Eckerson Group | Deep Dive

DataOps Deep Dive:

Different Approaches to the DataOps Platform

BY JOE HILLEARY

DECEMBER 2020

CUSTOM REPRINT PREPARED FOR UNRAVEL

unravel™

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

Unravel 9

Conclusion 14

About Eckerson Group 15

About the Reprint Sponsor 16

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.

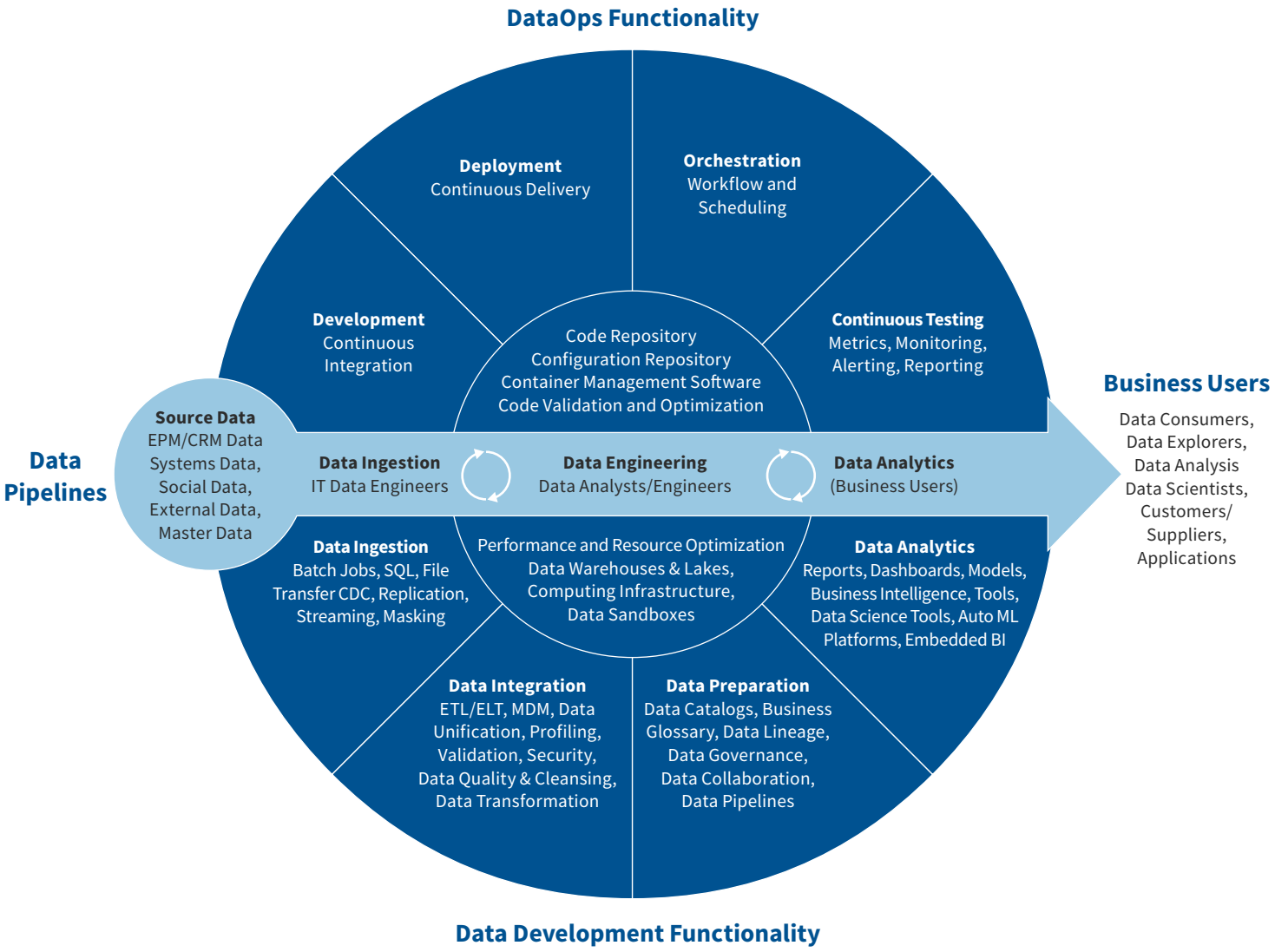
Testing. In the software world, nearly 50% of code and staff are dedicated to testing and quality. In the data world, 20% would be unusually high. DataOps seeks to change that. It encourages data engineers to bake tests into pipelines that check both data quality and pipeline functionality. The tests 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



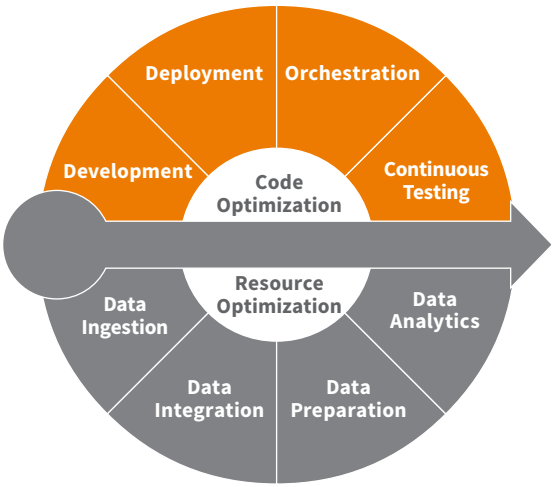
The pipeline itself runs from data sources to targets, passing through ingestion, engineering, and analysis on the way. The bottom half of the circle consists of the technologies used to build the data pipeline—including replication, ETL, data catalogs, data quality, lineage tracking, data science, and 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

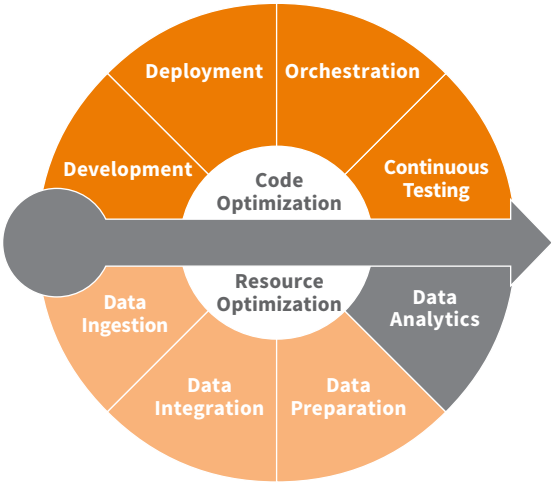
This report profiles four DataOps products that take divergent approaches to implementing DataOps principles and practices. Our goal is to help you better understand the range of DataOps capabilities available in the market today and identify products or categories of solutions best suited to your organization's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

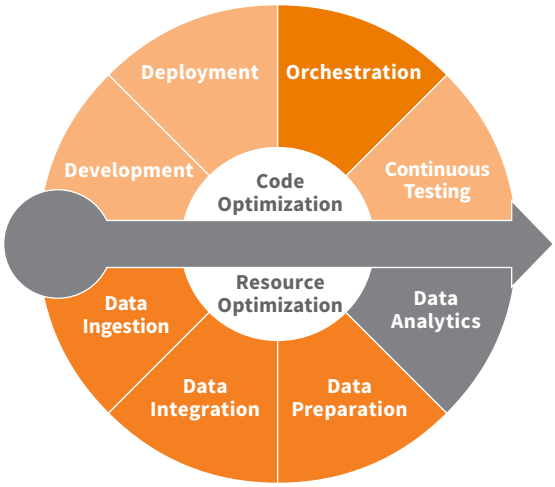


DataKitchen. For instance, the DataKitchen DataOps Platform focuses solely on the four key DataOps functionalities of orchestration, testing, CI/CD, and monitoring. It does not provide the components for a data pipeline. Instead, it layers DataOps functionality on top of existing data ecosystems. It 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.

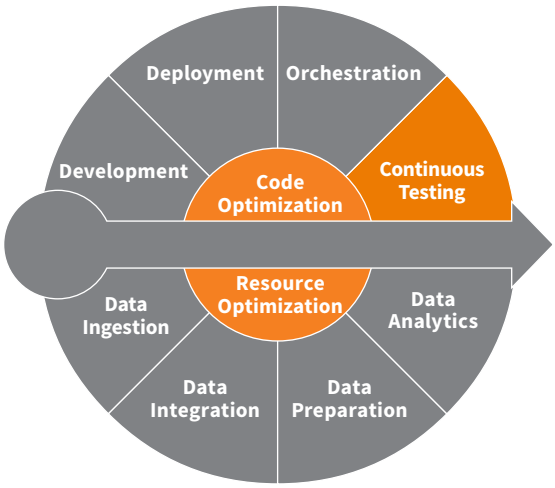


DataOps.live. DataOps.live represents more of a hybrid approach to the DataOps platform. It leads with DataOps orchestration functionality, but also provides select elements of the actual pipeline for ETL/ELT, modeling, and governance. DataOps.live has a novel approach to enabling CI/CD for data that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

Unravel

Founded: 2014
Product: Unravel Data Operations Platform 4.6.2
Initial Product Launch: 2018
CEO: Kunal Agarwal

Executive Summary

The Unravel Data Operations Platform is an AI-powered system for monitoring and managing data-driven applications and pipelines. It identifies performance bottlenecks in the underlying systems and software and uses artificial intelligence to detect anomalous behavior, suggest fixes, and remediate problems. It generates reports and alerts that enable developers to improve the code and tests they write; data architects and engineers to improve the throughput of data pipelines and workflows they build; and administrators to fine-tune system configurations to improve performance, reduce costs, and meet systems level agreements (SLAs). Unravel works on-premises and in the cloud, with in-depth support for cloud migration.

Unravel is designed for companies in data-intensive industries or where data is a central part of a product offering. It is a good fit for organizations that already have complex, developed data ecosystems but want to optimize application performance, automatically detect and remediate problems, and maximize customer satisfaction.

Background

Company

Kunal Agarwal founded Unravel in 2014 at Duke University with professor of computer science Shivnath Babu, who now serves as the CTO. Given the complexity of modern data environments, they wanted to create a simpler way to understand and optimize the performance of data applications. They built a team of employees with experience at companies such as Cloudera, Rocketfuel, Microsoft, Apple, and Appdynamics and, four and a half years later, released their central product offering—the Unravel Data Operations Platform. Unravel is growing quickly, and its solution is currently deployed across tens of thousands of machines. Now based in Palo Alto, Unravel's investors include M12 (formerly Microsoft Ventures), Data Elite, Menlo, GGV Capital, Harmony Partners, and Point72 Ventures. It also has AWS, Microsoft Azure, Google Cloud, Databricks, and IBM as technology partners and Accenture, ECS, and Logicalis as system integrators.

Customers

Unravel's typical customer today is data-mature but needs greater full-stack visibility to optimize the cost and development efficiency of its data applications. These companies see data as mission-critical and have already invested in complex data environments. They rely on data to run their business and

have customer-facing data-driven products. Unravel's customers are typically drawn from Fortune 2000 companies that use big data to create a competitive edge, or companies whose core offerings are built around data. Current clients come from a variety of industries and include organizations like Wells Fargo, Morgan Stanley, United Airlines, Kroger, Adobe, CVS, and Aetna.

Companies that spend a lot on compute power across a wide range of data technologies, but want greater clarity on where that money goes, would do well to consider Unravel. In particular, organizations that are on or moving to the cloud stand to save the most from Unravel's emphasis on improving code and run-time efficiencies. Additionally, its ability to extract usage information and metadata from across complex modern environments should make it especially appealing to organizations with decentralized data ecosystems.

Product

The Unravel Data Operations Platform is, in many respects, a next-gen application performance monitor tailored to the unique challenges of development in the modern big data context. Cloud and hybrid environments, coupled with the adoption of innovative technologies like Kafka, Spark, and Hadoop make monitoring data applications more complex than traditional applications. Unravel works to alleviate this complexity by providing full-stack visibility across the data ecosystem. It also reduces the time needed to manage these systems by way of automation and AI.

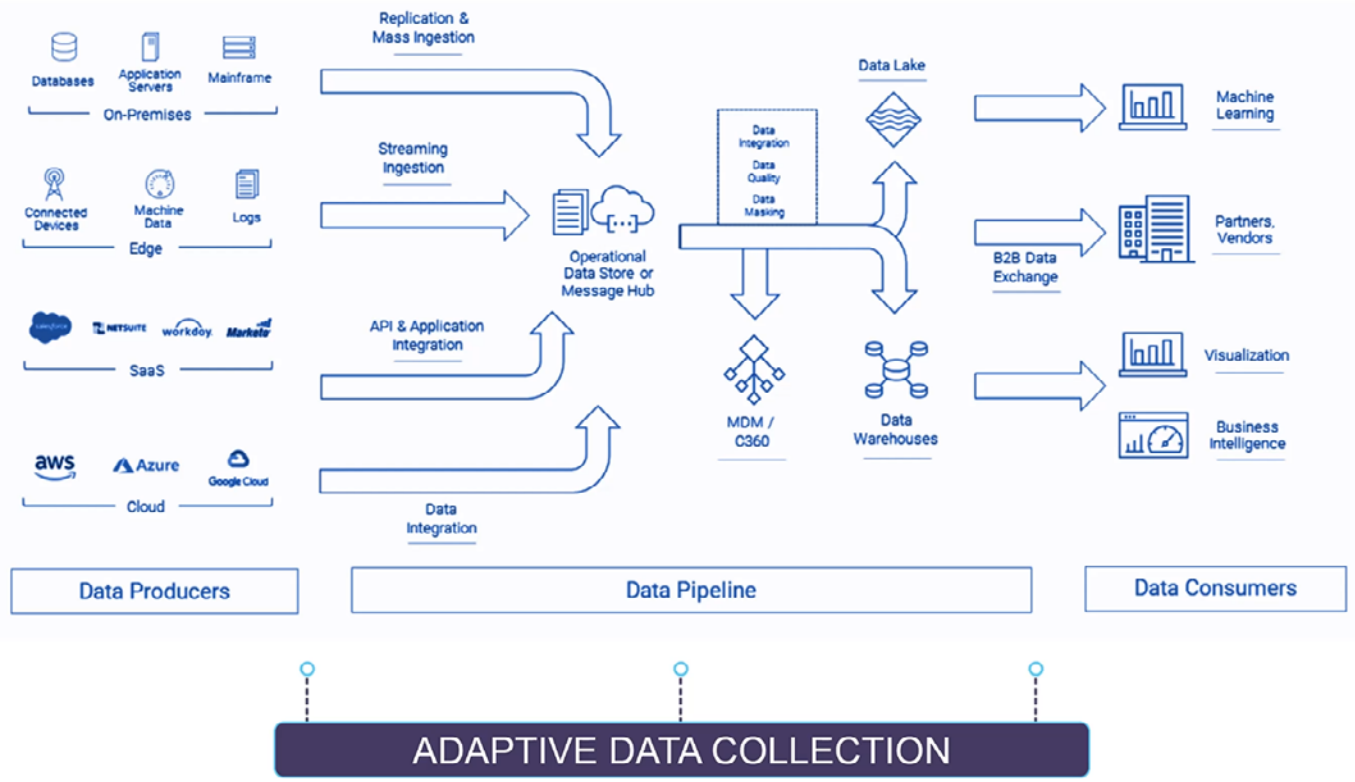
Unravel's features are grouped into toolsets for each of three primary personas: the developer, the architect/operations person, and the executive.

The Developer

The developer is an analyst, data scientist, or engineer who uses Unravel to write performance-optimized applications and data pipelines. The platform helps them debug problems and identify and reduce bottlenecks. The developer initially interacts with Unravel from their regular coding environment, sending code to Unravel by designating it in their configuration parameters when they submit the job. Unravel then imports the code and gathers execution information to provide insights back to the developer.

This information covers everything in the data ecosystem, from users and infrastructure to datasets, metadata, and environments. Unravel is agentless and uses APIs to gather information from a wide spectrum of third-party tools. It also uses a unique "sensor" technology, consisting of a small file that piggybacks on the code to gather information from the physical machine executing the application. This architecture makes Unravel lighter-weight than agent-based platforms, while still allowing customers to deploy Unravel on-premises, in a virtual private cloud, in the public cloud, or as a SaaS instance. Using its sensors and APIs, Unravel can effectively sit on top of the data pipeline while providing visibility across the entire stack. (See figure 2.)

Figure 2. Unravel's Data Collection Architecture



Because of its multi-pronged approach to information gathering, Unravel is able to provide a granular level of detail about how the code runs in the environment. It provides the user with a dashboard that shows the resource consumption of the code in terms of compute and memory demands down to the container level, the run times of each stage in the code, and a wide variety of metrics from the machines themselves. During the testing phase of development, users can directly compare these statistics to see how the same code runs in different environments.

In addition, Unravel uses AI to flag inefficiencies in the code and identify any errors in failed jobs. It suggests potential fixes, and through an "autotune" feature can actually implement those suggestions in the code. These features help developers, especially junior ones, learn best practices, while reducing the demands of applications on the compute environment.

The Architect

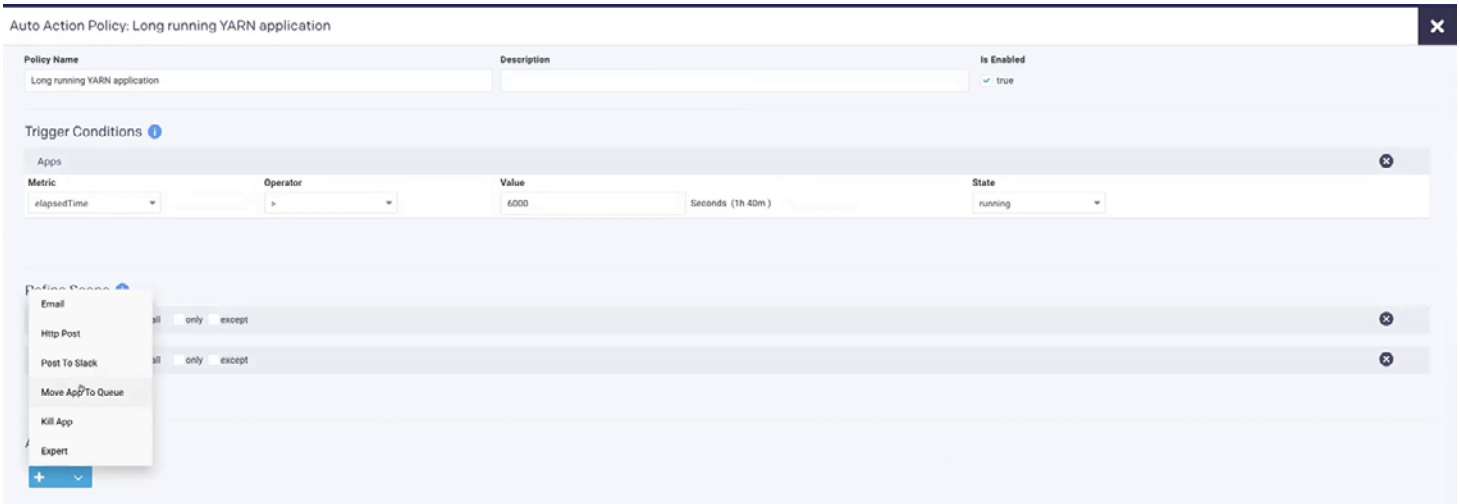
The architect or operations persona relies on the information Unravel gathers to appropriately purchase and allocate compute and memory resources and evaluate their overall platform architecture. They are concerned about the collective efficiency of the data team. This user could be a true data architect or, increasingly, a DataOps engineer who is responsible for managing the entire data ecosystem.

The architect uses Unravel to determine how to optimize resource allocation. For organizations that utilize queues, Unravel provides the number of jobs and compute and memory demands by queue. This dashboard shows which queues are maxed out and which have excess resources, so users can buy the right amount of cloud capacity for their applications, upscaling and downscaling capacity to fit currently running jobs. Unravel provides a heat map that shows cluster workloads by day and time. This visual helps users identify the times with the most jobs, so they can optimize the execution schedule. The third observational tool Unravel provides this persona is a system monitoring dashboard. This interface shows system-wide operational metrics from which users can drill down to identify the specific applications driving spikes in resource demand.

This information equips architects to identify rogue users who take up an unusual amount of resources and to select the right tools and environments for their team. For organizations moving to the cloud, Unravel even provides specific reports to help transition workloads. It analyzes on-premises workloads to understand the tables they touch and their resource usage, and then recommends the needed instances and services for each of the major public cloud platforms with estimated costs.

As at the developer level, at the architect level Unravel goes beyond observation. Through "auto-actions," the user can automate responses to certain scenarios. For example, if an application takes longer than a set amount of time to run, Unravel can automatically message the architect via Slack or email and/or kill the job. In fact, Unravel can provide a link within a Slack channel that allows the user to identify the precise source of the flag and start a conversation with the relevant developer. Users can build the rules for these auto-actions in two ways: either within a wizard, by selecting logical operators from dropdown menus, or by writing more complex rules as JSON files. (See figure 3.)

Figure 3. Auto-action Rule-Writing Wizard



The Executive

The final suite of tools is intended for a C-level executive who wants to understand how the company's data operations fit into larger corporate key performance indicators (KPIs). Using a "chargeback" dashboard, the executive can select a time range and a group (department, project, etc.) and then see compute and memory resource consumption for all of the group's applications. Users can also assign a price to memory and compute hours, allowing the executive to use Unravel as a kind of "phone bill" that shows the origin of cloud charges. This view can also double as a budget tracker, so users can see the return on investment (ROI) of data projects.

Pricing

Unravel has two pricing structures:

- **Pay-as-you-go.** The first option is a software-as-a-service (SaaS) offering. Within that paradigm, the price ranges from $0.10 to $0.30 per hour.
- **Pay-up-front.** The other option is a one to three-year subscription. This subscription includes all features of Unravel for an unlimited volume of data, users, and applications. The exact price of the subscription is determined by usage (in hours per year), but customers get discounts of up to 30% by volume. For example, 100,000 hours would be discounted 11%, 250,000 hours by 15%, and so on.

While the first model means customers have no upfront cost and only need to pay for what they use, the second rewards customers who are ready for a longer-term commitment and have a better understanding of their usage needs. Because, in both scenarios, usage drives the price, the actual price tag of Unravel varies significantly from client to client. That said, according to company officials, a typical customer on Databricks, AWS, Microsoft Azure, or Google Cloud Platform starts at about $25,000 per year.

Recommendation

Unravel separates itself from other DataOps platforms through its unique focus on monitoring and managing the performance of data-driven applications to optimize resource use. By providing insights into efficiency, from the level of KPIs down to lines of code, Unravel helps organizations understand and improve the ROI of their data initiatives. At the same time, the integration of AI to automate efficiency improvements saves developers valuable time in the DataOps lifecycle.

The Unravel Data Operations Platform is ideal for customers that want to:

- Optimize the use of data resources to reduce overhead costs especially when migrating workloads to AWS, Microsoft Azure, or GCP.
- Manage pipeline deployment efficiency.
- Track the ROI of their investments in data initiatives.
- Improve the quality of developers' code through automation.

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!**

Eckerson Group

RESEARCH CONSULTING ADVISORY

GET • MORE • VALUE • FROM • YOUR • DATA

About the Reprint Sponsor

Unravel Data is the leader in AIOps, AI-powered operations enablement, simplifying data operations for modern clouds and on-premises. Unravel powers monitoring, management, and AI-powered recommendations. Unravel also enables move-to-cloud, with comprehensive technology assessment and cost estimation.

Technology support on-premises includes Cloudera, Hortonworks, Spark, and Kafka on-premises. In the cloud, add Databricks, infrastructure as a service (IaaS), and AWS EMR, Azure HDInsight, Google Dataproc, and Google BigQuery.Partners, including AWS and Microsoft, are building Unravel Data into their processes to empower their users. And major customers such as Adobe, Deutsche Bank, Intel, and Wells Fargo are adopting Unravel across their on-premises and cloud infrastructures.