StreamSets

WHITE PAPER

The Dollars and Sense of DataOps

A Comprehensive Guide to Bottom Line Results from DataOps

Table of Contents

Data: Boundless Opportunity. Tremendous Pressure.

The world of data is moving *fast*. It makes it one of the most exciting frontiers in business in technology today, and you're right there in the middle of it (high five!).

It's your job as a data leader to get data into the hands of decision-makers. If you can, there are boundless opportunities for business impact: beat the competition, create new revenue channels, optimize operations, accelerate product development, and time to market...Is there anything data can't improve?

But delivering data is easier said than done. New applications – all with data to incorporate – are added, reconfigured, and retired regularly. Market changes requiring a shift in priorities and business operations happen faster than ever. And you needed to get everything to the cloud yesterday. The pressure is on, with the business clamoring for more and more data, faster, fresher, *now*.

Your teams, however, are stretched to a breaking point, with a seemingly interminable project backlog. New data platforms and technologies are tantalizing in their power and cost-effectiveness, but it can take months for your team to ramp up, and you can't afford to redo everything. Your data is scattered far beyond your data center and outside your control, floating in the cloud, owned by third parties, or hidden in rogue applications you're not even aware exist. Which means when:

- a column gets added to a data table,
- a new data source gets onboarded somewhere in the middle of a data supply chain without your knowledge or approval,
- or the applications producing the data are changed to meet new business requirements

and so on (and these things happen all the time — we call it data drift)...

...your data supply chain can break, causing irreparable corruption or data loss. And perhaps worst of all, you're often flying blind because it's nearly impossible to see all the data flowing through your organization with all its siloed systems and data platforms. Blind spots and control gaps = governance and compliance headaches.

> **data drift (noun):** unexpected, unannounced, and unending changes to data structure, infrastructure, and semantics

DataOps to the Rescue—But How?

The rapidly emerging practice of DataOps is heralded as the way to overcome all these challenges. But like any new discipline creating a key market shift, it can be hard to cut through all the hype. So what is it really? And is it just a trend, or can it impact the bottom line?

DataOps is the new way of thinking about working with data. It is a fundamental mindset shift that requires changes in people, processes, and supporting technologies. In this piece, we'll focus on the technology and how it can enable collaboration across your teams and streamline processes too.

> **dataops (noun):** a set of practices & technologies that operationalizes data integration to ensure resilience and agility in the face of data drift

DataOps is how you change your current data strategy and plan for unforeseen changes in the strategy. If you can make the mindset shift, DataOps delivers the continuous data needed to drive digital transformation, modern analytics, and real-time decision making across lines of business.

Continuous Data

Core to the mindset shift is a pivot to thinking about data as a continuous process rather than a static resource, as it may have been in the days of traditional business intelligence and reporting. Modern analytics – including real-time dashboards, data science, AI, machine learning and smart data applications – requires a whole different approach to integrating data. This is not just about serving up real-time data, although that is one pillar. It's about delivering the latest and freshest data continuously.

When a new data source becomes available, say a new genomic database or a new batch of clinical study results, it's about delivering that data within days or even hours to the business user, in this case, a pharmaceutical researcher developing a new cancer treatment. It's also about maintaining the high service levels needed to ensure continuous availability of that data to downstream users. Delivering continuous data is at the heart of modern analytics.

The 3 Principles of Continuous Data

DataOps consists of three continuous principles addressing the design, operation, and monitoring of modern data systems that together enable continuous data.

- **Continuous Design** lets your data team deliver data solutions on an ongoing basis rather than as discrete project events. Intent-driven design, a single user experience for all patterns and platforms, and full lifecycle automation are technical enablers in a data integration platform that shrink the delivery time by 10x to 50x, allowing for continuous design.

- **Continuous Operations** ensure that data can be delivered continuously by building pipeline and infrastructure resiliency, with high service levels to downstream users. Smart data pipelines built to handle data drift and platform-agnostic pipelines that can be easily ported to work with new data platforms are both critical for continuous operations. With continuous operations, 80% of breakages and maintenance work can be eliminated.
- **Continuous Data Observability** allows your data team to measure and monitor data pipelines and engines, easily making sense of the health of the overall data integration machine – no matter where it is running – and quickly performing root cause analysis of issues. By understanding the complex machinery that powers all data flows, you can eliminate blind spots, detect and prevent issues, understand the business impact of your data, and ensure you adhere to governance and compliance policies.

It all sounds great in theory, but does it work? Short answer: Yes!

DataOps in Action

Let's take a look at what we've seen putting DataOps in action with companies like IBM, GSK, Shell, and BT Group over the past seven years. Innovative data leaders from these companies and more have achieved impact in three broad areas by thinking differently about delivering data to the business.

Delivering 10x-50x Faster

One of the most obvious day-to-day challenges for data teams is simply keeping up with the growth in data, data sources, new systems/services/platforms, and the demand for the data.

> *As data and analytics teams become critical to supporting more diverse, complex and mission-critical business processes, many are challenged with scaling the work they do in delivering data to support a range of consumers and use cases.*
>
> *– Gartner, Introducing DataOps Into Your Data Management Organization*

Data teams that have found the right data integration platform have been able to deliver 10x faster, shifting from a project backlog to self-service, real-time delivery of new data. But every data integration vendor says their platform boosts productivity. What's different about a modern data integration platform that embraces DataOps? It's built for Continuous Design which allows for the following benefits.

Faster ramp up on new technologies

Data platforms and technologies used to have decades-long, stable lifespans. That has shrunk to years, with existing platforms making major changes to semantics and operations frequently and new groundbreaking technologies arising every year. They promise power, flexibility, and cost-effectiveness, but they're often difficult to learn, especially in the early days when a ton of coding savvy is required. These new platforms can take months to learn, and some require a level of coding sophistication that the average data engineer or developer doesn't possess.

In contrast, a modern platform with *intent-driven design* builds in the underlying details of these new platforms and technologies and abstracts them away from the data engineer. A data engineer can build a pipeline for a new platform in the same way they've built their existing pipelines, without the need for weeks of training and mounds of coding. This not only increases developer productivity it also accelerates the speed of new technology adoption.

Intent-driven design: a design approach for data integration based on the intended outcome instead of the full knowledge or understanding of the systems being integrated.

"We really wanted to get away from having to have specialized talent, we wanted to be able to take a data engineer, give them a tool, and have them move data in real time with minimal training."

– Jeff Currier, Director of Data Management and Analytics, Availity

Availity: A case study in fast ramp

For Availity, the largest health information network in the US, building a DataOps culture is core to their ability to deliver continuous data in real-time for advanced analytics and self-service data. StreamSets gives Availity's data engineers the ability to use the latest technologies without special skills, thanks to a simple and intuitive graphical interface. A self-service repository is helping to build a DataOps culture where data engineers develop and test data pipelines against best practices. Availity's Director of Data Management and Analytics, Jeff Currier, says, "Without StreamSets we're spending a lot of money on specialized skills and tools. With StreamSets we're streamlined and moving forward." Learn More

Pipeline builds in hours, not days

The more obvious way to deliver data faster is to be able to build data pipelines in less time. Data integration tools that let developers use a graphical UI to design pipelines have been around for decades. But most still require the developer to specify a lot of the

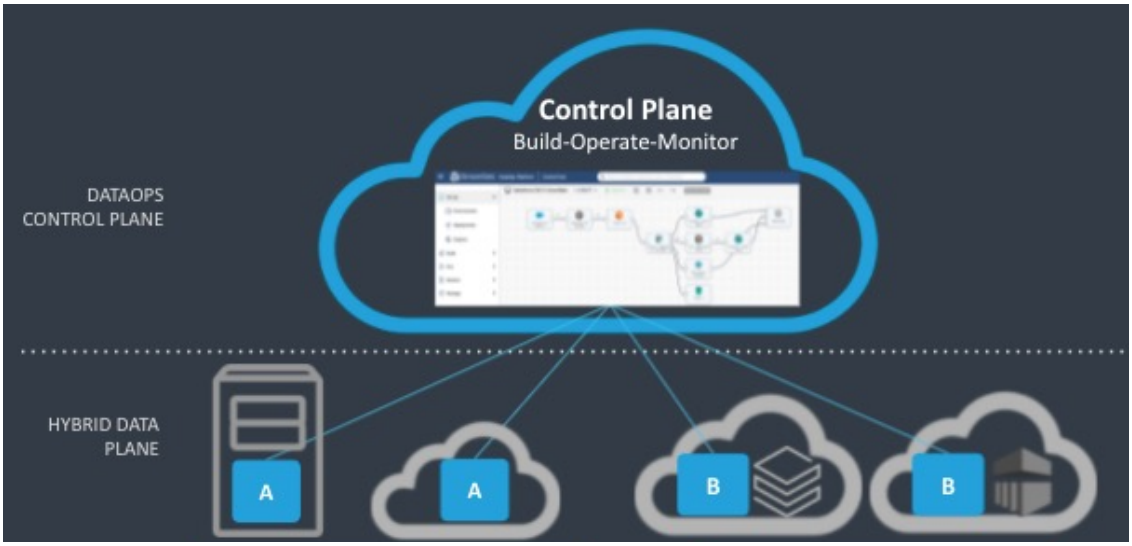
technical implementation details of the source and target systems, which takes time. An experienced ETL developer may take a day to build a pipeline with traditional tools.

Intent-driven design again changes the game. Intent-driven design abstracts away the vast majority of the underlying detail of the source and target system, and the processing technology, so the data engineer or developer has much less to specify as they design a pipeline. They need only focus on the what – i.e., the business logic – of the data flow, not all the "how" details tied to the technical minutiae.

Another game-changer for developer productivity is a single development and management experience for every data platform or pipeline pattern. Most data integration platforms cover this variety by offering a collection of different tools that are branded under one umbrella. So your data team has to swivel chair between different tools to tackle the gamut of pipelines your business demands, crushing productivity and leading to a lot of inconsistency.

A true single experience for all pipelines is key. It means engineers can design any pipeline from one place, no matter the pattern or execution technology. This is only possible with an architecture that separates the control plane (where users build, run, monitor, and manage pipelines) from the data plane (the engines which move and transform the data). This type of architecture design means that you can have different pipeline execution engines that:

- Are optimized for different types of workloads, such as streaming data ingestion vs. heavy-duty bulk data transformation vs. ELT to a data cloud
- Can be deployed in different platforms and environments, on-premises and in the cloud; as a dedicated VM, in a container or on a cluster; to optimize performance and keep the processing close to the data
- Are globally viewable and manageable by the control plane, no matter how many pipelines are running or how many engines are deployed



control plane: the system of communication for data infrastructure that consists of control messages, such as start/stop commands, and operational metadata, such as pipeline configurations/runtime notifications and interrupts. The control plane does not have direct or indirect access to any data that flows through the pipelines. The sole purpose of the control plane is to orchestrate and manage the smooth execution of the data plane.

data plane: the collection of various execution engines that execute data pipelines, thus being the agents that handle the acquisition, processing, and delivery of data in between various data platforms and applications. The data plane execution is managed by the control plane and can continue to operate even when disconnected from the control plane. The sole purpose of the data plane is to execute data pipelines, and report back the operational metrics to the control plane.

"From a productivity standpoint your ability to produce pipelines is at a much faster rate than your traditional coding tools like Informatica and Talend... The ability to have somebody who's not necessarily a top level ETL developer be able to reuse pipelines in a matter of minutes, or a matter of hours and test that and be able to deliver that is a lot more efficient than a lot more effective within our environment, especially given the rate at which we have to change and evolve."

– Joshua Picton, Sr. Information Architect, Availity

With intent-driven design and a single experience for all patterns, you can provide a 10x (or more) productivity boost just for the design phase of the data pipeline lifecycle.

- The lead data engineer at Generate Capital decreased pipeline build time from 1 day to 1 hour.
- A top U.S. financial services organization serving 13 million members was able to reduce the time to onboard new business units onto their cloud data platform by 6 months.
- The State of Ohio was able to integrate hundreds of data feeds in dozens of different formats from 88 counties literally overnight to put together its COVID-19 dashboard for the governor's daily briefings.

Delivering a COVID-19 Dashboard Overnight

The State of Ohio's data platform team supports dozens of state governmental agencies, with thousands of different data sources. Each agency acts as an isolated secure ecosystem, with skill sets, use cases, data sources, and infrastructure that vary widely between agencies.

To support this decentralized organizational structure, partner Avaap recommended providing agencies with a common data platform that was easy-to-use and could support any data source and infrastructure environment – on-premise, hybrid, or in the cloud.

The State's data platform team enabled each agency to use StreamSets to build data pipelines which meant they were ready to support the health department during the COVID-19 crisis. Using StreamSets, the team pulled together and ingested data from thousands of different data sources, in dozens of formats, into a cohesive dashboard for the governor's daily COVID press briefings – overnight. Learn More

88
Counties

1000s
Different data sources

1
Night to build the COVID dashboard

Full lifecycle automation

Because most tools vendors do, it's easy to forget that the design phase is just a small part of the overall data pipeline lifecycle. There's the testing. And debugging. And testing again. And deploying. And versioning. And redeploying. Those can take far more time than the initial design.

One key to ensuring continuous operations across the full lifecycle is automating testing, debugging, and deploying. Live data preview lets developers see the resulting data from their pipeline logic while they are still designing before they deploy, CI/CD capabilities automate testing and deployment, and version control lets you manage deployments and roll back when needed.

With full lifecycle automation:

- A top 10 European bank reduced manual testing efforts by 75%
- GSK reduced onboarding time for new data sources by 98% by fully automating the data ingestion pipeline build and deployment process

98%
Reduced time for onboarding new data sources

96%
Reduced time for new product discovery

3 years faster
Accelerated time to market for new drugs

How Self-Service Data Advances Drug Discovery

GlaxoSmithKline (GSK) is a science-led global healthcare company with a special purpose: to help people do more, feel better, and live longer.

Pharmaceutical companies spend years discovering, developing, and testing new drugs before bringing them to market. GSK set out to build a Data Center of Excellence to accelerate delivery of clean data from 1,000s of data sources to more than 10,000+ scientists involved in R&D around the world. Their goal? Accelerate time-to-market for life-changing healthcare solutions.

Using StreamSets DataOps Platform, the GSK team accomplished their mission with flying colors. Onboarding time for new data sources was reduced by 98%, new product discovery time was reduced 96% and they accelerated time to market for new drugs by almost 3 years! Learn More

Avoiding 80% of Breakages and Rework

The vast majority of business logic that drives the modern enterprise resides in integrating thousands of tiny, specialized applications across multiple platforms and multiple clouds. These integrations have become the most vulnerable points in modern business operations. Yet, traditional data integration tools ignore the simple fact that modern data semantics and structures change – frequently.

Data integration tools of yore assumed static environments. You built a pipeline from source to target, and it would just run. Nothing changed without going through a change management process. But in today's world, where changes are often unannounced and unexpected, those pipelines are built to break. Best case, the pipeline stops working, and your data engineers have to fix it. In a bad case, you lose data. In arguably the worst case, the pipeline works but delivers corrupt or incorrect data which takes weeks to detect and fix the problem with the business making the wrong decisions based on bad data in the meantime.

With data drift a constant and new data sources and platforms being introduced all the time, you need a data integration approach that *assumes* things *will* change. You need operational resilience built-in, or else your data engineers end up spending the bulk of their time on break-fix, maintenance, and rework. Your team burns out on the 2am calls and you lose credibility.

Operational resiliency for Continuous Operations is achieved at two levels.

Built-in data drift resiliency

Modern data integration platforms assume change is constant. They build and run smart data pipelines that detect and handle changes in schema, semantics, and infrastructure drift. Smart data pipelines are intent-driven, requiring minimal schema definition on the sources and destinations. They are highly decoupled, with each stage having minimal dependencies on other stages, so if one changes, you avoid the domino effect. Smart data pipelines are fully instrumented to detect changes in-flight, such as a new column added to a data stream.

Because of all this, they can automatically handle a vast majority of changes that can happen through the course of a pipeline's operation, eliminating 80% of breakages. These intent-driven, smart data pipelines can detect changes to the data in-flight and continue to apply the expressed intent to the flowing data, handling a vast majority of data drift automatically. For the small minority of changes where the pipeline cannot automatically handle a certain type of drift, it siphons off the affected records to an exception flow and raises the necessary alerts while the remaining data continues to flow uninterrupted. This minimizes the risk of unexpected and undetected breakages.

> **Smart data pipelines:** a data pipeline that abstracts away details and automates as much as possible, so it is easy to set up and operates continuously with very little intervention. Smart data pipelines create loose coupling and tight integration between their sources and destinations.

Generate Capital: A case study in pipeline resilience

Generate Capital eliminated the 10 to 20 hours of data engineering work that was triggered every time a new column was added to data (which happens all the time). This freed up the data engineer to spend the majority of their time on adding value and delivering new data rather than fixing breakages and reworking pipelines.

Cross-platform portability

Schema changes are one type of drift. A different level of drift is infrastructure drift – when the data infrastructure or underlying platform changes. For example, when a Hadoop data lake is migrated to AWS S3, a data source changes from a database table to a JSON file, or a SQL-based platform is replaced with a Spark cluster.

For the vast majority of tools that aren't designed for infrastructure drift, any pipelines that undergo an infrastructure change must be rewritten from scratch. That is because when the pipelines were originally built, the data engineer had to specify a lot of technical details specific to that platform or infrastructure, at each step of the pipeline. So replatforming triggers a mass rewrite.

Modern data integration platforms that are drift-ready and intent-driven enable you to replatform without rewrites. Existing pipelines can be ported with a simple change in the source or destination rather than a rewrite of the pipeline and its inner workings.

BT Group's Openreach: A case study in portability

BT, the top telecommunications network in the U.K., replatformed twice in the space of two years, from an on-premise Hadoop cluster to an AWS data platform and subsequently to a Google Cloud platform. They had thousands of pipelines that needed to be migrated. Because the pipelines were built to be platform-agnostic, BT was able to make simple updates in an automated manner to those pipelines to point to the new cloud data platform without rewrites.

Democratizing Data to Drive Business Results

BT Group's Openreach runs the UK's digital network. With pressure on to rollout Broadband Britain—an ambitious initiative to build full-fibre connections to 20 million premises by the mid to late 2020s—the Openreach team was running into several challenges they knew would delay their ability to deliver. A lack of data availability led to inaccurate reporting and analysis and inefficiency in resource allocation with their data integration infrastructure.

With siloed data scattered everywhere, they needed to make data sets available for analysis in a seamless and automated way. More than that, they wanted to deliver a democratized data platform that would serve everyone from technical data engineers to business end-users. The BT technology team and their partner TCS migrated their on-premises infrastructure to a cloud data lake to create "one true data source."

They used StreamSets to migrate data from on-premises to AWS – and then add Google Cloud Platform as a destination without rebuilding or duplicating pipelines. Not only did they democratize data access, but easy access got business users excited about building their own data pipelines. Bonus: it served perfectly as the basis of their DataOps practice. Learn More

250+
Platform users

16,000+
Different data sources

0
Pipeline rewrites needed to replatform 2x

Eliminating blind spots and control gaps

One of the most intractable problems with today's data, analytics, and cloud sprawl is ensuring global transparency and control. Data is scattered across thousands of sources, the business is implementing hundreds of small cloud applications, and project teams are spinning up new cloud data platforms at will.

Traditional data integration consoles are bound to a specific environment or a specific platform. It's nearly impossible to gain a global view of where all the data is flowing across your organization, much less manage it.

When you lack data observability, compliance and governance issues will inevitably arise.

- How can you enforce data security policies if you don't know where all the sensitive data is going?
- How can you adhere to SLAs when you can only see what happened yesterday?
- How can you know everything is running as expected when you have to swivel chair between a dozen different data integration consoles?

You need mission control for Continuous Observability: A single pane of glass for operating and monitoring all data pipelines, no matter where they are running: on-premises and in multiple clouds; streaming or CDC or batch; in a Linux VM, on a Hadoop cluster, on a Spark engine or in a cloud Kubernetes service; in the sales division or in finance; in Berlin, Vancouver or Bangalore. This single pane of glass delivers data observability and puts you in control.

Is this even possible in a complex global enterprise with dozens of business units, hundreds of project teams and data platforms, and tens of thousands of pipelines? Yes. If your data integration architecture separates the control plane from the data plane, you can achieve global visibility while each pipeline can be executed in the specific way and place dictated by the needs of that user and use case. The same architecture that gives data engineers a single design experience also provides you enterprise-wide manageability and visibility, across hybrid and multi-cloud architectures.

With this single pane of glass, you can reduce the number of resources required to monitor and manage your data pipelines and flows. You can observe your data as it behaves in the real world, eliminating blind spots and ensuring adherence to compliance policies and governance requirements. As discussed above, BT is scheduling and orchestrating tens of thousands of pipelines across three disparate data platforms, all from one management console. And IBM has transparent, real-time operational data for their entire global network across 1000 sites, with 20,000+ data pipelines and billions of streaming records.

1000
Sites Around the World

20,000+
Data Pipelines

1
Dashboard for Central Visibility

How Self-Service Data Supports Operational Excellence

With over 400,000 employees across the world, IBM has one of the largest corporate networks in existence. Charged with the health of the network, the CIO Network Engineering team delivers automation and tooling services that provide visibility into and reliable operations of the environment. To do this they need continuous, reliable, and transparent operational data that teams around the world can use in real-time.

Committed to DataOps, they realized their existing data pipeline solution wasn't cutting it. Hand coding made building pipelines exceedingly difficult and prevented non-data engineers from accessing data when they needed it. With 1000 sites worldwide and extremely high volume and velocity data, IBM turned to StreamSets for a solution that would truly operationalize continuous data management and integration and ensure resilience, agility, and visibility for a central team. Learn more

Continuous Data for Nonstop Business Results

No matter your ultimate goals – driving operational efficiency and excellence, reducing costs, helping new products get to market faster, or all of the above and more – the time for digital transformation is now. The market is in the midst of enterprise self-selection based on data and analytics transformation. Those who get it will survive and thrive. Those who don't will struggle with complexities and drive themselves out of the market.

For data teams, digital transformation is centered on operationalizing the delivery of data. To operationalize data flows, data engineers must be able to collaborate with different personas, reuse peer assets, support data pipelines in production and evolve quickly — and with confidence — as data platforms and business requirements rapidly change. Traditional ETL and point integration solutions cannot handle the complexity of fast-changing data pipelines, nor are they ready to scale up to meet the need for faster and better analytics.

By selecting a modern data integration tool built for continuous design, operations, and data observability, organizations can deliver continuous data for nonstop business. This translates to delivering business value from their data initiatives.

StreamSets is the only data integration platform that brings enterprise-grade capabilities for DataOps to modern hybrid and multi-cloud architectures. Only StreamSets empowers data engineers to build and run smart data pipelines that abstract away details and automate as much as possible, delivering continuous data for DataOps.

Next Steps

Take the guesswork out of modern data integration with smart data pipelines. Deliver continuous data to every part of your business with StreamSets.

Visit StreamSets.com | Get Started for Free | Request a Demo

About StreamSets

At StreamSets, our mission is to make data engineering teams wildly successful. Only StreamSets offers a platform dedicated to building the smart data pipelines needed to power DataOps across hybrid and multi-cloud architectures. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, AI/ML and smart applications. With StreamSets, data engineers spend less time fixing and more time doing.

To learn more, visit www.streamsets.com and follow us on LinkedIn.

TRY NOW

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

www.streamsets.com