### Introduction

DataOps is a collection of best practices, technologies, workflows, and principles. It aims to increase data reliability in an organization through the collaboration of data producers and consumers. DataOps also works to automate processes that were previously manual, and monitor data quality.

This guide will provide an overview of dataOps: why it's important, how it's implemented, the tools and processes involved and the signs your organization is running behind the times and need to implement dataOps.

### The Origins of dataOps

The term "dataOps" emerged in the late 2010s as a response to the challenges posed by the increasing volume, variety, and velocity of data in modern organizations. It aims to apply the principles of devOps - which focuses on improving collaboration and efficiency between software development and IT operations teams - to the realm of data management and analytics.

However, dataOps is not JUST devOps applied to data engineering and analytics. While the two concepts share some similarities in principles and practices, there is one key difference: the users and stakeholders involved. DevOps serves software developers and engineers, while dataOps serves data scientists, analysts, and business users. These groups have very different needs, skills, and priorities.

Software engineers think a lot about what the tech stack should look like to enable the correct data flow. On the other hand, data scientists and analysts think about the content of this data flow in order to understand the importance of each part of the data operation working together. Data professionals help software engineers improve the whole software development process.

### What is dataOps?

While different organizations and companies have defined dataOps differently, some of the widely agreed upon core pillars include:

1. **Emphasizing collaboration and communication:** DataOps encourages close collaboration between data professionals, data scientists, IT operations, and business stakeholders. Effective communication and collaboration are vital for aligning goals, sharing knowledge, and driving innovation.
2. **Prioritizing data quality and reliability:** DataOps recognizes that high-quality data is essential for making informed decisions. It emphasizes the need for data governance, data validation, and data quality monitoring throughout the data lifecycle.
3. **Automating data processes:** Automation is a key aspect of dataOps. By automating data pipelines, data integration, and data management tasks, organizations can reduce errors, improve efficiency, and accelerate the delivery of data-driven insights.
4. **Promoting self-service analytics:** DataOps aims to empower business users and data scientists with self-service analytics capabilities. By providing user-friendly tools, easy access to data, and promoting data democratization, organizations can foster a culture of data-driven decision-making.
5. **Enabling scalability and flexibility:** DataOps recognizes the need to handle diverse data sources and scale data processes efficiently. It promotes the use of cloud technologies, containerization, and scalable architectures to support the growing demands of data management and analytics.

More concretely, DataOps recommends treating “data as a product”. This means applying software engineering best practices like CI/CD and testing to data and the data pipeline, so it can become a well-governed, high-velocity asset. Capabilities such as data versioning, time travel, and sandboxing environments are enablers of DataOps, but the mindsets and disciplines around testing, monitoring, and continuous improvement are most important.

### Why is dataOps Important?

With the growth of mobile, IoT, and machine learning, more data is being generated from more sources than ever before. At the same time, business decision-makers increasingly want advanced data capabilities at the speed they've come to expect from software.

To meet these demands, data teams need to be able to quickly acquire and load new data sources, build proof-of-concepts in sandboxes, assemble production pipelines, and deliver correct data to the business for decision-making. The current landscape of disjointed, ungoverned data tools makes this difficult and inefficient.

Without dataOps, the cost, complexity, and risk involved in managing and processing data at scale is too high. Valuable data can easily become damaged or lost due to a lack of governance and automation. Data teams spend their days reactively hunting down data bugs or fixing data pipelines.

For most organizations, data is one of their most precious and high-value assets. But without the proper tools, processes, and mindsets to manage it, data loses its ability to drive business value. DataOps brings discipline and quality to data management so that organizations can tap into the potential of their data.

### How to Implement dataOps

Implementing dataOps requires organizational culture, processes, and technology changes:

### **Add data and logic tests.** Implement automated testing to verify data quality and validate new code changes. Start with simple tests and build over time. Tests provide statistical process control for your data pipelines and give you the confidence to release changes faster.

### **Use a version control system.** Maintain all code and configuration controlling your data analytics pipelines in a version control system like Git. This enables branching, merging, and collaboration.

### **Branch and merge.** Create branches in your version control system to work on new features or updates in isolation. Merge code back to the main branch when complete. Branching and merging boosts productivity by allowing parallel development.

### **Use multiple environments.** Provide team members with their own environments containing copies of data and tooling. This prevents conflicts and allows experimentation without impacting others. Use containers if environments are resource intensive to set up.

### **Reuse and containerize.** Break the data analytics pipeline into reusable components and containerize complex logic. This makes it easier for team members to understand, deploy, and build upon each other's work.

### **Parameterize your processing.** Design flexibility and options into the data analytics pipeline. Allow parameters to control which data sets, processing logic, outputs, etc. are used for different runs. This accommodates a range of use cases without rework.

1. **Develop a mindset of continuous improvement:** The biggest change to implement perhaps, is a mindset shift. “It's really about getting your development team together with building data pipelines or delivering dashboards and reports, whatever kind of data analytics artifact it is, and figuring out what are the bottlenecks that are preventing us from delivering faster, better, cheaper to meet the needs of our customers and keep them happy and satisfied and then knocking those down one by one,” said Wayne Eckerson, the founder and principal consultant at Eckerson Group.

### 

### DataOps Tools and Examples

The following tools and frameworks are the most common ones that organizations currently use to enable dataOps:

* **Data workflow orchestration engines:** Tools like Airflow, Prefect, dagster, and Kubeflow Pipelines help schedule, monitor, and improve data workflows.
* **Data quality monitoring tools:** Tools such as Great Expectations, DBT Tests, and Bigeye help define data quality metrics, set thresholds, monitor data, and trigger alerts.
* **Data catalogs:** Data catalog tools like Amundsen and Atlan and their incumbents, Colibra and Alation, allow you to discover and govern data assets. They serve as a single source of truth for data.
* **BI platforms:** Self-service BI tools give end-users a way to explore, visualize, and analyze data on their own using an intuitive interface. Examples include Power BI, Tableau, Looker, and Metabase.
* **Notebooks:** Jupyter notebooks provide a collaborative environment for data exploration, visualization, and analysis. Notebooks allow data teams and consumers to work together in one place.

Some companies known for their successful DataOps implementations include:

* **Netflix:** Netflix has [talked extensively about its Data Mesh architecture](https://netflixtechblog.com/data-mesh-a-data-movement-and-processing-platform-netflix-1288bcab2873) which aligns well with DataOps principles. They focus heavily on monitoring data quality, fostering a collaborative data culture, and building self-service tools for end users.
* **Spotify:** [Spotify has built an impressive data infrastructure and tooling ecosystem to support DataOps.](https://engineering.atspotify.com/2020/07/leveraging-mobile-infrastructure-with-data-driven-decisions/) “Back in 2017, Spotify put renewed emphasis on data and its importance to the overall competitiveness of the business. The goal was to use data and insights to drive the business forward and maximize the opportunity for data-driven innovations. The data infrastructure teams were tasked with making data generation, collection, and analysis easy and accessible,” the company wrote.
* **LinkedIn:** LinkedIn is an example of an early adopter of DataOps principles. They have been focused on data democratization, building self-service data tools, monitoring data health, and [building reliable and discoverable data products](https://engineering.linkedin.com/blog/2022/super-tables--the-road-to-building-reliable-and-discoverable-dat).
* **Uber:** Uber is another example of a company that was a pioneer in the DataOps space, building out tools like [Databook](https://www.uber.com/blog/metadata-insights-databook/), the in-house data catalog, Michelangelo, the in-house machine learning platform, and [WorkflowGuard](https://www.uber.com/blog/introducing-workflowguard/), the in-house workflow governance and observability platform.

.