



DataOps Deep Dive:

Different Approaches to the DataOps Platform

BY JOE HILLEARY

DECEMBER 2020

CUSTOM REPRINT PREPARED FOR DATAKITCHEN

About the Author



Joe Hilleary is a writer and a data enthusiast. He believes that we are living through a pivotal moment in the evolution of data technology and is dedicated to helping organizations find the best ways to leverage their information. With a background in both analytics and the liberal arts, he crafts clear, articulate narratives on technical topics that empower stakeholders to make informed decisions. Hilleary is a Research Analyst at Eckerson Group.

About Eckerson Group

Eckerson Group is a global thought leader that helps organizations get more value from their data. Our research and consulting experts think critically, write clearly, and present persuasively about data analytics.

They specialize in data strategy, data architecture, data management, data governance, data science, and data analytics. Organizations rely on them to demystify data and analytics and develop business-driven strategies that harness the power of data. **Learn what Eckerson Group can do for you!**



About This Report

To conduct research for this report, Eckerson Group interviewed numerous industry experts and practitioners. The report is sponsored by DataKitchen, DataOps.live, Zaloni, and Unravel who have exclusive permission to syndicate its content.

This is an abridged version of the full report. To read the complete report, [click here](#).

Table of Contents

Executive Summary	4
Introduction	5
DataKitchen	9
Conclusion	15
About Eckerson Group	16
About the Reprint Sponsor	17

This is an abridged version of the full report. To read the complete report, [click here](#).

Executive Summary

DataOps is an emerging methodology for building data analytics solutions. Drawing on DevOps and agile approaches to software development, it promises to reduce project times, decrease errors, reduce costs, and improve customer satisfaction. But just as DevOps requires a suite of tools to implement the methodology, so too does DataOps.

This report examines four leading DataOps platforms: DataKitchen, DataOps.live, Zaloni, and Unravel. It describes each product, highlights its key differentiators, and identifies target customers for each. From these profiles, readers will gain a better understanding of the range of DataOps offerings and discover which products are best suited to their needs.

This is an abridged version of the full report. To read the complete report, [click here](#).

Introduction

What is DataOps?

DataOps is an emerging methodology for developing and deploying data analytics solutions. Adapted from the DevOps and agile techniques for software development, DataOps takes a holistic approach to the people, processes, and technology required to build and automate data pipelines. It has four key pillars: continuous integration and deployment (CI/CD), orchestration, testing, and monitoring. These functions layer on top of the tools that make up data pipelines and help data teams deliver products faster, better, and cheaper.

Continuous Integration/Continuous Development. CI/CD requires a single source of truth for all the data and code that make up a pipeline. DataOps ensures that this source of truth remains untouched throughout the development process so pipelines already in production don't break. With the source of truth safely housed in a central repository, team-based development becomes possible and developers can innovate without fear, reducing development cycle times.

Orchestration. Modern data pipelines are complex. Data passes through numerous tools and storage locations on its journey from source to target. As a result, a functional DataOps strategy requires an orchestrator. Orchestrators connect to all the tools in the data workflow and automate the end-to-end journey of the data. Automation frees up developers to build new pipelines and enables one engineer to manage hundreds of pipelines in production.

Testing. In the software world, nearly 50% of code and staff are dedicated to testing and quality. In the data world, 20% would be unusually high. DataOps seeks to change that. It encourages data engineers to bake tests into pipelines that check both data quality and pipeline functionality. The tests run during both development and production. Although these tests may seem like extra labor, test-first development saves time because the pipelines deliver higher quality data so engineers don't have to constantly troubleshoot errors. And when pipelines do break, they are much easier to diagnose.

Monitoring. The final piece of the DataOps puzzle is monitoring the execution of code and data in production environments. Monitoring is critical for managing the underlying infrastructure of servers, CPUs, memory, and storage nodes that process data pipelines. It also aids in determining when and where bottlenecks and breakages occur. Finally, it helps engineers understand and optimize the impact of their pipelines on shared resources. These elements work in concert to improve performance, giving the team the information to optimize their pipeline execution. Monitoring is no easy feat in modern data ecosystems and, like orchestration, requires specialized tools that connect with and see across all of the component technologies. Once monitoring is in place, however, the increase in pipeline efficiency reduces overhead costs.

The Data Pipeline Framework

Eckerson Group uses the following framework to visualize the data pipeline within the DataOps paradigm. (See figure 1.)

Figure 1. Data Pipeline Framework



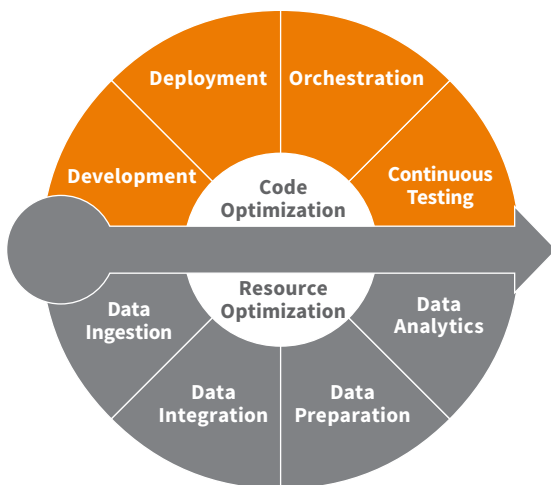
The pipeline itself runs from data sources to targets, passing through ingestion, engineering, and analysis on the way. The bottom half of the circle consists of the technologies used to build the data pipeline—including replication, ETL, data catalogs, data quality, lineage tracking, data science, and business intelligence (BI) tools. The top half represents the DataOps components used to manage and optimize the development and execution of data pipelines. Together they create a visual framework for modern data pipeline development.

Four Products, Four Approaches

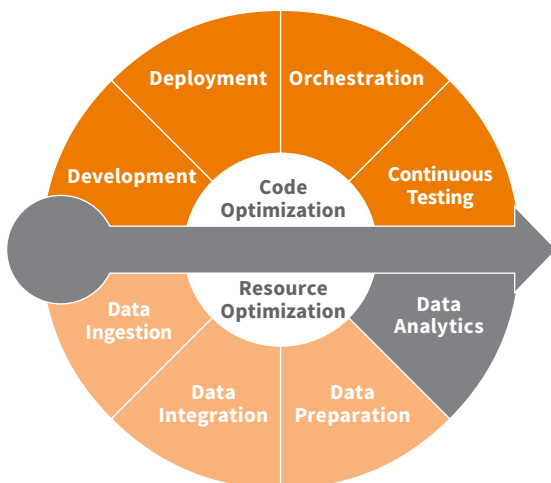
This report profiles four DataOps products that take divergent approaches to implementing DataOps principles and practices. Our goal is to help you better understand the range of DataOps capabilities available in the market today and identify products or categories of solutions best suited to your organization's needs.

The products come from four vendors: DataKitchen, DataOps.live, Zaloni, and Unravel. Each created a DataOps platform that facilitates key aspects of the DataOps methodology, but each takes a different approach, solves slightly different problems, and is geared to a slightly different target customer. Some provide both data development and DataOps functionality, while others offer purely the DataOps side. Most support all four pillars of DataOps (CI/CD, orchestrating, testing, and monitoring), but some specialize, focusing on a single pillar. Each vendor also possesses unique characteristics that further differentiate it from other tools in the same category of DataOps platforms.

Framework Key: ■ fully supported ■ partially supported ■ not supported

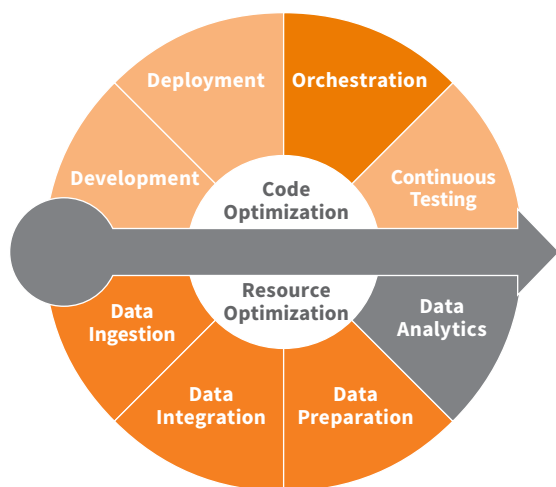


DataKitchen. For instance, the DataKitchen DataOps Platform focuses solely on the four key DataOps functionalities of orchestration, testing, CI/CD, and monitoring. It does not provide the components for a data pipeline. Instead, it layers DataOps functionality on top of existing data ecosystems. It serves as an overlay environment to existing data pipeline development tools—orchestrating jobs, managing development, building and running tests, and monitoring execution. Its target users have complex environments, are passionate about test-driven development, and want to continue using their current tools.

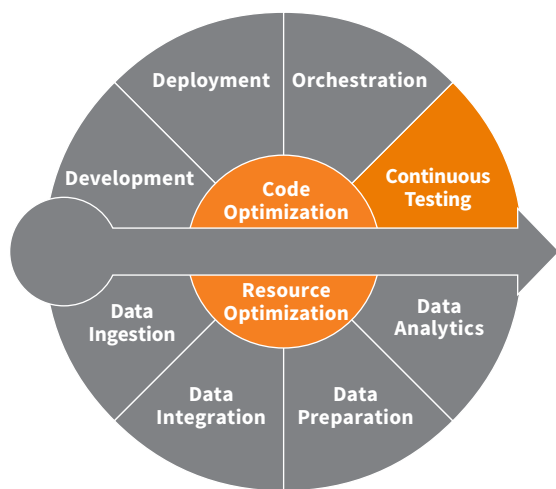


DataOps.live. DataOps.live represents more of a hybrid approach to the DataOps platform. It leads with DataOps orchestration functionality, but also provides select elements of the actual pipeline for ETL/ELT, modeling, and governance. DataOps.live has a novel approach to enabling CI/CD for data that allows for branching

databases, and its in-house tools help users quickly move pipelines into production, but its current release is dependent on customers using Snowflake. It is well suited to users who want both orchestration and pipeline tools out of the box, and it delivers additional benefits for those who use machine-generated data thanks to proprietary compression technology.



Zaloni. Zaloni’s Arena is an all-in-one enterprise development and execution environment with built-in DataOps features. While it can orchestrate other tools and provides the other necessary elements for DataOps, it focuses on delivering every component piece needed for the pipeline within a single platform. Arena provides a complete pipeline environment up to the point of analytics and is especially well adapted for companies with strict compliance and governance requirements.



Unravel. Unravel provides AI-driven monitoring and management to improve the performance, scalability, and reliability of data pipelines. It tunes and validates code and offers insights that improve testing, deployment, and resource optimization. Unravel is a great match for companies who are missing the observability, or monitoring, component of DataOps and want to focus on improving pipeline performance and the compute and memory efficiency of their data-driven applications across a complex data ecosystem.

While many other DataOps platforms exist, these four represent the major categories. The profiles below will delve further and provide details about the organization’s background and target customers, the product’s architecture, and the platform’s primary functionalities. Armed with a solid understanding of these offerings, you will be in a better position to vet any other DataOps products that interest you.

DataKitchen

Founded: 2013

Product: DataOps Platform 1.1.186

Initial Product Launch: 2016

CEO: Christopher Bergh

Executive Summary

The DataKitchen DataOps Platform is a highly-extensible tool that helps organizations implement a DataOps strategy by providing a graphical interface for developing, updating, managing, monitoring, and testing data workflows. It orchestrates all the software products and scripts in end-to-end data environments and allows data engineers to schedule and embed tests at each step of the workflow, ensuring functionality and data quality both during development and production. It is a good match for larger companies that want to keep their existing suites of tools, but need a better way to manage the development process to increase speed of delivery and decrease error rates. The company is particularly service-oriented and distinguishes itself by helping data teams holistically transform their development and operations procedures through hands-on consulting.

Background

Company

CEO Christopher Bergh and co-founders Gil Benghiat and Eric Estabrooks spent careers leading software development teams before meeting at LeapFrogRx, a data analytics firm that served the healthcare industry. There, they were continuously frustrated by the existing methodology for developing data solutions. They identified three key issues—source data was often low-quality, business users were intolerant of errors or delays, and the data science team didn't have the flexibility to innovate.

In response, they reimagined what data development could look like. Adopting techniques from the DevOps and agile methodologies, they honed a new development paradigm, which they called DataOps, borrowing a newly emerging industry term; wrote a [manifesto](#); and started DataKitchen to evangelize their new approach.

In 2016, after several years as a professional services group, the company launched its flagship product: the DataKitchen DataOps Platform. The platform helps companies better develop and manage data pipelines and solutions. For DataKitchen, DataOps is more than a product—it's a philosophy for transforming the data product development process.

DataKitchen currently has 41 employees, is profitable, and is growing rapidly. Its partners include tech companies such as IBM, Tamr, and Kinaesis, along with systems integrators including Capgemini, Wipro, and Cognizant.

Customers

DataKitchen's target customer is a large company that wants to move from a waterfall-based development methodology to an iterative one. Many of DataKitchen's current clients are pharmaceutical companies, including BMS and AstraZeneca, but it also has customers in other industries, like Catholic Relief Services.

According to Bergh, DataKitchen's customers all exhibited the following characteristics:

1. Slow delivery rates for building or changing data analytics solutions;
2. High error rates; and
3. A desire to transform team-based development processes.

The primary users fall into three categories. The first is the DataOps engineer. Like a DevOps engineer on a software development team, the DataOps engineer is a technical user who is focused on the process of building, changing, and managing workflows rather than writing code for the product itself. The platform helps them set up environments, manage integrations, and build workflows. The second user is a node writer, such as a data scientist or data engineer, who generates the code that runs within tools orchestrated by the platform. The platform gives these users a clear picture of how their part fits into the whole, facilitating better collaboration. The final core persona is the team manager who uses the reporting features of the platform to monitor their team's development efficiency.

DataKitchen is a good match for companies that need a better way to manage workflow development and deployment but want to keep the tools they have. The platform generates many useful dashboards that measure the productivity of development teams and processes to help managers gauge the impact of adopting the DataOps methodology.

DataKitchen has worked hard to define the term DataOps and evangelize the practice. It has made a name for itself through its blog and other publications, which provide in-depth explanations of the DataOps philosophy and practices. These resources are an excellent starting point for anyone considering a DataOps transformation.

DataKitchen believes that software is just one of many critical success factors. It provides education and consulting services to help customers get started with DataOps and address the people, process, and operational issues required to succeed. This holistic approach to DataOps distinguishes DataKitchen from its competitors.

Product

The DataKitchen DataOps Platform provides functionality that spans the entire DataOps lifecycle:

- 1. Deployment.** Users can create new data workflows and migrate code across development, testing, and production environments.
- 2. Execution.** The platform executes these data workflows by orchestrating all software in an organization's data analytics environment whether on premises or in the cloud.
- 3. Management.** The platform monitors the execution of data workflows, runs tests, and reports on the results.

Deployment

The first step for new users create workspaces called “Kitchens” for production, development, and any other environments. These Kitchens isolate the environments, helping engineers avoid disrupting production while building new solutions. Once a new data solution is ready, users can automatically and continuously deploy it into production or another environment.

Within a Kitchen, the platform's graphical interface allows users to construct visual workflows, which DataKitchen calls “Recipes.” These development or production workflows consist of nodes which represent code that acts on data inside the tools the data team already uses. For example, a node might execute data transformations or calls for a machine learning (ML) model. (See figure 2.)

Figure 2. Sample DataKitchen “Recipe”



The nodes can interface with anything from BI tools, ETL tools, or databases, to additional orchestrators. Data scientists can also upload their models from Jupyter notebooks or other development environments directly into nodes and the platform will store the code. Engineers can save commonly reused parts of the workflows as “Ingredients”— reusable bits of code that would otherwise be copy-and-pasted over and over again. Teams can easily share these Ingredients, saving data solution developers valuable time and increasing collaboration.

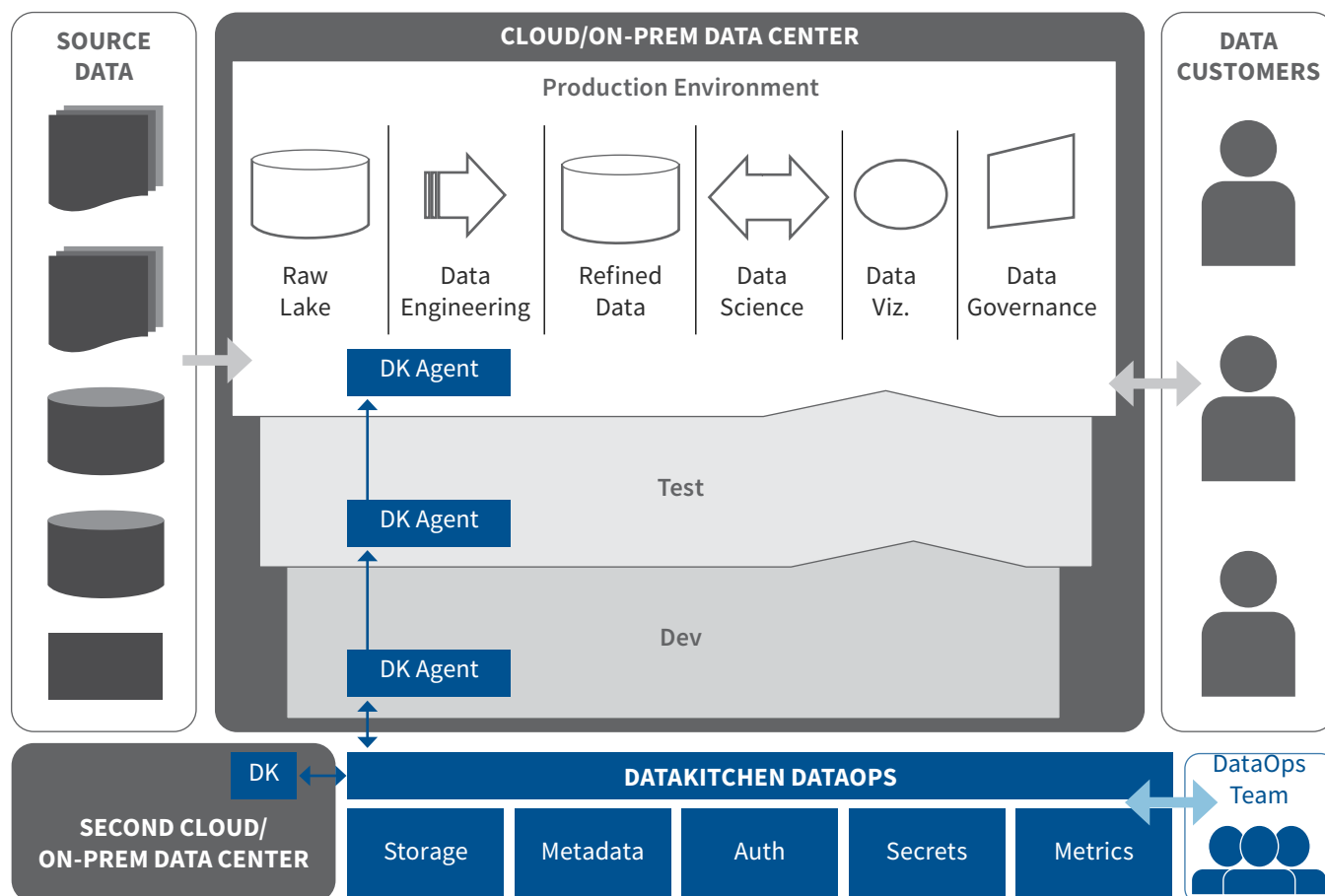
Once the user is satisfied with the workflow’s structure, they can write tests and embed them at each stage in the workflow. Tests are critical for DataKitchen, which believes at least 20% of a data team’s effort should be devoted to testing. These tests vary by domain and can involve everything from evaluating models to checking data quality. During development, the data is fixed and the code varies, while during production the code is fixed and the data varies, but in both instances, tests are vital for catching errors. As a result, the data team can deal with any errors in the data, the timing of the workflow, or the artifacts created from data such as models and dashboards before the solution reaches the end user.

Execution

With its orchestration capabilities, DataKitchen never touches the actual data. Instead, it directs the tools that do, allowing users to manage every step of the process from source to customer in one place. When a data workflow runs, DataKitchen uses an agent to push code, configurations, and tasks to the appropriate tools at the appropriate times and receive run information in return. These agents sit in Kubernetes clusters in the cloud or on premises, enabling the platform to work across hybrid environments. This feature is a boon to organizations that manage production, testing, and development in different environments or have teams operating on different clouds.

The platform sits in DataKitchen’s own cloud and acts as a hub for all the agents, with which it communicates constantly. Despite this continual back-and-forth, the data itself never enters DataKitchen’s cloud and remains in the user’s environment at all times. The workflows, along with any uploaded code, are stored in Git, the results in MongoDB, and any keys or passwords are stored in a secrets vault, all of which again can be on premises or in the customer’s cloud environment. (See figure 3.)

Figure 3. DataKitchen Architecture



Management

When a user runs the workflow, the DataKitchen DataOps Platform extracts test and log data via APIs. This information allows it to keep track of data quality and workflow functionality despite its lack of direct contact with the data. Any tests the user has written into the workflow trigger automatically, and the results are reported back to a dashboard. The visibility of these metrics makes it easier to track how changes in the development process actually affect the quality of the data solution. After all, you can't manage what you can't measure.

Pricing

The default DataKitchen package is a two-year subscription. Customers are charged an average of \$150,000 per year per "use-case," which is a project, team, or application (depending on the company's structure). Within a given use-case, there are no additional limits on usage—in contrast to other products that charge by user, cores, machines, or data volume.

Recommendation

DataKitchen sells more than software. It offers a methodology designed to help data teams transform the way they build data workflows. Its target customers are companies that are tired of band-aid solutions and are ready to fundamentally change their operating principles.

DataKitchen is ideal for customers that:

- > Have a complicated data analytics environment.
- > Would like to keep using their existing tools.
- > Struggle with poor data quality and slow data delivery times.
- > Want additional consulting services to help implement a holistic DataOps strategy.

Conclusion

DataOps is a new practice that is gaining adoption as organizations recognize the importance of transitioning from artisanal- to industrial-scale processes for developing and running data pipelines. DataOps enables organizations to evolve from slow, one-off development efforts to a team-based development approach that can build, change, and manage thousands of pipelines with high speed and accuracy.

We're still in the early years of DataOps, but the shape of DataOps platforms is moving out of the shadows into the clear light of day. This report outlines four approaches to DataOps using leading DataOps vendors as examples. The first exclusively provides the four DataOps pillars of orchestration, testing, CI/CD, and monitoring. This category of products focuses on wide extensibility and layers DataOps functions onto complex data ecosystems. The second group is a hybrid class of platforms. These tools lead with DataOps but also provide select elements of the actual data pipeline. The third is the all-in-one enterprise development and execution environment with built-in DataOps features. This type of product natively provides all the pieces of a data pipeline within a single platform and then layers some DataOps functionality on top. The final classification showcased in this report is the niche player which specializes in a select element of the DataOps methodology and provides best-in-breed functionality for that component.

Recommendations

- > DataKitchen is an excellent option if you have a complicated data ecosystem and just want to layer DataOps functions on top of your existing development toolset. It's also a good fit for companies that are passionate about a test-driven development strategy.
- > DataOps.live is a great choice for companies that need orchestration of complex data pipelines around Snowflake, especially if they rely on a lot of IoT data. Its hybrid approach means you can use native development and testing tools for some data pipeline processes, and orchestrate any 3rd party tools as required.
- > Zaloni is ideal for companies that want to consolidate their data landscapes by using an all-in-one enterprise development and execution environment with built-in DataOps functionality. Its security and governance features are also beneficial for organizations with lots of regulatory and governance requirements.
- > Unravel, as a specialist tool focusing on monitoring, is an excellent choice for organizations that want to add the monitoring capability component to their DataOps strategy but have complex, high-volume data environments—be they on-premise, hybrid, or in to the cloud. It provides insights from across the full stack of modern data applications and is especially helpful during cloud migrations.

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 expert guidance during every step of their data journey.

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 field.
- > **Our consultants** listen carefully to craft tailored solutions that translate your business requirements into compelling strategies and solutions.
- > **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global thought leader that helps organizations get more value from their data. Our research and consulting experts think critically, write clearly, and present persuasively about data and analytics. They specialize in data strategy, data architecture, data management, data governance, data science, and data analytics. Organizations rely on them to demystify data and analytics and develop business-driven strategies that harness the power of data.

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

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



About the Reprint Sponsor

DataKitchen's mission is to make DataOps happen in the world.



DataKitchen's DataOps Platform simplifies complex toolchains, environments, and teams, so your entire data analytics organization can quickly innovate, seamlessly collaborate, and instantly deliver the kind of error-free, on-demand insight that leads to one successful business decision after another. The platform automates the key functions of a DataOps program—orchestration, testing, monitoring, environment creation and management, and deployment of new analytics. Process metrics provide unprecedented visibility into the state of your data operations.

Yet, a DataOps transformation involves more than just tools and technology. DataKitchen's DataOps Transformation Advisory Service helps organizations plan and execute the other critical elements of a DataOps program - people and process—that ensure DataOps success.

With DataKitchen by your side, you and your team can stop worrying about everything that went wrong and start winning respect and appreciation for everything that goes right.