Nexla

WHITE PAPER

An Introduction to DataOps

How DataOps helps data teams move faster

TABLE OF CONTENTS

Nexla

AN INTRODUCTION TO

DATA OPERATIONS

The term “DataOps” or “Data Operations” is having a moment. With startups, incumbents, and analysts all using this term, you might be left scratching your head, wondering what it means. In this whitepaper, we explain what DataOps is, why it's relevant now, and how to build the DataOps capability within your organization.

What you can expect from this whitepaper:

- Understand the term DataOps and where it comes from
- How DataOps can benefit organizations
- Why people are talking about DataOps now
- The ins and outs of DataOps: what exactly does it control?
- The key factors for a successful DataOps function

> **"DataOps tools and processes drive the integration and automation that make data flows scalable, repeatable, and predictable for data engineers, data scientists, and business users.**

What is DataOps?

WHAT DOES "DATAOPS" MEAN?

When we hear the word "Operations", it brings to mind concepts like process, efficiency, automation, reliability, and scale. At its very core, the idea of operations is to take a functioning unit and make it scale.

While definitions continue to emerge, we believe that the essence of DataOps lies in scalability and repeatability.

At Nexla, we define DataOps as an organization-wide data management practice that controls the flow of data from source to value, with the goal of speeding up the process of deriving value from data. The outcome is scalable, repeatable, and predictable data flows for data engineers, data scientists, and business users. DataOps is as much about people as it is about tools and processes.

A DataOps practice can open data access to more stakeholders within an organization, further increasing capacity for scale. This collaboration can help data teams move faster. Tactically speaking, DataOps takes care of the grunt work typically placed on IT or data engineers. This includes integrating with data sources, performing transformations, converting data formats, and writing or delivering data to its required destination. DataOps also encompasses the monitoring and governance of these data flows while ensuring security.

Ops roles centered around infrastructure have been common starting with Network Operations in 1960s, Security Ops, and more recently Devops is helping companies ship better software faster. However, when it comes to data, the stakeholders that care about and interact with data in a company are far broader than those who interact with infrastructure or servers.

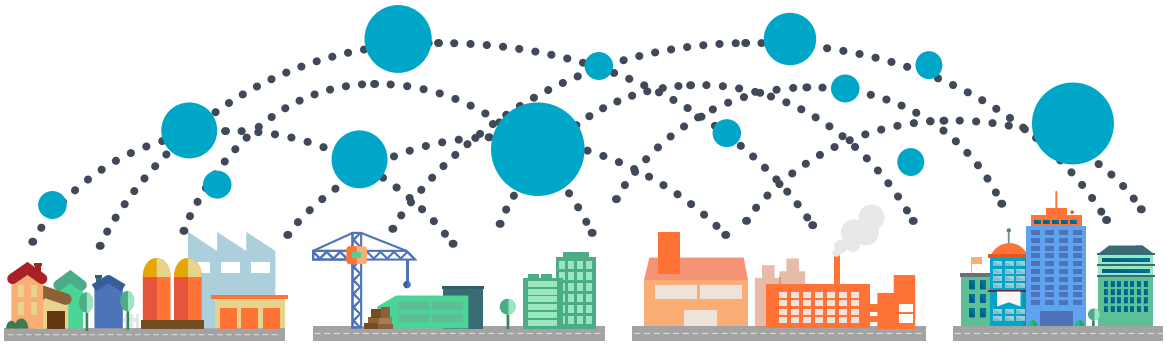
Another example of Ops success is found in the world of online advertising. At companies like Google and Facebook, advertising operations (AdOps) specialists help make sure ad campaigns are delivering results for large advertisers. Through deep subject matter expertise, data fluency, and analytical chops, AdOps managers can often maximize campaign performance and troubleshoot issues. They are not data engineers, but they must process data nonetheless. Because tools have been built to help scale their work, they are able to solve problems before they land on product or engineering's plate. Online advertising would not be able to scale to millions of advertisers and billions of dollars without the AdOps function making sure things run smoothly.

At Nexla we believe that restricting DataOps to the purview of the engineering team or data team can be a narrow view, for reasons we'll explore in the next section. The more "data leverage" you can create in an organization, the more likely you are to be successful.

Ultimately, DataOps is not just about tools and processes. It represents a greater cultural shift that breaks down the silos between what has traditionally been viewed as "data backend" that produces usable data and "data frontend" that derives value from data." Only by enabling more users within their data systems can companies realize the economic benefits of becoming data-driven.

Why now?

WHY IS NOW THE TIME FOR DATAOPS?



In today's environment, the number of data flows a company has to manage is never-ending. Because the volume, velocity, and variety of data are increasing, companies need a new way to manage this complexity. In order to maximize data efficiency and value creation, scalability and repeatability of data work are essential.

The tasks of data prep, integration, monitoring, and governance have been done for decades. What is new now? We believe there are five transformational forces that are changing the game.

1. CONSUMERIZATION OF ENTERPRISE AND THE RISE OF DATA FLUENCY

Everyone is an expert in some data

Then: Can IT get this data for me?
Now: Why can't I do this myself?

Enterprise software got a user experience makeover. It's harder and harder for data and analytics software vendors to ship software that requires weeks-long trainings in order to operate. Today's business software is more user-friendly, and line of business users have come to expect a level of self-service that would have been unthinkable ten years ago.

The expectation for self-service has grown in conjunction with the rise of data fluency. More and more employees within organizations are now data experts—they are the subject matter masters, the partner relationship owners, and the analytics power users. Organizations have made huge investments in becoming data-driven, and more and more stakeholders now have the ability to drive value with data. They won't wait for IT-centric data solutions.

2. RISE OF THE DIGITAL NATIVES

To disrupt a sector, software needs to connect with data in that ecosystem

Then: Sell technology to non-tech sector
Now: I will leverage data in this sector to change how things work

Software is still eating the world. Marc Andreessen's seminal essay still rings true, with tech startups continuing to disrupt everything from groceries to shipping logistics. The backbone of much of this disruption is data. Instacart can let you shop from thousands of grocery stores on your phone, Convoy can help you find space on a semi truck from your browser. For companies to compete, they need to be able to quickly and efficiently integrate with more and more partners. The threat of disruption is real, and it's increasingly driven by access to data.

3. BUY, NOT BUILD

Build your core competency, buy everything else means more B2B partnerships, more B2B data

Then: Can I get a report?
Now: I need raw data, along with report.

The rise of B2B software has been a productivity boon, but a data nightmare. Organizations have critical data held hostage by myriad SaaS providers, including customer data, marketing data, advertising and attribution data, sales productivity data… the list goes on.

While relying on best-in-class software for the task at hand makes companies more productive, it can make leveraging or analyzing that data cumbersome. Practitioners don't want to wait for IT to integrate feeds into the data lake (see force #1) and engineers don't want to integrate or maintain these service APIs. For evidence of the business value of this data access, see Salesforce's $6.5 billion acquisition of Mulesoft.

> **"DataOps is not a feature or a tool, but a broader practice of making data work an infinitely scalable and repeatable process.**

4. BIG DATA MATURITY

Many companies have built and operationalized their analytics stacks

Then: What analytics system should I use?
Now: I need to feed more data into my AI

Over the last ten years companies have been building out their analytics stacks and migrating to the cloud. For most organizations, analytics platforms are established and teams can turn their focus to optimization and continuous improvement.

As most companies invest in machine learning and AI, the question is no longer what analytics system should I use but how can I feed and tune my hungry models? Since quality data is essential for any ML or AI success, access to more data sources will only increase.

5. MORE DATA, MORE PROBLEMS

Data doubling every 18 months

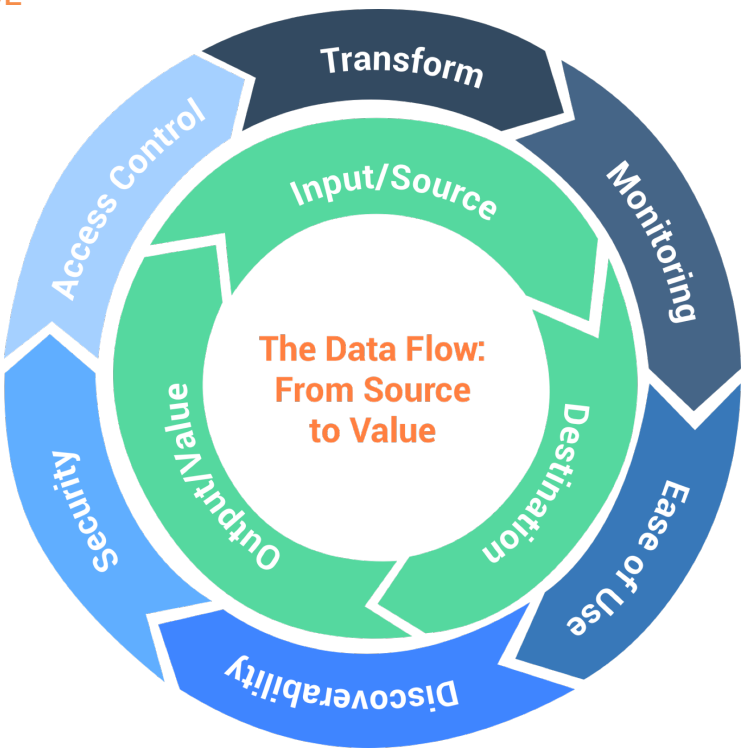
Then: I have a few data sources
Now: I have a long roadmap of data I want to use. How can I get to that faster?

According to the latest Definitive Data Operations Report, the average company is managing close to 5,000 data sets, has a front-end data user to data engineer ratio of 5 (with some companies closer to 30!), and integrates with more than 10 third party partners. Companies are processing data in a variety of formats with varying frequency, and the heterogeneity of data shows no sign of slowing.

These five forces are pushing a need for faster data movement, across thousands of data flows, with a high degree of variation. DataOps is a practice that ensures all aspects of these tasks are carried out efficiently to deliver data to where it can add value. DataOps is not a feature or a tool but a broader practice with the goal of making data work an infinitely scalable and repeatable process.

The Data Flow

FROM SOURCE TO VALUE



We illustrate the data journey with the above graphic. Starting with the input, we have data sources (or data creation—we can include IoT sensors or connected cars here). These sources can be internal or external. This number will only increase as machine learning models require ever-more data.

Data is then connected to a data pipe. We believe this can be a "smart pipe" that helps prep, clean, and transform data before it is delivered to the destination. This smart pipe can span across organizations, allowing both the sending and receiving company to transform data to suit their needs. The ultimate destination could be another database, an API, or even a simple CSV. At the end of the data journey, data should be ready for advanced analytics, machine learning, partner consumption, or another use case. This is the ultimate goal of the data: to provide business value.

Throughout the data journey, DataOps must monitor the data, secure and manage access, all while providing ease of use and discoverability. Let's take each of these requirements in turn.

MONITOR:

Monitoring the entire data journey is of critical importance. It means keeping track of the data streams that are operational, being alerted immediately when a schema changes, or an abnormality is detected. As the number of integrated data streams increases, manual monitoring becomes untenable.

SECURE AND MANAGE ACCESS:

Especially when ingesting or sending data from or to third parties, security is important. Attribute-level access management is required, and data governance must be controlled. With many data sources, the need for a central "command and control" becomes clear.

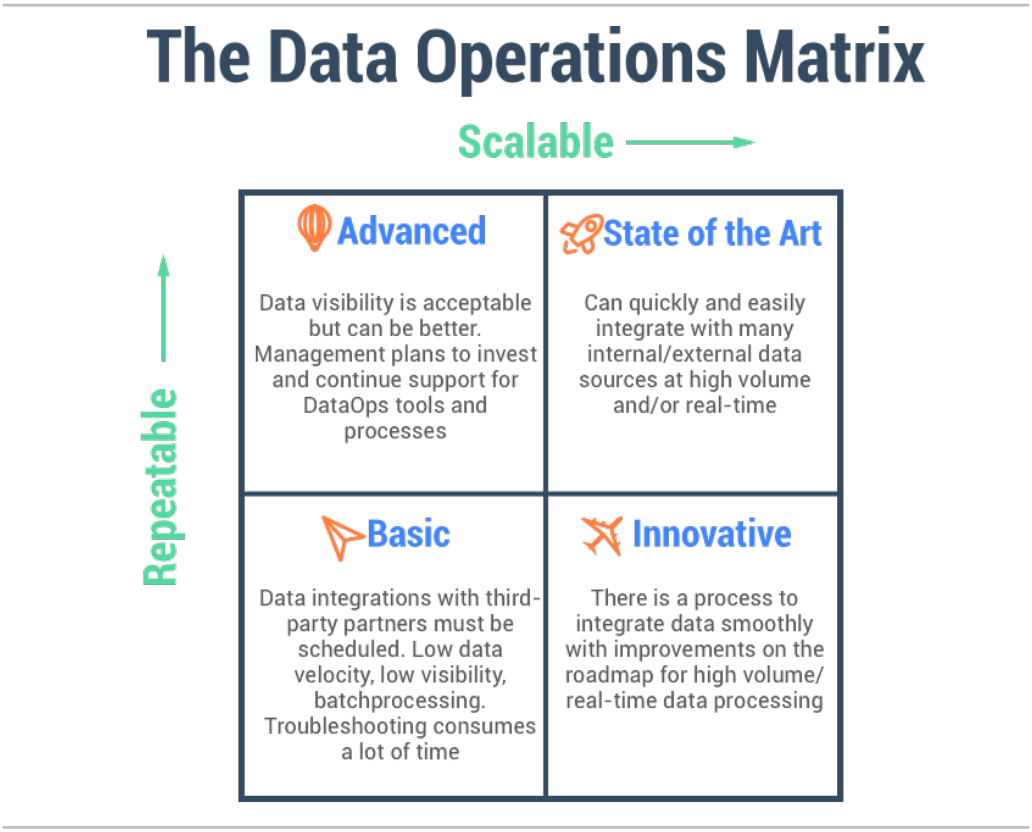
EASE OF USE AND DISCOVERABILITY:

Since the ultimate goal of DataOps is to drive business value, it is important for business line users to have access to data. Beyond proper access, they need to easily be able to retrieve the data they need to analyze it with their preferred tools. This can mean exporting to another database to query with SQL or even Excel. It can also mean feeding the right data to an analytics platform like Tableau or Looker. Finally, discoverability and data mapping is critical as data becomes democratized. Understanding what data is available and its schema is the first step to business user analysis.

Assessing your DataOps

SCALABILITY & REPEATABILITY

Whether you're a DataOps guru or you're just getting started, it's important to know how scalable and how repeatable your DataOps are. To assess the current state of a company's Data Operations, we developed the DataOps Matrix. This is a helpful way to rate your capabilities across what the two most important DataOps vectors: scalability and repeatability. Once you've determined where your organization lies on these axes, you'll be able to understand your current DataOps capabilities.



FACTOR ONE: SCALABILITY

Scalability in this context is a measure of how easily a DataOps system can grow the volume of data, the number of data users, and operational complexity.

A data operations infrastructure that is highly scalable can handle high volumes of data and process it in near real-time. How you define "high volume" will be dependent on the industry. For example, if we consider online advertising technology, high volume can be measured in terabytes (or even petabytes) per hour. If we consider retail product data, volumes could be in gigabytes per day. While the definition of "high volume" is relative, the ability to elegantly process data at max volumes is critical.

In addition to scaling with data volumes, a DataOps infrastructure needs to scale with people. As companies ingest and send more data, the number of people who need to work with the data will only grow. Business analysts, data analysts, technical support teams, implementation teams, partnership teams and more all have people who understand the data very well but don't have programming skills. Empowering them with tools is essential to the

scalability of DataOps. If they can handle 80% of use cases with the right tools, that frees up data engineers to focus on the most complex problems.

This creates organizational leverage around data. The right tools, processes, and people as part of the DataOps solution can have a force multiplying effect.

> **"A DataOps infrastructure needs to scale with people. Empowering them with the right tools is essential.**

FACTOR TWO: REPEATABILITY

Repeatability in this context is a measure of how easily a system can automate or repeat tasks.

In DataOps, rule-based automation is table stakes. Standard processes like ETL (extract, transform, load) would fit into this category. For DataOps to move data quickly and easily, these rules must be followed. When errors occur, DataOps needs to provide alerts and potentially recommendations on how to correct and reprocess errors. Errors can happen for any number of reasons—a partner changes a schema, unexpected values appear, or a colleague updates a downstream business process to name a few.

Sophisticated DataOps systems will be able to maintain repeatability despite heterogeneous data types and sources. The average company is processing data from flat files on FTP servers, APIs, and file sharing services like Box. The ability to easily process data from multiple sources is key to repeatability.

Because data flows are constantly being added, DataOps needs to continuously add new pipelines. Systems that can support duplicating pipelines, making edits, and pausing and activating flows provide the most repeatability. As the number of data flows increase, orchestration becomes more complex and important. DataOps is responsible for managing the interdependence of data flows and must provide a mechanism to ensure changes to an upstream process don't break things downstream.

Systems need to allow for development and testing before pushing pipelines to production, where monitoring becomes critical. DataOps is at its most repeatable when it can be smart about the source connections and easily integrate and transform data. DataOps platforms that allow for shared transformations across an organization can put the power of repeatability in more hands.

For businesses that need to ingest data from and send data to an increasing number of partners, repeatability might even be more important than the ability to scale.

The Future of DataOps

WHERE WILL DATAOPS GO FROM HERE?

As volume, velocity, and variety of data increases, new tools and processes are needed to extract insight. IDC expects the volume of data created to grow to 163 zettabytes by 2025, with 36% of it structured data. Today's tools, processes, and organizational structures aren't equipped to handle this massive increase in data inputs, and the increasing value expected from its output. As more of the workforce requires access to this data to perform their jobs, a philosophical shift is needed to breakdown the cultural and organizational barriers to provide scalable, repeatable, and predictable data flows.

This shift is happening because of the DataOps revolution. Companies would be wise to put the processes and tools in place now to prevent data heartache down the road. Nexla is the data operations platform that helps teams create scalable, repeatable, and predictable data flows for any data use case. Nexla helps data engineers, analysts, and business users to integrate, automate, and monitor data flows. The end result is predictable, reliable data access inside and outside the organization, for today and the next 163 zettabytes.

About Nexla:
Nexla is a data operations platform that helps teams create scalable, repeatable, and predictable data flows for any data use case. Analysts, business users, and data engineers across any sector including e-commerce, insurance, travel, and healthcare can use Nexla to integrate, automate and monitor their incoming and outgoing data flows. The end result is predictable and reliable data access inside and outside the organization.

Nexla, Inc. 475 El Camino Real Suite 305, Millbrae, CA 94030 USA | nexla.com | Copyright ©2018 Nexla, Inc.