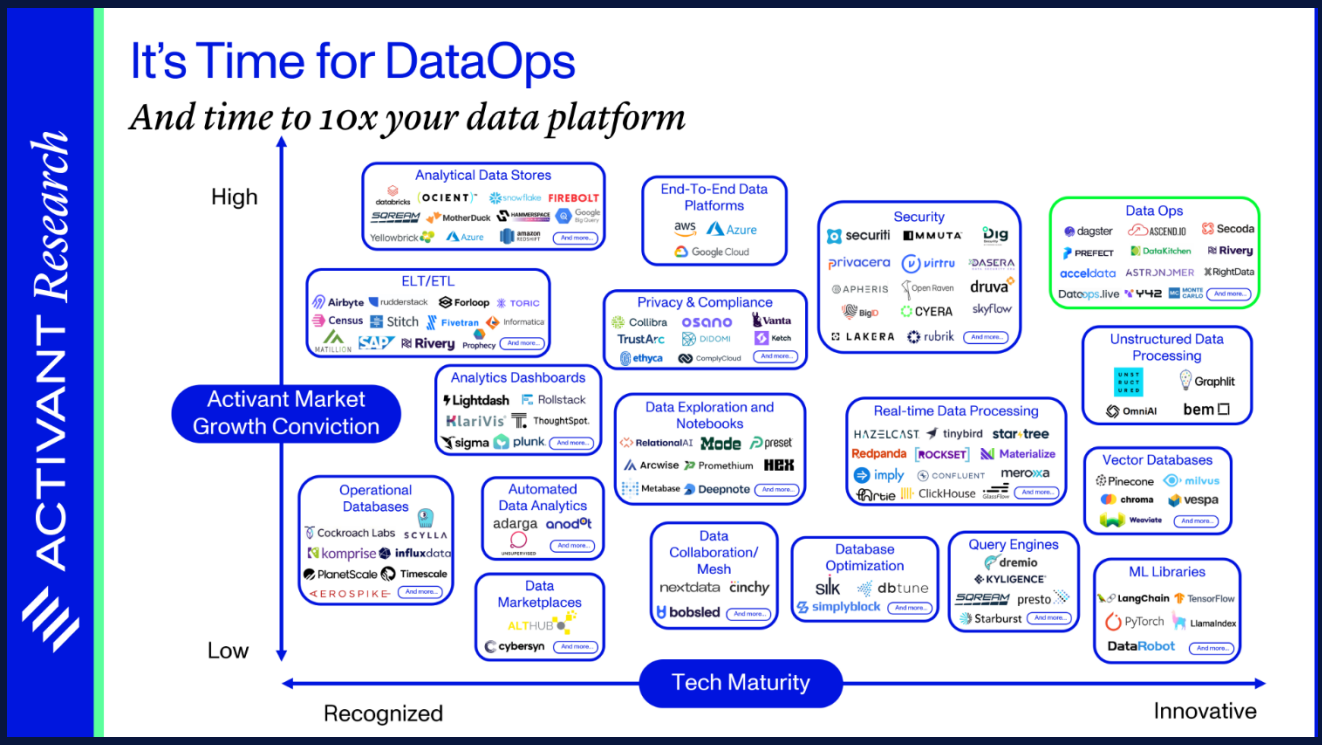
ACTIVANT RESEARCH

It's Time for DataOps

And time to 10x your data platform

Jonathan Vickery, Nina Matthews



Q3 2024

Introduction

In the race to outpace competitors, companies across industries are turning to data to speed up and improve decision making, roll out better products, and – most importantly – leverage modern AI models and tools. Meta, for example, generates ~$130 billion in advertising revenue with dynamic, targeted advertising powered by data from ~4bn users.[1] Similarly, Uber broke into the saturated food delivery sector thanks to the insights from ~9 billion annual trips, where more accurate delivery estimates keep Uber Eats customers happy and make drivers more efficient.[2] AI models also power music recommendations on Spotify, protect against fraud when consumers tap Visa credit cards, and power your first article draft through ChatGPT – all originating from the troves of data used to train those models.

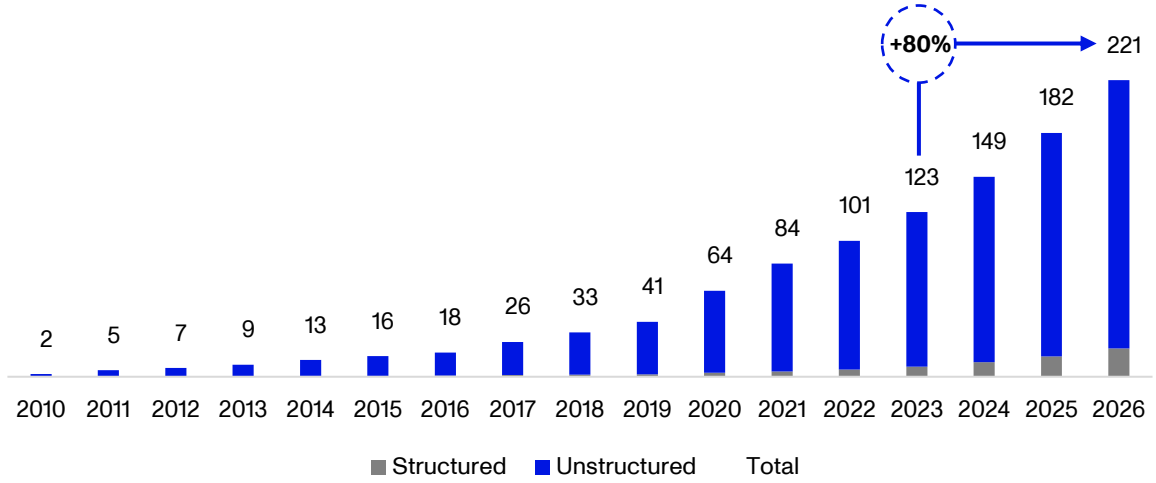
While these may just be anecdotes, the global economy is pouring investment into the infrastructure that underpins data analytics and AI, creating a burgeoning $100 billion market.[3] This is a mature market that has brought us scaled public companies like Snowflake, leading outsiders to think that data teams operate in a highly sophisticated, flawless manner, but that couldn't be further from the truth.

Most investments into data are a failure. In fact, 87% of data analytics projects never make it to production.[4] Data platforms are chaotic, manually operated, and highly error-prone. In this article, we dive into these issues and introduce a new approach – DataOps – which is ready to not just solve those problems, but provide a 10x improvement.

It's a Data-driven World

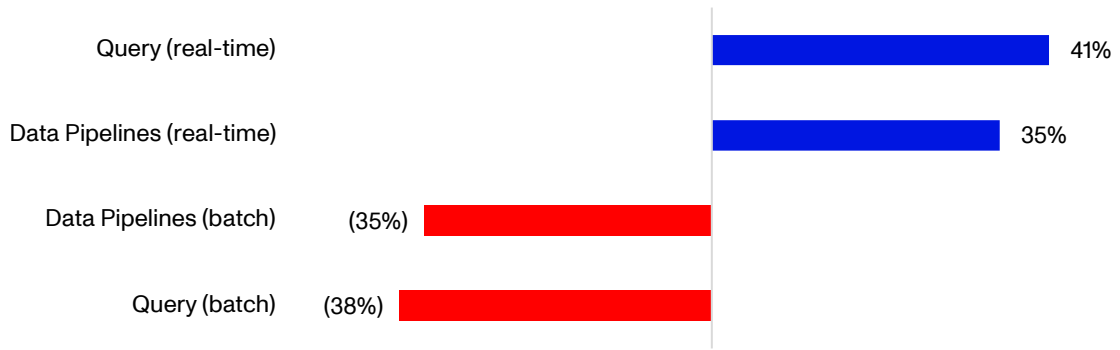
The amount of data that we're generating is exploding. New data volumes are **expected to grow by 80% in just the next three-year period** and will soon be nearly 100x their 2010 level. As we generate and capture more data, the opportunities for data-driven companies like Meta and Uber naturally expand, but so does the pressure on data teams and their infrastructure. Data teams can run into processing bottlenecks, scaling limitations, and skyrocketing costs, with these issues made all the worse if their data platform was architected ***before*** a 100x jump in data volumes. For data-driven business models, this volume growth is the gift that just keeps giving, but for data teams, it's the hamster wheel that just keeps spinning.

Worldwide Annual Data Generation Forecast, in Zettabytes[5,6]



Increasingly, companies expect and need data to be pushed through their systems **in real time**. Executive teams want live dashboards of their key business metrics, and technology companies like Uber need their decisioning systems to be updated for real-time data including traffic, driver capacity, and more. The shift from batch to real-time processing is both a significant trend and a challenge for teams already struggling with batch infrastructure.

Expected Change in Data Budget over the next 5 – 10 years[7]



And it's not only executives demanding access to more data - **data democratization** means that employees at all levels are expected to leverage enterprise data in their day-to-day jobs. But that also means that data infrastructure needs to be more accessible, data more user-friendly and collaboration fostered without requiring years of technical expertise.

Generative AI adoption, arguably the strongest driver of new data use, is forecast **to jump from its current 5% to 80% by 2026** as companies race to unlock the value promised by large language models.[8] However, 70% of organizations have already found data to be a challenge to capturing value from Generative AI tools and cite it as their most common issue.[9]

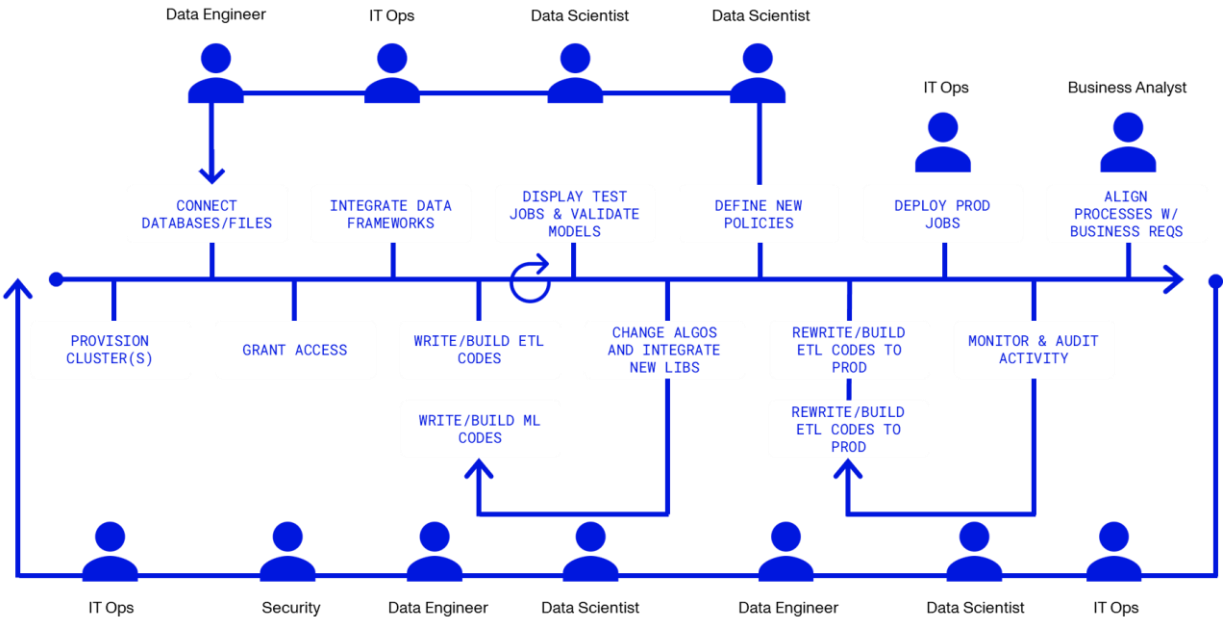
Enterprises are using more data, more timeously, with increasingly sophisticated AI systems. These trends are here to stay and data teams will have to ensure that their infrastructure is up to the challenge. It likely isn't. In a recent study **96% of analytics leaders reported being held back by data management challenges** - dealing with the storage, access, quality, state, and flow of the data.[10]

Data Management's Dilemmas

Practically all organizations are struggling to manage their data, as reflected in the following commonly occurring themes in the IT landscape:

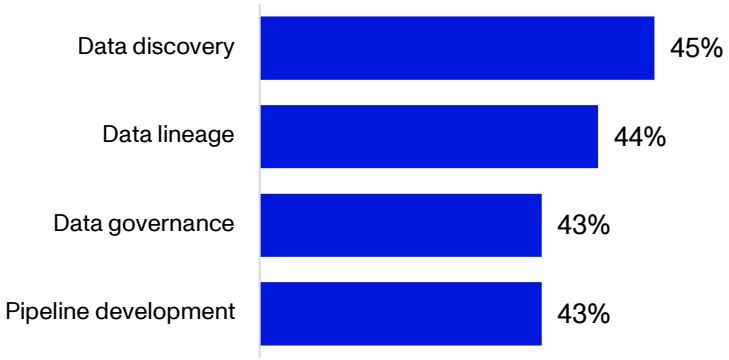
1. **Data and teams are siloed and fragmented:**
   Organizations use an average of 187 different software systems. For large enterprises, this number can reach 2,000+.[11] The data in these systems is trapped unless teams can build and manage the pipelines that connect all these systems. However, with on-premises, legacy, and closed systems in the mix, building and maintaining these pipelines can be a highly complex and resource-intensive task. Compounding this issue, each system might have its own departmental owner (a data security challenge), creating a bureaucratic nightmare that could delay executing on data projects by as much as 12 to 18 months.[12]

Illustrative data science lifecycle[13]

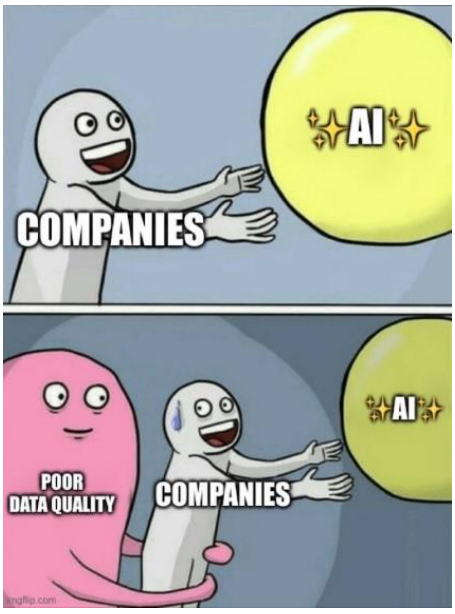


2. **The "Modern Data Stack" itself is fragmented and complex:**
   Companies like Fivetran and Snowflake exemplify the modern data stack (MDS) – a modular network of best-of-breed tools built in the mid to late 2010s to empower data engineers by being easy to set up and use. Unfortunately, these characteristics meant that teams quickly become inundated with numerous, sometimes duplicative MDS tools, aggregating into an infrastructure that is costly and complex to manage. Worse still, MDS tools targeted at business users might have thin APIs, making it tough to integrate them into one single control plane for data platform visibility and automation. **Hence, the MDS has no single source of truth**.
3. **As a result, almost half of all data workflows are still manual:**
   The fragmented nature of data necessitates significant intervention by data teams for tasks such as data discovery, lineage tracking, governance, quality checks, and ETL (extract, transform, load) development. These processes are time-consuming and error-prone, introducing data quality and reliability concerns. As we can see below, nearly half of these are still handled manually using pieced-together code to build and maintain pipelines.

Percentage of processes that are manual (moderately and highly) [14]



4. **Data outputs are error-prone and teams lack visibility for root-cause analysis**:
   Consider an AI project that aims to enrich the quarterly financial reporting for a CEO's board meeting. The project would need to pull data from various sources such as SAP, Salesforce, and ADP Workforce Now. When this is done by a data engineer in an ad-hoc and manual fashion, errors often arise and when the CEO says that one of the figures "looks off," it can take days or even weeks to trace through the system to test validity and correct any errors. If this sounds like an exception, consider that 33% of business users are not confident in the quality of data provided to them and 44% don't have visibility into their data pipelines.[15]



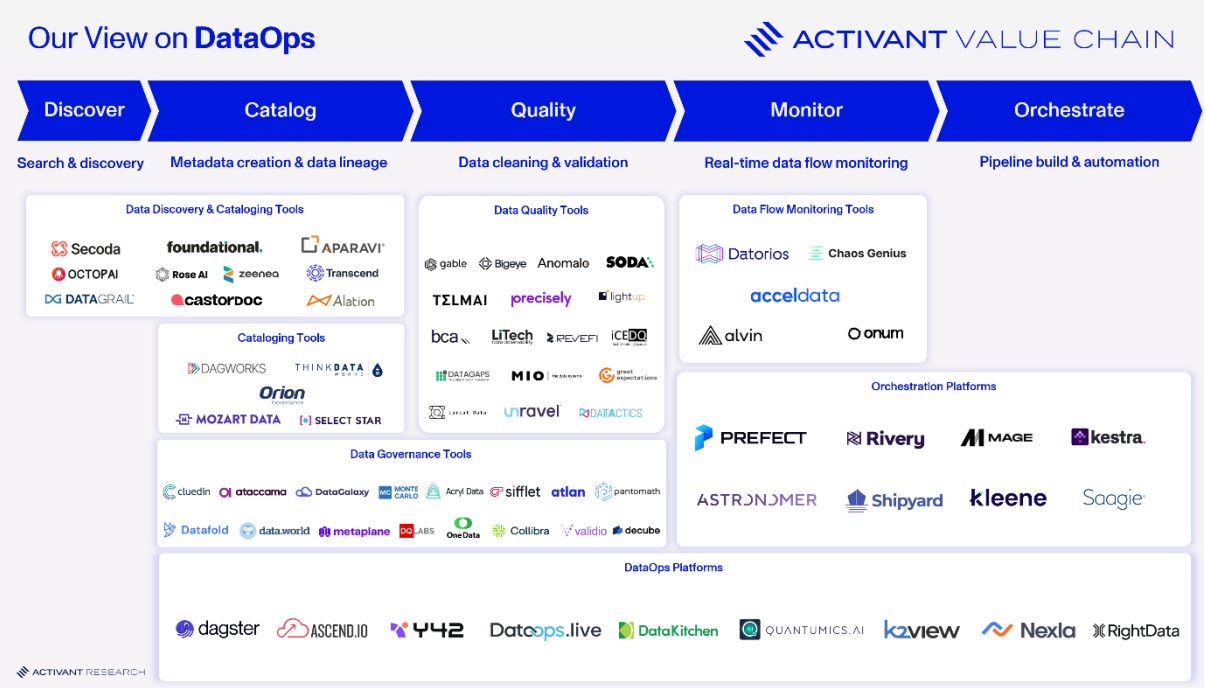
Moving from the status quo to the infrastructure that our future requires will drive a massive upgrade in both data tooling and practices. We think that DataOps presents that upgrade.

> Data engineering has evolved from a niche specialty, managed by PhDs, to a widespread necessity. As every company becomes data-driven, we need DataOps to allow teams to use these extremely powerful data tools efficiently and safely, much like how DevOps revolutionized software development.
>
> **Sean Knapp, CEO, Ascend.io**

DataOps is the Key

DataOps, like DevOps, is a *methodology*, but also encompasses several key technologies. At its core, DataOps is about making data teams more efficient - breaking down silos between themselves and IT staff and then accelerating data analytics and the AI lifecycle. Applied correctly, these tools improve the quality and reliability of final data outputs.

From a technology perspective, DataOps products augment the data infrastructure to achieve these goals. Using our Activant Value Chain below we define five core capabilities in the context of the data pipeline infrastructure and summarize how we view the DataOps market:



- **Data discovery** capabilities enable automatic systems searches to discover new or previously overlooked data across all sources. This capability is rarely seen as a stand-alone solution but is rather coupled with data cataloguing capabilities.
- **Data catalogs** serve as a single source of truth for exploring datasets across fragmented infrastructure and are essential for facilitating data lineage. Lineage visualizes and traces data flow, recording usage details throughout. Secoda enhances cataloging capabilities by using AI for data tagging and has a built-in AI copilot, focusing on creating and analyzing metadata to categorize datasets and present data in a user-friendly way. Vendors like Octopai, however, are more focused on lineage - making use of the catalog metadata. Users are able to understand how data is being leveraged and can assess how changes may impact downstream systems, preempting potential disruptions.
- **Data quality** tools implement automated validation/testing at various points along the pipeline to ensure the accuracy and reliability of users' data. Soda is an innovator in this space, providing out-of-the-box quality checks and empowering users at all technical levels by implementing AI to convert natural language processing (NLP) into RegEx/SQL

for generative AI quality checks. Comprehensive quality tools such as these quantify reliability levels, reduce time spent identifying errors, and automate manual testing procedures.

- **Data flow monitoring** tools collect and analyze logs and performance metrics across the entire pipeline to flag errors/failures in the execution of pipeline jobs. Providers collate and present the metrics in dashboards to provide visibility into data pipelines. Acceldata's solution stands out in this segment thanks to their dual focus on the monitoring of data pipelines as well as the data itself, creating a unified assessment of the health of a customer's data lifecycle. They also offer a built-in co-pilot, as well as platform audit reports for vendors such as Databricks, Hadoop, Kafka, and Snowflake.
- **Data orchestration** automates the pipeline end-to-end, triggering workflows based on schedules, events, or other preconditions. Orchestration platforms act as an authoritative source of truth by defining workflow execution and implementing optimizations such as parallel processing to increase data throughput. Orchestration, traditionally a tech-heavy task, is moving towards more accessible tools that incorporate user-friendly low-code options. Prefect is a leader that strikes a good balance – providing user-friendliness while still maintaining high-code capabilities to offer flexibility to data engineers.

In addition to the key areas of our value chain outlined above, there are a few key trends that drive the market direction. These include notable platforms, the proliferation of open-source and the rise of data products.

Platforms

Platform offerings such as Dagster and Ascend.io stretch across the value chain with the most full-featured products on the market. Dagster adopts an asset-orientated approach (rather than a task/workflow-oriented approach), which enhances debugging, reusability, visibility, flexibility, and data management. Following Ascend.io's recent partnership with Wizeline, the company is implementing AI-enabled solutions that offer cost optimization, discovery of security vulnerabilities, improved architecture, and AI-assisted orchestration.

> Dagster is an asset-oriented data orchestrator featuring built-in global lineage within a multi-tenant framework. While deployable across various environments, it offers data teams unmatched visibility into their entire data stack through a single pane of glass and serves as a strong foundation for a company-wide data platform.
>
> **Pete Hunt, CEO, Dagster Labs**

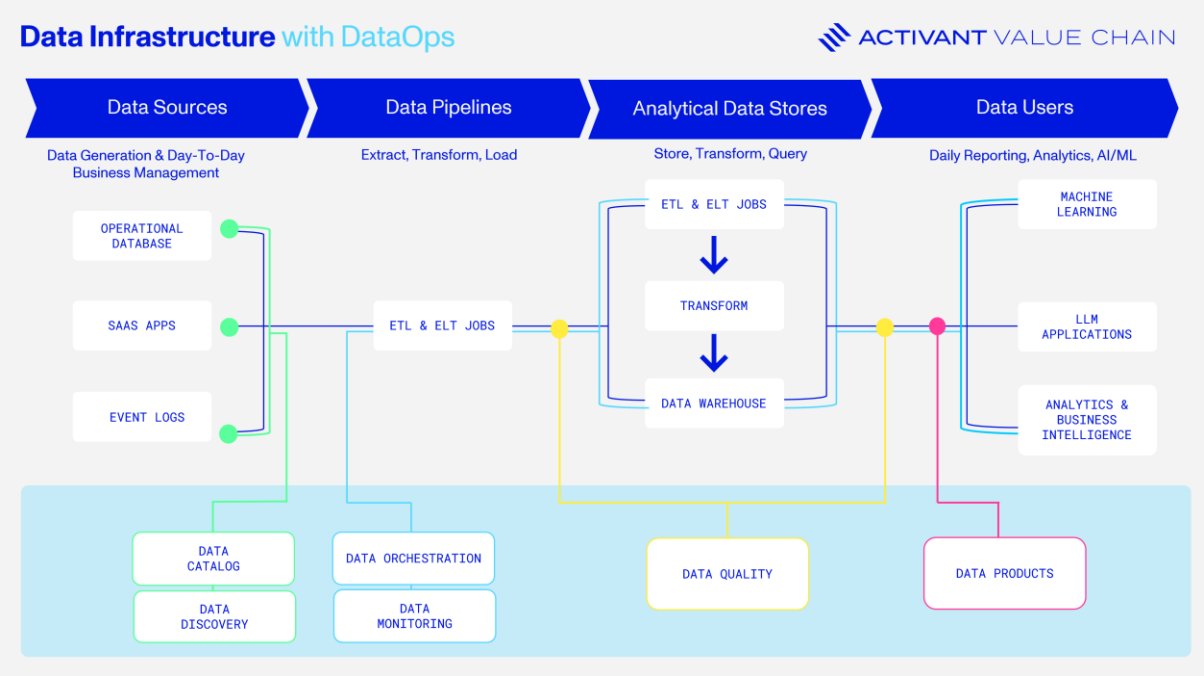
Open Source

Throughout the value chain (notably in orchestration), open-source software (OSS) tools (like Airflow) have found significant traction, but increasing adoption of managed orchestration platforms is underway as they decrease labor overheads and provide more advanced capabilities than their OSS counterparts.

Data Products

Providers are not only embracing core DataOps capabilities but also integrating data products and low-code/no-code (LCNC) functionality into their strategies as a way of liberating data for non-technical users and achieving data democratization. Vendors like RightData, DataOps.live and Ascend.io are working hard to create data hubs and marketplaces, offering a user-friendly "online shopping" experience for data to be accessed by all teams.[16]

As we can see below, DataOps augments and improves the existing flow of data. Rather than surfing thousands of data sources, data users simply go to their **Data Catalog** to explore data, their ELT & ETL jobs are executed autonomously by **Data Orchestration** and any issues are picked up by **Data Monitoring.** As data flows through to its end location, it can be tested for accuracy by **Data Quality** tools both before it is loaded in the analytical store and when it is queried by the target system. Finally, data consumers can discover analytics-ready datasets as **Data Products,** without worrying about all the integration and transformation that took place before that. **Previously, data teams were suffering from significant manual work, a lack of visibility, and poor-quality data. Now, their workflows are automated, monitored, and any data quality issues are detected automatically.**



While the traditional data flow can exist without these DataOps tools, the value proposition that DataOps brings to the data team is becoming clear.

Early Adopters are Reaping the Benefits

Gartner believes that data teams who make use of DataOps tools and practices will be 10x more productive.[17] Given the scale of issues that data teams are facing, such an immense gain *feels* achievable, and the research backs that up:

1. **Faster time to market** is achieved by teams who use DataOps tools to reduce time spent on manual work like pipeline builds and error troubleshooting. Users are reporting an 80% reduction in the time taken to develop data pipelines, equating to a **5x ROI through workflow automation**.[18,19] Data consultancy Dataengine found that implementing DataOps practices for a customer had the potential to reduce time to market from 100 days to 10.[20] The Enterprise Strategy Group found that through these improvements, engineers using DataOps tools are 5x to 7x more productive.[21]
2. **Reduced infrastructure costs and greater scalability** are the primary benefits of automation. In some cases, users have reported a 10x increase in pipeline throughput and a significant reduction in processing times, from 8 hours per client to 3 hours for all clients.[22] In the Dataengine case referenced above, they also found a **90% reduction in data processing time**, and where customers leveraged cloud computing, they were able to achieve approximately **30% savings on infrastructure costs**.[23]
3. **Improved data quality and reliability** can save organizations up to an average of $12.9 million annually.[24] By automating data quality solutions, companies are rebuilding trust in AI models, dashboards, charts, and statistics that are relied on for decision-making.

With such immense gains available on the table, we see DataOps tools becoming a mission-critical element of every enterprise tech stack, driving efficiency, scalability, and accuracy that modern data teams can't afford to overlook.

Final Thoughts

With the rapid adoption of generative AI and the increasing need for data on demand, DataOps will undoubtedly shift from a "nice to have" to a necessity. Its ability to streamline operations, automate workflows, and ensure data quality will make it essential for modern enterprises aiming to stay competitive in a data-driven world. Reflecting on all that technology has achieved with broken and inefficient data practices, it's incredibly exciting to look forward to what we can achieve with modern analytics, generative AI and **10x better data platforms**.

If you have a different view or are building in this space, we'd love to hear from you.

End Notes

[1] The Securities and Exchange Commission, Meta Platforms, Inc. FORM 10-K, 2023

[2] Uber, How Uber Accomplishes Job Counting At Scale, 2024

[3] IDC, Worldwide Big Data and Analytics Software Market, 2023

[4]Gartner, Gen AI Seminar, 2024

[5] IDC, IDC WW Global Datasphere Structured & Unstructured Data Forecast, 2022 - 2026

[6] Exploding Topics, Amount of Data Created Daily, 2024

[7] Redpoint, Cloud Infrastructure, 2022; Survey of 60 data leaders within organizations of 500+ FTEs

[8] Gartner, Hype Cycle for Generative AI, 2023

[9] McKinsey, The state of AI in early 2024, 2024

[10] Forrester, DataOps Can Build The Foundation For Your Generative AI Ambitions, 2024

[11] Okta, Businesses at Work, 2022

[12] Saagie, The ultimate guide to DataOps, 2019

[13] Ibid

[14] Forrester, DataOps Can Build The Foundation For Your Generative AI Ambitions, 2024

[15] Ibid

[16] IDC, Unlock Data Value by Enabling Data Product Sharing, 2024

[17] Gartner, Market Guide for DataOps Tools, 2022

[18] Enterprise Strategy Group, Analyzing the Economic Impact of Ascend Data Automation, 2023

[19] Senior Data Engineer at Clearcover, Tegus Interview, 2022

[20] Dataengine, A Leading Economic Hub Optimises Operations with Data-Driven Solutions, 2024

[21] Enterprise Strategy Group, Analyzing the Economic Impact of Ascend Data Automation, 2023

[22] Ibid

[23] Ibid

[24] Gartner, How to Improve Your Data Quality, 2021