Eckerson Group

Best Practices in DataOps

How to Create Robust, Automated Data Pipelines

By Wayne Eckerson

June 2019

Research Sponsored by

StreamSets

About the Author

Wayne W. Eckerson has been a thought leader in the data 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 This Report

To conduct research for this report, Eckerson Group interviewed numerous DataOps practitioners from both user and vendor organizations. This report is sponsored by DataKitchen, Infoworks, Unravel, and StreamSets who have exclusive permission to syndicate its content.

Table of Contents

Executive Summary

DataOps promises to take the pain out of managing data for reporting and analytics. In most companies, data travels a tortuous route from source systems to business users. Behind the scenes, data professionals go through gyrations to extract, ingest, move, clean, format, integrate, transform, calculate, and aggregate data before releasing it to the business community.

These "data pipelines" are inefficient and error prone: data hops across multiple systems and is processed by various software programs. Humans intervene to apply manual workarounds to fix recalcitrant transaction data that was never designed to be combined, aggregated, and analyzed by knowledge workers. Reuse and automation are scarce. Business users wait months for data sets or reports. The hidden costs of data operations are immense.

DataOps promises to streamline the process of building, changing, and managing data pipelines. Its primary goal is to maximize the business value of data and improve customer satisfaction. It does this by speeding up the delivery of data and analytic output, while simultaneously reducing data defects—essentially fulfilling the mantra "better, faster, cheaper."

DataOps emphasizes collaboration, reuse, and automation, along with a heavy dose of testing and monitoring. It employs team-based development tools for creating, deploying, and managing data pipelines. This report explains what DataOps is, where it came from, what it promises, and how to apply it successfully.

Key Takeaways

- DataOps applies the rigor of software engineering to data development.
- DataOps practices borrow from DevOps, Agile, Lean, and Total Quality Management (TQM) methodologies.
- DataOps makes it possible to scale development and increase the output of data teams while simultaneously improving the quality of data output.
- The core mantras of DataOps are: faster, better, cheaper; collaborate, iterate, automate; and standardize, reuse, refine.
- DataOps requires a culture of continuous improvement.

Recommendations

This report recommends 10 steps to DataOps success. On the surface, most of these recommendations seem obvious but collectively they provide a powerful strategy for maximizing the value of data in an organization.

1. Assess your data environment
2. Start small
3. Create a data operations department
4. Align with the organization
5. Educate your team
6. Create cross-functional teams
7. Build for reuse and automation
8. Implement data development tools
9. Apply quality checks
10. Create an enterprise data platform

BONUS: Continuously improve

Understanding DataOps

Shakeeb Akhter's data warehousing and analytics team was mired down in a continuous flood of request tickets. Most "new" requests were nearly identical to prior tickets, just tweaks to existing reports or data extracts. The volume of these requests made it impossible for the team to focus on higher-value projects, such as self-service and predictive analytics, much to their dismay. Customers, too, were frustrated by the slow delivery and lack of transparency about their requests.

"We wanted to flip the model and look at different ways of delivering value to our customers," says Akhter, director of enterprise data warehousing at Northwestern Medicine, a leading academic medical center based in Chicago. "DataOps gave us a streamlined, customer-centric process that we needed."

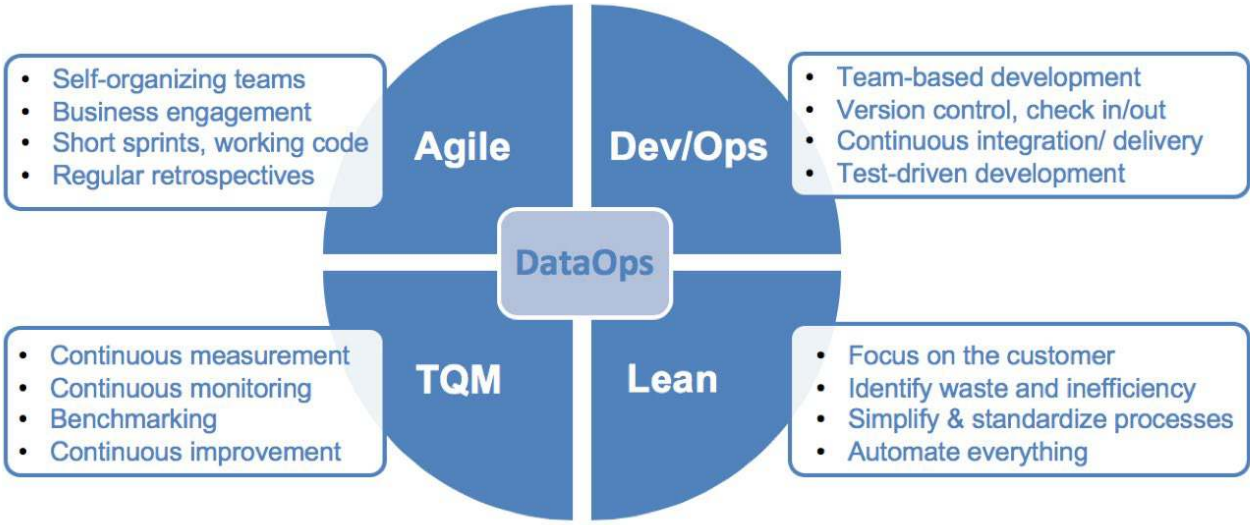
DataOps defined. Short for data operations, DataOps is a set of practices, processes, and technologies for building analytic solutions, including reports, dashboards, self-service analytics, and machine learning models. It applies the rigor of software engineering to the development and execution of data pipelines, which govern the flow of data from source to consumption. The purpose is to accelerate the delivery of data and analytics while simultaneously improving quality and lowering costs. By delivering data "faster, better, cheaper," data teams increase the business value of data and customer satisfaction.

[DataOps] applies the rigor of software engineering to the development and execution of data pipelines.

DataOps is inspired by the DevOps movement in software engineering that uses code repositories, testing frameworks, and collaborative development tools to scale development, increase code reuse, and automate deployments. DevOps bridges the gap between development, QA, and operations teams so organizations can shrink cycle times while reducing 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, collaborative fashion.

DataOps also borrows heavily from Agile, Lean, and Total Quality Management. Like Agile, DataOps emphasizes the use of 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 eliminate 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



To each his own. DataOps means different things to different teams. Some embrace agile concepts and methods, while others implement DevOps tools to better 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.

For the data team at Northwestern Medicine, DataOps first meant creating agile, cross-functional teams dedicated to individual business groups. Each agile team consists of a data architect, a data engineer, a report developer, and a business representative (i.e., product manager), who are cross-trained in each other's skills (except the business person). The team consolidates and prioritizes requests and building end-to-end solutions for their client in an incremental fashion. Says Akhter, "This approach has improved customer satisfaction. There is greater communication and transparency, and the teams have delivered a series of quick wins."

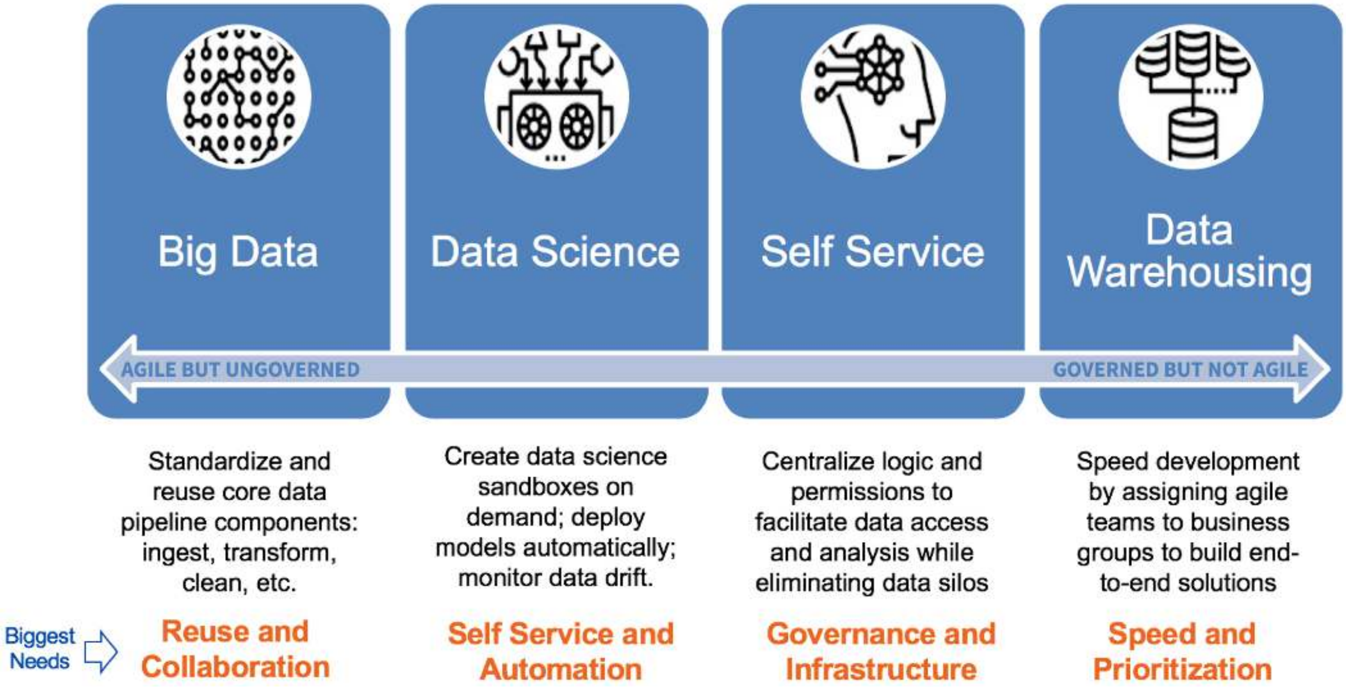
To increase team productivity, Northwestern Medicine invested in a number of tools to foster collaboration and automation.

To increase team productivity, Northwestern Medicine invested in a number of tools to foster collaboration and automation. The data teams now use Git as a source control repository for data integration code; Jira to coordinate Scrum processes and manage user stories; TeamCity to facilitate code integration in a team-based development environment; and Octopus to deploy code from test into production. "Historically, data teams have been loose in how they build things," says Akhter. "We now follow DataOps principles where we segregate duties and environments and apply automation wherever possible."

DataOps Use Cases

Data warehousing is just one area where DataOps can make an impact. It can be applied to any business process or analytic solution that involves extracting, ingesting, cleaning, moving, storing, transforming, integrating, or aggregating data. DataOps first arose in the big data space to address the complexity and scale of those environments. Since many big data environments promote self-service and data science applications, companies now apply DataOps to those use cases as well. (See figure 2).

Figure 2. Primary DataOps Use Cases



Standardize and reuse core data pipeline components: ingest, transform, clean, etc.

Create data science sandboxes on demand; deploy models automatically; monitor data drift.

Centralize logic and permissions to facilitate data access and analysis while eliminating data silos

Speed development by assigning agile teams to business groups to build end-to-end solutions

Other areas for DataOps include artificial intelligence (“AIOps”), cloud migration (“CloudOps”), digital transformation, and Customer 360 projects—essentially, any business activity that requires agile manipulation of complex data to support or create business applications.

Big data. A data lake will turn into a data swamp without robust controls around data or an easy way to operationalize and manage applications built there. DataOps helps big data teams create reusable components, automate data pipelines, and monitor operations. The DataOps mantra here is: “standardize, reuse, collaborate.” For example, Intel applies DataOps to its multi-petabyte data lake to create a “lights out” data processing environment that “minimizes waste and redundancy and fosters a culture of continuous improvement," according to Greg Martinez, enterprise analytics engineer manager at the company.

Data science. Data scientists are often hamstrung by a lack of access to production data and sufficient computational processing to run their models; they are forced to work with sample data on laptop computers. Conversely, many have become dependent on data engineers to create working data sets and application engineers to deploy their models in operational environments. DataOps enables data scientists to provision temporary data sandboxes and create simple data pipelines as well as deploy models with minimal IT or engineering assistance.

Self-service reporting. Data analysts armed with self-service visualization tools have an endless appetite for data sets to feed their analytical inquiries. DataOps enables data analysts to service their own data needs within a curated data environment facilitated by a data catalog and data preparation tools. The data department creates a self-service data infrastructure that balances speed and standards and fosters a culture of governance that accelerates the delivery of data without spawning data silos.

Data warehousing. Data warehouses provide the underpinning for standard reports and dashboards, but are notoriously slow and costly to build and change. Many companies have implemented data warehousing automation tools to reduce the time required to create and change schema and rules. Others, like Northwestern Medicine, apply DataOps principles to better partner with the business and accelerate delivery cycles.

Signs That You Need DataOps

Most data teams can benefit from DataOps, some more than others. If there is “data pain,” DataOps can help. Following is a list of symptoms that indicate whether your data team is a good candidate for DataOps.

1. Your data team is flooded with minor request tickets and is burning out.
2. Business users don’t trust the data because it contains too many errors.
3. You are too busy putting out “data fires” to implement predictive analytics.
4. Source system changes keep breaking your data integration jobs and data pipelines.
5. Business users don’t understand why it takes so long to get a new data set.
6. You have difficulty meeting service level agreements (SLAs) for critical applications.
7. It takes months to deploy a new analytics use case.
8. You rely on business users to debug data quality issues, much to their dismay.
9. Data analysts recreate existing data pipelines with minor variations.
10. Data scientists wait for months for data and computing resources.
11. It is difficult to migrate to the cloud because your data environment is too complex.
12. Your self-service initiative has spawned hundreds of data silos.
13. Your data lake is more of a data swamp.
14. It takes months to deploy a single predictive model.

Best Practices

DataOps represents a broad set of principles, practices, and technologies. The DataOps Manifesto describes core DataOps principles, many of which are pulled directly from Agile, Lean, DevOps, and TQM methodologies. It includes principles such as the following:

- Continually satisfy your customer
- Self-organize
- Reduce heroism
- Reuse
- Monitor quality and performance

To put some meat on these principles, we talked to a number of DataOps practitioners from user and vendor organizations. Following is a compilation of best practices gleaned from these conversations.

1. Assess Your Data Environment

You can't manage what you don't measure. Before starting a DataOps initiative, it's best to conduct an inventory of your existing data environment and processes. The goal is to create a benchmark that you can use to evaluate the impact of DataOps practices. For example, you might want to measure the cycle times for key data processes, such as pulling data from a new source, adding a new column to a database table, populating an OLAP cube, deploying a machine learning model, or creating a sandbox for an individual data scientist.

Identify gaps. Then, you should identify inefficiencies, manual workarounds, and error-prone jobs that prevent the free flow of data from source to target. Also, evaluate how efficiently code moves through each step in the development lifecycle, from development to test and production. "We apply Lean techniques to measure bottlenecks, then we adjust our processes to remove constraints and minimize or eliminate the waste," says Martinez from Intel. Another data leader adds: "It's important to recognize where you are wasting time, effort, and money."

Map processes. It may not be feasible to map all your data pipelines if they are overly complex and messy. But a process map of data flows is a powerful tool to display the waste and inefficiency in a data operation. A picture is worth a thousand words. The map can create a visceral or emotional response that convinces recalcitrant executives to invest more in data operations.

2. Start Small

DataOps is a broad discipline for optimizing the interplay of people, process, and technology to generate data and analytic solutions. Consequently, most experts recommend starting small to avoid getting overwhelmed by possibilities. As they say, a journey starts with a single step. Most recommend starting with the single biggest bottleneck choking the delivery of analytic output, gleaned from your data operations assessment (above).

There are many types of bottlenecks. Most require individuals to serve as “data heroes” sacrificing nights and weekends to avoid delays or embarrassing errors. Without heroes, data teams simply hope and pray when they deploy new functionality that it doesn’t implode or break something downstream. The most common types of bottlenecks are the following:[1]

- Changes to database schema
- Data errors that create unplanned work and disrupt schedules
- Adding a data feed from a new data source
- Deployment processes that frequently break downstream systems
- A slow-moving impact review board
- Provisioning new development environments
- Long test cycles spanning unit, integration, and system testing
- Manual data flows that require human intervention
- Overly cautious development and testing cycles
- Lack of teamwork among data engineers, scientists, analysts, and business users

A process is only as fast as its slowest link. Therefore, focus on the biggest bottleneck in your data operations and devise plans to break the logjam. This may involve reengineering processes or applying new technology to automate steps. Measure the improvement from your actions and then tackle the next bottleneck. Establish a regular cadence (i.e., process) for identifying, addressing, and monitoring the elimination of key bottlenecks.

3. Create a Data Operations Department

It’s much easier to address process and systems bottlenecks if all data and analytics professionals work together in the same department. A key to DataOps success is to create a data team, ideally outside of the IT department and headed by a chief data officer (CDO).

[1] See the online article, “Eliminate Your Data Analytics Bottlenecks,” May 16, 2019.

The IT department excels at managing infrastructure, but is less skilled with data. "IT has technology experts, not data experts," says James Royster, senior director of commercial analytics at Celegne, a global biopharmaceutical company. Royster is creating a data department whose mission is to "structure data and unlock its value."

But even if data operations remain within IT, it's best to carve out a separate identity for the team and populate it with data specialists: data architects, business analysts, data engineers, data scientists, and business intelligence (BI) developers. Most companies have already done this in the data warehousing space, but not in areas such as big data and data science, where the IT department oversees the Hadoop infrastructure or cloud environment.

Carving out a separate data team can be difficult without executive support. MoneySuperMarket, a British price comparison web site, created a data team outside of IT to accelerate the company's data science initiatives, which were stalled in a cloud migration project. "IT was in charge of the data infrastructure, and we were just one of their competing priorities," says Harvinder Atwal, head of data strategy and advanced analytics. "We told IT, 'We can't continue like this. We have to start from scratch.'" Subsequently, the company's CDO orchestrated a reorganization that pulled in data specialists from IT and other areas to create a team dedicated to data science. This "massively reduced friction" for getting things done, says Atwal.

4. Align with the Organization

Scrum. The primary caveat of DataOps is to align with the business. This means putting the customer first and continuously delivering value. (See DataOps Manifesto #1: Continually satisfy your customer.) There are many ways to align with the business. Scrum bakes business engagement into the methodology. Scrum teams, for example, must have a business representative (i.e., product manager) who reprioritizes user stories after every sprint.

Quarterly consolidation. Some companies go a step further and regularly gather Scrum teams with their business counterparts to identify, consolidate, and prioritize cross-functional requirements for analytic solutions. For example, each month, Northwestern Medicine brings together business, IT, and analytics representatives to discuss and prioritize requirements for a particular operational area. The beauty of this approach, according to Akhter, is that "we prioritize work in the line of business as well as across lines of business to ensure we are working on the right things."

According to the Scrum methodology, a business group must provide a "product owner" to the Scrum team to review output and reshuffle priorities. But many businesses balk at this requirement, endangering their Scrum initiatives. Northwestern Medicine draws a line: "If [the business] doesn't assign a certain percentage of a business user's time to the Scrum process, we won't give them bandwidth," Akhter says.

Cascading alignment. Atwal from MoneySuperMarket goes a step further. He says data teams need to align at the strategic level and then cascade requirements to individual projects. "Agile maps really well to business strategy." He says that business objectives should map to agile themes, business strategies to agile initiatives, business tactics to epics, and business actions to user stories. "Every business has a hypothesis of what creates value, and that's their strategy. That's all we work on and nothing else."

ROI metrics. DataOps practitioners say it's critical for data teams to measure business outcomes, not just output. Data teams get so focused on measuring cycle times for producing data sets or data models that they neglect the business impact. Did their efforts move the needle for the business? Did it add revenues, lower costs, or minimize risks? Business metrics that calculate the return on data investment should ultimately guide the data team's work.

DataOps practitioners say it's critical for data teams to measure business outcomes, not just output.

5. Educate Your Team

Expect resistance to DataOps initiatives from data and analytics professionals. Most have worked independently without much structure, process, or controls. They will predict that DataOps will "slow us down" and that the new regimen is "better suited for software development, not data development." Says Akhter, "The controls we put in place felt burdensome because we were the wild west before and people could do whatever they wanted."

To overcome resistance, it's important to educate the team about DataOps. Bring in a consultant or vendor to train the team about DataOps concepts or build an internal curriculum and reading list that explains DataOps principles, practices, and technologies. This gives everyone baseline knowledge and shared terminology that helps the team decide how best to apply DataOps practices.

After learning about DataOps, the data team at Northwestern Medicine decided to create distinct development, test, pre-production, and production environments and implement controls around the creation and promotion of code and data from one environment to another. It also implemented a suite of DataOps tools, including a code repository and continuous integration, continuous delivery, and collaboration products. These tools helped enforce the segregation of environments and enabled the team to scale up development capacity. "If you are building 30 data marts at a time with 30 data architects, you need a streamlined, automated process to protect the production environment and save you from making mistakes," says Akhter.

After the shock of adjustment, the data teams at Northwestern Medicine embraced DataOps. "Our data architects now love DataOps because it provides a framework to deploy code without worrying about breaking things in production. They now say, 'I can't believe we lived without this process.' And it frees them up to tackle other things, such as predictive analytics, non-relational data, and the cloud."

"Our data architects now love DataOps.... They say, 'I can't believe we lived without this process.'"

6. Create Collaborative, Cross-Functional Teams

To scale up operations—deliver more output with fewer people—data teams often think they need to create an assembly line of specialists. This approach is baked into the waterfall method of developing software and does not work well in the analytic world where business users often don't know what they want until they see it.

End-to-end. To better serve its customers, Northwestern Medicine creates cross-functional development teams dedicated to individual business groups. Their task is to build complete end-to-end solutions—currently, dimensional data marts and associated dashboards. Each team has a senior data architect, a data engineer, a BI developer, and a product manager from the business. "DataOps helped us bring resources together across our separate [data and analytics] teams and dedicate them to a particular customer," says Akhter. The data team now runs 30 parallel development projects serving nearly every group in the organization, he adds.

Cross-training. To strengthen teams, Northwestern Medicine cross-trains each team member so when one person is out, another can pick up those tasks. That means a BI developer needs to learn how to architect a system, which can be a tad scary and vice versa, says Akhter. The company uses a buddy system to cross-train individuals, who also take courses to beef up their skills in different areas. As a result, each team now has the skills to "support the customer end to end and provide real value and quick wins," Akhter says.

Incentives. Likewise, MoneySuperMarket brings together data scientists, data engineers, BI developers, and software engineers to deliver end-to-end data science solutions for the business. There are no hand-offs to specialists since teams are organized around the complete data science lifecycle. Team members also share bonuses, which encourages collaboration and helps optimize solutions. "It used to take us a few weeks to create a model, now it takes hours," says Atwal.

"It used to take us a few weeks to create a model, now it takes hours," says Atwal.

7. Build for Reuse and Automation

The best way to improve operational efficiency is to maximize reuse. Unfortunately, most data developers build things in isolation and duplicate efforts. Before long, a company has multiple, redundant ingest mechanisms, data extracts, tests, and data transformations. Without collaboration tools and a shared repository, data architects, engineers, analysts, and data scientists continually reinvent each other's work.

"We create reusable design patterns that enable us to create a data pipeline quickly, change it as needed, and maintain reliability and consistency of the data output."

Reuse is critical in a large data environment. Intel has 30 development teams working in a petabyte-scale Hadoop environment that pulls data from more than 150 sources. To operate efficiently at that scale, the team has standardized numerous data constructs and components that developers can reuse or tweak to accelerate the development of new data pipelines. "We create reusable design patterns that enable us to create a data pipeline quickly, change it as needed, and maintain reliability and consistency of the data output," says Martinez.

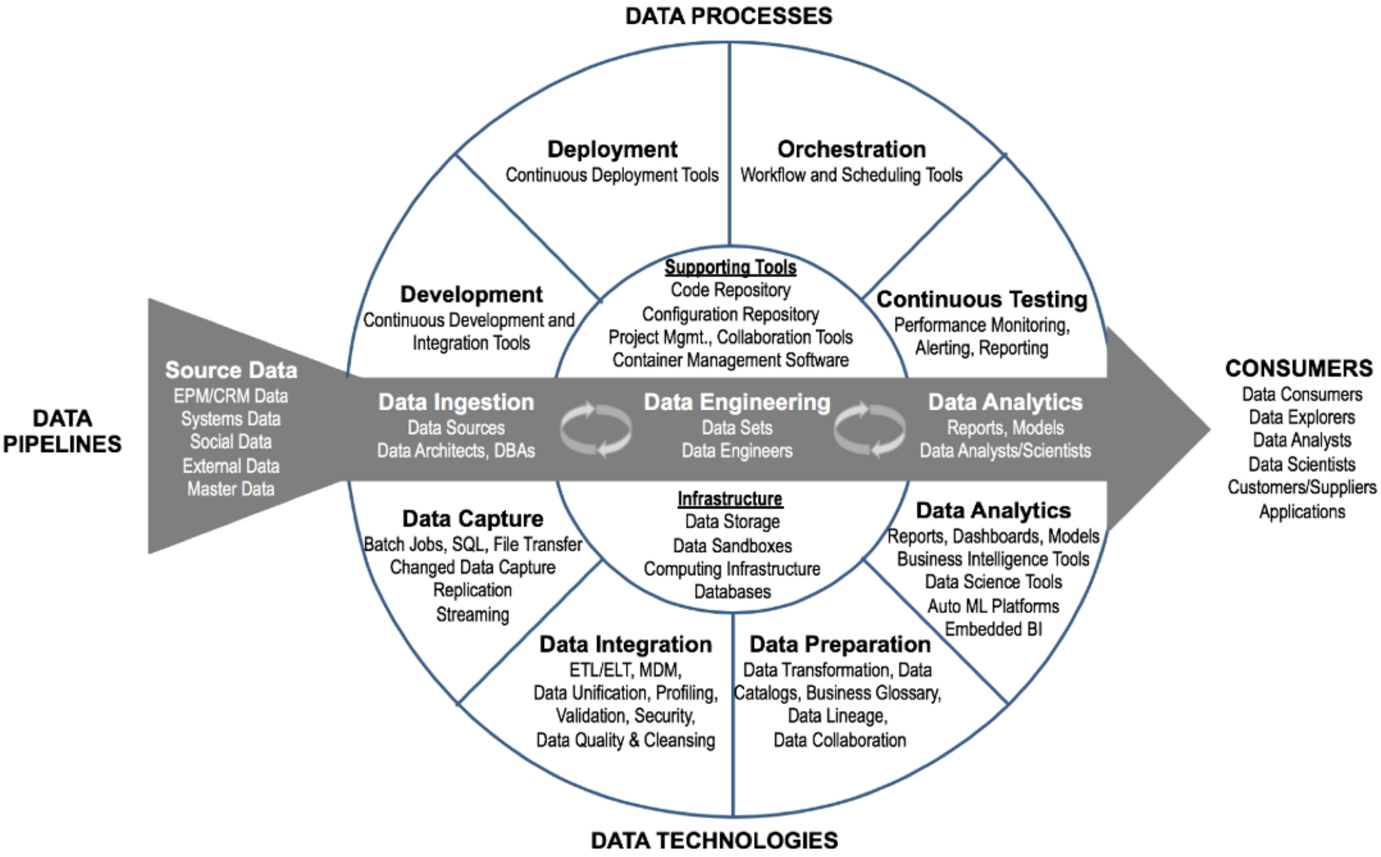
Intel's big data team also has worked hard to automate every aspect of its data operations, from ingest to deployment. In fact, it has developed a "Schema Evolution Framework" that detects changes in source systems and automatically updates target schema and transformations to handle the changes without manual intervention. "We use metadata to automate our data pipelines; we want to focus our engineering resources on innovation rather than rudimentary tasks, such as adjusting data models and transforms every time there is a schema change," says Martinez. Intel plans to offer its framework, along with its automated build and test framework, as an open source project next year.

8. Implement Collaborative Data Development Tools

DataOps tools foster collaboration that is critical for maximizing reuse and automating processes. They enable DataOps teams to scale, increase development capacity, accelerate cycle times, reduce errors, and improve data quality. In fact, DataOps, like its DevOps cousin, is often most associated with a portfolio of tools that can be used either for application or data development.

Figure 3 depicts a DataOps framework that presents a list of DataOps components and technologies. (For the complete description, see “DataOps Explained: A Remedy for Ailing Data Pipelines.”)

Figure 3. DataOps Technical Framework



Case-driven solutions. Some DataOps tools are geared to specific use cases. For instance, data warehouse automation tools are geared to creating small data warehouses and data marts. AIOps tools focus on data science implementations, and CloudOps tools help organizations migrate from on-premises to cloud platforms or support hybrid and multi-cloud environments.

Specialized DataOps solutions. DataOps startups, such as Infoworks, now offer end-to-end solutions for creating, operationalizing, and managing complex data pipelines that span both on-premises and cloud platforms. In contrast, DataKitchen, another DataOps startup, takes a best-of-breed approach, orchestrating the flow of data through existing systems rather than providing an all-in-one solution. Another DataOps vendor, StreamSets, provides a GUI-based tool that makes it easy for data scientists and data engineers to leverage codeless design and manage batch and streaming data pipelines running on-premises, in the cloud, or in edge environments, while adhering to data privacy policies.

Component solutions. Other DataOps tools focus on a single component of the data lifecycle. For instance, Unravel offers a performance management and monitoring tool (see below) geared explicitly to DataOps. It uses machine learning to automatically troubleshoot performance issues afflicting business applications and automatically recommends or executes fixes to comply with SLAs.

Horizontal tools. The bulk of DataOps tools, however, are horizontal in nature and are borrowed directly from the DevOps world. Together, they create a development platform that unites all contributors, fostering reuse and collaboration. The most common categories of DataOps tools are the following:

- **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

The heart of DataOps applies the listed tools in a governed environment to support large-scale development in complex, distributed computing environments. Organizations should allocate plenty of training hours to get team members up to speed on the tools as well as the processes governing their use.

9. Apply Quality Checks

It’s one thing to speed up delivery, it’s another to maintain quality. As Atwal says, “A car needs brakes to go fast.” In the world of DataOps, tests are the brakes that developers create when building code. Those tests are applied not just in unit and integration tests during the development phase, but also during production to ensure that data drift hasn’t altered the accuracy of analytic output, and that changes to software configurations and data schema don’t break production jobs.

Tests are the bedrock of automation. Without tests, automation is a runaway freight train that inevitably crashes.

Tests are the bedrock of automation. Without tests, automation is a runaway freight train that inevitably crashes. With tests, data teams can sleep well at night knowing they have built all the safeguards necessary to keep the train on the track. And if the train starts to deviate, they are proactively notified and can take action before business users experience problems.

"Test automation is a huge part of what we do," says Intel's Martinez. "Without it, we can't maintain a high level of quality at the scale and speed with which we operate. We have more than 1,000 tests in our test automation framework, and we keep adding tests all the time. We continually measure our progress over time, both individually and as a group. We are only as good as our test practices, and we strive to improve here."

Developer tests are supplemented by performance management tools (see above) that monitor system performance and its effect on business applications and users. These systems-level tests enable administrators to optimize performance and ensure compliance with SLAs. For example, a performance management tool will identify "noisy neighbors" on a cluster and detect long-running queries and improperly configured virtual machines.

10. Create an Enterprise Data Platform

DataOps requires a robust, enterprise data platform to succeed. The platform should serve the enterprise, not an individual department or line of business. An enterprise data platform makes it easier to build reusable components and automate data pipelines. It also simplifies governance, security, lineage, auditing, and monitoring because everything runs in one place.

Portability. Ideally, the platform abstracts underlying components, enabling data administrators to swap pieces of the infrastructure or change providers without affecting business applications. This is required for hybrid and multi-cloud strategies where data pipelines span multiple data platforms from different vendors.

Security. DataOps practitioners emphasize the need for an enterprise data platform that simplifies data access while securing data from unauthorized use. Users should be given access to different points of the data environment based on their roles and skill sets. (See my 2016 report titled "A Reference Architecture for Self-Service Analytics: Balancing Agility and Governance.") The infrastructure should automatically detect and mask sensitive data, such as social security numbers.

Centralized logic. A data infrastructure should also centralize business logic used in multiple applications. This removes the temptation for individual developers to embed custom logic into their own reports and data preparation jobs. Business logic can be many things: calculations for core metrics, such as net sales; statistical models for key measures, such as customer attrition; definitions of key business entities, such as "active" and "lost" customers; master data that uniquely defines each product, customer, supplier, and partner; and reference data that defines things like corporate hierarchies, regions, and currency conversions.

"We persist business logic in our data marts so when our BI developers write reports, they don't have to recreate that logic which otherwise would vary from report to report," says Akhter of Northwestern Medicine. Royster of Celegne agrees. His teams select the best layer in the architecture for each type of business logic. Sometimes it goes in the data model, other times in the data integration code, and occasionally in a report, especially when it's a local calculation that isn't shared widely. The company also uses a DataOps tool to track and automatically propagate rule changes to dozens of dashboards that use various rules.

Data catalogs. Many companies use a data catalog to store business logic, including data pipelines, queries, metric calculations, reports, and workflows. This makes it easy for data developers to find and reuse logic instead of starting from scratch, which would lead to a proliferation of data silos and conflicting data. A data catalog is fast becoming a required component in a modern data architecture because it can be used to consolidate and curate business logic for analytic applications.

A data catalog makes it easy for data developers to find and reuse logic instead of starting from scratch.

Self-service. Data scientists also benefit from a data platform that centralizes logic and abstracts the underlying complexity of data. Rather than rely on data engineers to fetch data for them, data scientists should be able to build their own data pipelines, according to Jeff Magnusson, vice president of data platform at Stitch Fix. They should also be able to deploy predictive models without engineering assistance. The only way to accomplish this is to create a robust data platform that simplifies these tasks.

"I'd rather focus good, strong engineers on building tools and abstractions to make ETL, data movement, and data science easier versus having those folks engineering each specific data pipeline that needs to get developed. And so, by creating those tools, that in turn empowers data scientists to take full ownership of their pipelines from data acquisition to production, and then they can control their iteration cycles, and that often increases velocity."

An enterprise data infrastructure with centralized rules and permissions makes it possible to support self-service without creating data silos and spreadmarts. A good data infrastructure builds governance into the fabric of the data environment, balancing governance and self-service, agility and architecture, and speed and standards.

Buy, don't build. Rather than build a data platform for internal use like Stitch Fix, Atwal from MoneySuperMarket prefers to buy it from a dedicated platform vendor. "It's silly to build your own data platform today," he says, especially when vendors specialize in building scalable, elastic, open, and services-oriented data platforms. MoneySuperMarket selected Google as its analytics cloud provider and Domino Data Lab as its data science platform. "We just bring our data and code; there is nothing for us to manage. We run everything on the same platform, which makes it easy to govern."

Bonus: Continuously Improve

DataOps is a journey. The goal is to create a culture of continuous improvement where every team member works to identify and eliminate waste, maximize reuse and automation, and accelerate cycle times to deliver greater value to the business. “We are on a journey of continuous innovation,” says Intel’s Martinez. “We focus on business outcomes and continually experiment with new approaches to optimize data operations.”

To make progress, it’s important to periodically stop and review what you’ve done. Scrum teams finish each sprint with a half-day “retrospective” where the team reviews what went well, what didn’t, and how it can improve. Intel goes a step further and dedicates every third or fourth sprint to examining ways it can improve data operations. The team learns and applies new techniques so that it might work faster and more efficiently.

Intel has doubled development capacity and tripled its output without adding new people or overhead.

Stopping midstream to review and reflect when there are so many pressing business requirements to deliver takes courage and vision. But Intel’s commitment to continuous improvement and buy-in from top executives makes this possible. The results are impressive: Intel has doubled development capacity and tripled its output without adding new people or overhead.

Conclusion

As data pipelines become more complex and development teams grow, organizations need to apply standard processes to govern the flow of data from source to consumption. The goal is to improve agility and cycle times while reducing data defects, giving business users greater confidence in data and analytic output. This is the vision of DataOps.

DataOps is a full-throated strategy for maximizing the business value of data.

Most people associate DataOps with either agile principles or team-based development tools. But it's much more than that. DataOps is a full-throated strategy for maximizing the business value of data. New development tools and processes won't deliver much value unless they are backed by enlightened leadership that recognizes the power of data to transform organizations and fuel new data-centric strategies, such as digitalization, Customer 360, artificial intelligence, and the Internet of things.

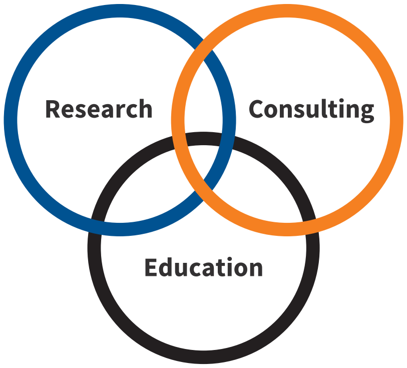
Strong leaders build new organizations to match their vision. DataOps requires a dedicated data organization that spans all data-related activity: data warehousing, data lakes, data science, and self-service analytics. Finally, DataOps requires a data-driven culture that validates decisions with facts and seeks to continuously improve the processes by which it delivers data to business users.

About Eckerson Group

Wayne Eckerson, a globally known author, speaker, and advisor, formed Eckerson Group to provide data-driven leaders like you a cocoon of support during every step of your journey toward data analytics excellence.

Today, Eckerson Group has three main divisions:

- **Eckerson Research** publishes insights so you and your team can stay abreast of the latest tools, techniques, and technologies in the field.
- **Eckerson Consulting** provides strategy, design, and implementation assistance to meet your organization's current and future needs.
- **Eckerson Education** keeps your data analytics team current on the latest developments in the field through three- and six-hour workshops and public seminars.



We Help Analytics Leaders Succeed

Unlike other firms, Eckerson Group focuses solely on data analytics. Our veteran practitioners 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 share 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 serve—we see ourselves as a family of continuous learners, interpreting the world of data for you and others.

Accelerate your data journey. Put an expert on your side.
Learn what Eckerson Group can do for you!

Contact Us

Schedule a Call

About Streamsets

StreamSets built the industry's first multi-cloud DataOps platform for modern data integration, helping enterprises to continuously flow big, streaming and traditional data to their data science and data analytics applications. It uniquely handles data drift, those frequent and unexpected changes to upstream data that break pipelines and damage data integrity. The platform combines the award-winning, open source StreamSets Data Collector™ for execution of any-to-any pipelines (the data plane) with a cloud-native StreamSets Control Hub™ for the continuous automation and monitoring of multi-pipeline topologies (the control plane).

Learn more at www.streamsets.com