Becoming a DataOps Expert:

Putting DataOps Strategies into Practice

Tamr

Successful businesses are data-driven businesses.

And today, businesses are accelerating the pace of their investments in data because they recognize the strategic value their data provides. By becoming data-driven, businesses empower their decision-makers to make better, more confident decisions. And these decisions lead to higher returns on investment and shorter time to value.

By harnessing the power of their data, businesses are driving greater value and improving outcomes across all parts of their organizations. They're becoming more competitive and more resilient to the constant changes of today's dynamic business environment.

In fact, in a recent survey of data and AI executives, NewVantage Partners found that 97% of businesses have invested in data initiatives to help improve their business outcomes and 92% reported that the pace of investments is accelerating. (source)

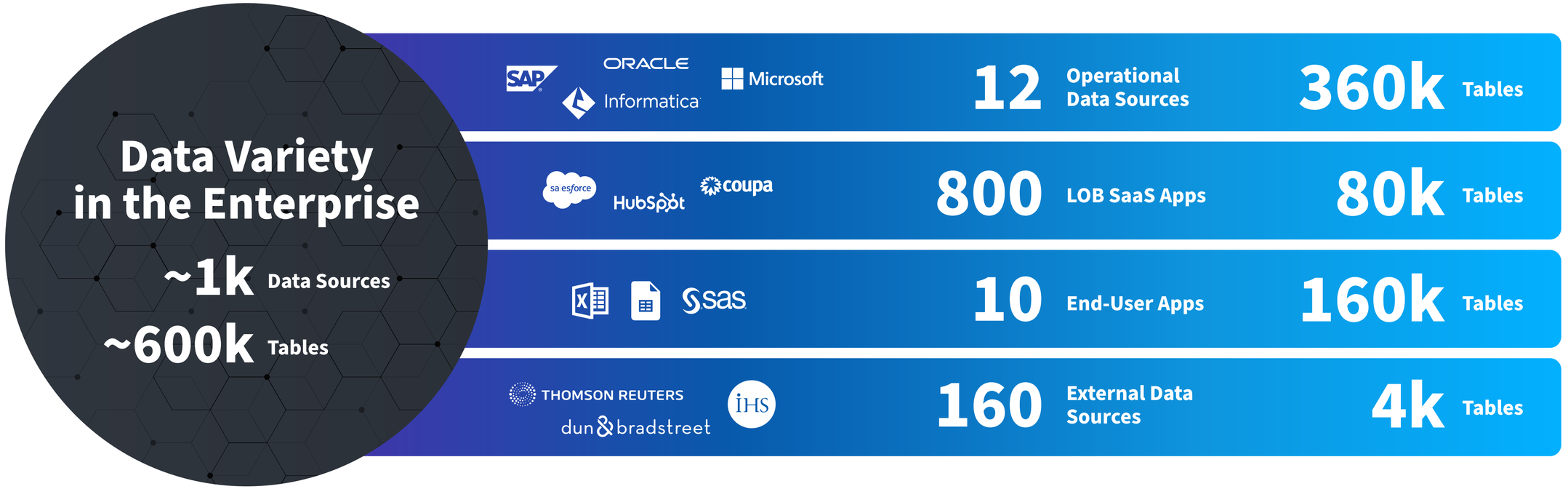
Furthermore, IDC forecasts that the spend on digital transformation will reach $2.5 trillion in 2025, over double the amount allocated in 2020. They went on to estimate that from 2020 to 2025, direct digital transformation investment spending will surpass $10 trillion. (source)

As the spend in digital transformation grows, so too does the volume of data. Businesses are experiencing exponential growth in data volume with estimates showing that the average Global 2000 enterprise has thousands of data sources and hundreds of thousands of tables.

And that growth only accelerated during the COVID-19 pandemic, forcing many enterprises to expedite their digital transformation. But it's not just data volume that is growing. Data variety is growing, too. And the speed at which new data is arriving and changing is occurring at an unprecedented velocity.

Businesses struggle with converting this immense volume and variety of data into value. It's a challenge, and one that leaves a large chasm between accessing data and the ability to drive meaningful insights from it.

But there's a new discipline emerging, fueled by the accelerating pace at which businesses are realizing the strategic value of becoming a data-driven organization. It's DataOps.



What is DataOps?

DataOps may have felt like a data buzzword just a few years ago, but that's certainly no longer the case today. In fact, DataOps is on its way to becoming mainstream. Why? Because today, many companies are citing DataOps as the reason for the success of their digital transformation and their ability to become more data-driven.

Finding the exact definition of DataOps can be a challenge as the definition often depends on the person defining it.

But today, many define DataOps by a series of data management principles organized into a loosely agreed upon manifesto, similar to the Agile manifesto made popular by software development.

Regardless of your definition, we believe there are certain key principles that define a DataOps ecosystem. And as an aspiring DataOps expert, it's important that you take note.

In 2015, Tamr's Co-founder and CEO, Andy Palmer, defined DataOps as: **a data management method that emphasizes communication, collaboration, integration, automation and measurement of cooperation between data engineers, data scientists and other data professionals.**

9 Key Principles that Every DataOps Expert Should Know

- **Start in the cloud:** it's imperative that any new, big data project start on the cloud. And there are a few reasons why. First, when you use cloud-native infrastructure, you reduce project times significantly. Second, when you use modern cloud database systems for large quantities of data, you're able to scale out natively and simplify operations and maintenance. Finally, large cloud data platforms allow you to scale up or down as required with little to no capital investment.

- **Data will change:** enterprise data sources are dynamic and change rapidly. That's why it's critical that your data infrastructure enables a continuous flow of data and treats data updates as the norm - not the exception. By designing for operations, repeatability, and automated testing and release of data, you're able to keep up with the dramatic pace of change. These principles are similar to the ones that drove the automation of software build/test and release in DevOps. And it's a primary reason why we call our approach DataOps.

- **Best-of-breed is best:** gone are the days when all solutions came from a single vendor and "one throat to choke" was the mantra. Instead, modern DataOps ecosystems should mimic DevOps, where there are many, best-of-breed and proprietary tools that interoperate via APIs. By adopting this approach, you'll select the tools and technologies that are built for purpose, giving you the best possible solution and the most reasonable cost. And because you've embraced an open ecosystem, if one vendor doesn't work out, you can easily replace them with minimal disruption to your business.

- **Tables in, tables out:** if you take our advice and adopt a best-of-breed approach, the next logical question is "how will these systems communicate?" Focus on the lowest common denominator - tables. And that includes both individual tables and collections of tables. When you adopt a tables in/tables out method of integrating these various tools and software artifacts, you can easily share or move tables using many different data services methods.

- **Lineage/provenance is essential:** properly managing lineage metadata to ensure reproducible data production for analytics and machine learning is a critical part of your DataOps ecosystem. When done right, reproducibility increases, along with confidence in the data. But it's important to note that lineage/provenance is not absolute. There are subtle levels when it comes to provenance and lineage and it is important to embrace the spectrum and appropriate implementation.

- **Feedback is critical:** most data flows from sources to consumers. But many times, there are no mechanisms in place to collect feedback from data consumers. And because the core of DataOps is about harnessing quality data, it's important to find a way to embed "feedback services" in all analytical consumption tools such as visualization tools, models, spreadsheets, etc. It's ok if, initially, this is simple. Over time, it can become more intelligent and automated, helping to automatically curate data as appropriate.

- **People are critical, too:** it's tempting to rely on traditional, deterministic approaches when it comes to engineering the alignment of data with rules or ETL. But when dealing with data at scale, you need to shift your thinking. The only viable method of bringing data together successfully is to use machine-based models (probabilistic) + rules (deterministic) + human feedback (humanistic).

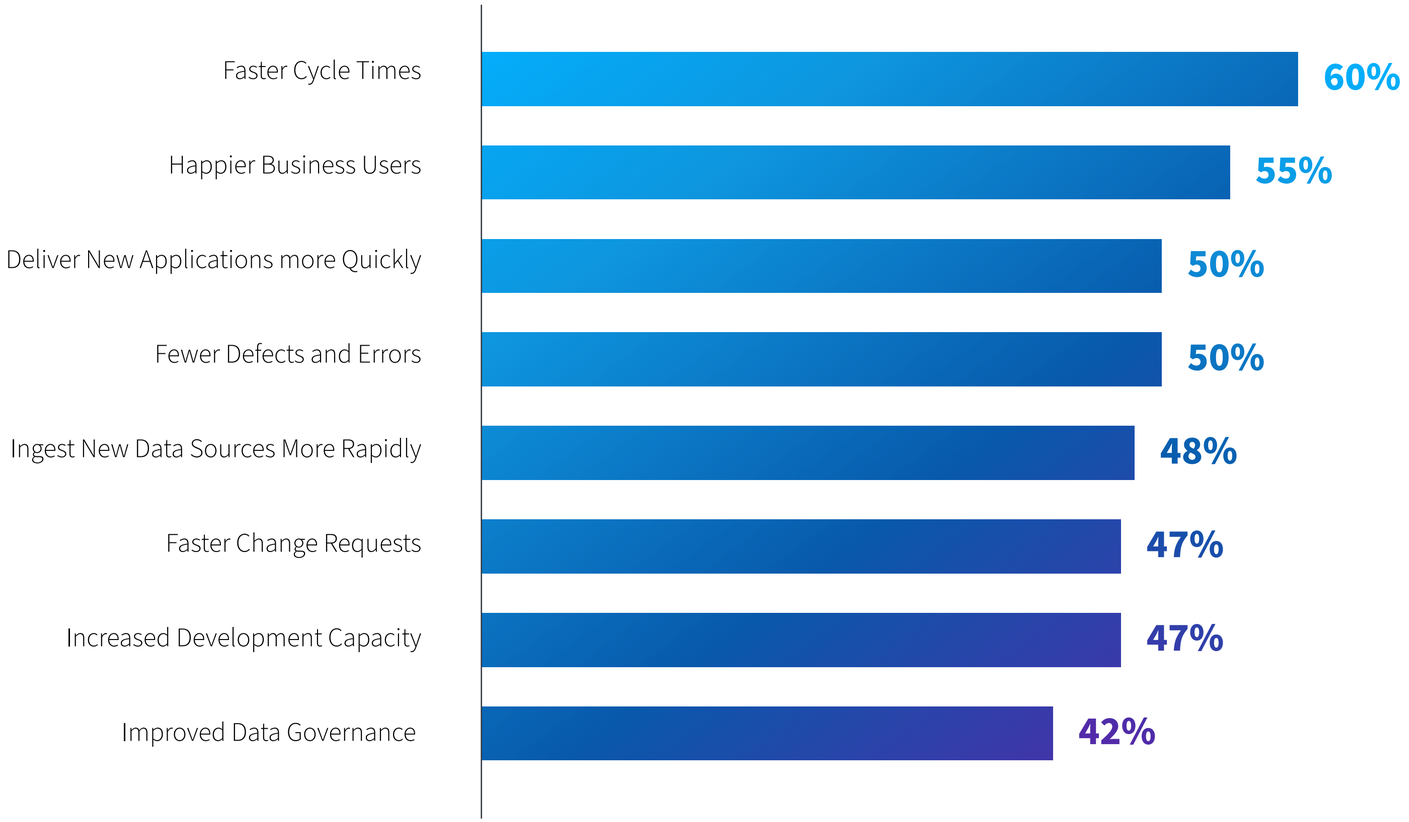
- **Embrace aggregated and federated data:** modern enterprises need an overall architecture where sources and intermediate storage of data are a combination of both aggregated and federated data. There are always tradeoffs of performance and control when you aggregate vs. federate. But over and over, we find that workloads across an enterprise require both.

- **Simultaneously process in both batch and streaming modes:** a healthy next-gen data ecosystem includes the ability to simultaneously process data from source to consumption in BOTH batch and streaming modes. These design patterns can give you the best of both worlds: the ability to process batches of data as required and also to process streams of data that provide more real-time consumption (with all the usual caveats about consistency).

The Benefits of DataOps

DataOps acknowledges the interconnected nature of data engineering, data integration, data quality, and data security/privacy. And it aims to help organizations rapidly deliver data that not only accelerates analytics but also enables analytics that were previously deemed impossible. It also provides a myriad of benefits ranging from “faster cycle times” to “fewer defects and errors” to “happier customers.”

DataOps helps data teams streamline the process of deploying code, without the worry of breaking what’s already in production. And because the size and complexity of production data pipelines varies widely – from simple exports to complex flows consisting of moving, merging, and aggregating multiple sources and fields and generating personalized dashboards – having defined processes in place is critical to helping data teams avoid burnout and realize benefits such as the ability to deliver new applications, faster.

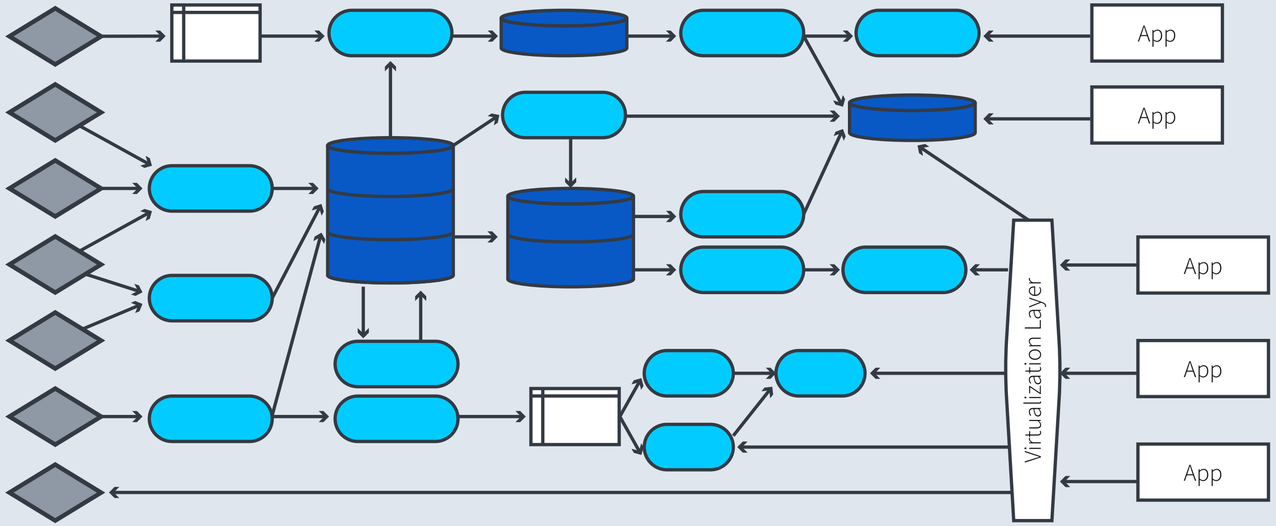


Types of Data Pipelines

The DataOps mindset also has a laser focus on continuous improvement -- Continuous Integration (CI) and Continuous Deployment (CD). DataOps practitioners continuously scan the data processing environment, looking for constraints and bottlenecks. Once they find them, they work as a team to address them.

By becoming a DataOps expert, you'll introduce your data organization to the practices, processes, and technologies needed to accelerate the delivery of analytics. You'll bring rigor to the development and management of data pipelines, enabling CI/CD across your data ecosystem.

Simple - "Export this data into a CSV file and place into this file folder."



Complex - "Move tables from 10 sources into a target database, merge common fields, array into a dimensional scheme, aggregate by year, flag null values, convert into an extract for a BI tool, and generate personalized dashboards based on the data."

What every DataOps expert knows

Savvy DataOps experts know a lot. But most importantly, they know that the key to becoming data-driven is having clean, curated, comprehensive data. And they know that DataOps is critical to reaching this goal.

Here are a few other things every DataOps expert knows.

1 Strategy always comes first

While many DataOps experts are eager to begin applying the principles of DevOps to their data organizations, the most astute experts know that they need to start with strategy.

Why? Because getting the organization on-board requires more than just an idea or a directive. It requires stakeholder buy-in. And that requires you to develop a business case, define the architecture scope, create a deployment plan, and document how you'll maintain it.

The most successful DataOps experts prioritize potential use cases for DataOps and next-gen MDM and develop a plan that starts with the most important one. They build relationships with key stakeholders, and partner with the business to ensure the project is a success. And, they ensure that the next-gen MDM initiative fits within the company's overall digital transformation strategy so that it accelerates the effort to derive value from data.

Putting it into practice:

As tempting as it is to jump straight to execution, take a breath and start by developing a comprehensive next-gen MDM strategy. Trust us, it's time well spent.

Start by documenting your goals. Then build relationships and open the lines of communication with your business stakeholders and their enterprise data architecture teams. And finally establish benchmarks so you can measure your progress – and the company's progress – against the goals you define.

2 Adopt the right technology

The backbone of every good next-gen MDM strategy is the right technology. And DataOps experts know that the right technology needs to include five important capabilities:

- **Cloud-native technologies:** Look for cloud-native capabilities that utilize the built-in elastic and ephemeral cloud and compute benefits of cloud technology.
- **Machine learning first approach:** Ensure that machine learning is at the core of the solution so it can handle increases in data volume and complexity.
- **Human feedback:** Machine learning is important, but so is human feedback. Make sure your solution has the right balance.
- **Near real-time reading & writing:** Make sure the solution provides integrations and endpoints to seamlessly connect operational processes to your company's master data.
- **Open and interoperable architecture:** Look for solutions that support a best-of-breed approach and are complementary through RESTful APIs and robust integration capabilities.

Putting it into practice:

When evaluating next-gen MDM solutions, it's critical that you choose one that ticks all the boxes. Ask tough questions like "where is your solution hosted" and "is machine learning at its core?" Be sure that you understand how the solution scales. The total investment cost. And when you can expect to realize results.

To ensure you're asking all the tough questions, read our e-book entitled "6 Questions to Ask When Evaluating Next-Gen MDM: A CDO's Checklist." Here, you'll find a full list of things to ask when evaluating next-gen MDM solutions.

Technology

3 Data mastering is the key to data-driven decisions

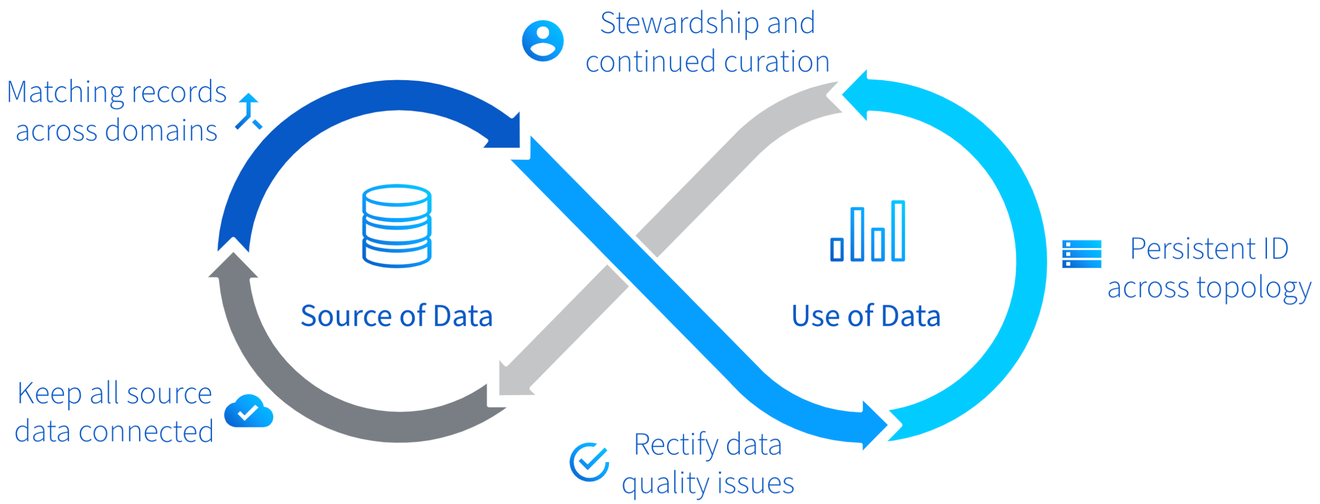
DataOps experts recognize that change is a steady state. And that for years, companies have employed traditional master data management (MDM) to define and manage their critical data.

But there are flaws with traditional MDM. It assumes that a steady state of data supply and demand was within sight. And that with just one more integrated source, one more optimized set of integration rules, or one more multi-million dollar, multi-year contract their business would reach data utopia.

DataOps experts know this simply isn't true. And that embracing a DataOps approach to data mastering is key to driving value from their data.

That's why DataOps experts employ the next-generation of data engineering tools and processes that rapidly and proactively deliver high quality data from many sources to many systems and users in an agile way. They create frameworks to consume, integrate, and deliver high quality data products around core entities to downstream systems – and incorporate feedback from those systems into its data models.

DataOps experts also know that agility, flexibility, and iteration are foundational for every data engineering process and tool. They acknowledge the scale of data supply and data demand, and allow for an iterative approach to improvement. And, they recognize the need to work with other best-of-breed technologies in order to track, manage, and use data strategically across the organization.



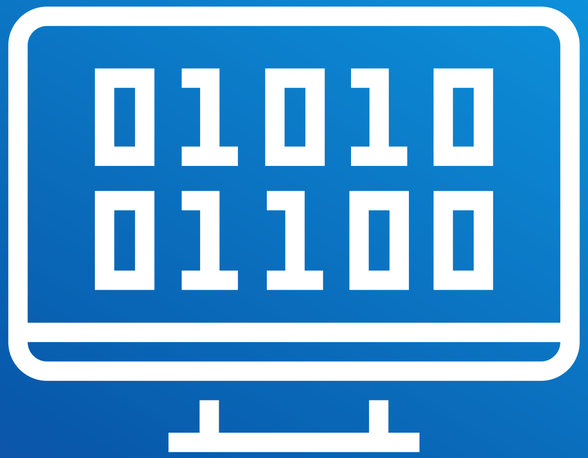
Data Mastering

Putting it into practice:
As a DataOps expert, it's your job to help your organization embrace a DataOps approach to mastering data. Put into practice principles from DevOps, Agile, Lean, and Total Quality Management (TQM).

In a DataOps environment, data is considered a shared asset, so any data models must follow the end-to-end, design thinking approach. The team should have the mindset to start small and expand their technology stack alongside their growing business needs. And remember, a DataOps engineering process must be agile, driven by collaboration and the rapid use of technology to automate repeatable processes.

Apply your expertise to bring together the right tools, processes, and people your organization needs to drive value from your data. Create a culture of continuous improvement, where data teams are empowered to spot – and fix – issues in the data pipeline. So your organization can finally unlock the full potential of your data.

To dig deeper into all the ways that DataOps can fuel your data organization, read "The Ultimate Guide to DataOps: Product Evaluation and Selection Criteria" by The Eckerson Group.



Data Mastering

4

Humans are still required

We cannot say this enough. While ensuring next-gen MDM solutions embrace the latest cloud and machine learning capabilities is important, DataOps experts know that they need humans, too.

First, leaders should create a DataOps team that includes experts with a variety of technical skills and backgrounds. DataOps methodology encourages communication and collaboration between data engineers, developers, and operations personnel.

Next, leaders should foster collaboration across all stakeholders. From engaging business users and subject matter experts in the remediation process to soliciting feedback from them on the accuracy and relevance of the data, their role is critical to next-gen MDM success. After all, clear, constant communication and common metadata, particularly when changes are being introduced into a pipeline, is essential to speed of delivery.

By combining human expertise with machine learning, you can integrate datasets from a variety of sources, allowing scale without sacrificing accuracy.

Putting it into practice:

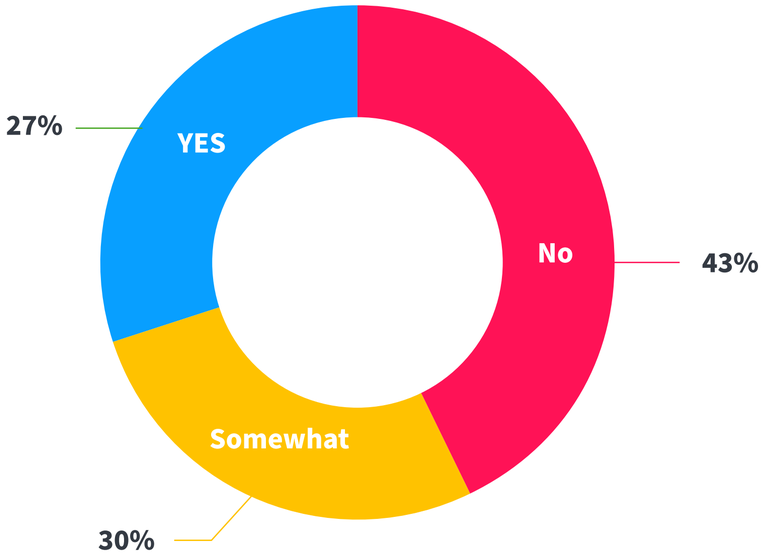
By engaging those who know the data best, not only can you improve the machine learning models quickly, but you can also drive tighter alignment between the data and business outcomes that require curated data. That's why it's critical that you build relationships and secure buy-in early on in the process.

Humans Required

How to get started

Many of you reading this guide are likely already on your way to becoming a DataOps expert. But given that just many organizations are yet to fully implement DataOps, there is still plenty of room for you to grow. Just as your data is constantly changing, so are the best practices, strategies, and case studies related to DataOps.

Does Your Organization Have a DataOps Initiative?



To stay on top of the latest thinking, we suggest following some of the leaders in DataOps.

- **Tamr**: resources that help organizations accelerate the digital journey by enabling continuously curated and consumable clean data
- **DataKitchen**: providing the software, service, and knowledge that makes it possible for every data and analytics team to realize their full potential with DataOps
- **data.world**: the largest collaborative data community that's free and open to the public
- **NewVantage Partners**: strategic advisors in data-driven business transformation to Fortune 1000 companies and industry leaders

You should also invest your time in attending in-person and online events where like-minded data professionals gather. **A few of our favorites** include:

- **Tamr DataMasters**
- Gartner Data & Analytics Summit
- Chief Data Officer Magazine's CDO & Data Leaders Global Summit

Becoming a DataOps expert doesn't happen overnight. It takes time and dedication. But the work you put in to bring DataOps to your organization will pay off in dividends when your business realizes the true potential of its data and finally becomes a truly data-driven organization.

To learn more, visit **tamr.com**.