ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/Mahesh-deshpande/94

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 4

Research Article

Rise of DataOps: Streamlining Data Pipelines and Workflows for Agile Data Management

Mahesh Deshpande*

Mahesh Deshpande, San Jose, California, USA

Citation: Deshpande M. Rise of Data Ops: Streamlining Data Pipelines and Workflows for Agile Data Management. *J Artif Intell Mach Learn & Data Sci 2023*, 1(4), 338-342. DOI: doi.org/10.51219/JAIMLD/Mahesh-deshpande/94

Received: 03 October, 2023; **Accepted:** 28 October, 2023; **Published:** 30 October, 2023

***Corresponding author:** Mahesh Deshpande, San Jose, California, USA

ABSTRACT

DataOps has emerged as a transformative approach to data management, combining the best practices of DevOps, Agile, and Lean methodologies to streamline data workflows and accelerate the delivery of data-driven insights. This article explores the key principles and practices of DataOps, including collaboration, automation, continuous integration and delivery, monitoring, and agile methodologies. It highlights the need for DataOps in the face of growing data complexity and the limitations of traditional approaches. The article also discusses the benefits of adopting DataOps, such as faster time-to-value, improved data quality, increased agility, enhanced collaboration, and better data governance.

Keywords: DataOps, Data Management, Data Pipelines, Automation, Collaboration, Continuous Integration, Continuous Delivery, Monitoring, Agile Methodologies, Data Quality, Data Governance, Data-Driven, DevOps, Lean, Data Workflows

Introduction

In today's data-driven landscape, organizations across various industries are grappling with the challenges of managing and leveraging vast amounts of data effectively. The exponential growth of data volumes, coupled with the increasing complexity of data pipelines and workflows, has made traditional data management approaches insufficient to meet the demands of modern businesses[1]. As a result, organizations are seeking innovative solutions to streamline their data operations and enable agile data management practices. Enter DataOps, a transformative approach that combines the best practices of DevOps, Agile, and Lean methodologies to optimize data workflows and accelerate the delivery of high-quality, reliable data to stakeholders[1].

DataOps has emerged as a game-changer in the realm of data management, offering organizations a comprehensive framework to tackle the challenges of data silos, inconsistent data quality, and inefficient data pipelines[3]. By fostering collaboration, automation, and continuous improvement, DataOps empowers data teams to work more efficiently and effectively, ultimately driving better business outcomes[4]**.** The rise of DataOps can be attributed to the growing recognition of data as a strategic asset and the need for organizations to harness its full potential to gain a competitive edge in the market[5].

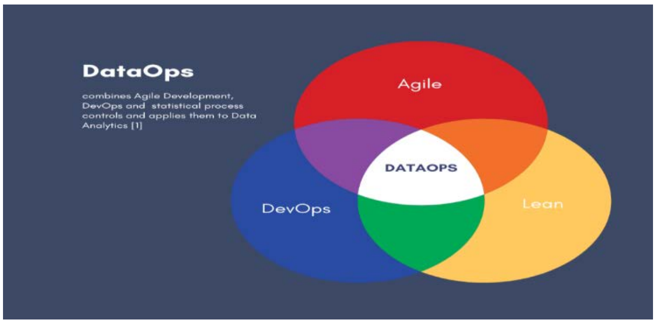

DataOps
combines Agile Development, DevOps and statistical process controls and applies them to Data Analytics [1]
Agile
DATAOPS
DevOps
Lean

Figure 1: What is DataOps.

The concept of DataOps draws inspiration from the success of DevOps in the software development industry. DevOps revolutionized the way software is developed, tested, and deployed by breaking down silos between development and operations teams, emphasizing automation, and fostering a culture of collaboration and continuous improvement[6]. Similarly, DataOps aims to bridge the gap among data engineers, data scientists, and business stakeholders, enabling them to work together seamlessly to deliver data-driven solutions faster and with higher quality[7].

As organizations embark on their digital transformation journeys, the adoption of DataOps has become a critical enabler for success. By streamlining data pipelines and workflows, DataOps allows organizations to extract valuable insights from their data assets more quickly and efficiently[5]. This agility is essential in today's fast-paced business environment, where the ability to make data-driven decisions in real-time can be the difference between success and failure[1].

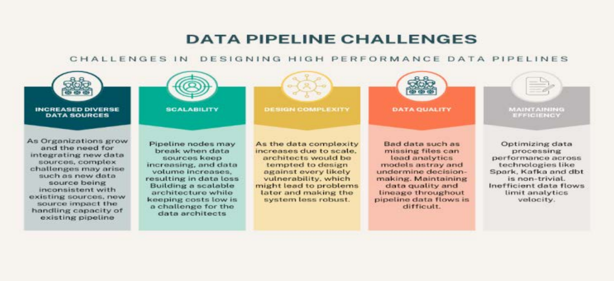
This article delves deep into the need for DataOps, its key principles and practices, the benefits of adoption, real-world success stories, and the challenges and considerations organizations should keep in mind when implementing DataOps. By the end of this article, readers will have a comprehensive understanding of how DataOps is transforming the landscape of data management and why it is crucial for organizations to embrace this approach to stay competitive in the data-driven era[11].

2. The Need for DataOps

Traditional data management approaches have been the backbone of organizations for decades, but they are struggling to keep pace with the rapidly evolving data landscape. The limitations of these conventional methods have become increasingly apparent, hindering organizations' ability to extract maximum value from their data assets[1].

The key limitations of traditional data management are:

- *Lack of Agility and Flexibility.* The traditional data management approaches often rely on rigid, monolithic architectures that make it difficult to adapt to changing business requirements and data needs[2]. As a result, organizations find themselves trapped in a cycle of *lengthy development cycles, manual processes, and delayed time-to-value* for data-driven initiatives[3]. A recent survey by Hitachi Vantara found that 67% of IT executives believe that traditional data management approaches are too slow to meet the needs of modern businesses[3].
- *Creation of Data Silos:* Traditional data management approaches often lead to data silos, where data is isolated within different departments or systems[4]. These silos hinder collaboration, data sharing, and the creation of a unified view of the organization's data assets. Consequently, *decision-makers struggle to access the right data at the right time,* leading to suboptimal business outcomes[4]. According to a study by Forrester, 72% of organizations struggle with data silos, leading to incomplete insights and delayed decision-making[15].
- *Managing Growing Complexity of Data Pipelines and Workflows:* With the explosion of data sources, formats, and volumes, managing data has become an increasingly intricate task[6]. Data pipelines have evolved from simple, linear flows to complex, multi-stage processes that involve various tools, technologies, and stakeholders[6]. **Figure 2** illustrates the challenges in designing modern data pipelines.



As Organizations grow and the need for integrating new data sources, complex challenges may arise such as new data source being inconsistent with existing sources, new source impact the handling capacity of existing pipeline

Pipeline nodes may break when data sources keep increasing, and data volume increases, resulting in data loss Building a scalable architecture while keeping costs low is a challenge for the data architects

As the data complexity increases due to scale, architects would be tempted to design against every likely vulnerability, which might lead to problems later and making the system less robust.

Bad data such as missing files can lead analytics models astray and undermine decision-making. Maintaining data quality and lineage throughout pipeline data flows is difficult.

Optimizing data processing performance across technologies like Spark, Kafka and dbt is non-trivial. Inefficient data flows limit analytics velocity.

Figure 2: Data Pipeline Design Challenges.

This complexity makes it difficult to ensure data quality, maintain data lineage, and troubleshoot issues when they arise[13].

- *Lack of Advanced Data Observability:* Traditional data management approaches often lack the necessary automation, monitoring, and observability capabilities to effectively manage these complex pipelines[7]. As a result, data teams spend a significant amount of time on manual, repetitive tasks, leading to inefficiencies and reduced productivity[7]. A study by Forbes Research found that data professionals spend up to 80% of their time on data preparation and cleansing tasks, leaving little time for strategic initiatives[8].

DataOps emerges as a solution to address these limitations and enable agile data management. By adopting DataOps practices, organizations can overcome the challenges posed by traditional approaches and unlock the full potential of their data assets[5].

3. Key Principles and Paractices of Data Ops

DataOps is a transformative approach that combines a set of key principles and practices to streamline data management and enable agile, efficient, and reliable data workflows. While DataOps draws inspiration from DevOps, it is more than just applying DevOps principles to data. DataOps addresses the unique challenges of data management and encompasses a broader set of principles and practices[2].

A. DataOps vs. DevOps

DevOps revolutionized software development by breaking down silos between development and operations teams, emphasizing automation, and fostering a culture of collaboration. Similarly, DataOps aims to break down silos between data teams and stakeholders, automate data workflows, and promote collaboration. However, DataOps goes beyond DevOps by addressing the specific needs of data management, such as data quality, data governance, and data security[2].

Table 1: The key differences between DataOps and DevOps.

Aspect	DevOps	DataOps
Focus	Software development	Data management and analytics
Goals	Faster software delivery	Faster insights and data-driven decisions
Automation	CI/CD pipelines	Data pipelines and workflows
Collaboration	Dev and Ops teams	Data teams and stakeholders
Quality	Code quality	Data quality and governance
Version Control	Source code	Data and pipeline configurations
Testing	Unit and integration tests	Data validation and quality checks

B. Principles of DataOps

DataOps, being a collaborative data management practice, emphasizes on communication, integration, automation, and cooperation among data engineers, data scientists, and other data professionals. The DataOps principles, derived from the Agile Manifesto and adopted for data management, provide a framework for organizations to implement DataOps practices effectively and drive the success of their data initiatives[9].

The 18 DataOps Principles, as outlined by Ereth[9], are summarized in the following table:

Table 2: DataOps Principles from the DataOps Manifesto.

Principle	Description
1. Continual Satisfaction	Deliver value to customers and stakeholders continuously
2. Value Working Analytics	Prioritize working analytics over comprehensive documentation
3. Embrace Change	Adapt to changing requirements and feedback
4. Daily Interaction	Encourage daily collaboration between data teams and stakeholders
5. Motivated Individuals	Build projects around motivated and trusted individuals
6. Face-to-Face Conversations	Prioritize face-to-face conversations for effective communication
7. Working Analytics Measure Progress	Use working analytics as the primary measure of progress
8. Sustainable Development	Maintain a constant and sustainable pace of development
9. Continuous Attention	Pay continuous attention to technical excellence and good design
10. Simplicity	Maximize the amount of work not done and keep things simple
11. Self-Organizing Teams	Rely on self-organizing teams for the best results
12. Regular Reflection	Reflect on performance regularly and adjust accordingly
13. Analytics is Code	Treat analytics as code and manage it in version control systems
14. Orchestration, Not Scripting	Use orchestration tools instead of scripting for managing workflows
15. Disposable Environments	Create disposable environments for experimentation and testing
16. Integrated Lifecycle	Integrate the entire data lifecycle, from ingestion to delivery
17. Expose Data Early	Make data available to stakeholders as early as possible
18. Manage Data, Not Systems	Focus on managing data effectively rather than just managing systems

These principles emphasize the importance of collaboration, automation, agility, and a focus on delivering value through working analytics. By adhering to these principles, organizations can foster a culture of continuous improvement, develop more reliable and efficient data workflows, and ultimately drive better business outcomes through data-driven decision-making.

C. Putting the principles to practice

The principles outlined in the DataOps Manifesto provides goals for all data stakeholders can work towards to achieve successful DataOps adoption[11]. Putting the principles to practice involves focusing on these 7 key pillars:

1. Collaboration and self-service

Collaboration and self-service are the most important pillar of DataOps, which can be viewed from two perspectives. *Firstly,* at the development and operations level, coordination among team members (data engineers, analysts, scientists, etc.) and their specific tools must be extremely smooth similar to a well-orchestrated industrial manufacturing line, for delivering high-quality data products, Secondly, at the customer and stakeholder level, the challenges of data accessibility, sensitivity, and discoverability need to be addressed by providing a user-friendly data catalog, anonymization techniques for safe data sharing, and lightweight processes for ingesting simpler data from end-users. These features enable organizations to become truly data-driven, allowing users to create and discover insights that were previously unattainable due to technological and privacy silos[10].

2. Orchestrate through ELT instead of ETL

Extract, load, transform (ELT) is a data pipeline model that offers several benefits over the traditional extract, transform, load (ETL) approach, particularly when used with data lake implementations. ELT stores information in its original raw format, enabling faster loading times and eliminating the need to define queries and schemas in advance. By loading data before transformation, no details are lost, maximizing future data processing and analysis potential. This approach supports data governance, auditability, and lineage principles, as the ingested data is almost exactly what was extracted. The Orchestrating through ELT takes this concept further, advocating for the avoidance of all actions that remove potentially useful or valuable data, including changes in data [10]. Although solutions to this problem are well-understood, they are time-consuming and expensive, requiring the cost of configuration and implementation to be made trivial for widespread adoption.

3. Agility and continuous integration/delivery (CI/CD)

DataOps integrates the principles of Continuous Integration/Continuous Delivery (CI/CD) from software development into data pipeline management. It emphasizes frequent integration of code changes, building, testing, and the ability to deploy on demand. DataOps applies to the entire data lifecycle, from model generation and orchestration to governance and business metrics. Like DevOps, DataOps encourages prototyping within the same system using branches, which include configuration, code, and data. This approach allows for iterative development and testing until stakeholder requirements are met, without compromising data integrity or quality[10]. DataOps also aims to streamline data lifecycle management, with some cloud data platforms, such as Snowflake Data Cloud, making it nearly invisible by leveraging cost-effective storage options and automatic data compression.

4. Modular component design and maintainability

DataOps takes advantage of the virtually infinite resources provided by cloud computing technologies, shifting the focus from CPU time to developer productivity. Optimizing code prematurely can hinder developers' ability to process stakeholder changes efficiently. DataOps emphasizes maintainability, which can be improved by following principles such as building modular programs, writing readable, simple, and transparent code, using composition, and creating robust and extensible programs. DataOps favors small, reusable components that reduce overall development time. By prioritizing configuration over code, using low-code solutions, and creating small, reusable components, it becomes easier to refine, optimize, and replace sections of the system without impacting the end-user experience.

5. Branching and version control

Just like in software development, DataOps needs branching and version control practices to data. Data needs to be treated as code, enabling teams to track changes, collaborate on data workflows, and maintain a clear history of data transformations. Version control systems like Git are used to manage data pipeline code and configurations[5].

While web development organizations have mastered the creation of multiple long-lived environments (PROD, QA, DEV) and dynamic feature branches for individual engineers' integration testing, even the most advanced organizations rarely have more than 2 or 3 manually created long-lived environments, leading to significant time spent managing differences between them[10].

Effective environment management in DataOps requires automatic building, changing, and destruction of environments, as well as the ability to quickly replicate production-like environments, which may involve duplicating terabytes of data. Cloud data platforms like Snowflake, with capabilities such as Zero Copy Cloning, enable DataOps environment management facilities today, and more will follow. These platforms have the potential to revolutionize data analytics in the same way Kubernetes has transformed microservices. Properly automated environment management is a significant advancement for an organization's data platform development, reducing costs by automating manual steps for creating, destroying, and cross-checking environments.

6. Governance, monitoring, and change audits

DataOps emphasizes the importance of Governance by Design and Privacy by Design, ensuring that strong governance and privacy models are included in the core of the platform from the outset. To address the risk of substantial fines for improper data use, DataOps subjects every change to automated testing and requires manual checks by two pairs of eyes before each pull request and code merge. It also establishes the code repository as the definitive source of truth for applications, similar to how software development relies on the application's supporting code repository. Furthermore, DataOps strengthens governance by automatically creating audit trails that document every change, test, and approval, allowing organizations to demonstrate how data has been and is being used.

7. Automated test-driven data development

DataOps advocates for a test-driven approach to data development. It involves writing tests for data transformations and pipeline components before implementing them. Test-driven data development ensures the correctness and reliability of data workflows, catches errors early in the development process, and enables faster debugging and troubleshooting[5].

When data teams struggle to keep up with stakeholder requests, automated data testing becomes the solution to identify issues and relieve the team's burden. In fast-paced agile development environments, automated testing and monitoring provide a way to counter optimism and detect problems caused by it. World leaders in rapid testing and deployment, such as Amazon, Netflix, and Etsy, rely on automated testing to successfully deploy new updates quickly with near-perfect availability, even when managing millions of lines of code[10].

DataOps monitoring goes beyond confirming system availability; it also checks the production quality of data and artifacts in analytics processes and tests new code during deployment. By detecting unexpected variations and generating operations statistics, data teams can gather insights to further enhance pipeline and platform performance. DataOps expands the definition of data availability to include the ability to return valid data for decision-making, not just the ability to run queries. Even with high uptime, broken pipelines can compromise data and negatively affect the accuracy of analytics without being detected, making automated testing crucial for ensuring actionable insights.

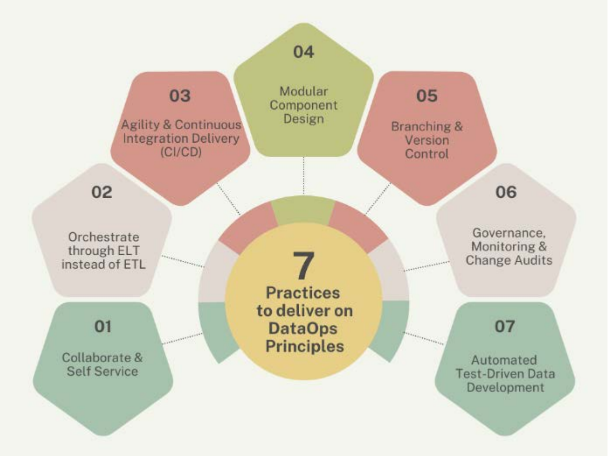

04
Modular Component Design
03
Agility & Continuous Integration Delivery (CI/CD)
05
Branching & Version Control
02
Orchestrate through ELT instead of ETL
06
Governance, Monitoring & Change Audits
7
Practices to deliver on DataOps Principles
01
Collaborate & Self Service
07
Automated Test-Driven Data Development

Figure 3: Summary of the Seven DataOps practices.

By embracing these principles and practices, organizations can establish a strong foundation for DataOps and drive the successful implementation of data-driven initiatives. DataOps enables organizations to overcome the challenges of traditional data management approaches and unlocks the full potential of their data assets[10].

4. Benefits of Adopting DataOps

Adopting DataOps practices offers numerous benefits to organizations seeking to optimize their data management processes and drive better business outcomes. Firstly, by embracing DataOps, companies can experience *faster time-to-value for their data-driven initiatives*, as the streamlined workflows and automated processes enable quicker delivery of insights and solutions[8]. A case study by DataOps.Live demonstrated that implementing DataOps automation led to a 70% reduction in data pipeline deployment time and Streamlines maintenance of code, which can reduce the TCO of data products by over 60%[15].

Secondly, DataOps *improves data quality and reliability* by implementing robust testing, monitoring, and validation mechanisms throughout the data lifecycle. This ensures that the data being used for decision-making is accurate, consistent, and trustworthy[7]. DataOps *also increases agility and flexibility in data management*, allowing organizations to adapt quickly to changing business requirements and market conditions[10]. It ensures that data initiatives are closely tied to business objectives and that data insights are readily accessible to decision-makers.

Thirdly, another key benefit of DataOps is *enhanced collaboration and productivity among data teams.* By fostering a culture of collaboration and breaking down silos between data engineers, data scientists, and business stakeholders, DataOps enables teams to work together more effectively and efficiently. This collaborative approach leads to better alignment between data initiatives and business objectives, ultimately driving more impactful outcomes.

Furthermore, DataOps helps reduce data silos and improves data governance. By establishing clear processes, roles, and responsibilities for data management, DataOps ensures that data is properly managed, secured, and compliant with regulatory requirements[12]. This holistic approach to data governance enables organizations to leverage their data assets more effectively while minimizing risks associated with data misuse or breaches. A survey by DataKitchen found that organizations adopting DataOps reported a 60% improvement in collaboration between data teams and stakeholders[14].

Lastly, DataOps adopts agile methodologies to enable iterative and incremental development of data pipelines and workflows[16]. By breaking down data initiatives into smaller, manageable chunks, teams can deliver value faster and adapt to changing requirements more effectively. A study by Gartner found that organizations with mature DataOps practices reported a 56% reduction in data team's operations management time and a 45% improvement in data quality[14].

In summary, adopting DataOps practices offers a wide range of benefits, including faster time-to-value, improved data quality, increased agility, enhanced collaboration, and better data governance. As organizations continue to recognize the critical role of data in driving business success, embracing DataOps becomes a strategic imperative for staying competitive in today's data-driven landscape.

Table 3 summarizes the key differences between Traditional Data management and DataOps:

Table 3: Comparing traditional data management and DataOps.

Aspect	Traditional Data Management	DataOps
Agility	Rigid, inflexible	Agile, adaptable
Collaboration	Siloed, isolated	Cross-functional, aligned
Automation	Manual, error-prone	Automated, streamlined
Monitoring	Reactive, ad-hoc	Proactive, continuous
Time-to-value	Lengthy, delayed	Rapid, iterative

The need for DataOps is evident from the growing adoption rates and the tangible benefits organizations are realizing. According to a survey by TechTarget, 90% of organizations have either adopted or plan to adopt DataOps practices in the next 12 months[3]. Furthermore, organizations that have implemented DataOps report significant improvements in data quality (66%), data pipeline efficiency (63%), and data team productivity (57%)[3].

5. Conclusion

DataOps has emerged as a transformative approach to data management, enabling organizations to overcome the limitations of traditional methods and unlock the full potential of their data assets. By adopting the key principles and practices of DataOps, such as collaboration, automation, continuous integration and delivery, monitoring, and agile methodologies, organizations can streamline their data workflows, improve data quality, and accelerate the delivery of data-driven insights.

The need for DataOps is evident, as the complexity of data pipelines and the growing volume and variety of data have made traditional approaches insufficient. DataOps addresses these challenges by fostering a culture of collaboration, breaking down silos, and automating processes to ensure data reliability and consistency.

The benefits of adopting DataOps are numerous, including faster time-to-value for data initiatives, improved data quality, increased agility, enhanced collaboration, and better data governance. As organizations continue to recognize the critical role of data in driving business success, embracing DataOps becomes a strategic imperative.

Looking ahead, the future of DataOps is promising, with the continued evolution of tools, technologies, and best practices. As more organizations adopt DataOps and share their experiences, the community will continue to refine and enhance the approach, driving even greater value from data. By embracing DataOps, organizations can position themselves to thrive in the data-driven era, leveraging their data assets to drive innovation, improve decision-making, and gain a competitive edge in their respective markets.

6. References

1. www.humansofdata.atlan.com/2022/08/the-rise-of-dataops/
2. www.em360tech.com/sites/default/files/2020-08/DataOps Cookbook 2nd Edition FINAL.pdf
3. www.ciosummits.com/Online_Assets_Hitachi_Vantara_DataOps_Unlocks_Value_of_Data.pdf
4. www.datakitchen.io/what-is-dataops/
5. www.oreilly.com/library/view/creating-a-data-driven/9781492049227
6. www.eyer.ai/blog/challenges-of-building-high-performance-data-pipelines-for-big-data-analytics/
7. www.store.hbr.org/product/competing-on-analytics-updated-with-a-new-introduction-the-new-science-of-winning/10157
8. www.dataversity.net/survey-shows-data-scientists-spend-time-cleaning-data/
9. www.poweradmin.com/blog/operation-data-the-18-key-principles-of-dataops/
10. www.truedataops.org/7-pillars
11. www.dataopsmanifesto.org/en
12. www.researchgate.net/publication/355107036_Discovering_DataOps_A_Comprehensive_Review_of_Definitions_Use_Cases_and_Tools
13. www.streamsets.com/blog/data-pipeline-architecture-deep-dive/
14. www.datakitchen.io/the-business-case-for-dataops/
15. www.dataops.live/what-is-dataops