StreamSets

WHITE PAPER

Modern Data Integration for DataOps

Bringing Speed, Flexibility, Resiliency, and Reliability to Analytics

Table of Contents

Overview

Modern companies need to be able to pivot on a dime in today's world. Data is the one thing that can tell us what happened to our sales and marketing channels, operations, and online security. It lets us make informed, real-time decisions in a world of constant change. Data feeds the data science and machine learning that lets us become better at predicting future outcomes.

In order to reap the benefits of these analytic insights you first need to design data delivery to a growing number of analytics teams, all with different data requirements. Adding to the complexity, enterprise and cloud data ecosystems are changing all the time, breaking data pipelines and mis-guiding insights.

DataOps is a set of practices and technologies that operationalizes data management and integration to ensure resiliency and agility despite constant change. It combines the DevOps principles of continuous delivery with the ability to harness **data drift** (unexpected and undocumented changes to data).

DataOps is not just a technology platform, it requires a change in mindset and often a change in the way you put together teams and processes. You have to think differently about the data management and integration value chain. If you can make the mindset shift, DataOps delivers the continuous data needed to drive modern analytics and digital transformation.

Core to delivering on DataOps practice is a **modern data integration platform** that provides users **speed, flexibility, resilience,** and **reliability**. DataOps systems embrace change by allowing their users to easily adopt and understand complex new platforms in order to deliver the business functionality they need to remain competitive.

What Is DataOps?

In today's challenging environment, delivering data to make decisions as fast as possible is more critical than ever. The only way to keep pace with all the changes happening in the broader world, in your organization, and across all the data you need to make decisions is DataOps. Modern analytics requires the free movement of data to all corners of the business to ensure you can make the best decisions possible on a moment's notice, using the freshest information. DataOps can help reduce the cycle time of data analytics in close alignment with business objectives.

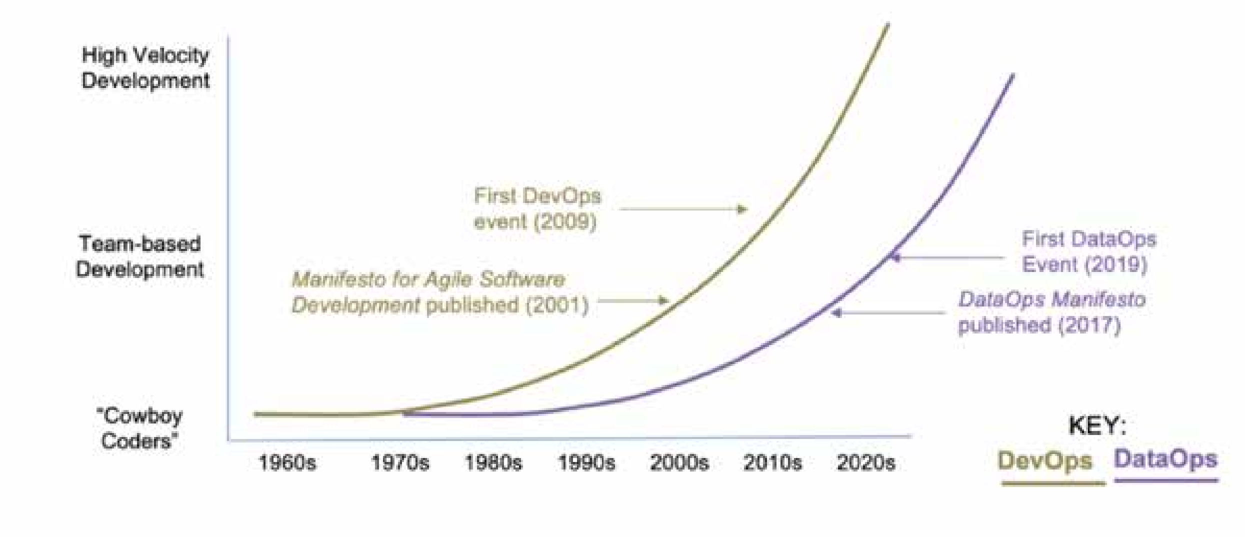
> *DataOps is a set of practices and technologies that operationalizes data management and integration to ensure resiliency and agility despite constant change. It combines the DevOps principles of continuous delivery with the ability to harness data drift (unexpected and undocumented changes to data).*

DataOps is the new way of thinking about working with data, it provides practitioners (architects, developers) an ability to onboard and scale data projects quickly while giving operators and leaders visibility and confidence that the underlying engines are working well. It is a fundamental mindshift that requires changes in people, processes, and supporting technologies. DataOps is not only the way you change your current data strategy, but also how you plan for unforeseen changes in the strategy. It draws inspiration from the philosophy of agility that DevOps brought to the software development lifecycle, and adapts it to the unique challenges of working with modern data.

DataOps is not just DevOps for data, however. Where the enterprise application workbench is fairly static across the development and deployment lifecycle, DataOps is increasingly complex due to the constant rate of change called **Data Drift**. DataOps is complementary to and sits alongside philosophies such as MLOps (management of the full lifecycle of machine learning). It lets practitioners feed data into analytics and machine learning systems continuously and reliably.

DataOps has been gaining traction since it entered the Gartner Hype Cycle in 2018. Eckerson Group tracked the emergence of DataOps in the **below diagram**.The industry's first Guide to DataOps, tips and trends in DataOps, and the first ever DataOps Summit were all unveiled in 2019. Also in late 2019 John Schmidt and Kirit Basu released DataOps: The First Authoritative Edition. While still a growing practice, the focus and development of best practices in the last year has shown notable progress. As many companies begin to formalize their DataOps team and processes, new roles and standardized architectures will further enforce the enterprise adoption of DataOps.

Figure 1. The Trajectory of DevOps and DataOps



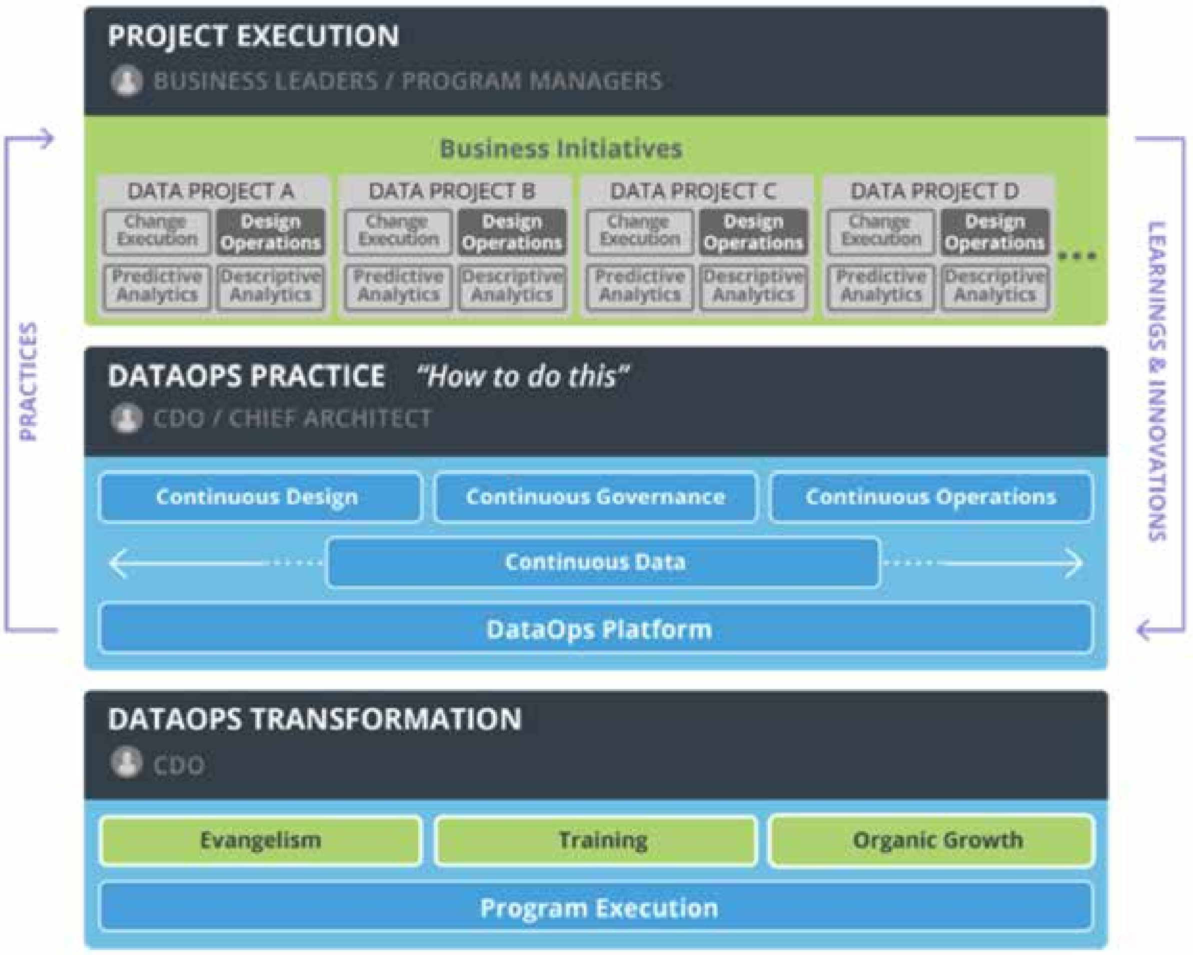
Data source: Eckerson Group

4 Ways You Need to Think Different: The DataOps Mindset

DataOps can provide many optimizations and accelerate value to a variety of analytics initiatives. However, legacy processes and platforms do not provide the speed, resilience, and flexibility needed to perform modern data integration, which can help you better leverage streaming and unstructured data sources. When scoping how to implement and measure DataOps, you will have to fundamentally change the way you think about these four things.

- **Self Service:** What aspects of the data lifecycle can be automated? Can we give end-users more power to find, understand, ingest, wrangle, and transform data without sacrificing security and governance? The previous world where business users filed request tickets and waited weeks for IT to implement their goals is no longer acceptable. Self service data access should be implemented wherever safe and possible. This gives your company the speed and flexibility to deliver parallel analytics activities.
- **Visibility and Scale:** This means visibility that spans systems and data sets. Visibility into your pipelines, your transformations, and what other users are doing with data. DataOps helps you discover design patterns and implement optimizations based on usage patterns. This holistic visibility ensures that data systems scale to meet current and future needs.
- **Real-time Governance:** Gone are the days of implementing governance with good intentions. Costly regulations, such as GDPR and CCPA, and the reality that data must pass through multiple systems with varying degrees of visibility and governance. These gaps in governance introduce new areas of risk. DataOps requires a continuous data governance approach where policies are enforced at runtime as well as in trusted analytics zones, including while data is in motion, rather than through onerous, lengthy governance processes that impede business agility.
- **Designing for Change:** Things change all the time: new data sets and schemas are introduced, existing ones are changed or used in new and unexpected ways, ML models drift, data quality levels fluctuate, ownership and stewardship responsibilities change. The goal of DataOps is to detect and manage changes so things do not break. That requires a high degree of monitoring, sophisticated change detection, and automated change handling. DataOps architectures and processes are resilient to unavoidable changes and provide reliability as you scale.

Executing data projects is not only a matter of ensuring data delivery happens successfully, DataOps is a philosophy that challenges the practitioner to constantly try and achieve a state of continuous everything with all aspects of data.



Continuous Everything

Core to the mindset shift is a pivot to thinking about data not as a static resource but a continuous process. DataOps includes four continuous principles that are considered in the design, deployment, and operation of modern data systems.

- **Continuous Design:** The continuous design function enables delivering data solutions on an ongoing basis rather than as discrete project events.
- **Continuous Operations:** DataOps encourages a holistic view where operators are able to see a living map of all data working together to serve the data needs of higher order business functions.
- **Continuous Governance:** Continuous governance is responsible for establishing an information governance framework, a methodology, and standards for enterprise information management.
- **Continuous Data:** Continuous data is responsible for the publication of data in a united hub maintaining data services, service levels and performance for both externally sourced and internally generated information and their use by users or application systems.

By embedding these principles, DataOps makes it possible to deliver the continuous data needed to drive modern analytics and digital transformation.

Continuous Data

Modern enterprise systems have to work with a very wide variety of data: IoT sensors emit readings, web applications produce events and messages, customers or partners send binary files and every flavor of database systems (relational, nosql, graph, time series, etc.) contain different types of data. Real-time insights or decisions generated by operational and analytical systems are fueled by events and data from these varied systems along with historical and contextual data.

Continuous data converges the paradigms of stream and batch processing. Streaming data is imperative for real-time predictive and preventive analytics. Batch processes are required to enrich or provide training data for these forward looking analytical systems, as well as drive business critical descriptive analytics.

Continuous Design

Modern data architectures are often highly distributed and complex. Applications and use cases built up with a number of these complex systems have a high degree of interconnectedness.

Continuous design is a paradigm where data integration is always performed in context. The 'big picture' is kept in view and is often the starting point of the design process. Data architecture diagrams are not just images on a slide deck, but the living map of what is real in the system. When change happens to the underlying systems, e.g. data sources or destinations are added or removed, it is immediately reflected in that living map. A dataflow pipeline designed to move data from a source to destination shows up in that map and is globally visible to anyone interacting with the environment. When the next project wants to access the same data, developers look to the living map or topology to reuse existing pipelines.

Continuous Operations

Large interconnected dataflows are often monitored in isolation; an operator monitoring a pipeline in one area of the application will not know how it relates to another area of the app. The cascading effects of one failure are impossible to understand when the operators' only view is a tabulated list of individual pipelines.

DataOps encourages a holistic view of the entire architecture. Operators who are able to see a living map of all the pipelines working together to serve the data needs of the higher order business function are able to better manage the system as a whole. Yet, they still need to be able to drill down into problem spots as required.

Continuous Governance

Modern architectures are often hybrid. Data originates, is processed or stored in any number of systems - on edge, within the data center or in the cloud. Data often makes several hops, getting processed on the fly or within ephemeral compute systems as it travels. Traditional data lineage solutions were built for point-to-point dataflows, and are not able to capture or interpret end-to-end lineage. A DataOps practice would use systems that are designed for these architectures, and expect the data itself to provide fine grained metadata and end-to-end lineage about itself.

DataOps calls for a framework of constant vigilance and protection. Continuous governance is about defining a security policy and automatically enforcing it whenever data flows through the enterprise.

The promise of continuous everything as a core design and implementation principle gives us the latitude to use data in the way that different teams need to truly enable a self-service and DataOps culture. Once this continuous solution is built, automation can be used to ensure that continuous data is something that the business can feel confident in assuring service level agreements to their stakeholders.

So if continuous everything is so ideal, why are many companies still so far away from realizing this state? The truth is that there are hidden complexities to executing on continuous design and development that are related to limitations to traditional approaches to data management and integration.

The Hidden Complexity of Data Operations

As companies target and acquire new data sources and aim to deliver new form factors of data analytics, a fair amount of creative capital is spent designing these new patterns and solutions. Not surprisingly, rich tools have arisen that give visual control over analytics and more recently data movement. Data engineers now have highly intuitive tools to control the integration of data across their business. A savvy engineer can boost their internal brand immensely by bringing a forward-thinking, new capability into reality. However, how can they find the time when they are so often mired in the task of keeping existing ideas healthy, modern, and in production?

The truth is that they often spend a great deal of time handling the operations that keep data pipelines running and functioning to meet the requirements of downstream projects. It's understood widely among data engineers, but not broadly acknowledged, as an activity that will be rewarded with praise and accolades. But when things go wrong the pain that data engineers feel is real. Whether it's a data science model that an analyst convinced an engineer to support or the daily dashboard supporting the sales team, when these capabilities break, friendship and an impeccable record of new ideas will only win you so many graces.

However, data operations doesn't have to be a zero-sum game. You can build a system that provides resilience and flexibility even at scale. Thinking smart about how you scale and react to the operational needs of your data pipelines and data processing can be cumbersome at first, but will pay dividends as workloads and data projects amass.

> *"In a static data world, upfront developer productivity matters more than operations. In a continuous data world, operations are everything."*
>
> - *Kirit Basu, author of* DataOps: The Authoritative Edition

So how do we build operations for a continuous data world? The following components are key to delivering DataOps functionality.

Automation

As companies scale and accelerate their data practice, automation and integration with automation tooling becomes paramount to move with speed. No aspect of self-service can be delivered without some level of automation. When engineers are able to work on automation tasks the impact can be amplified across multiple workloads. In the DataOps ecosystem, users will want to automate anything they can, while remaining reliable. Proprietary systems often offer poor extensibility making them complicated to automate and integrate with automation tools. This is why DataOps is often focused on open solutions or platforms that provide API extensibility and programmable integration with infrastructure and computing platforms.

Visibility

Think of a single data pipeline as a tab on your internet browser. At a small scale, toggling between tabs on your browser is relatively manageable (though not ideal). Depending on the size of your organization you likely have multiple tools (legacy and modern) to build data pipelines. These tools will all have a different degree of control and granularity to understand both the progress and health of your data pipelines. Data teams often spend a good deal of their time managing to the limitations of the tools.

But what about when you have 30, 100, 1K pipelines? Diving through 100 internet browser tabs with differing degrees of helpfulness will likely only produce delays and cause an increasing amount of pain as connections grow. With DataOps, the goal is to have a comprehensive and living map of all of your data movement and data processing jobs. When pipelines or stages of a pipeline break down, the errors aggregate into a single point of visual remediation. That way data engineers understand the issue and react in a manner that doesn't harm downstream projects, and take a blow to the engineer's brand. This map should be as useful, and as reactive, for today's data workloads, as it is for handling tomorrow's workloads.

Monitoring

A big component of resiliency and reliability is active monitoring. Many tools aim to monitor the operations inside their product portfolio, but DataOps demands a level of monitoring that persists outside of a single system or workload. Continuous monitoring requires that systems share metadata and operational information so users can see a broader scope of challenges. These monitoring capabilities also help companies define, refine, and deliver on downstream data SLA's. This ensures that analytics can be delivered with confidence and the company can evolve the art of self-service, which further removes the data engineer from risk. Monitoring should not only tackle alerting a team when something breaks, but, in a DataOps scenario, it should actively monitor for the precursors to potential problems. Monitoring in DataOps should also be comprehensive, not allowing for data systems and silos to become operational black holes.

Managing an Evolving Landscape

What is the cost of having a rigid, brittle data system? Will it cost you your competitive stance? Today's data landscape is evolving at a feverish pace and the toll this takes on data professionals is unforgiving. Your data strategy must not only be competitive with today's requirements, but also be future leaning to consider your company's transformation over the next five to ten years.

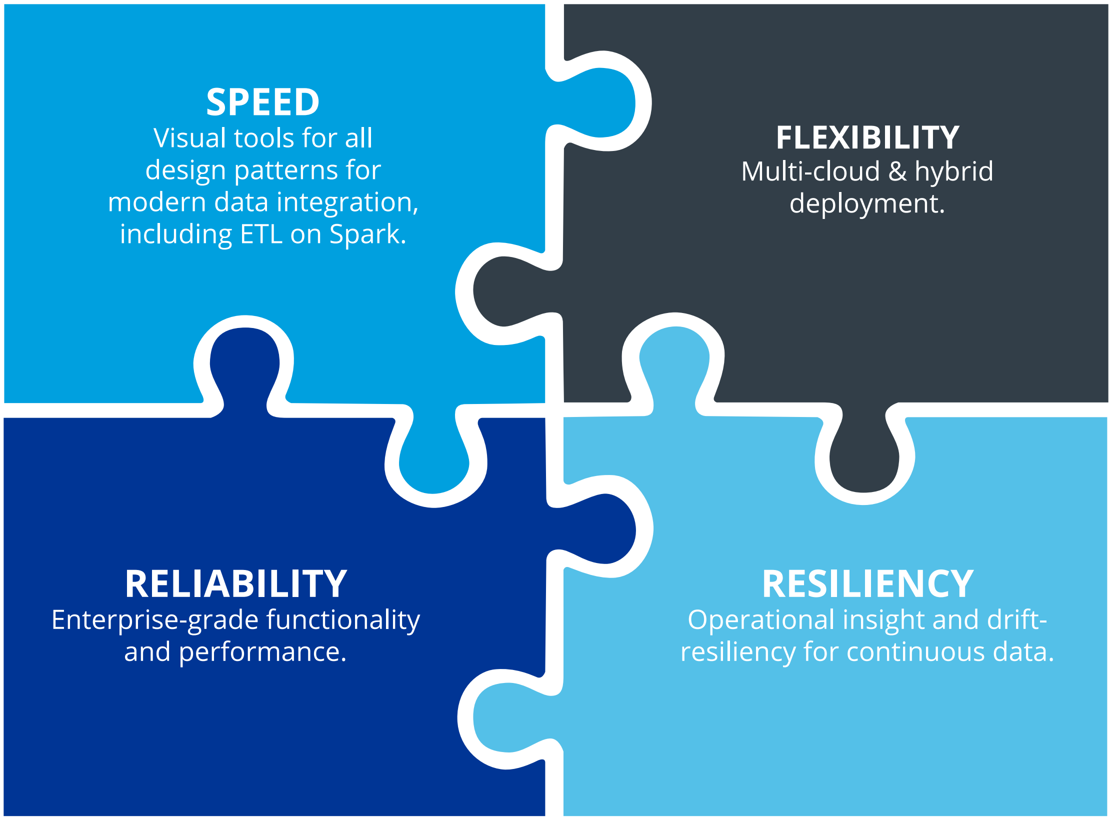
Managing and evaluating this fast-moving landscape requires flexibility, architected upon tools and platforms that can abstract away from reliance on a single data platform or analytics solution. Logic for designing pipelines should be transferable, no matter the source and destination, allowing for change based on the business requirements vs. managing to the limitations of the system. In DataOps, change is a given. DataOps systems embrace change by allowing their users to easily adopt and understand complex new platforms in order to deliver the business functionality they need to remain competitive. DataOps embraces the cloud and helps companies build hybrid cloud solutions that may someday live natively in the cloud.

Architecting For Change

Data is no longer a static resource. Data values mutate over time, semantics drift, and architectures evolve. A process or platform that assumes that data will act the same way every time it's moved or processed is destined to fail. Shockingly, the best possible scenario is that a data pipeline or process will break, the effect of that small change can wreak havoc on downstream systems. However, sometimes the pipeline doesn't break and ultimately just corrupts the downstream data and everyone working on that data is blissfully unaware. For years, data professionals managed around data drift aiming to mitigate the impact to downstream teams. DataOps embraces data drift and assumes that data, semantics, and infrastructure will change. DataOps pipelines and processes are loosely coupled and flexible to schema variations. DataOps systems work for delivering on today's use cases while also being resilient to unforeseen changes, giving companies the agility to change strategies as their industry and customer demands dictate.

Overcoming the mounting complexity of data operations is highly dependent on choosing tools for integration and ingestion that provide the same degree of speed, flexibility, reliability, and resilience that is essential in making this key mindset shift.

A Modern Data Integration Platform for DataOps



SPEED
Visual tools for all design patterns for modern data integration, including ETL on Spark.

FLEXIBILITY
Multi-cloud & hybrid deployment.

RELIABILITY
Enterprise-grade functionality and performance.

RESILIENCY
Operational insight and drift-resiliency for continuous data.

DataOps is a practice that involves people, processes, and technology, and central to delivering on the notions of continuous everything and building a foundation for your DataOps practice is a modern data integration platform. The emphasis on modern signifies a wide range of data formats and interoperability between enterprise systems is required. The system should handle both streaming and batch semantics and run optimally where your data lives.

4 Evaluation Considerations

When evaluating a modern data platform you should consider the strengths and weaknesses of a solution based on its ability to deliver on speed, flexibility, reliability, and resilience. This section identifies the functionality recommended to meet these requirements.

Speed

Speed can be measured both in finite terms, as well as used as a term to describe high level acceleration. For instance, to a developer, speed may mean easy to use tools for designing and operating pipelines of all patterns, including ETL on Spark, streaming ingestion and CDC. These capabilities can help developers design pipelines in minutes and operate continuously without much difficulty.

For data team managers, speed might be the ability to use a single visual tool for all design patterns (streaming, CDC, batch) with no coding required--even on Spark. This would help teams reuse artifacts and apply skills to a wide variety of use cases.

For IT, speed may be a term used to describe the ability to remediate any problems. You might appreciate a single tool for operating data pipelines of all types, across all platforms that would provide simplified operations management and reduce operational overhead and cycle time for new projects. And for a CDO or executive data leader, speed may mean the ability to roll out Apache Spark development for the masses, accelerating AI and machine learning projects for the business.

Flexibility

Speed and flexibility are often multiplying forces that can provide big business impact when applied across teams and roles. When making data pipelines and scheduling large batch and ETL processes having Apache Spark processing available anywhere (on-premises, public cloud or virtual private cloud) gives designers the ability to select the platform with the best performance.

When tools have pre-built support for cloud destinations and run natively on Apache Spark managed services, it means that teams get faster delivery of their core use cases. An additional added benefit is when you have portability and reusability across these platforms, both on-premises and cloud. This can give the operations team the flexibility to change platforms quickly and avoid vendor lock-in. For the business, you can feel confident in getting hybrid and multi-cloud support and keep options open to shift with the momentum of your industry and requirements. The ultimate flexibility is achieved when unforeseen events are not able to hamper the company's core directive.

Resiliency

Resilience in the face of change is the ultimate goal of DataOps. However resilience has a lot of hidden complexity. The details involved in building large, interconnected enterprise systems that are truly resilient requires a great deal of thought and coordinated effort.

In order for data pipelines to be resilient, they must be fully instrumented so that users can understand what is happening at every stage of the pipeline. This level of visibility and instrumentation leads to faster debugging. If data drift is always present, then drift resilient, fully-instrumented pipelines should lead to decreased risk and easier troubleshooting. This can prevent outages for analytics systems.

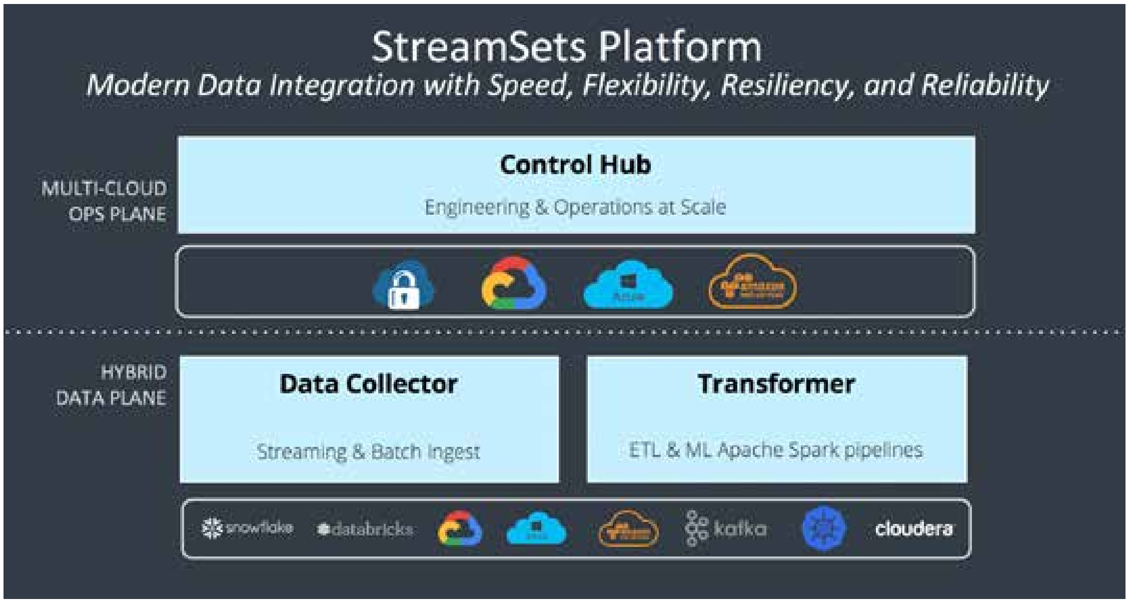
For managing pipelines continuously, an operational view of real-time data flows gives the IT team the ability to see the full picture and drill down where needed. If, at a company level, you can have a single view of data flows, you can ensure visibility and control, further decreasing data silos and redundant operations. True resilience requires intent-driven design. That means even though sources and destinations change, the intent of the pipeline is retained.

Reliability

Delivery of enterprise-grade functionality and performance in the face of constant change is almost insurmountable. Too often, companies are forced to make tradeoffs between leveraging the fastest solution to a problem (i.e. cloud or a managed service) versus the solution that the company is confident in supporting. In order to mitigate the risk of new platforms, data engineers must have advanced enterprise-grade features. This helps reduce the time to configure, tune and maintain data pipelines. Companies should seek out DataOps vendors that have proven enterprise success. This makes them a safe bet to standardize on as you build and scale your systems. Reliability means that you can say “yes” at the executive level to new data requests, feeling confident that developers, data teams, and IT teams can execute the vision.

The StreamSets Platform

StreamSets has built the first Modern Data Integration Platform for DataOps. The platform is built to fulfil the capabilities of **continuous everything** and allows companies to **go fast and be confident**. The platform gives data professionals **speed, flexibility, resiliency, and reliability** at the single pipeline level and over the entire enterprise data fabric. As pipelines scale, they are fully instrumented giving users a high degree of visibility and allowing users to share pipeline components. They actively detect for **Data Drift** and their intent-driven design means that when things change, pipelines don't break.



StreamSets Platform

Modern Data Integration with Speed, Flexibility, Resiliency, and Reliability

MULTI-CLOUD OPS PLANE

Control Hub

Engineering & Operations at Scale

HYBRID DATA PLANE

Data Collector

Streaming & Batch Ingest

Transformer

ETL & ML Apache Spark pipelines

At the highest level, the StreamSets DataOps Platform is a modern data integration platform, combining high-performance execution engines and visual, full-lifecycle tools for designing, operating, managing and optimizing data pipelines across your enterprise. With end-to-end instrumentation and visibility, the platform provides real-time operational insight across all your pipelines, no matter where they are, on premises or in the cloud.

The platform's drift-resilient pipelines reduce the risk of outages or data loss due to data drift. And the solution is architected to be platform-agnostic, enabling easy portability of pipelines across platforms, whether on-premises or in the cloud, ensuring flexibility to meet changing needs. The DataOps Platform is a key technology foundation for your DataOps practice, letting you rapidly deliver data on a continuous basis to the business, in a world of ceaseless change.

The platform consists of two powerful execution engines and a management hub.

StreamSets Data Collector

StreamSets Data Collector is an easy-to-use modern execution engine for fast data ingestion and light transformations. Data Collector's easy-to-use visual tools let you design, deploy and operate streaming, change data capture (CDC) and batch data pipelines without hand coding. The full variety of data sources such as Kafka, S3, Snowflake, Databricks, JDBC, Hive, Salesforce, Oracle and more are available out-of-the-box. Fully instrumented "smart" data pipelines let you monitor data in flight, and are designed to handle data drift with built-in detection and handling. Data Collector pipelines are designed to be platform agnostic, so you can adapt as needed and avoid vendor lock-in.

Data Collector gives users the ability to quickly load data into enterprise systems with pre-built sources and destinations and an intuitive graphical interface that lets them design data pipelines in minutes. Pipelines can be ported across multiple environments on premises or in the cloud without hefty redesign. Pipelines are by default drift resilient, fully-instrumented "smart" pipelines helping users reduce risk with easier troubleshooting and fewer outages. Data Collector balances enterprise functionality and performance in a single tool for all types of ingestion patterns.

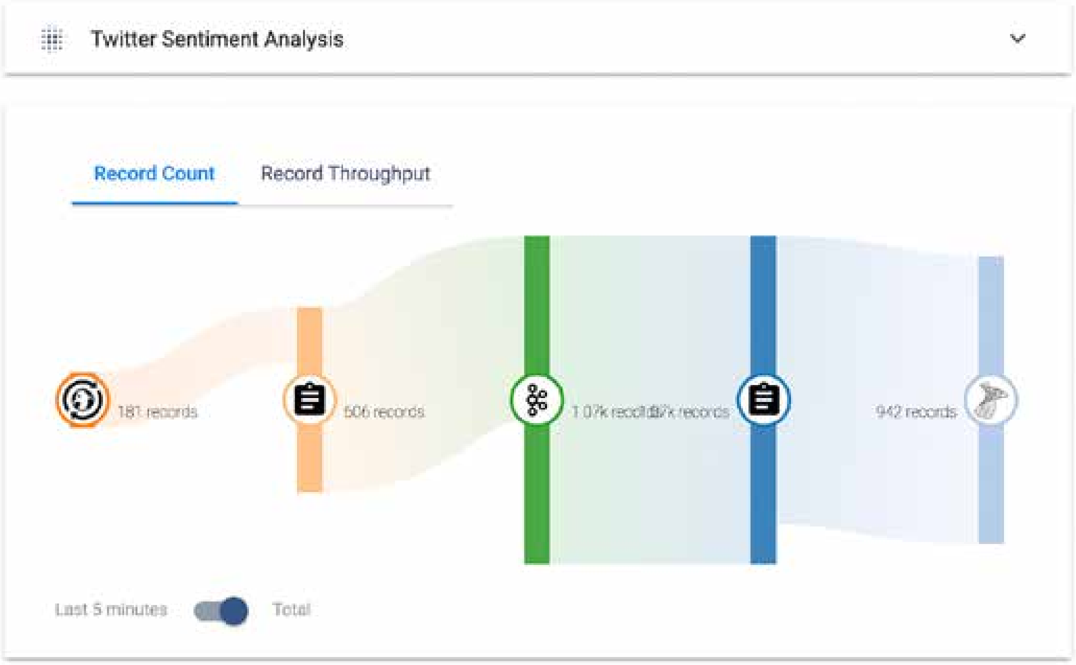
StreamSets Transformer

StreamSets Transformer is a modern transformation engine designed for any developer or data engineer to build data transformations that execute on Apache Spark. Using a simple, drag-and-drop UI, users can create pipelines for performing ETL, stream processing, and machine learning operations. It allows everyone, not only the savvy Spark developer, to fully utilize the power of Apache Spark without having to code in Scala or PySpark. Transformer pipelines are instrumented to provide unparalleled visibility into the execution of Spark applications with built-in previews and easy trouble-shooting. And Transformer is designed to run on all the major Spark distributions to ensure you have the flexibility to run on your platform of choice, or switch platforms when your needs change.

Transformer helps bring Apache Spark to the hands of every type of data engineer. Using Transformer, companies can accelerate their adoption of Apache Spark without heavy skills investments which require zero ramp up for ETL and machine learning projects.

Transformer runs where your data lives and can natively execute on platforms like Hadoop YARN, EMR, HDInsight, Databricks, and in containerized Spark environments such as Microsoft SQL Server 2019 Big Data Cluster and Kubernetes Cluster. The engine provides deep visibility into Spark execution allowing users to troubleshoot at the pipeline level and at each stage in the pipeline progress. Transformer gives users the enterprise features and agility they get from legacy ETL tools, while revealing the full power and opportunity of Apache Spark.

StreamSets Data Collector and Transformer pipelines can be deployed on premises, across public clouds, and on cloud managed services. All engine instances can be deployed, managed, and viewed directly in the central hub for system wide control. This hub organizes pipelines into graphical maps called topologies.



Control Hub

StreamSets Control Hub is a single hub for designing, deploying, monitoring, managing and optimizing all your data pipelines and data processing jobs. The central nervous system of the DataOps Platform, Control Hub lets your team collaborate to manage data pipelines and jobs running on Data Collector and Transformer, enables pipeline re-use, and gives you a real-time, end-to-end view of all data flows across your enterprise.

Control Hub also simplifies and centralizes the management of the StreamSets Data Collector and Transformer engines themselves to optimize your overall StreamSets environment. And finally, Control Hub's hybrid/multi-cloud architecture provides centralized monitoring and management across on-premises and cloud data sources and platforms, so you can add or change data sources or data platforms without losing visibility or control.

Control Hub provides a unified console for collaboration and visibility across all lifecycle stages, all design patterns, all engines. This helps speed development by agile use and reuse of skills and assets. The hub gives centralized monitoring and management across on-premises and cloud data sources and platforms, and gives you the flexibility to add or change sources or platforms without losing visibility or control. Live data maps called topologies give real-time operational insight to de-risk with broad visibility to detect and prevent issues.

StreamSets Data Collector, StreamSets Transformer, and Control Hub work in concert to deliver the requirements of DataOps to a wide variety of common enterprise use cases, including stream processing, ingestion, ETL, machine learning, and powering real-time applications.

Conclusion

DataOps is a practice that is well staged to become the blueprint of how forward-thinking companies design data systems that handle change and the complexity of evolving ecosystems. In close alignment with business objectives these practices shift to tackle the needs of today and grow to address the needs of the next platform, analytics environment, and industry trend. In building a DataOps system, you need to enable continuous design, continuous data, continuous operations and continuous governance. A modern data integration platform built for DataOps should deliver these capabilities and give your company speed, flexibility, resilience, and reliability.

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. The platform uniquely handles data drift, those frequent and unexpected changes to upstream data that break pipelines and damage data integrity. The StreamSets DataOps Platform allows for execution of any-to-any pipelines, ETL processing and machine learning with a cloud-native operations portal for the continuous automation and monitoring of complex multi-pipeline topologies.

TRY NOW

Get up and running with StreamSets in minutes. Visit us at:

www.streamsets.com