DataOps: Industrializing Data and Analytics

Strategies for Streamlining the Delivery of Insights

By Julian Ereth and Wayne Eckerson

June 2018

Research sponsored by:

About the Authors

Wayne W. Eckerson has been a thought leader in the business intelligence 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.

Julian Ereth is a researcher and practitioner in business intelligence and data analytics. In his role as researcher he focuses on new approaches in the area of big data, advanced analytics, and the Internet of Things. He is author of multiple internationally accepted research papers and is currently working toward his Ph.D. at the University of Stuttgart (Germany). He is also co-founder of pragmatic_apps, a company that builds custom business software and analytics solutions.

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 25 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.

About This Report

The research for this report is made possible by Unravel.

Table of Contents

Executive Summary

DataOps is an emerging set of practices, processes, and technologies for building and enhancing data and analytics pipelines to meet business needs quickly. As these pipelines become more complex and development teams grow in size, organizations need better collaboration, development and operations processes to govern the flow of data and code from one step of the data lifecycle to the next – from data ingestion and transformation to analysis and reporting. The goal is to increase agility and cycle times, while reducing data defects, increasing application reliability and giving developers and business users greater confidence in data analytics output.

DataOps builds on concepts popular in the software engineering field, such as agile, lean, and continuous integration/continuous delivery, but addresses the unique needs of data and analytics environments, including the use of multiple data sources and varied use cases that range from data warehousing to data science. It relies heavily on test automation, code repositories, collaboration tools, orchestration, monitoring frameworks, and workflow automation to accelerate delivery times while minimizing defects.

DataOps requires cultural shift. It is not something that can be implemented all at once or in a short period of time. DataOps is a journey. Leaders use productivity metrics to gauge their progress and impel them and their teams to continuously search for new ways to cut wasted effort, streamline steps, automate processes, increase output, and get it right the first time. For large organizations with big development teams, DataOps is an antidote to many of the woes that beset IT and development organizations.

The Conditions for DataOps

Immature Data and Analytics Pipelines

Although data is a critical business asset, most organizations still don't have mature processes for converting data into insights that drive business value. The development of data and analytics pipelines, both simple and complex, is still a handcrafted and largely non-repeatable process with minimal reuse, managed by individuals working in isolation with different tools and approaches. The result is both a plodding development environment that can't keep pace with the demands of a data-driven business and an error-prone operational environment that is opaque and slow to respond to change requests.

Drawing Swords. Immature development and delivery processes force business users to take matters into their own hands. They buy self-service tools and build their own pipelines that result in an ever-expanding universe of data silos with conflicting data, logic errors and complex interdependencies that go unnoticed until a major decision backfires. The lack of efficiency on the development and operations side leads to a lack of effectiveness on the business side. Business and IT draw swords and go to war instead of partnering.

New Approach

These problems beg for a new approach to building data analytic solutions. As data and analytics pipelines become more complex and development teams grow in size, organizations need to apply standard processes to govern the flow of data from one step of the data lifecycle to the next – from data ingestion and transformation to analysis and reporting. The goal is to increase agility and cycle times, while reducing data defects, increasing application reliability and giving developers and business users greater confidence in data analytics output.

DataOps applies rigor to developing, testing, and operating code that manages data flows and creates analytic solutions.

This is the vision of DataOps, an emerging methodology for building data analytic solutions that deliver business value. Building on modern principles of software engineering, DataOps applies rigor to developing, testing, and operating code that manages data flows and creates analytic solutions. The goal is to foster greater collaboration among development, test, operations, and business teams and create a culture of continuous improvement.

This Report. This report defines DataOps, explains its origins and core principles, and describes three major use cases: data warehousing and data management; dashboards and reporting; and data science. Finally, the report outlines concrete steps for implementing DataOps and common pitfalls to avoid.

What is DataOps?

DataOps is an integrated approach for delivering data analytic solutions that uses automation, testing, orchestration, collaborative development, containerization, and continuous monitoring to continuously accelerate output and improve quality. The purpose of DataOps is to accelerate the creation of data and analytics pipelines, automate data workflows, and deliver and operate high-quality data analytic solutions that meet business needs as fast as possible.

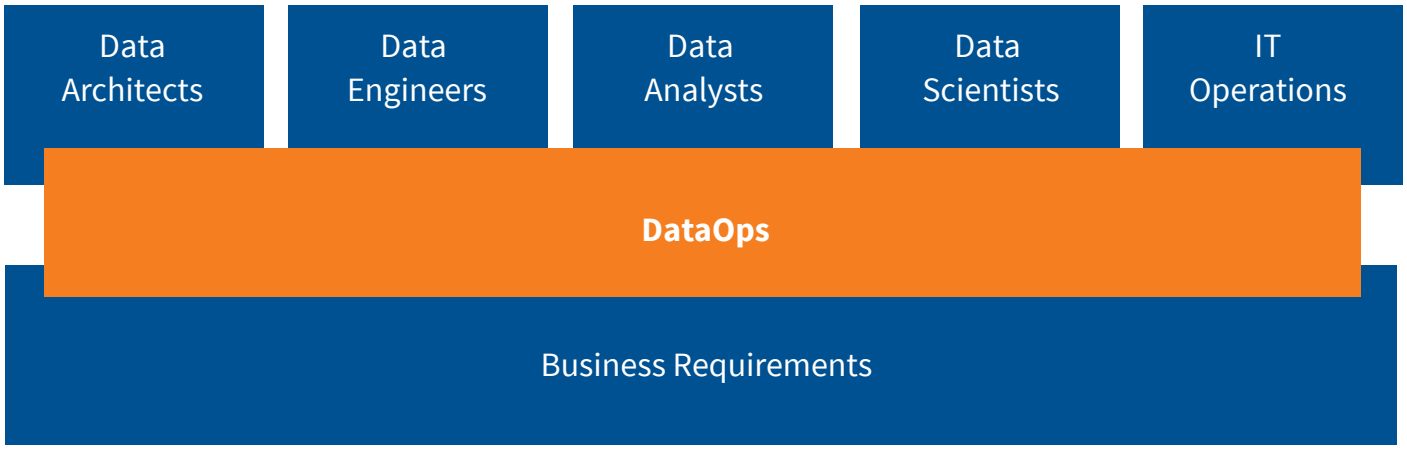
"DataOps consists of a stream of steps required to deliver value to the customer." DataOps practitioner

As one DataOps practitioner from a Fortune 50 company says, "DataOps consists of a stream of steps required to deliver value to the customer. We automate those steps where possible, minimize waste and redundancy, and foster a culture of continuous improvement."

Roots in DevOps. The underlying idea comes from the DevOps movement in the software engineering world. DevOps bridges the traditional gap between development, QA, and operations so the technical teams can deliver high-quality output at an ever faster pace. Rather than separate teams working at cross purposes (speed or quality), the goal of DevOps is to foster a culture of collaboration and trust between these parties and improve both speed and quality.

Data Stakeholders. Similarly, DataOps brings together stakeholders throughout the data landscape. These stakeholders span traditional roles, such as data architects and data engineers, newer roles, such as data scientist, and IT operations people who build and maintain the data applications and infrastructure (See figure 1.). DataOps builds a bridge between these stakeholders and aligns them with business requirements.

Figure 1. DataOps Unites Data Stakeholders Around Business Requirements



Best Practices. Since the activities and interests of these stakeholders can diverge, a successful implementation of DataOps requires the following:

- **Culture.** The core of DataOps is a culture of collaboration and trust. All stakeholders must work together and feel responsible for the entire process. Awareness of the business requirements in all stages is essential.
- **Processes.** DataOps requires well-defined processes, roles, guidelines, and metrics to reinforce DataOps principles. Consequently, many companies establish testing and certification programs to educate and train staffers.
- **Technology.** DataOps requires tools and infrastructure to support automation, testing, monitoring, and orchestration, as well as collaboration and communication among all stakeholders.

Home-Grown Tools. Unlike DevOps, the tools required to support DataOps are in their infancy. For example, testing automation plays a major role in DevOps, but most DataOps practitioners have had to build or modify testing automation tools to adequately test data and analytics pipelines and analytic solutions. “We had to build a test platform that gets data from Hadoop, SQL Server, Excel, external files, and so on,” says one DataOps practitioner.

Commercial Tools. An emerging class of DataOps vendors are building capabilities to support data analytics pipelines and workflows, which tend to be more complex and diverse than those in the DevOps world. Also, there is more potential for reuse in the DataOps world so providing easy access to past output is a key part of many DataOps implementations.

Principles and Benefits of DataOps

Core Principles

Disciplines. As a practice that stems from software engineering, DataOps borrows heavily from Agile, Lean, DevOps, and Total Quality Management disciplines.

Like Agile, DataOps emphasizes the use of self-organizing teams with business involvement and short (i.e., 2-3 week) development sprints that deliver fully tested code. Like Lean, DataOps focuses on efficiency, using version control systems and code repositories that foster parallel development and code reuse. And like Total Quality Management, DataOps espouses continuous testing, monitoring, and benchmarking to detect issues before they turn into major problems.

Given its software engineering heritage, DataOps adheres to a common set of principles that it has adapted from established methodologies. Here are some of the more prominent principles:

- **Business Value.** Data is not an end in itself, but a means to deliver insights that add value to the business and satisfy the customer.
- **Continuous Improvement.** Learn from mistakes, review processes continuously, and adapt to changing circumstances.
- **Collaboration and Communication.** Share knowledge, simplify communication, and provide feedback at every stage of the data analytics lifecycle.
- **Reuse and Automate.** Automate wherever possible and reuse existing artifacts to avoid unnecessary rework and repetition.
- **End-to-End Processes.** Avoid data silos and consider analytics an enterprise endeavor. Orchestrate data, schema, tools, code, and stakeholders throughout the data landscape.
- **Short Cycles and Incremental Change.** Avoid "big bang" releases and bloated processes. Iterate in short cycles so you can adapt quickly to new and changing needs.
- **Analytics is Code.** Look at data artifacts, like models and visualization, as code and adopt software methods like version control, automated testing, and continuous deployment to them.
- **Quality and Testing.** Make quality and testing a top priority and ensure that no untested artifact reaches production.
- **Full-stack Monitoring and Data-Driven Improvement.** Continuously monitor applications down to infrastructure and use the resulting insights to enhance performance and reliability.
- **Keep It Simple.** Whenever an easier solution appears, it is most likely also a superior one.

This collection of core principles represents experiences from companies that have used DataOps to better manage their development processes and generate greater business value more quickly. The list is not meant to be comprehensive but provides a starting point for understanding DataOps concepts.

Benefits of DataOps

There are many benefits of DataOps but the most important fall in to the following four buckets:

1. **Improve Collaboration and Communication.** DataOps comes with a change in culture that promotes collaboration, trust, and responsibility. The goals are to blur the lines between departments and functions, encourage the exchange of knowledge, reduce conflicts, and eventually increase productivity.
2. **Accelerate Time to Production.** A major driver for DataOps is speed. The idea of streamlined and largely automated analytics pipelines helps deliver new features and insights quickly and reduces manual effort. Moreover, the short feedback and testing cycles help speed up reactions to changing business requirements and increase flexibility.

3. Increase Quality, Reliability and Visibility. Well-defined analytics pipelines enhance both speed and robustness. For instance, multiple stages of automated and manual tests prevent the deployment of flawed updates. Besides, DataOps also includes monitoring of production environments to identify bottlenecks or potential issues and thereby improves efficiency and stability of infrastructure and applications. Lastly, the convergence of different roles helps align changes throughout various stages, such as when a data engineer is informed about the later cleansing issues encountered by a data scientist or the lack in performance of an ETL process in production. A way to achieve this, can be a Self-Service Application Performance Management Platform that allows all stakeholders to understand and rationalize the performance of analytic applications. (See "Why is APM a Must-Have for Big Data Self-Service Analytics?")

4. Enables Self Service. With greater automation and machine learning algorithms that simplify development, deployment and performance management tasks, organizations need fewer experts to build and manage data and analytics pipelines. Business users with some degree of technical savvy can build their own pipelines or move code into production. For instance, Stitch Fix has a data infrastructure that enables data scientists to deploy models into production. (See "Interview with Jeff Magnusson: How to Create a Self Service Platform for Data Scientists.")

DataOps is about more than speed and quality. With a culture of continuous improvement, organizations can deliver data analytics solutions more efficiently, releasing valuable team members for more valuable activities, such as building innovative new products.

Data and Analytics Pipelines

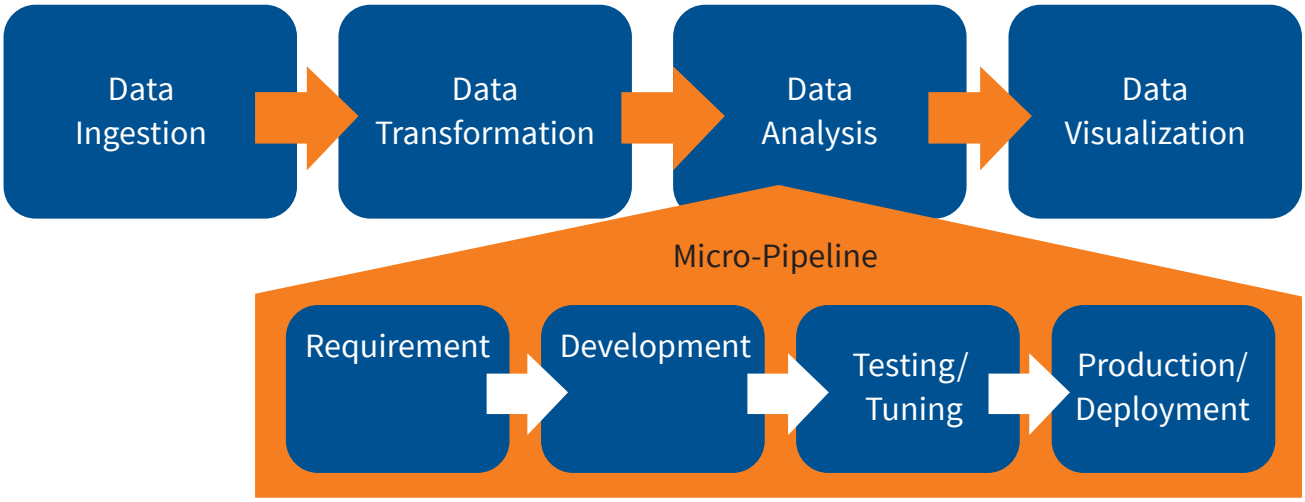
Major Pipelines

Assembly Line for Data. The purpose of DataOps is to manage data and analytics pipelines. Like an assembly line on a plant floor, a data analytics pipeline turns raw data from one or more sources into an output consumed by business users or another application.

Outputs and Use Cases. The most common outputs are reports, dashboards, charts, or data sets. But outputs can also be data marts, data science sandboxes, or even data warehouses and data lakes. Each output is designed to support a specific use case. And each use case might support a single individual or application or entire department or enterprise. Not surprisingly, organizations can have hundreds, if not thousands, of data analytic pipelines.

Simple and Complex. Each pipeline consists of a number of steps that convert data as a raw material into a usable output. Some pipelines are relatively simple, such as "Export this data into a CSV file and place into this file folder." But many are complex, such as, "Move select tables from ten sources into the target database, merge common fields, array into a dimensional schema, aggregate by year, flag null values, convert into an extract for a BI tool, and generate personalized dashboards based on the data". Figure 2 depicts the most common steps in a data analytics pipeline.

Figure 2. A Generic Data Analytics Pipeline



Although most data analytics pipelines are fairly complex, with dozens or hundreds of steps, most consist of the following stages:

1. **Data Ingestion.** Data from various sources is extracted, validated, and loaded into downstream systems.
2. **Data Transformation.** Data is cleansed, enriched, integrated, and modeled to support the target application or data structure.
3. **Data Analysis.** Data is analyzed refined and packaged in data models to provide certain insights or fulfill a task.
4. **Data Visualization.** Data is visualized in a suitable manner, e.g., a static report or an interactive dashboard.

Data analytics pipelines often span multiple functions supported by different specialists who use a variety of tools and technologies. For instance, a data engineer uses ETL or data preparation tools to extract data from a source and load it to a central data warehouse or data set. A data scientist might then use Python or R to build a predictive model that provides data for a report built by a BI developer using a visualization tool and a chart embedded in a business application managed by a software developer.

Micro Pipelines

To manage this complexity, it helps to divide individual stages into micro-pipelines to ensure that changes can be made to each step without disrupting the entire flow. The micro-pipelines consist of the following steps:

1. **Requirement.** Agile methods, such as Scrum and Kanban, elicit requirements from stakeholders

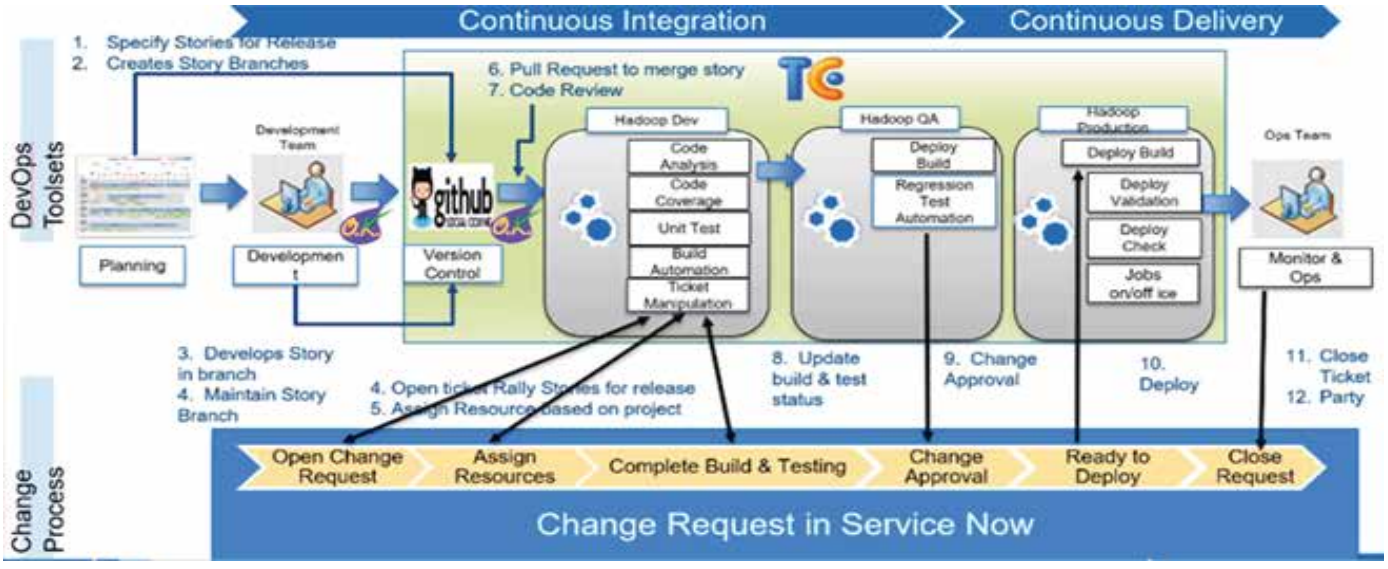
in an iterative manner. A requirement could be the integration of a new data source in an ETL process or the addition of a new field to a database table.

2. Development. In data and analytics, development does not necessarily mean programming; it can comprise tasks like changing a data model or reconfiguring a dashboard. Here, DataOps requires a high level of automation, e.g., automatically generating database queries when a model is changed, to avoid repetitive manual work.

3. Testing/Tuning. Testing is often neglected in data and analytics, as building an adequate testing environment in complex analytics solutions can be difficult, and many tools simply lack testing functions. In DataOps, however, testing is essential. The goals of this stage are to check functionality and quality and to ensure a seamless deployment and operation in a production environment. This is especially important as there might be fundamental differences between the development infrastructure (e.g. a local database) and the eventual production environment (e.g. a cloud-based cluster). Accordingly, this phase should involve automated tests, such as programmed validations (e.g., comparing sums after database migrations), manual reviews and acceptance tests, as well as performance tests that verify configuration settings, resource utilization, and other factors that are relevant to make sure that an application scales as required, and no other data pipelines are affected.

4. Production/Monitoring. After testing, changes are deployed to production. This process should be a fully automated, documented, and easily reversible. Reversibility, in particular, can require more than just reverting to an older codebase; it may require reconfiguring reports and downstream systems to account for new or deleted tables or columns in a database. After a successful deployment, DataOps staff must continuously monitor all changes to identify bottlenecks, measure adoption and usage of resources, and identify new requirements.

Figure 3. DataOps Orchestration Using Hadoop



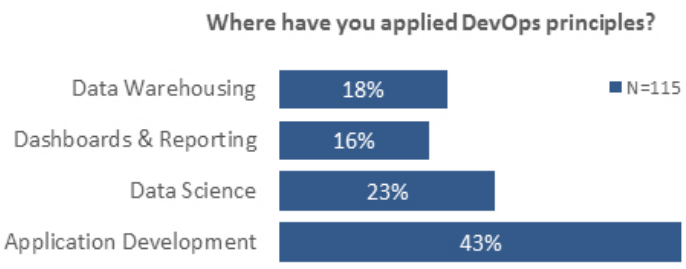
Orchestration and Automation. A key role of DataOps is to orchestrate and automate the flow of data and code between people and tools in an efficient manner that ensures clean handoffs and minimal errors and disruptions. With complex pipelines, this can be challenging, making orchestration and automation key facilities in any DataOps implementation.

Figure 3 shows a DataOps flow designed by a high-tech company with dozens of development and operations teams that it uses to guide its delivery of Hadoop-based solutions.

Continuous Iteration. The stages and steps that must be orchestrated in a data analytics pipeline are not always serial; often multiple steps happen in parallel and once completed, a step might be repeated as part of an iterative, agile workflow to refine output until it gains user acceptance. This approach increases quality, because it ensures that no untested change makes it to production. It also improves orchestration and collaboration, as different actors in the pipeline rely on another and work together in a fluid process.

DataOps Use Cases

Figure 4. DataOps Orchestration Using Hadoop



The one-question poll was promoted via Twitter by Eckerson Group and its social media partners.

Although DevOps originated in the software engineering field, organizations are beginning to adapt the principles to the world of data and analytics. Figure 4 shows that other than application development, the three primary DataOps use cases are data science, data warehousing, and dashboards and reporting, which are being applied by roughly 20% of our respondent base. The rest of this section examines the three major DataOps use cases.

Data Warehousing and Data Management

Data warehousing and data management are core disciplines of data and analytics. A data warehouse (DW) is a central repository for enterprise data. The concept dates back to the 1980s and has evolved such that today's DWs are complex structures that manage many tools and data management workflows to cleanse, check, and aggregate data to meet the ongoing needs of the information lifecycle. Despite the maturity of data warehousing and data management, they currently face two major challenges:

- **Increasing Heterogeneity and Complexity.** Today, the DW needs to support many

heterogeneous sources of data with different characteristics for quality, speed, and structure. At the same time, the range of products for transforming and storing data is ever-increasing. This heterogeneity of sources and tools leads to more complex data landscapes that are hard to operate and often lack in performance.

- **Rapidly Changing Business Requirements.** DWs usually come with rigid data structures and processes. Adding a field to a DW can take several weeks due. On one hand, providing a stable and consistent data platform is desirable; on the other hand, this is insufficient to keep pace with changing business requirements.

These issues indicate a need for a new approach to tame the complex data landscapes and adjust data warehousing for today's business reality. This is where DataOps can reduce complexity and improve manageability of complex solutions. (See Table 1.)

Table 1. DataOps for Data Warehousing and Data Management

Current Challenges:	DataOps' Answers:
• Complex data landscapes and processes with heterogeneous data sources and tools that are hard to operate and lack in performance. • A lack of agility makes it hard to respond to rapidly changing business requirements.	• Analytics pipelines that streamline processes, improve collaboration, and establish a culture of continuous improvement. • Use of automation to increase agility and speed and free resources for more valuable activities. • Use monitoring to get insights about application performance and resource usage to continuously improve reliability and efficiency.

Improve Manageability and Collaboration. A major driver for DataOps is its holistic approach that enhances manageability even in complex scenarios, and collaboration is key here. For instance, if a data engineer discusses upcoming features with the DW administrator he/she can get insights about performance and resource usage at an early stage and thereby avoid unwanted consequences like insufficient scaling. Moreover, thinking in analytics pipelines can help reveal the big picture and streamline particular processes to raise efficiency across the entire data landscape and promote a culture of continuous improvement.

Increase Agility and Automate the Data Warehouse. The DataOps response to insufficient agility and speed has two sides: First, agile methods and a streamlined analytics pipeline help to quickly elucidate, prioritize, and communicate business requirements. Second, with automation eliminates manual work wherever possible. For instance, data architects need not code database queries anymore; an

automation tool can generate all necessary structures from the data model, validate them automatically, and then deploy them to production without any manual steps. Speed increases and resources are free for more important activities.

In summary, data warehousing can heavily benefit from the holistic DataOps approach that brings together all stakeholders, streamlines processes, and emphasizes a culture of continuous improvement.

Dashboards and Reports

A data and analytics initiative provides value through insights that support business decision making, but is not an end in itself. These insights usually emerge when data is visualized in interactive dashboards or static reports. However, as data is becoming more important, the requirements for dashboards and reporting are also changing in two major ways:

1. **Proliferating Reports and Dashboards.** Reports and dashboards generally proliferate as organizations become more data-driven. Consequently, managing the lifecycle of reports and dashboards becomes increasingly complex. This is especially problematic as the underlying changes over time and certain reports or dashboards lose validity or even provide conflicting information. In addition, business stakeholders become more data-hungry and expect higher-quality insights to arrive ever-faster.
2. **Relevant Data and Visualizations.** A core challenge for a data engineer is to interpret what data may be relevant to a business stakeholder and how it should be visualized. For instance, a developer might visualize an unfiltered dataset in a large table, where a business stakeholder would prefer a simple visualization of selected numbers. This issue is not new, but it becomes more prominent as available data grows.

Creating and managing numerous dashboards and reports is clearly difficult and complex, and companies have widely adopted agile methods to better understand what the business expects. DataOps completes this approach by linking these ideas with the rest of the development processes and thereby creating an end-to-end responsibility for all stakeholders. Moreover, it introduces several technical concepts that help scale the processes and thereby manage even large numbers of reports and dashboards. (See Table 2.)

Table 2. DataOps for Dashboards and Reports

Current Challenges:	DataOps' Answers:
• Manage a growing number of reports and dashboards over the entire lifecycle. • Understand what the business needs and quickly respond to changing requirements.	• Establish an end-to-end responsibility to bridge the gap between data and business. • Ensure relevancy with short cycles, fast feedback, self-service monitoring and data-driven improvements. • Improve speed, quality, and scalability with defined analytics pipelines.

Culture of Continuous Improvement and Responsibility. A core principle of DataOps is to bring together different stakeholders, such as data engineers and business people. The goal is to establish a culture where reports do not arise from arbitrary requirements that were thrown over the fence, but where data engineers understand the business and feel responsible for subsequent decisions.

Short Cycles, Fast Feedback, and Data-Driven Improvements. The highly iterative approach of DataOps improves relevancy and agility. Stakeholders are involved early in the analytics pipeline, providing feedback on prototypes so developers can refine the solution before it's released. DataOps also puts metrics in place and conducts planned experiments (e.g., A-B testing) to measure how dashboards and reports perform and spur insights about the relevancy, resource usage and suitability of visualizations. Here, a self-service monitoring platform can help to increase the visibility in the data landscape, e.g. when a data engineer can immediately see how a changed report affects performance or acceptance.

Streamline Processes and Automated Tests. Last, the creation of reports and dashboards can heavily benefit from a defined analytics pipeline that automates steps and ensures that all reports and dashboards pass certain tests, such as that the correct aggregation rules were used or that a report complies with style guidelines. This increases overall quality and can simplify managing even large numbers of reports and dashboards, e.g., with automated lifecycle tests that regularly confirm data validity.

Operationalize Data Science

Data science describes the use of sophisticated statistical methods and algorithms to explore data sets and extract valuable insights. The idea of applying statistical models to generate deeper insights is not entirely new, and data mining has been around for years. However, advancements in big data, machine learning, and artificial intelligence enable many new applications. The emergence of this novel sector also poses two major challenges to data and analytics practices:

1. Operationalization of Data Science. Due to its novelty, data science lacks standards and best practices. Consequently, data scientists often have custom-tailored processes and use exotic tools. This rarely becomes an issue until an isolated data science approach must be converted to a repeatable, efficient, and flawless application. For instance, a statistical model might behave different when it is run in a large big data cluster and it can be hard to tell why. However, this often difficult operationalization is necessary to get ROI from data science.

2. Aligning Data Scientists with Other Stakeholders. A key to data science is the combination of business knowledge, statistical experience, programming, and data skills. These different skillsets are rarely found in one person and usually require different experts to work together. However, it can be difficult to align the various mindsets and interests of data scientists, product owners, and data engineers.

Getting the most out of data science requires demystifying the discipline by establishing defined structures and processes. The hard part here is to put data science into a frame without infringing the necessary autonomy of data scientists. DataOps can provide three starting points. (See Table 3.)

Table 3. DataOps for Data Science

Current Challenges:	DataOps' Answers:
• Transform statistical models from an experimental stage to production so they deliver ROI. • Alignment among data scientists, product owners, and data engineers.	• Establish analytics pipelines that keep the data science process flexible. • Use tools to orchestrate stakeholders and technologies and establish monitoring to assure a scalable operation in production. • Combine different skillsets with cross-functional thinking.

1. Bring Data Science to Production with Analytics Pipelines. An analytics pipeline for data science can help structure the process of moving statistical models from the lab into the field. Here, models need to pass different tests that measure their accuracy and scalability, bringing together the knowledge of data engineers, data scientists and the operations team. A streamlined deployment process with comprehensive monitoring can help to refine statistical models, ensure scalability and thereby prevent one-off models, and deliver continuous business value.

2. Orchestration and Reuse. DataOps can discover common ground where various stakeholders and technologies can act in concert. Here, orchestration tools enable automatic combinations of various technologies. Such a tool, for instance, allows a data scientist to check in her custom Python code, and then ensures that it automatically gets tested and converted so it can be used

subsequently by other ETL or visualization tools. Therefore, it puts an end to technology discussions and shifts focus to actual business problems.

3. Collaboration and Cross-Functional Thinking. Communication is key to bridging the gaps among the magic of data scientists, the daily work of a data engineer, and the business. The DataOps way of work includes cross-functional thinking and short feedback cycles to increase relevancy of data science, as the business always has a seat at the table. Similarly, integrating the knowledge of data engineers early simplifies the scaling of statistical models for production.

DataOps Technologies

DataOps Components

There are many technologies and tools that DataOps teams use to build and maintain pipelines. Perhaps the most important are test automation tools that build testing into the development process, improving the quality of code. Code repositories are also critical, especially in team-based development environments that require rigorous check in/out procedures and continuous integration processes. Beyond pre-production, continuous deployment tools that move code from development to testing and production as well as monitoring and performance intelligence platforms are also critical to increase performance and visibility and guarantee a stable and efficient operation of complex data landscapes.

The major DataOps technical components are:

- Test automation
- Code repositories
- Orchestration frameworks
- Collaboration and workflow management
- Metadata management
- Lineage and impact analysis
- Database management systems
- Data integration, preparation, and automation tools
- Analytics and visualization tools
- Monitoring and performance intelligence platforms

Categories of DataOps Tools

Many commercial vendors have begun shipping tools to facilitate some or all of the DataOps lifecycle. Commercial DataOps products offer a reliable alternative to open source libraries. Commercial tools are typically more robust and comprehensive, and the vendor can serve as a partner that offers ongoing support and service to address thorny integration, orchestration, or automation issues.

There are several major categories of DataOps products: orchestration and operations platforms, data warehouse automation tools, data engineering tools, and data science platforms.

- **Orchestration and Operations Platforms.** There are orchestration tools that automate the flow of data and code across multiple tools, applications, and people. They act as a digital control room where all data sources and processes are managed and tuned. This reduces the complexity of managing complex data pipelines in a heterogeneous environment. Often these tools go beyond pre-production and provide monitoring capabilities to increase the visibility of performance and resource consumption for all stakeholders.
- **Data Warehouse Automation (DWA).** These metadata-driven tools enable the automatic generation and deployment of data structures in a data warehouse, including staging areas, target databases, BI databases, and documentation. They are ideal for accelerating change management requests. Some DWA vendors are now extending their products to work with big data (Hadoop) and the cloud and handle more generic data-centric design, testing, and operations workflows.
- **Self-Service Data Preparation.** These business-centric tools enable data analysts and other business users to build their own data and analytics pipelines, so they are no longer dependent upon the IT department. These tools facilitate a handoff from corporate IT, which uses data integration and extract, transform, and load (ETL) tools to ingest, clean, and lightly integrate data, while data analysts take the IT output and use data preparation tools manipulate the data to support local or individual use cases.
- **Data Science Platforms.** Data science platforms are designed to accelerate, integrate, and automate the entire data science lifecycle, from data preparation and model creation to model deployment, monitoring, and management. Some platforms focus more on model development, others on model deployment, while some tackle the entire lifecycle.

Putting DataOps to Work

Since DataOps is a new concept, there is a dearth of DataOps blueprints and best practices that others can follow. This section outlines guidelines and advice and identifies some pitfalls to consider when implementing DataOps.

Tips of the Trade

A small team working on a single project or a handful of projects can probably work efficiently and effectively to meet business needs. But as the scale and complexity of a development environment increases and the number of developers, QA staff, and operations people expand, it's imperative that organizations establish a DataOps practice.

"We need a DataOps culture … if we are going to go as fast as the business and deliver the quality they want."

One early adopter of DataOps has 40 development teams building data and analytics pipelines for several major departments in a large manufacturing company. "With this many teams, we have to make sure we are aligned with each other and the business. We need a DataOps culture and techniques if we are going to go as fast as the business and deliver the quality they want."

Many people associate DataOps (or DevOps) with test automation software, code repositories, such as GitHub, and continuous integration/continuous delivery methods. These tools and practices are powerful when applied, but it is challenging to get developers and operations staff to change behavior established over years.

"Test automation is only as good as our testing practices," says another DataOps practitioner. "Are people logging what they are testing? Are they logging results? We spent a large amount of time educating people and getting teams to adopt the right practices." Besides testing, it is also important to increase visibility of the data landscape. One practitioner reported about the struggle to track down unexpected performance collapses due to a lack communication and data between operations and developers. "We need a platform where all developers can get insights about how their products are behaving in the wild".

Role of Metrics. DataOps leaders rely heavily on metrics to track productivity (not performance) and cultivate a culture of continuous improvement. For example, one team tracks the number of automated tests executed over time, among dozens of other metrics. "We're on a journey, and the metrics help us gauge our progress and strive to deliver faster, better, cheaper," says a DataOps leader.

"We're on a journey, and the metrics help us gauge our progress and strive to deliver faster, better, cheaper."

To this end, DataOps leaders continuously look for new ways to free up developers to work more productively. This might be as simple as automating manual change request forms or implementing new methods to capture storyboards. "Our [DataOps] managers don't control resources and tasks. They serve the team, seeking ways to make them more productive by removing constraints and future uncertainty," says a DataOps practitioner.

Best Practices

Companies that want to implement DataOps should focus their efforts in three areas: culture, organization, and technology.

Culture

To realize the true promise of DataOps, practitioners need to build a culture of continuous improvement where each team member takes accountability for the whole process, works across boundaries, and embraces automation. To effect cultural change, team leaders need to:

1. Clearly communicate the concepts and principles of DataOps and avoid positioning it as a technology initiative.
2. Formulate the expectation of end-to-end responsibility and knowledge transfer between all stakeholders.
3. Encourage cross-functional communication and collaboration.
4. Embrace continuous improvement and use productivity metrics to drive further efficiencies and process improvements.

Organization

To reinforce DataOps principles, team leaders need to establish agile and lean processes, augmented by testing and automation tools, to support the fast creation of data analytics pipelines. More specifically, leaders should:

1. Encourage everyone to think in analytics pipelines—even non-technical stakeholders.
2. Empower people and give them roles, so they feel responsible. Designate DataOps engineers to promote DataOps thinking throughout the organization.
3. Establish iterative processes with short feedback loops.
4. Bring together diverse stakeholders and experts to support collaboration. Leaders can do this formally in DataOps teams, and informally through goals and meetings.
5. Identify repetitive tasks that can be automated.

Technology

DataOps uses technology to help data stakeholders collaborate and align interests. DataOps teams should strive to:

1. Think of analytics as code and use software engineering tools, such as version control systems, to keep track of change; virtualization to build disposable environments; and test automation to rapidly validate changes.
2. Use automation tools and scripts whenever possible.
3. Use tools to support the orchestration of complex analytics pipelines that involve many stakeholders, tools, and technologies.
4. Establish a good collaboration platform that digitally supports cross-functional communication and knowledge transfer.
5. Use monitoring tools in production to increase visibility of the data landscape and derive insights

and recommendations for possible improvements.

Common Pitfalls of DataOps

The holistic character of DataOps makes an implementation challenging, and even a successful adaptation can encounter a rocky road. The following common pitfalls should be considered on the way to a DataOps culture.

Acceptance: Don't make it technical. DataOps has the same burden as other related ideas in software engineering, which are often characterized as technical knick knacks. It is important to show that DataOps is not a technology trend, but a business trend. Communicate the holistic idea of DataOps, get everybody on board early, and show that business value is the goal, rather than having a fancy tech stack.

Engineering: Keep it simple. DataOps combines many concepts which can be extended almost endlessly, running the risk of over-complication. Always take care to balance effort and benefit. This applies to both technical solutions, such as unnecessary tests and overly complex deployment processes, and to organizational structures, which can easily become burdened with too many roles, feedback loops, and overly rigid processes.

Evolution not revolution. DataOps aims at a fundamental mind-shift that concerns many processes and structures. Consequently, it is not a good idea to introduce DataOps as a "big bang" approach. Preferably, the change is incremental and DataOps gains momentum as people perceive the benefits. Start with a few key processes, gather quick wins, and strive for a slowly-baked cultural change.

Conclusion

This report has defined DataOps and illustrated its holistic approach of bringing stakeholders together to improve quality, increase speed, and establish a culture of continuous improvement. DataOps is not a particular method or tool but rather a collection of principles and a way of doing things on a cultural, organizational, and technological level. To make this more concrete, the report introduced core principles and the idea of analytics pipelines as the heart of DataOps. Additionally, it discussed selected use cases coming from data warehousing and data management, dashboards and reporting, and data science, and outlined concrete steps and common pitfalls for implementation.

DataOps can help to deal with complex data landscapes and analytic solutions that require the coordination of a broad range of stakeholders and technologies. More precisely, DataOps can contribute in the following areas:

- Speed up processes and increase quality by providing streamlined data analytics pipelines via deep levels of automation and testing.
- Increase the value proposition of data and analytics by industrializing processes.
- Establish a culture of continuous improvement and collaboration.
- Support the management and orchestration of heterogeneous technologies and stakeholders.
- Increase visibility in complex data landscapes and assure a stable and efficient operation of applications and infrastructure.
- Operationalize data science to provide more value to the business.

In summary, DataOps is a reasonable potential solution for the new role of data in modern organizations. It requires fundamental changes on many levels indeed, but eventually pays off by making data and analytics more efficient and paving the way to the next level of maturity.

About the Research Sponsor

Unravel Data provides the only Application Performance Management (APM) solution for Big Data. Unravel doesn't just monitor and unify system-level data, but tracks, correlates, and interprets performance data across the full application stack in order to optimize, troubleshoot, and analyze from a single console. Machine Learning capabilities enable autonomous remediation to ensure Big Data application are production ready and fully optimized. Customers include leading Big Data practitioners such as Kaiser Permanente, Leidos, Deutsche Bank and Autodesk.