlock the full value of their data and maintain competitiveness in today's data-driven world.

These outcomes align closely with the core research questions addressed in this study. **(RQ1)**: Our analysis identified several key innovations driving the advancement of DataOps practices, including the integration of automated data pipelines and continuous integration/continuous delivery (CI/CD) methodologies. These technologies streamline the data lifecycle, enhance team collaboration, and enable more efficient data processing. Tools like Jenkins for orchestration and Kubeflow for automation allow organizations to scale their data operations with minimal manual intervention, effectively addressing bottlenecks in traditional data management and positioning DataOps at the forefront of enterprise data strategies. **(RQ2)**: The analysis also revealed that DataOps significantly improves data integration, delivery, and governance through robust frameworks, as discussed in the Policy and Governance section. Continuous testing and monitoring enhance data quality and ensure compliance with governance standards. Frameworks such as the European Industrial Data Space (EIDS) provide structured approaches for managing data at scale while preserving data sovereignty and ensuring compliance. These practices, rooted in automation and agile methodologies, reduce the time required to integrate and govern complex datasets, as demonstrated in the empirical case studies. **(RQ3)**: Empirical evidence from case studies shows the direct impact of DataOps on operational efficiency and decision-making. For example, implementing a DataOps pipeline in the aviation industry improved passenger flow predictions by 60 %, enhancing resource planning and customer satisfaction. Similarly, in the healthcare sector, optimized data pipelines for cardiovascular research achieved higher predictive accuracy. These cases highlight the tangible benefits of DataOps, demonstrating how these practices improve resource allocation, reduce operational overhead, and enable more informed decision-making processes.

5.2. *Unified DataOps strategy framework*

After comparing existing frameworks, it is evident that a comprehensive solution is required—one that leverages the strengths of current DataOps practices while addressing their limitations. To this end, we propose the comprehensive framework illustrated in Fig. 26. This framework encapsulates the core components of a successful DataOps strategy, emphasizing structured methodologies, policy and governance, and innovative tools essential for efficient DataOps implementation.

A transformative element that has not been sufficiently explored in existing contributions is the integration of Generative AI into DataOps. This integration introduces significant advancements and revolutionary changes, primarily driven by the key benefits that Generative AI can offer. Similar to its demonstrated impact in Agile methodologies—where

Table 5
Statistical impact of DataOps adoption.

Impact Area	Statistical Impact
Time-to-Insight	Without DataOps, 50 % of companies need over 3 days to generate insights.
Business Success	81 % say improved DataOps positively impacts success
Staff Efficiency	48 % saw better infrastructure utilization, 25 % reduced staff search time for data.
Data Management	25 % of staff spend 50 % of time on data access without DataOps, 30 % need over 1 week to generate insights
Data Governance	44 % reported reduced compliance risk, 35 % faster legal response times
Data Analytics	42 % of companies deployed machine learning, 31 % saw faster customer info requests
Automation	84 % expect process changes, 16 % of companies freed up IT staff for higher-value tasks.
Investment	86 % plan to increase their DataOps investment, and 40 % are investing in self-service analytics due to the perceived benefits of DataOps.
Cloud Support	87 % of companies believe that multi-cloud support is a critical or extremely critical part of their DataOps strategy.
	55 % of companies use 3 or more cloud providers, meaning they face a real need for systems (like DataOps) that can manage data across these environments.

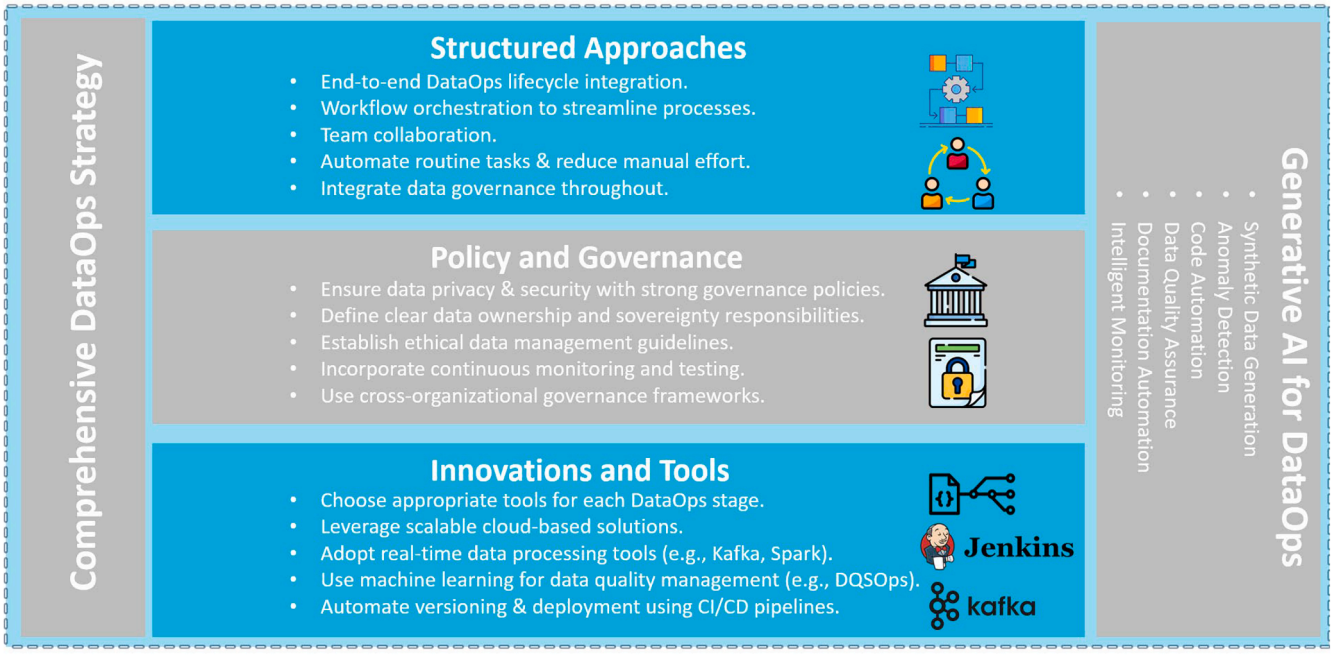

Fig. 26. Framework for unified DataOps strategy.

it enhances adaptability, streamlines workflows, and fosters innovation (Bahi et al., 2024)—Generative AI can also bring agility to DataOps by automating complex tasks, enabling faster decision-making, and improving overall data pipeline efficiency. Key capabilities of Generative AI for DataOps, such as synthetic data generation, anomaly detection, code automation, data quality assurance, documentation automation, and intelligent monitoring, align seamlessly with the principles of agility. These capabilities not only support dynamic data operations but also facilitate the continuous improvement and rapid adaptation required in data-driven environments. By leveraging this framework, organizations can optimize operational efficiency, enhance scalability, and maintain competitiveness in an increasingly data-centric landscape.

To measure the effectiveness of the Unified DataOps Strategy Framework, it is essential to define Key Performance Indicators (KPIs) that align with its primary objectives. These KPIs provide measurable benchmarks for evaluating the framework's success in improving data operations, governance, and technological innovation.

Ø Workflow Efficiency: Workflow efficiency measures the optimization of processes, focusing on task automation, collaboration improvement, and the reduction of manual efforts.

$$Workflow\ Efficiency\ (\%) = \frac{Baseline\ Time - Improved\ Time}{Baseline\ Time} * 100$$

Ø Compliance Rate: Compliance rate evaluates adherence to regulatory and governance standards, ensuring secure and ethical data management practices.

$$Compliance\ Rate\ (\%) = \frac{Compliant\ Activities}{Total\ Activities} * 100$$

Ø Processing Speed (TB/hour): Processing speed measures the efficiency of data processing tools by calculating the amount of data processed per hour.

$$Processing\ Speed\ (TB/hour) = \frac{Total\ Data\ Processed\ (TB)}{Total\ Time\ Taken\ (hours)}$$

Ø Anomaly Detection Accuracy: Anomaly detection accuracy assesses the ability of tools to identify outliers and inconsistencies in data pipelines. This is crucial for maintaining data quality and operational reliability.

$$Anomaly\ Detection\ Accuracy(\%) = \frac{True\ Positives + True\ Negatives}{Total\ Predictions} * 100$$

Before the implementation of the framework, surveys such as (451 Research 2020), including one conducted by 451 Research, analyzed 300 companies across North America spanning industries such as financial services, healthcare, technology, manufacturing, retail, government, and hospitality. The respondents held diverse leadership roles, with 30 % serving as CIOs, 24 % as IT operations directors, and 9 % in business intelligence and analytics. The survey revealed that 49 % of the companies had between 1500 and 4999 employees, while 51 % had 5000 or more. Regarding data volume, 38 % managed 2–5 PB (petabytes) of data, 37 % handled 5–10 PB, and 25 % managed over 10 PB. Additional studies (Lo Giudice, 2024; Gillen, 2023; IBM, n.d.; Snowflake, n.d.) indicated that workflow efficiency improved by approximately 50 %, highlighting the effectiveness of streamlined processes and task automation. Similarly, compliance rates reached 80 %, reflecting robust governance practices and adherence to regulatory standards. Furthermore, organizations utilizing advanced DataOps tools reported processing speeds ranging from 10 to 50 TB/hour, emphasizing the scalability and performance of modern data pipelines. High-performing systems for anomaly detection achieved an average accuracy of 75 %, underscoring their critical role in maintaining data integrity and operational reliability. However, by adopting this **holistic DataOps framework**, which integrates advanced methodologies, policy-driven governance, and innovative technologies, the potential for improvement becomes evident. The framework is projected to enhance workflow efficiency, processing speed, compliance rate, and anomaly detection by an additional improvement of **40 %**, calculated as:

$$Improvement(\%) = \frac{New\ Value - Baseline\ Value}{Baseline\ Value} * 100$$

To fully realize this improvement, the successful adoption of a strategy is required that leverages its architecture, tools, and methodologies to drive tangible outcomes. Securing executive sponsorship is a critical starting point to align organizational goals with resource allocation and to support the adoption of essential tools such as Jenkins for automation, Terraform and Ansible for Infrastructure as Code, and Kubeflow for orchestrating machine learning workflows. Demonstrating the framework's value through modular pilot projects, such as implementing the data quality module using AVH for historical data validation or DQSOps for real-time data quality scoring, can showcase measurable improvements in efficiency and build stakeholder confidence. To optimize resources, organizations must invest in advanced tools, including GitLab for version control, Spark for distributed data processing, and Tableau for analytics and visualization, which reduce manual tasks and enhance scalability. Ensuring seamless cross-team collaboration, regulatory compliance, and robust security is vital. The governance layer of the framework integrates tools like Apache Atlas for metadata management and Collibra for data governance, while security tools such as Vault for secrets management, Okta for authentication, and Fortinet for advanced threat protection ensure data and system integrity. Continuous improvement is supported by Generative AI capabilities, enabling tasks such as anomaly detection, pipeline monitoring, synthetic data generation, and automated documentation, which foster agility and adaptability. Additionally, targeted training programs are essential to upskill employees in tools like Spark for data processing, Tableau for visualization, and Vault for managing secure access, building the competencies required for effective implementation. To ensure success, organizations should monitor progress using Key Performance Indicators (KPIs) such as anomaly detection accuracy, workflow efficiency, compliance rates, and security breach prevention metrics, which provide actionable insights for optimizing efforts. By following these strategies, organizations can overcome adoption challenges, integrate the Unified DataOps Strategy Framework effectively, and drive operational efficiency, innovation, and long-term value.

5.3. Limitations and future directions in DataOps research

While this study provides a robust synthesis of the current DataOps landscape, several limitations must be acknowledged:

- **Rapidly evolving field:** The number of contributions in the field of DataOps increased significantly from 2015–2019 to 2020–2024, with a percentage increase of approximately **254.55 %**. In terms of proportion, the contributions from 2015–2019 account for **22 %** of the total contributions, while those from 2020–2024 account for **78 %**. This data clearly demonstrates the rapid evolution and growing interest in the field of DataOps.

- **Sector focus**: While this research spans multiple sectors, including aviation, healthcare, and finance, a closer analysis reveals notable disparities in sector-specific contributions. Among the reviewed studies, **15 % focus on aviation, 30 % on healthcare, and 24 % on finance**, highlighting a disproportionate emphasis on certain industries. This uneven distribution suggests the need for further exploration of sector-specific applications, especially in underrepresented areas. Addressing these gaps can provide deeper insights and tailored solutions to the unique challenges faced by each sector, ultimately fostering a more comprehensive understanding of DataOps practices across diverse operational contexts.
- **Impact Assessment**: Future research should quantify the real impact of DataOps adoption using statistical analysis to provide concrete evidence of its effectiveness on efficiency and performance.
- **Framework application**: The proposed framework developed, provides valuable insights, but further studies are needed to confirm its effectiveness and adaptability across diverse operational contexts.

6. Conclusion

This systematic review has explored the evolving field of DataOps, highlighting its significance in enhancing data management practices across large-scale enterprises. The key findings of this research reveal that DataOps is a transformative methodology that integrates people, processes, and technology to streamline the data lifecycle, from acquisition to analysis and reporting. Through the adoption of Agile, DevOps, and CI/CD principles, DataOps fosters better collaboration among data professionals, accelerates data delivery, and improves data quality, all of which are critical for informed decision-making and business innovation. The empirical evidence gathered from various case studies demonstrates the tangible benefits of implementing DataOps practices. These include significant improvements in operational efficiency, enhanced data governance, and the ability to respond swiftly to market dynamics. Furthermore, the integration of advanced technological tools and innovative frameworks has shown to be instrumental in overcoming the challenges associated with big data management, such as data silos, quality control, and scalability.

In summary, the significance of DataOps lies in its capacity to transform data operations from a fragmented and error-prone process into a cohesive, efficient, and agile system. By embedding DataOps within organizational structures, enterprises can unlock the full potential of their data assets, driving innovation, and maintaining a competitive edge in an increasingly data-driven world. As the field of DataOps continues to evolve, future research should focus on addressing its challenges and expanding its application across diverse industries, ensuring that the benefits of this methodology are fully realized.

CRediT authorship contribution statement

Aymen Fannouch: Writing – original draft. **Jihane Gharib:** Writing – review & editing, Validation, Investigation, Supervision, Formal analysis, Conceptualization. **Youssef Gahi:** Validation, Supervision, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

451 Research. (2020). *DataOps unlocks the value of data*. 451 Research.

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25*(3), 443–448.

Al-Quraishi, T., Al-Quraishi, N., AlNabulsi, H., Hussein, A. Q., & Ali, A. H. (2024). Big data predictive analytics for personalized medicine: Perspectives and challenges. *Applied Data Science and Analysis, 2024*, 32–38.

Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing, 23*, 839–859.

Arigbabu, A. T., Olaniyi, O. O., Adigwe, C. S., Adebiyi, O. O., & Ajayi, S. A. (2024). Data governance in AI-enabled healthcare systems: A case of the project nightingale. *Asian Journal of Research in Computer Science, 17*(5), 85–107.

Artac, M., Borovssak, T., Di Nitto, E., Guerriero, M., & Tamburri, D. A. (2017). DevOps: introducing infrastructure-as-code. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)* (pp. 497–498). IEEE.

Atwal, H. (2019). *Practical DataOps: Delivering agile data science at scale*. Apress.

Avram, M. G. (2014). Advantages and challenges of adopting cloud computing from an enterprise perspective. *Procedia Technology, 12*, 529–534.

Bahaa, S., Ghalwash, A. Z., & Harb, H. (2023). DataOps lifecycle with a case study in healthcare. *International Journal of Advanced Computer Science and Applications, 14*(1).

Bahi, A., Gharib, J., & Gahi, Y. (2024). Integrating Generative AI for Advancing Agile Software Development and Mitigating Project Management Challenges. *International Journal of Advanced Computer Science & Applications, 15*(3).

Barbierato, E., Gribaudo, M., & Iacono, M. (2014). Performance evaluation of NoSQL big-data applications using multi-formalism models. *Future Generation Computer Systems, 37*, 345–353.

Bass, L., Weber, I., & Zhu, L. (2015). *DevOps: A software architect's perspective*. Addison-Wesley Professional.

Bayram, F., Ahmed, B. S., Hallin, E., & Engman, A. (2023). DQSOps: Data quality scoring operations framework for data-driven applications. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering* (pp. 32–41).

Bergh, C., Benghiat, G., & Strod, E. (2019). The dataOps cookbook. *DataKitchen Hqrs*.

Berners-Lee, T., & Shadbolt, N. (2011). There's gold to be mined from all our data. *The Times*.

Bosch, J., & Olsson, H. H. (2023). Maturity assessment model for industrial data pipelines. In *2023 30th Asia-Pacific Software Engineering Conference (APSEC)* (pp. 503–513). IEEE.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data. *McKinsey Quarterly, 4*(1), 24–35.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314–347.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. *Frontiers of Computer Science, 7*, 157–164.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*, 171–209.

DataKitchen. What is a DataOps Engineer. url: https://datakitchen.io/what-is-a-dataops-engineer/.

DeBellis, M., Pinera, L., & Connor, C. (2023). Interoperability frameworks: Data fabric and data mesh architectures. *Data Science with Semantic Technologies* (pp. 267–286). CRC Press.

De Haes, S., Van Grembergen, W., & Debreceny, R. S. (2013). COBIT 5 and enterprise governance of information technology: Building blocks and research opportunities. *Journal of Information Systems, 27*(1), 307–324.

Dingsøyr, T., Nerur, S., Balijepally, V., & Moe, N. B. (2012). A decade of agile methodologies: Towards explaining agile software development. *Journal of Systems and Software, 85*(6), 1213–1221.

Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. *Future generation Computer Systems, 37*, 267–281.

Elouataoui, W., El Mendili, S., & Gahi, Y. (2023). Big data quality anomaly scoring framework using artificial intelligence. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)* (pp. 87–92). IEEE.

Elouataoui, W., El Mendili, S., & Gahi, Y. (2023). An automated big data quality anomaly correction framework using predictive analysis. *Data,, 8*(12), 182.

Elouataoui, W., El Mendili, S., & Gahi, Y. (2024). Active metadata and machine learning based framework for enhancing big data quality. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security* (pp. 1–8).

Ereth, J. (2018). DataOps-Towards a Definition. *LWDA, 2191*, 104–112.

Fannouch, A., Gahi, Y., & Gharib, J. (2024). Unified data framework for enhanced data management, consumption, provisioning, processing and movement. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security* (pp. 1–7).

Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., … Moser, B. (2020). Ai system engineering—key challenges and lessons learned. *Machine Learning and Knowledge Extraction, 3*(1), 56–83.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future, 2007* (2012), 1–16.

Garriga, M., Aarns, K., Tsigkanos, C., Tamburri, D. A., & Heuvel, W. V. D. (2021). Dataops for cyber-physical systems governance: The airport passenger flow case. *ACM Transactions on Internet Technology (TOIT), 21*(2), 1–25.

Gillen, A. (2023). *Generative AI likely to impact ops and developer staff first*. June. International Data Corporation. Retrieved from https://www.idc.com/getdoc.jsp?containerId=US50827423.

Gür, I., Möller, F., Hupperz, M., Uzun, D., & Otto, B. (2022). Requirements for DataOps to foster dynamic capabilities in organizations-A mixed methods approach. In *, 1. 2022 IEEE 24th Conference on Business Informatics (CBI)* (pp. 166–175). IEEE.
Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.
Hashim, W., & Hussein, N. A. H. K. (2024). *Securing Cloud Computing Environments: An Analysis of Multi-Tenancy Vulnerabilities and Countermeasures, 2024* pp. 8–16). SHIFRA.
Heck, P. (2024). What About the Data? A Mapping Study on Data Engineering for AI Systems. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* (pp. 43–52).
Hoda, R., Noble, J., & Marshall, S. (2012). Self-organizing roles on agile software development teams. *IEEE Transactions on Software Engineering, 39*(3), 422–444.
IBM. (n.d.). DataOps Framework: 4 Key Components & How to Implement Them. Retrieved from https://www.ibm.com/think/topics/dataops-framework.
International, D. (2017). *DAMA-DMBOK: Data management body of knowledge*. Technics Publications, LLC.
Jabbari, R., bin Ali, N., Petersen, K., & Tanveer, B. (2016). What is DevOps? A systematic mapping study on definitions and practices. In *Proceedings of the scientific workshop proceedings of XP2016* (pp. 1–11).
Jakobsen, A. S. (2022). *Master's thesis, uis*.
Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government information quarterly, 37*(3), Article 101493.
Jiang, H., Chen, Y., Qiao, Z., Weng, T. H., & Li, K. C. (2015). Scaling up MapReduce-based big data processing on multi-GPU systems. *Cluster Computing, 18*, 369–383.
Josey, A. (2016). *TOGAF® version 9.1-A pocket guide*. Van Haren.
Jukić, N., Sharma, A., Nestorov, S., & Jukić, B. (2015). Augmenting data warehouses with Big Data. *Information Systems Management, 32*(3), 200–209.
Kannan, R., & Jain, V. (2023). Automated data and ML pipelines to accelerate subsurface digitalization. In *SPE/AAPG/SEG Latin America Unconventional Resources Technology Conference*. URTEC (p. D031S027R002).
Kim, G., Humble, J., Debois, P., Willis, J., & Forsgren, N. (2021). The DevOps handbook: How to create world-class agility, reliability, & security in technology organizations. *It Revolution*.
Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*(3), 78–85.
Kontostathis, K. (2017). Collecting Data. *The DevOps Way* [Retrieved: 2021-07-30]. [Online]. Available: https://insights.sei.cmu.edu/devops/2017/11/collecting-data-the-devops-way.html.
Kumar, A., Niu, F., & Ré, C. (2013). Hazy: making it easier to build and maintain big-data analytics. *Communications of the ACM, 56*(3), 40–49.
Liu, X. Y., Xia, Z., Yang, H., Gao, J., Zha, D., Zhu, M., & Guo, J. (2024). Dynamic datasets and market environments for financial reinforcement learning. *Machine Learning, 113*(5), 2795–2839.
Lo Giudice, D. (2024). *Generative AI and TuringBots in Forrester's top 10 emerging techs*. August 1). Forrester. Retrieved from https://www.forrester.com/blogs/generative-ai-turingbots-win-again-in-forrester-top-10-emerging-technologies/.
Lwakatare, L. E., Kilamo, T., Karvonen, T., Sauvola, T., Heikkilä, V., Itkonen, J., … Lassenius, C. (2019). DevOps in practice: A multiple case study of five companies. *Information and Software Technology, 114*, 217–230.
Mainali, K., Ehrlinger, L., Matskin, M., & Himmelbauer, J. (2021). Discovering DataOps: a comprehensive review of definitions, use cases, and tools. In. In *DATA ANALYTICS 2021 The Tenth International CCookbookUnlocks the Value of Dataonference on Data Analytics*.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.
Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
Mertens, C., Alonso, J., Lázaro, O., Palansuriya, C., Böge, G., Nizamis, A., & Poulakidas, A. (2022). A framework for big data sovereignty: the European industrial data space (EIDS). *Data Spaces: Design, Deployment and Future Directions* (pp. 201–226). Cham: Springer International Publishing.
Munappy, A. R., Mattos, D. I., Bosch, J., Olsson, H. H., & Dakkak, A. (2020). From ad-hoc data analytics to dataops. In *Proceedings of the International Conference on Software and System Processes* (pp. 165–174).
Nemoto, R. H., Ibarra, R., Staff, G., Akhiiartdinov, A., Brett, D., Dalby, P., … Piebalgs, A. (2023). Cloud-based virtual flow metering system powered by a hybrid physics-data approach for water production monitoring in an offshore gas field. *Digital Chemical Engineering, 9*, Article 100124.
Pelluru, K. (2022). Unveiling the Power of IT DataOps: Transforming Businesses across Industries. *Innovative Computer Sciences Journal, 8*(1), 1–10.
Pinkel, C., Schwarte, A., Trame, J., Nikolov, A., Bastinos, A. S., & Zeuch, T. (2015). DataOps: seamless end-to-end anything-to-RDF data integration. In *The Semantic Web: ESWC 2015 Satellite Events: ESWC 2015 Satellite Events, 12* pp. 123–127). Portorož, Slovenia: Springer International Publishing. May 31–June 4, 2015, Revised Selected Papers.
Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*, 1–10.
Rodriguez, M., de Araújo, L. J. P., & Mazzara, M. (2020). Good practices for the adoption of DataOps in the software industry. In *, 1694. Journal of Physics: Conference Series*. IOP Publishing, Article 012032.
Saimler, M., İckin, S., Bernini, G., Toumi, N., Diamanti, M., Papavassiliou, S., … Khorsandi, B. M. (2024). The Role of AI Enablers in Overcoming Impairments in 6G Networks. In *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)* (pp. 913–918). IEEE.
Savitz, E. (2012). Gartner: Top 10 strategic technology trends for 2013. URL http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years.
Shojaee Rad, Z., & Ghobaei-Arani, M. (2024). Data pipeline approaches in serverless computing: a taxonomy, review, and research trends. *Journal of Big Data, 11*(1), 1–42.
Snowflake. (n.d.). DataOps for Data Speed and Quality. Retrieved from https://www.snowflake.com/guides/dataops-data-speed-and-quality.
Soares, S. (2010). *The IBM data governance unified process: driving business value with IBM software and best practices*. MC Press, LLC.
Tamburri, D.A., Van Den Heuvel, W.J., & Garriga, M. (2020, August). Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching. In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 391–394). IEEE.
Tamburri, D. A., van Mierlo, V. R., & van den Heuvel, W. J (2022). Big data for the social good: The drought early-warning experience report. *IEEE Transactions on Big Data, 9*(3), 773–791.
The DataOps Manifesto, "The DataOps Manifesto," [Retrieved: 2021-08-19]. [Online]. Available: https://www.dataopsmanifesto.org/.
Thomas, G. (2006). *The DGI data governance framework*. The Data Governance Institute.
Thusoo, A., & Sarma, J. (2017). *Creating a Data-Driven Enterprise with DataOps*. O'Reilly Media. Incorporated.
Torre-Bastida, A. I., Gil, G., Miñón, R., & Díaz-de-Arcaya, J. (2022). Technological perspective of data governance in data space ecosystems. *Data Spaces: Design, Deployment and Future Directions* (pp. 65–87). Cham: Springer International Publishing.
Tse, D., Chow, C. K., Ly, T. P., Tong, C. Y., & Tam, K. W. (2018). The challenges of big data governance in healthcare. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 1632–1636). IEEE.
Tu, D., He, Y., Cui, W., Ge, S., Zhang, H., Han, S., & Chaudhuri, S. (2023). Auto-validate by-history: auto-program data quality constraints to validate recurring data pipelines. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 4991–5003).
Underwood, M. (2023). Continuous metadata in continuous integration, stream processing and enterprise DataOps. *Data Intelligence, 5*(1), 275–288.
Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & society, 12*(2), 197–208.
Voruganti, K. K. (2023). Leveraging DataOps principles for efficient data management in cloud environments. *Journal of Technological Innovations, 4*(4).
Ward-Riggs, S., "The difference between DevOps and DataOps – Altis Consulting," [Retrieved: 2021-08-22]. [Online]. Available: https://altis.com.au/the-difference-between-devops-and-dataops/.
Xu, J., Naseer, H., Maynard, S., & Fillipou, J. (2022). Leveraging data and analytics for digital business transformation through DataOps: An information processing perspective. arXiv preprint arXiv:2201.09617.
Xu, J., Naseer, H., Maynard, S.B., & Filippou, J. (2023). Managing and making sense of data to drive digital transformation: A Case Study.
Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: Architecture and challenges. *IEEE Network, 28*(4), 5–13.
Yin, Z., Zhou, S., Zhou, J., Tian, M., Lin, M., & Liu, S. (2023). Research on DataOps Capability-Practice and Development. In *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (pp. 2170–2174). IEEE.
Yu, S., Chen, T., Han, L., Demartini, G., & Sadiq, S. (2022). Dataops-4g: On supporting generalists in data quality discovery. *IEEE Transactions on Knowledge and Data Engineering, 35*(5), 4668–4681.
Zahid, H., Mahmood, T., & Ikram, N. (2018). Enhancing dependability in big data analytics enterprise pipelines. In *, 11. Security, Privacy, and Anonymity in Computation, Communication, and Storage: 11th International Conference and Satellite Workshops, SpaCCS 2018* (pp. 272–281). Springer International Publishing. December 11-13, 2018, Proceedings.
Zhang, F., Liu, M., Gui, F., Shen, W., Shami, A., & Ma, Y. (2015). A distributed frequent itemset mining algorithm using spark for big data analytics. *Cluster Computing, 18*, 1493–1501.