Get more value from your data

Trends in DataOps

Bringing Scale and Rigor to Data and Analytics

By Wayne W. Eckerson

August 2019

Research Sponsored by

About the Author

Wayne W. Eckerson has been a thought leader in the data and analytics field since the early 1990s. He is a sought-after consultant, noted speaker, and expert educator who thinks critically, writes clearly, and presents persuasively about complex topics. Eckerson has conducted many groundbreaking research studies, chaired numerous conferences, written two widely read books on performance dashboards and analytics, and consulted on BI, analytics, and data management topics for numerous organizations. Eckerson is the founder and principal consultant of Eckerson Group.

About Eckerson Group

Eckerson Group helps organizations get more value from data and analytics. Our experts each have more than 25+ years of experience in the field. Data and analytics is all we do, and we're good at it! Our goal is to provide organizations with a cocoon of support on their data journeys. We do this through online content (thought leadership), expert onsite assistance (full-service consulting), and 30+ courses on data and analytics topics (educational workshops).

Get more value from your data. Put an expert on your side. Learn what Eckerson Group can do for you!

About This Report

This report is a sequel to Best Practices in DataOps: How to Create Robust, Automated Data Pipelines published in June. This report is based on a survey of 175 respondents from Eckerson Group's member list. The report is sponsored by [list sponsors separated by commas] who [has/have] exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Most organizations are plagued by data silos, poor data quality, slow development processes, and a huge gulf between business and IT. DataOps promises to address these challenges and enable data teams to develop data pipelines "faster, better, cheaper." DataOps therefore has the potential to heal the rift between business and IT and help organizations get more value from their data.

This report is a sequel to our June 2019 report Best Practices in DataOps: How to Create Robust, Automated Data Pipelines, which profiled numerous DataOps pioneers and examined the keys to their success. This report surveys data and analytics professionals and provides an overview of the trends in DataOps, including adoption rates, benefits, challenges, use cases, and data processing environments. Both reports should be read together.

Key Takeaways

- A majority of companies have yet to fully implement DataOps. More than a quarter (27%) have a DataOps initiative, 43% do not, and 30% have been experimenting.
- The most popular DataOps tool is a code repository (77%), followed by continuous integration/continuous delivery (CI/CD) tools (45%), performance monitoring tools (44%), orchestration tools (37%) and configuration repositories (34%).
- Organizations with large data environments (thousands or tens of thousands of sources and targets) are more likely to have a performance monitoring tool (65%) and orchestration software (53%) than organizations with less complex data environments.
- The biggest benefit of DataOps is "faster cycle times," selected by 60% of respondents.
- The biggest challenge of DataOps is "establishing formal processes," selected by 55% of respondents.
- The most common use case for DataOps is "data warehouses and marts" (66%), followed by "reporting and dashboards" (60%), and "self-service analytics" (56%). However, organizations with complex data environments (thousands or tens

of thousands of data sources and targets) apply DataOps first and foremost to support data science projects (64%).

- Almost half of respondents (48%) experience more than six errors a month, defined as incorrect data, broken reports, late delivery, and customer complaints.
- Half (50%) of all respondents say they take weeks or months to build a new data pipeline from scratch (17% take months), while almost a third (31%) can deploy one in minutes or hours.
- The DataOps mindset has a laser focus on continuous improvement.
- DataOps is also associated with a set of technologies, borrowed from DevOps, and adapted for data processing.

Overview of DataOps

Understanding DataOps

Definition. DataOps, short for data operations, is a set of practices, processes, and technologies for building analytic solutions, including reports, dashboards, self-service analytics, and machine learning models. It applies the rigor of software engineering to the development and execution of data pipelines, which govern the flow of data from source to consumption.

Purpose. The purpose of DataOps is to accelerate the delivery of data and analytics while simultaneously improving quality and lowering costs. DataOps helps data teams fulfill the promise of delivering data solutions "faster, better, cheaper."

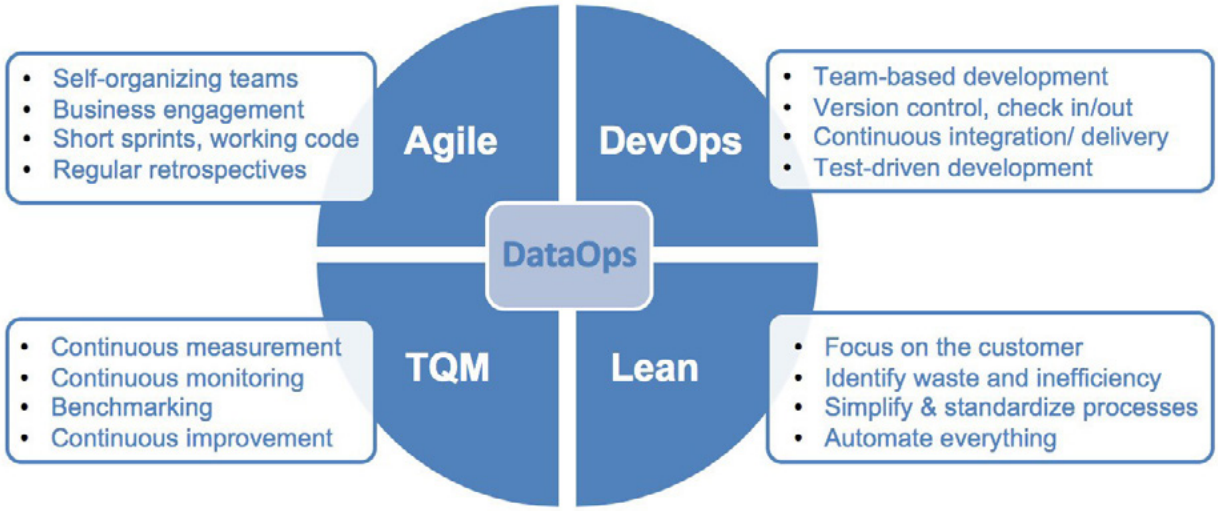
Origins

DevOps. DataOps is an extension of the DevOps movement in software engineering that uses code repositories, testing frameworks, and collaborative development tools to scale software development, increase code reuse, and automate deployments. DataOps uses the same tools but applies them to data development. Like DevOps, data developers write code (e.g. SQL, Python) but unlike most software development, they also manage the flow of data from source to target. Hence, the term "DataOps."

Another goal of DevOps is to bridge the gap between development, QA, and operations teams so organizations can shrink cycle times and reduce defects. Likewise, DataOps brings together data stakeholders—data architects, data engineers, data scientists, data analysts, application developers, and product owners (i.e., business people)—to build end-to-end solutions in an agile, efficient, and collaborative fashion.

Other Influences. Besides DevOps, DataOps borrows heavily from Agile, Lean, and Total Quality Management. Like Agile, DataOps emphasizes self-organizing teams with business involvement, short development sprints that deliver fully tested code, and regular process reviews. Like Lean, DataOps requires a laser-like focus on the customer and the creation of simple, standardized, automated processes that minimize waste, redundancy, and cost. And, like Total Quality Management, DataOps espouses continuous testing, monitoring, and benchmarking to detect issues before they turn into major problems. All three methodologies espouse a culture of continuous improvement. (See figure 1.)

Figure 1. Dimensions of DataOps



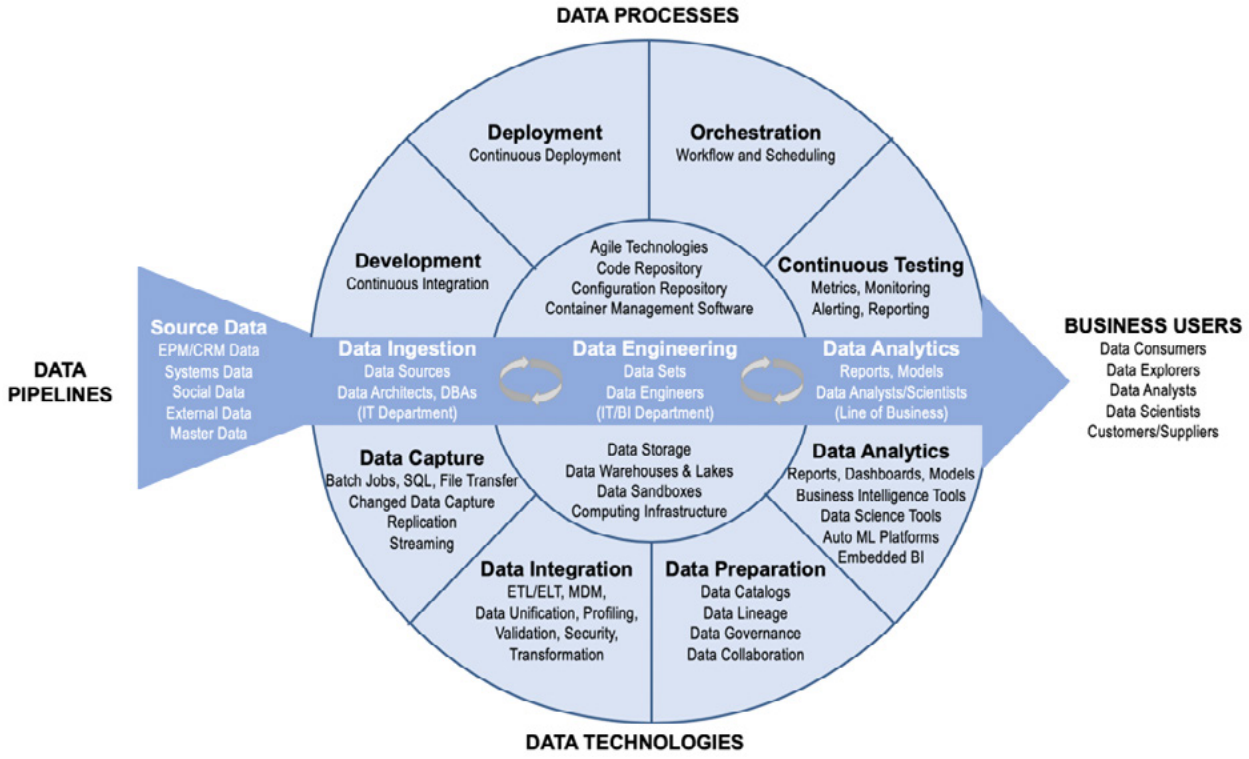
DataOps means different things to different teams. Some emphasize agile concepts and methods, while others embrace DevOps tools to streamline and govern development processes. Others focus on testing to improve quality and create a "lights out" data operating environment. However, once teams experience the benefits of DataOps, they often embrace the complete package of DataOps techniques and tools to deliver faster, better, cheaper data products.

Keys to DataOps

DataOps Mindset. DataOps is both a mindset and a set of technology. The DataOps mindset has a laser focus on continuous improvement. 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. Most data teams are too busy putting out fires to spend valuable time examining practices and processes. But DataOps practitioners flip the equation; they say you can't permanently put out fires until you spend time reviewing what happened in the past and understand how to improve.

Core DataOps Toolset. DataOps is also associated with a set of technologies, borrowed from the DevOps movement and adapted for data processing. The core DataOps toolset consists of a code and configuration repository, continuous integration/continuous delivery (CI/CD), test management, configuration management, orchestration software, and performance management. Figure 2 shows Eckerson Group's DataOps framework, which consists of processes and tooling above a generic data pipeline and traditional data integration and analytics tools below it. The next section of this report provides a short description of each type of DataOps tool.

Figure 2. DataOps Framework



From Wayne Eckerson, "Diving into DataOps: The Underbelly of Modern Data Pipelines," Data Strategy Insider, *December 19, 2018. See below for definitions of various DataOps technologies.*

Report Scope

This report is a sequel to a June 2019 report published by Eckerson Group, Best Practices in DataOps: How to Create Robust, Automated Data Pipelines. That report profiles numerous DataOps practitioners, examining what motivated them to implement DataOps, the benefits it's providing, and the challenges they faced. The report defines 10 best practices for implementing DataOps, ranging from "build for reuse and automation" and "apply quality checks" to "start small" and "create an enterprise data platform." I highly recommend that you read both reports together.

This report is based on a survey of 175 respondents conducted in April 2019 by Eckerson Group, which promoted it through social media and email to its member list. It also was promoted by two partners, Bloor Group and Infoworks.

The respondents are mostly BI directors and managers from large companies. About a quarter (26%) work at companies with more than 10,000 employees and a third at companies with less than 500 employees. (See figure 3.) About a third (32%) are BI or IT directors, and 29% are BI or IT architects, engineers, analysts, or administrators. (See figure 4.) So, almost two-thirds of the respondent pool work in BI environments, which is the primary audience for Eckerson Group.

Figure 3. Company Size

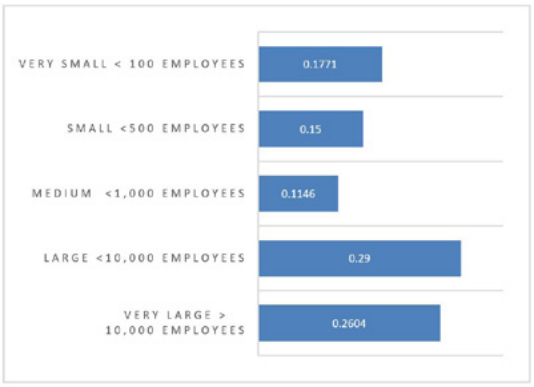
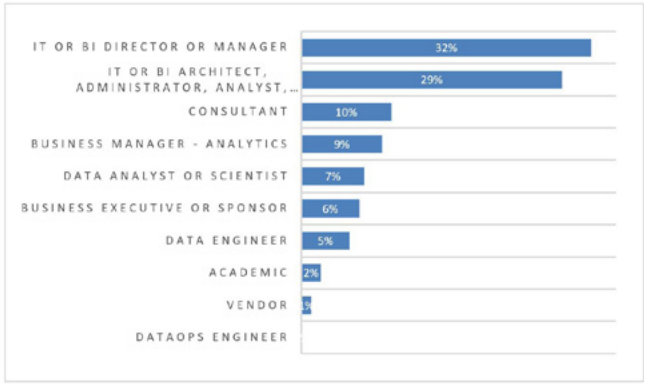


Figure 4. Respondent Roles



Eckerson Group coordinated its survey with DataOps vendor, DataKitchen, which ran the same survey at the same time, but promoted it to its audience of customers and prospects. While both groups had about the same number of active DataOps practitioners (31% to 27%), the DataKitchen respondents were much more likely to be considering or experimenting with DataOps (52% to 30%, respectively, answered "somewhat" to the question, "Does your organization have a DataOps initiative?"). This is not surprising, given the nature of DataKitchen's business.

The DataKitchen audience is also more advanced in some respects. They are more likely to run data environments "mostly" or "entirely" in the cloud (43%) and more likely to use DataOps for data science projects (62%). They are also more likely to take weeks or months to deploy data pipelines (55%) and struggle to "build rigorous tests upfront" (64%) than our group.

Survey Results

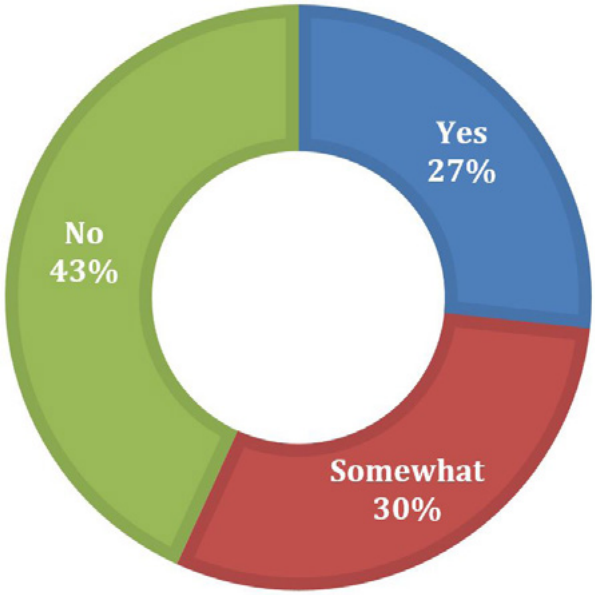
Adoption

It is still early days for DataOps. A majority of companies have yet to really embrace the practice. Forty-three percent have no DataOps initiative, and 30% have "somewhat" of a DataOps initiative. From our experience, this "somewhat" category includes organizations that may have implemented a Git repository or experimented with agile techniques but haven't made a full-fledged commitment to DataOps practices. This includes implementing the core DataOps toolset or consistently scheduled review sessions to examine how to optimize processes and create a culture of continuous improvement.

On the flip side, more than a quarter (27%) of organizations have implemented a DataOps initiative. (See figure 5.) For data warehousing teams, this largely means implementing agile development techniques with self-organizing teams and short sprints. For big data and

data science teams, this means creating a collaborative environment with CI/CD tools and a container-based data infrastructure to automate deployment of analytic models.

Figure 5. Does Your Organization Have a DataOps Initiative?



Components

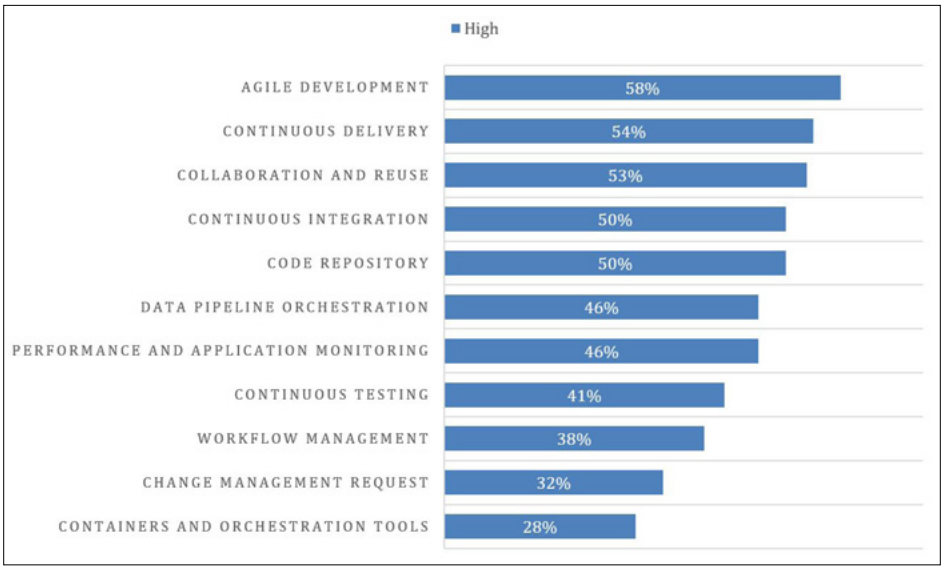
When asked to rate the importance of various DataOps components, our respondents put "agile development" at the top (58% gave it a "high" rating). (See figure 6.) This is not surprising given the makeup of our respondent list, which skews heavily toward data warehousing. Agile has become a standard practice among data warehousing teams, although most teams tailor agile methodologies (e.g., Scrum) to meet their particular needs.

Continuous delivery (54%) and continuous integration (50%) both rated over 50%. Larger teams with many developers working in the same area typically deploy these types of tools to support parallel development efforts. These tools merge the work of multiple developers into a single code stream and automatically deploy it into a production environment, respectively. CI/CD tools have been popular among software developers for a while but are now gaining traction among data developers.

Collaboration and reuse (53%) and code repositories (50%) go hand-in-hand with CI/CD tools. The repositories store code, and along with CI/CD tools, foster collaboration and reuse, although some companies also deploy separate collaboration platforms or use the collaborative features of data catalogs and server-based data preparation and data visualization tools.

Dropping below the 50% mark are orchestration tools (46%), performance and application monitoring (46%), and continuous testing (41%). These largely complete the DataOps toolkit, giving data developers a complete set of tools to work efficiently in a team-based environment.

Figure 6. DataOps Components

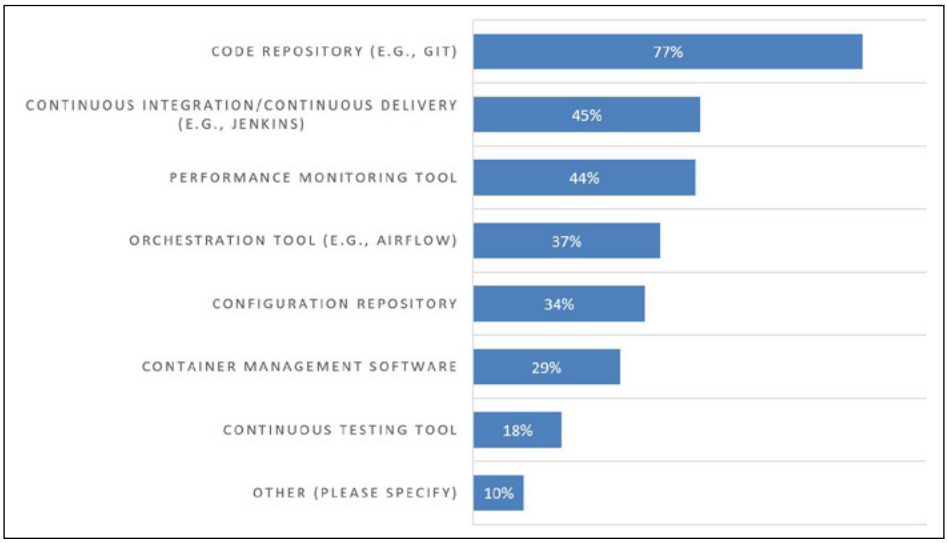


Tools

Use of DataOps tools aligns with how respondents rated the importance of DataOps components. No surprise, a code repository is far and away the most commonly used DataOps tool (77%), followed by CI/CD tools (45%), performance monitoring tools (44%), orchestration tools (37%), and configuration repositories (34%). Organizations with large data environments (thousands or tens of thousands of sources and targets) are more likely to have a performance monitoring tool (65%) and orchestration software (53%). This indicates a greater emphasis on execution rather than development, a sign of more mature DataOps implementations. (See figure 7.)

Testing. Surprisingly, continuous testing tools only registered 18%. Testing is a critical part of DataOps. Most DataOps practitioners use a test-first development strategy where they build tests before writing code. Then, they run these tests during every step of the development process and then through and beyond deployment. The tests become the foundation of automated pipelines. They serve as an early warning system that something has changed that could impact data availability or quality. I expect interest in continuous testing tools to grow in the near future.

Figure 7. DataOps Tools



The following descriptions of DataOps tools come from our report on DataOps best practices:

- **Data preparation tools** enable data engineers to build data pipelines to query, clean, transform, and enrich data to support a specific analytic solution.
- **A code repository** provides one place for data engineers to store their code, such as Spark or ETL code. Most code repositories, such as GitHub, offer check in/out and version control and integrate with most types of development tools on this list.
- **A configuration repository** that stores configurations and settings for all systems in the data environment, spanning development, test, and production systems. A configuration repository manages software releases and ensures error-free deployments.
- **Agile project management tools**, such as Jira, enable agile teams to plan, track, and report on their activities and collaborate around user stories.
- **Continuous integration tools**, such as Jenkins and TeamCity, automatically branch and merge code from multiple developers to support large team-based development environments.
- **Continuous delivery tools** deploy finished code to production environments in a secure, error-free manner. Most continuous integration tools now support continuous delivery and vice versa.
- **Automated testing software** supports test-first development methods, helping developers create and run tests in all phases of the data lifecycle, including production environments, and manage the results, alerting users when failures occur.

- **Orchestration software**, such as AirFlow, coordinates the execution of jobs throughout a data pipeline to automate the flow of data.
- **Performance management tools** monitor underlying systems and pinpoint the cause of performance issues and outages affecting business applications. The tools notify administrators of issues and recommend actions to ensure compliance with SLAs.
- **Data catalogs** create a marketplace of data assets, making it easy for data analysts, data engineers, and data scientists to find and profile relevant data assets prior to creating new data pipelines or data sets.
- **Business glossary** is a data dictionary that contains business descriptions of data entities and attributes. Data glossaries make it easy for business users to discover data definitions, data owners, and data lineage, building greater trust in the data.
- **Containers** virtualize the minimum code required to run a service, making it easy for software engineers to build applications from component parts without worrying about underlying hardware and software configurations, making applications portable. Containers are often used to embed predictive models in operational applications

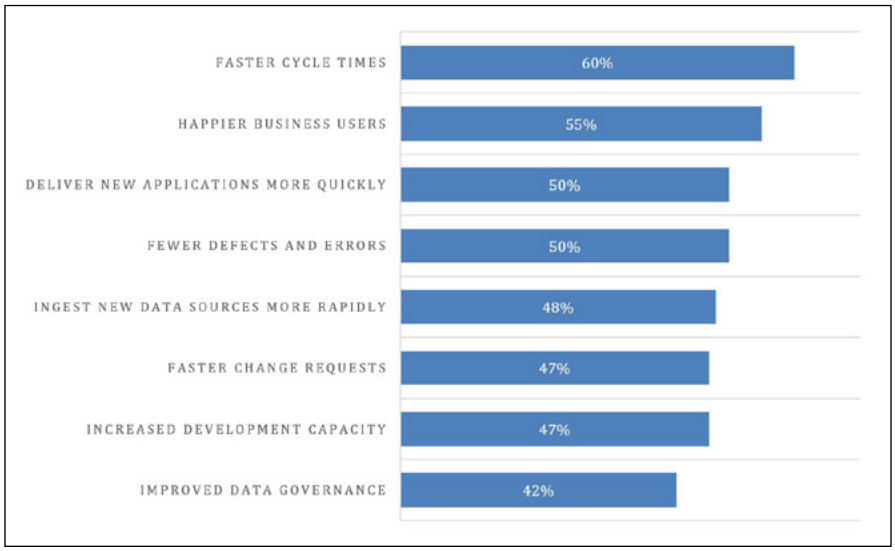
Benefits

The biggest benefit of DataOps is "faster cycle times," checked by 60% of respondents. Faster cycle times is a classic win-win situation. It makes customers happier because they get solutions and fixes faster, and it makes developers happier because it frees them to focus on more value-added activities, such as predictive analytics and cloud solutions.

In the same vein, 50% of respondents checked "deliver new applications more quickly," 47% checked "faster change requests," and 48% checked "ingest new data sources more quickly," all of which are variants of faster cycle times. Other benefits cited include "fewer defects and errors" (50%), "increased development capacity" (47%), and "improved data governance" (42%). Of course, if all those things improve, the result is "happier customers" cited by 55% of respondents. (See figure 8.)

DataOps fundamentally attacks long-standing bottlenecks that have plagued both IT and BI development teams for decades. Those bottlenecks have poisoned the relationship between business and IT, causing the business to build data silos rather than wait for IT. DataOps promises to break this cycle, enabling IT to finally build solutions "faster, better, cheaper." A key DataOps strategy is to identify and prioritize constraints facing development teams and address them one by one.

Figure 8. Benefits of DataOps



Challenges

For companies that have embarked on a DataOps initiative, it's not all smooth sailing. They have to surmount many challenges before they achieve the benefits described above. Chief among these are "establishing formal processes" marked by 55% of respondents. (See figure 9.) It's easy to purchase DataOps tools, but it's hard to use them effectively. That requires defining policies and procedures for how developers use the tools, training developers in those new processes, and establishing permissions for who gets to see which data and code releases.

One respondent wrote, "Make sure your team is prepared to go slow before it can go fast. Less regular work is going to get done. Tools need evaluating and learning. Processes need research and documenting. Things will speed up, but it takes time."

Resistance to DataOps. A related issue is that it's difficult to persuade developers to adopt new practices, which may seem quite rigid and limiting to them at first. Developers who have worked independently without much process or controls will push back, saying it will "slow us down" and that the new regimen is "better suited for software development, not data development." However, after the initial shock of adjustment, developers usually embrace DataOps because it provides a framework to deploy code without worrying about breaking things in production. More importantly, it frees them up to tackle other things, such as predictive analytics, non-relational data, and the cloud.

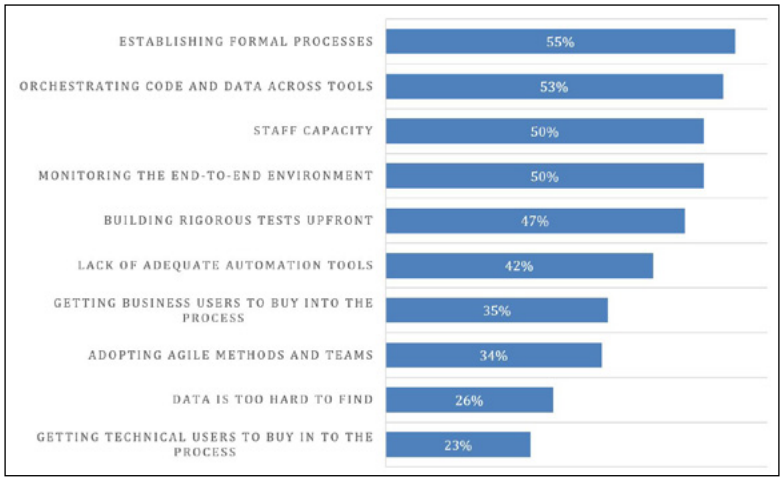
Orchestration. Another major challenge is "orchestrating code and data across tools" (53%). (See figure 9.) Most data pipelines operate in a heterogeneous environment: Data is pulled from source systems running on one application and platform and moved through a series of ingestion, transformation, and delivery applications running on multiple platforms, both on premises and in one or more clouds. Coordinating this movement at scale and speed is not easy.

Other challenges include “staff capacity” (50%), which cuts both ways. This implies the capacity of developers to learn and apply new processes and technologies as well as subject matter experts in the business to work on agile teams to guide their work and prioritize their activities. Pertaining to the latter, 35% of respondents said it’s difficult to get “business users to buy into the process.”

Performance Monitoring. “Monitoring the end-to-end environment” (50%) is another technical challenge since DataOps is essentially data middleware, sandwiched between business applications and data infrastructure. Installing a robust performance management stack that can correlate underlying errors to application service-level agreements (SLAs) is critical. Only a handful of performance management vendors focus on data-driven applications and infrastructure explicitly.

Other challenges include “building rigorous tests upfront” (47%), “lack of automation tools” (42%), and “adopting agile methods and teams” (34%).

Figure 9. DataOps Challenges



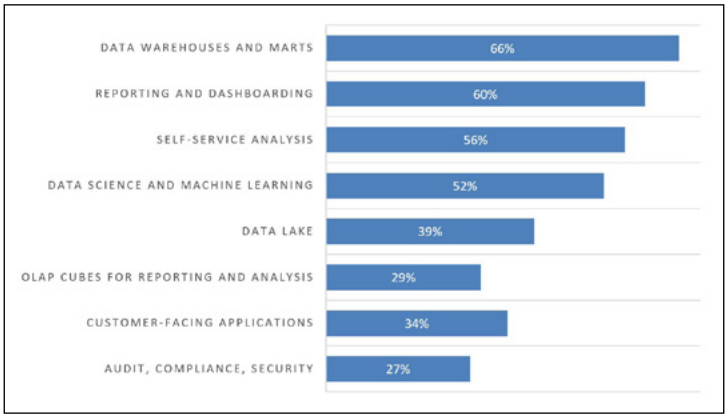
Use Cases

Given the audience for our survey—largely data and analytics leaders—it’s not surprising that the biggest use case for DataOps is “data warehouses and marts” (66%), followed by “reporting and dashboards” (60%), and “self-service analytics” (56%). (See figure 10.)

However, among organizations with complex data environments (thousands or tens of thousands of data sources and targets), the responses were flipped. These large organizations are applying DataOps to support data science projects (64%) first and foremost, followed by data warehouses (59%), and then data lakes (41%) and customer-facing applications (41%).

The bottom line is that DataOps is a general practice that can be applied to any data-driven task or activity. Smaller organizations apply it first to data warehousing, while bigger ones apply it more frequently to data science projects.

Figure 10. DataOps Use Cases



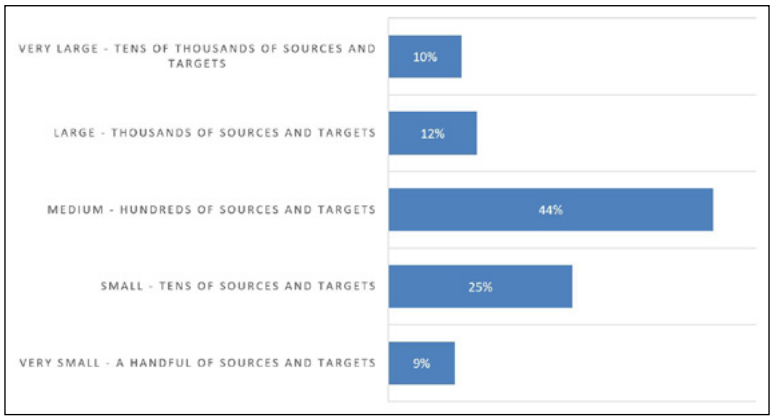
Data Environments

The second part of our survey explored users' data environments to understand the degree of difficulty that they have managing data pipelines.

Scope

A good indicator of the scale and complexity of a data environment is the number of data sources used and the condition of the data they produce. Companies with more source applications typically have more complex DataOps environments. The largest percentage of respondents (44%) have hundreds of sources and targets, but about another quarter (22%) have thousands of sources and targets (12%) and tens of thousands (10%). Only a quarter (25%) have tens of sources and targets and 9% have just a handful. (See figure 11.)

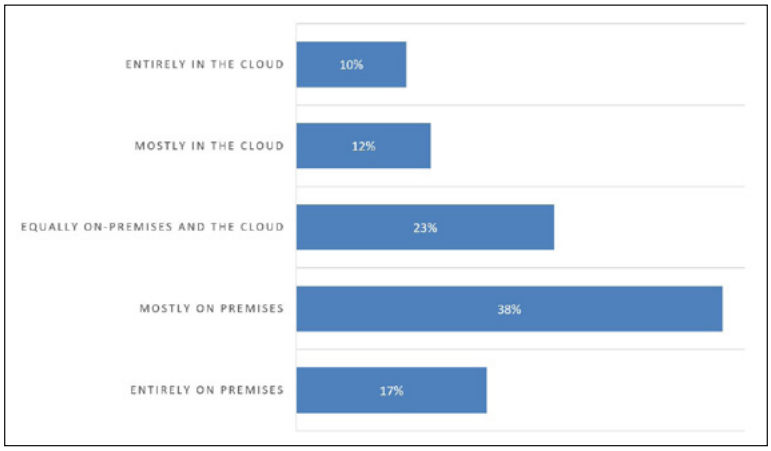
Figure 11. Number of Sources and Targets



Platform

Given the traditional data warehousing audience of this survey, it's not surprising that more than half (51%) of respondents run their data pipelines either entirely (17%) or mostly (38%) on premises. Almost a quarter (23%) run them equally in the cloud and on premises, while almost a quarter (22%) run entirely (10%) in the cloud or mostly in the cloud (12%). These percentages didn't vary much by the scale and scope of the data environment. (See figure 12.)

Figure 12. Data Platforms

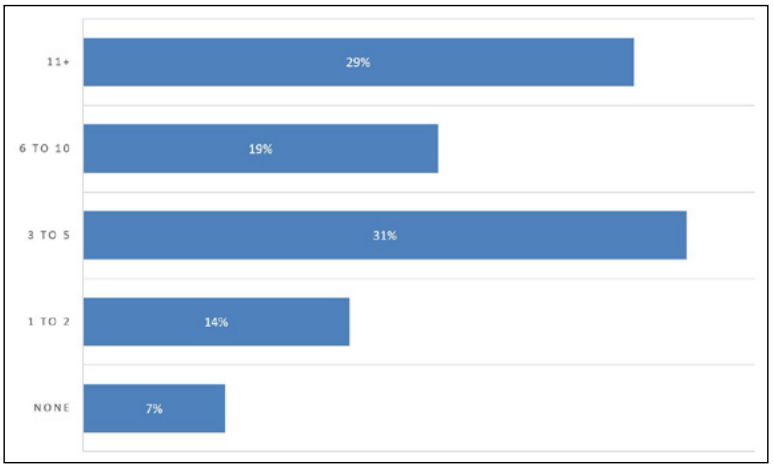


Errors

Data errors negatively impact the productivity of the analytics teams in several ways: they flood Kanban boards with new tasks; they cause developers to switch contexts abruptly to deal with the problem, eroding productivity; and it makes data teams more cautious, causing them to work more slowly. In short, data errors are a major bottleneck that affects the entire workflow of new analytics development.

Almost half (48%) of respondents experience more than six errors a month, defined as incorrect data, broken reports, late delivery, and customer complaints. Almost a third (31%) experience three to five errors a month. These percentages didn't vary much by the scale and scope of the data environment. (See figure 13.)

Figure 13. Number of Data Errors Per Month

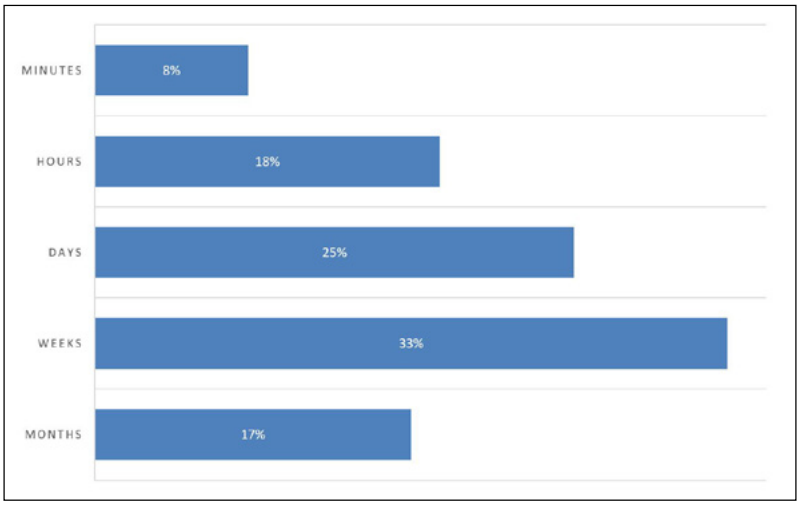


New Pipelines

In addition, half (50%) take weeks or months to build a new data pipeline from scratch (17% take months). The other half were much speedier: 25% take days, 18% hours, and 8% minutes. These percentages didn't vary much by the scale and scope of the data environment. (See figure 14.)

Of course, these answers depend on how you define "data pipeline"—some refer to it as just the data ingestion portion of the stack and others just the data transformation. Our definition, which I think represents the vast majority of opinion today, encompasses the entire workflow, from ingestion to transformation to delivery to business users. If that's true, then delivery in "days" is exceptional, while "hours" and "minutes" seems a tad preposterous for even one portion of a data pipeline, unless it's an extremely simple application or it's simply the auto-provisioning of a new data environment for testing or analysis.

Figure 14. Time to Create a New Data Pipeline

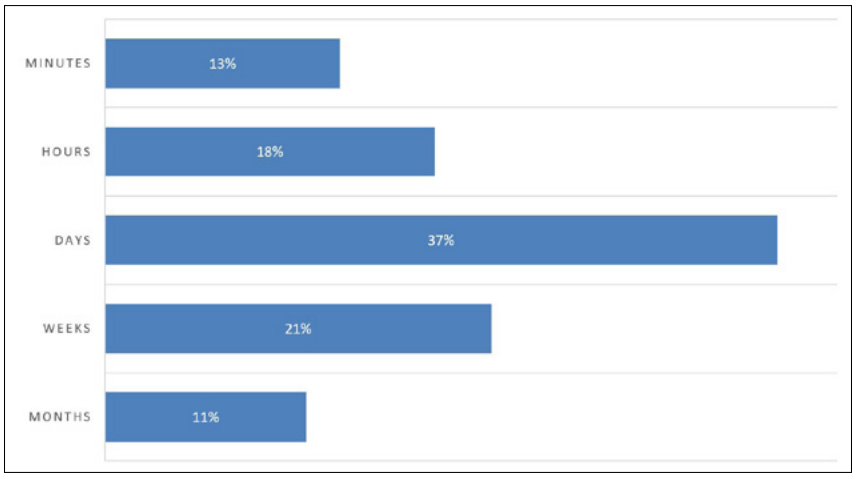


Deployment

Cycle time includes the process of creating new analytics (and their development environments) and deploying them into production. A short cycle time enables an analytics team to respond quickly to requests for new analytics. When analytics are produced quickly, the data team can keep pace with the endless stream of requests from the business units. A short cycle time fosters close collaboration with business users. In our experience, this unlocks an organization's creativity. However, the cycle time in most data organizations is plagued with inefficient manual processes, bureaucracy, lack of task coordination, and dependencies on bottlenecks. The ability to rapidly produce and deploy analytics is at the heart of a data team's ability to add value.

According to our survey, almost a third (31%) can deploy a data pipeline in minutes or hours, but 37% require days, and a whopping 32% take weeks or months. Companies with large-scale data environments (thousands or tens of thousands of data sources and targets) were more likely to take months to deploy pipelines into production (19%) and less likely to take days (19%). (See figure 15.)

Figure 15 Putting a Data Pipeline into Production



DataOps Is the Remedy for What Ails Us

DataOps has potential to turn around the sorry state of data processing. Most organizations are plagued by data silos, poor data quality, slow development and deployment processes (e.g. cycle times), and a huge gulf between business and IT. DataOps makes it possible to develop data pipelines "faster, better, cheaper," which is music to the ears of business. Therefore, DataOps has the potential to heal the rift between business and IT and help organizations get more value from their data.

However, the same requirements for succeeding with DataOps are the very things missing from organizations with shoddy data operations: discipline and time. DataOps requires a

relentless focus on continuous improvement. And continuous improvement requires time—and I daresay—a bit of courage to ignore the fires for an hour or a day or two in order to create a fire prevention system that will eliminate fires once and for all.

We are at the beginning of the DataOps movement. There are many challenges ahead. But the promise of DataOps is too large to ignore. Data teams need to learn about DataOps and begin implementing its practices and tools. No step is too small; but the biggest and most important is the first one—make a commitment to improve data operations.

About Eckerson Group

Wayne Eckerson, a globally known author, speaker, and advisor, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with a cocoon of support during every step of their data journeys.

Today, Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps you abreast of the latest trends, techniques, and tools in the data analytics field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate your business requirements into compelling strategies and solutions.
- **Our educators** share best practices in more than **30 onsite workshops** that align your team around industry frameworks.

Get More Value From Your Data

Unlike other firms, Eckerson Group focuses solely on data analytics. Our experts each have more than 25+ years of experience in the field. They specialize in every facet of data analytics—from data architecture and data governance to business intelligence and artificial intelligence. Their primary mission is to help you get more value from data and analytics by sharing their hard-won lessons with you.

Our clients say we are hard-working, insightful, and humble. We take the compliment! It all stems from our love of data and desire to help you get more value from analytics—we see ourselves as a family of continuous learners, interpreting the world of data and analytics for you and others.

Get more value from your data. Put an expert on your side.
Learn what Eckerson Group can do for you!

About

Automatically deliver and operate all your analytic data pipelines — simply, collaboratively, and at enterprise scale — with DataKitchen's DataOps Platform. DataKitchen provides the world's first DataOps platform for data-driven enterprises, enabling them to support data analytics that can be quickly and robustly adapted to meet evolving requirements utilizing their existing tools. DataKitchen is leading the DataOps movement to incorporate Agile Development, DevOps, and manufacturing based statistical process control into analytics and data management. DataKitchen is headquartered in Cambridge, Massachusetts.