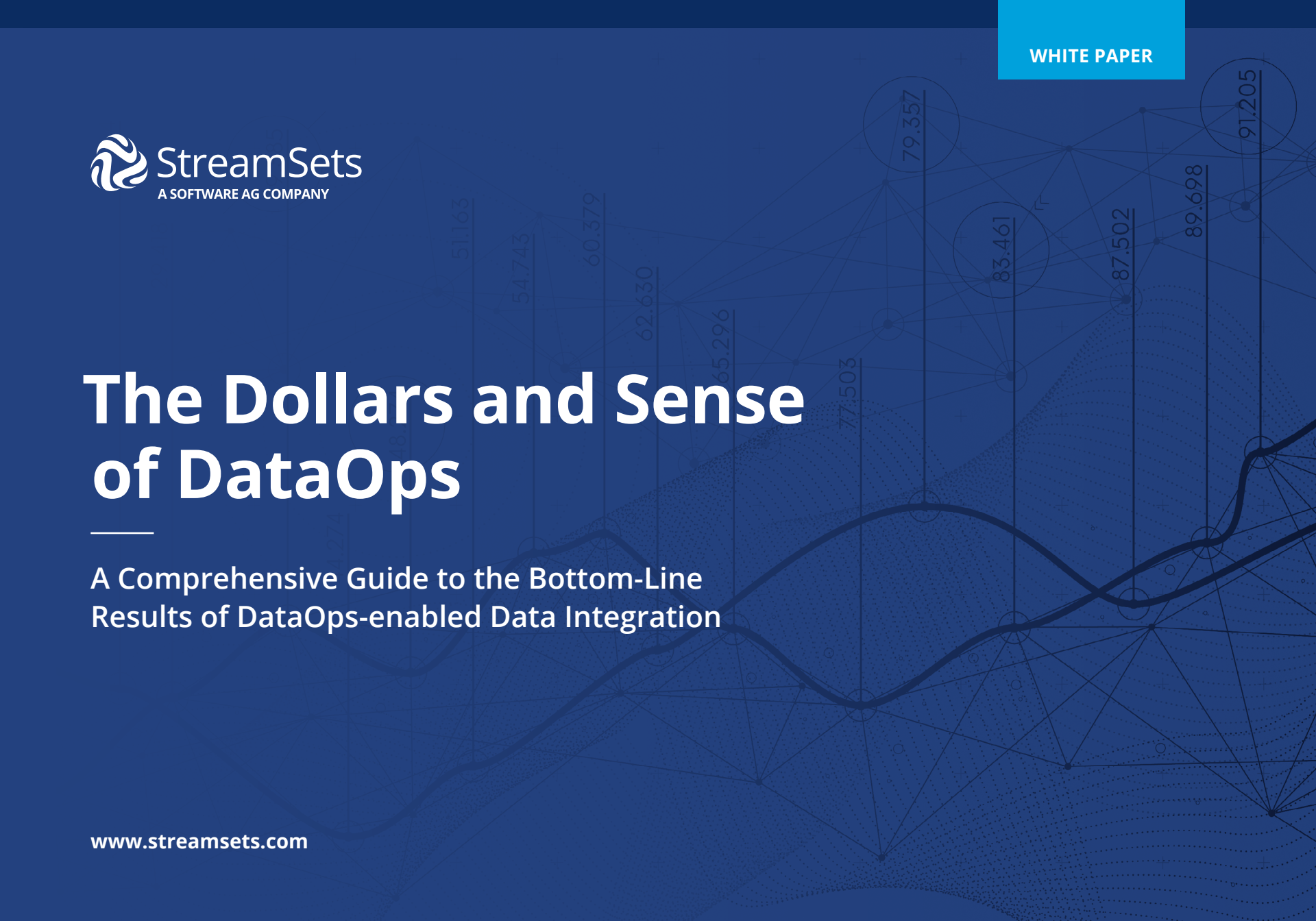
WHITE PAPER

StreamSets
A SOFTWARE AG COMPANY

The Dollars and Sense of DataOps

A Comprehensive Guide to the Bottom-Line Results of DataOps-enabled Data Integration



www.streamsets.com

Table of Contents

Data: Boundless Opportunity. Tremendous Pressure.

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

The world of data is moving *fast*. It makes it one of the most exciting frontiers in business and 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 your line of business partners 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 don't even know 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. Perhaps worst of all, you're too often flying blind because it's nearly impossible to see all the data flowing through your organization with all the siloed systems and data platforms. Blind spots and control gaps lead to governance and compliance headaches.

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, cutting through all the hype can be tricky. So what is it? And is it just a trend, or can it impact the bottom line?

As the word suggests, DataOps is a practice that operationalizes data. Gartner® defines it as "a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and data consumers across an organization."

As its predecessor DevOps did for software development, DataOps aims to knock down siloes, speed delivery, and automate for continuous integration and continuous deployment (CI/CD).

In this piece, we'll focus on how technology — specifically data integration built on DataOps principles — can facilitate collaboration across your teams and streamline processes with automation to deliver value to your business faster with fewer resources.

> **dataops** (noun):
> an engineering methodology and set of practices designed for rapid, reliable, and repeatable delivery of production-ready data and operations-ready analytics and data science models... DataOps supports business operational agility with the ability to meet new and changing data and analysis needs quickly. It also supports portability and technical operations agility with the ability to rapidly redeploy data pipelines and analytic models across multiple platforms in on-premises, cloud, multi-cloud, and hybrid ecosystems.
>
> **Eckerson Group**

The Key Is Continuous

Digital transformation, modern analytics, and real-time decision-making across lines of business require a fundamental shift in how organizations think about data. It's no longer the static resource it was in the days of traditional business intelligence and reporting. Modern analytics — including real-time dashboards, data science, AI, machine

learning, and smart data applications — require a completely different approach to integrating data. This is not just about serving up real-time data, although that is one pillar. It's about delivering fresh, accurate data continuously.

When a new data source becomes available, your team must be prepared to deliver that data within days or even hours to the business user — and to maintain 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

There are three principles of data integration systems that, together, enable continuous data delivery.

- **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, pre-defined processors, full lifecycle automation, and more are technical enablers in a data integration platform that allows your team to design continuously, shrinking data delivery times by 10x to 50x.
- **Continuous Operations** ensure that data can be delivered continuously by building pipeline and infrastructure resiliency, with high service levels to downstream users. Resilient and repeatable 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, you can eliminate 80-90% of breakages and maintenance work.
- **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 remove blind spots, detect and prevent problems, understand your data's business impact, 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

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

"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

Let's look at what we've seen helping companies like IBM, GSK, Shell, and BT Group build their DataOps practices with StreamSets data integration at the center. Innovative data leaders from these companies and more have achieved impact in three broad areas by thinking differently about delivering data to the business.

Support Diverse Lines of Business, Faster with Fewer Resources

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.

Our customers' data teams have been able to deliver data 10x-50x 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 on Continuous Design principles which allow for the following

Faster Ramp On New Technologies

Data platforms and technologies used to have decades-long, stable lifespans. That's 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 coding sophistication that the average data engineer or developer doesn't have.

> **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.

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 like they've built their existing pipelines without the need for weeks of training and mounds of coding. This increases developer productivity and accelerates the speed of new technology adoption.

Case Studies In Fast Ramp

AVAILITY

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 allows Availity's data engineers to use the latest technologies without special skills, thanks to a simple and intuitive graphical interface. A self-service repository of reusable pipelines helps them 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

AON

Before adopting the StreamSets platform, lead time to acquire data from various sources for projects could take months, as the teams that needed that data would have to go through a project intake process, buy and install ETL tools, spin up SQL databases, and then begin their work. With StreamSets, not only has Aon's time for data ingestion significantly reduced, but the organization now also has a team onboarding process that takes weeks or less. Aon has removed the need to provision dedicated infrastructure to teams, which enables data engineers to land data quicker. This benefit is passed downstream as the time from ideation to first exploration of data by data scientists and data analysts is dramatically improved.

Pipeline Builds in Minutes or Hours, Not Days

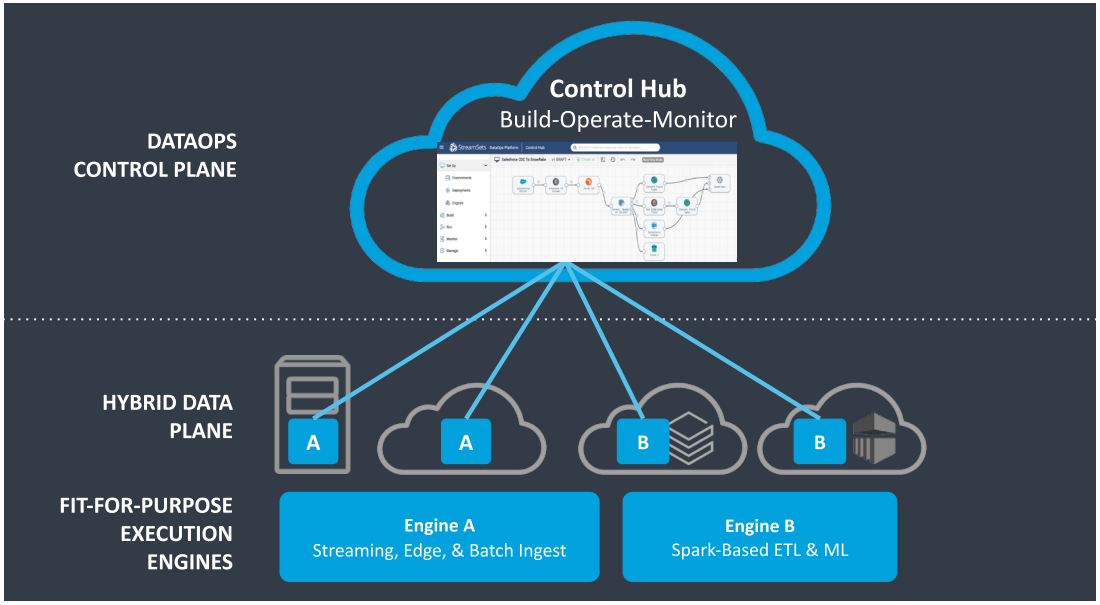
The more obvious way to deliver data faster is to build data pipelines in less time. Data integration tools that let developers use a graphical UI to design pipelines have been around for decades. But most still require the developer to specify a lot of the technical implementation details of the source 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. With StreamSets, data engineers and developers can learn once to create many different data integration pipelines. This reduces the required tools and skill sets, maintenance, and other related technology overhead to drive out complexity.

And it allows business users — those who understand the data best — to build data pipelines. Reusable pipelines and pipeline fragments mean you don't need to start from scratch every time (but you still have the flexibility to add custom code if needed).

Another game-changer for developer productivity is a single development and management experience for every data platform or pipeline pattern (streaming, batch, CDC, ETL, ELT, and ML). Most data integration platforms cover this variety by offering a collection of different tools that are branded under one umbrella. 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 essential. It lets engineers design any pipeline from one place, no matter the pattern or execution technology. But it's 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



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 between various data platforms and applications. The control plane manages the data plane execution 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 the operational metrics back to the control plane.

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, hybrid, legacy 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

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

- 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 reduced the time to onboard new business units onto their cloud data platform by 6 months.
- The State of Ohio integrated 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.

"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 and 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

Delivering a COVID-19 Dashboard Overnight

88
Counties

1000s
Different data sources

1
Night to build the Covid dashboard

The State of Ohio's data platform team supports dozens of state government 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

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.

Automating testing, debugging, and deploying is one key to providing continuous operations across the full lifecycle. Live data preview lets developers see the resulting data from their pipeline logic while still designing *before* deployment, 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

GSK

How Self-Service Data Advances Drug Discovery

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

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 the 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, 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

Keep Up with Constant Change

The vast majority of business logic that drives the modern enterprise lives in integrating thousands of tiny, specialized applications across multiple platforms and 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 just ran. 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. In the best-case scenario, the pipeline stops working, and your data engineers must fix it. In a bad case, you lose data. In the worst case, the pipeline works but delivers corrupt or incorrect data, which takes weeks to detect and fix, 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 constantly, 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 2:00am calls, and you lose credibility.

We help clients achieve operational resilience for Continuous Operations with three data integration platform features.

Built-In Data Drift Resiliency

StreamSets DataOps-focused data integration platform assumes change is constant. It helps you build and run data pipelines that detect and handle schema, semantics, and infrastructure drift changes. These resilient data pipelines are intent-driven, requiring minimal schema definition of sources and destinations. They're highly decoupled, with each stage having minimal dependencies on other stages; if one changes, you avoid the domino effect. Resilient 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 the vast majority of changes that might happen during a pipeline's operation,

eliminating 80-90% of breakages. These intent-driven, resilient data pipelines detect changes to data in-flight and continue to

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

Generate Capital: A Case Study in Pipeline Resilience

Generate Capital cut the 10 to 20 hours of data engineering work 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 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 you migrate from a Hadoop data lake to AWS S3 — or from one cloud platform to another — or 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 two years, from an on-premise Hadoop cluster to an AWS data platform and then to a Google Cloud platform. They had thousands of pipelines they needed to migrate. Because the pipelines were built to be platform-agnostic, BT was able to make simple updates to those pipelines in an automated manner to point to the new cloud data platform without rewrites.

Democratizing Data to Drive Business Results

250+
Platform users

£310M
Saved since implementation

16,000+
Pipelines in use

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 for their DataOps practice.

Learn More

Enable Innovation, Prototyping, and Experimentation With Centralized Guardrails

One of the most intractable problems with today's data, analytics, and cloud sprawl is ensuring global transparency and control. When 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, it's hard to understand, govern, and manage the data supply chain.

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 management console lets you see how systems are connected and data is flowing across the enterprise.

This is 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. 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 enterprise-wide manageability and visibility across hybrid

and multi-cloud architectures. It provides visibility into data connections and flows, including volume and throughput of data, as well as exactly what data is moving between components. You can automatically see when a new integration point is being created, as well as if there is a more direct route for the data. And you can know where data comes from to help understand and explain outcomes, for example, in AI/ML models. Finally, you can expose hidden problems in your data flows with data SLAs and rules by creating guardrails and quality checks and then managing by exception.

With StreamSets' hybrid deployment with centralized engine management, you can bridge the gap between new and legacy environments easily and securely. With a data "mission control" across all your environments, you can easily move between clouds and on-premises.

You'll extract maximum value from your data faster while lowering the cost and effort of managing your disparate tool chains.

With this single pane of glass, you can reduce the resources required to monitor and manage your data pipelines and flows. You can observe your data as it behaves in the real world, removing blind spots and ensuring adherence to compliance policies and governance requirements.

IBM: A Case Study in Global Transparency

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.

How Self-Service Data Supports Operational Excellence

1,000
Sites around the world

20,000+
Data pipelines

1
Dashboard for central visibility

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 volumes and velocity of data, IBM turned to StreamSets for a solution that would truly operationalize continuous data management and integration and provide resilience, agility, and visibility for a central team.

Learn More

Continuous Data for Nonstop Business Results

No matter your goals — driving operational efficiency and excellence, reducing costs, helping new products get to market faster, or all of the above — the time for digital transformation is now. The market is undergoing 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 with confidence as data platforms and business requirements rapidly change. Traditional ETL and point integration solutions can't 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 to support DataOps practices, organizations can deliver continuous data for nonstop business. This translates to delivering business value from 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 resilient data pipelines that abstract away details and automate as much as possible, delivering continuous data across the enterprise.

Next Steps

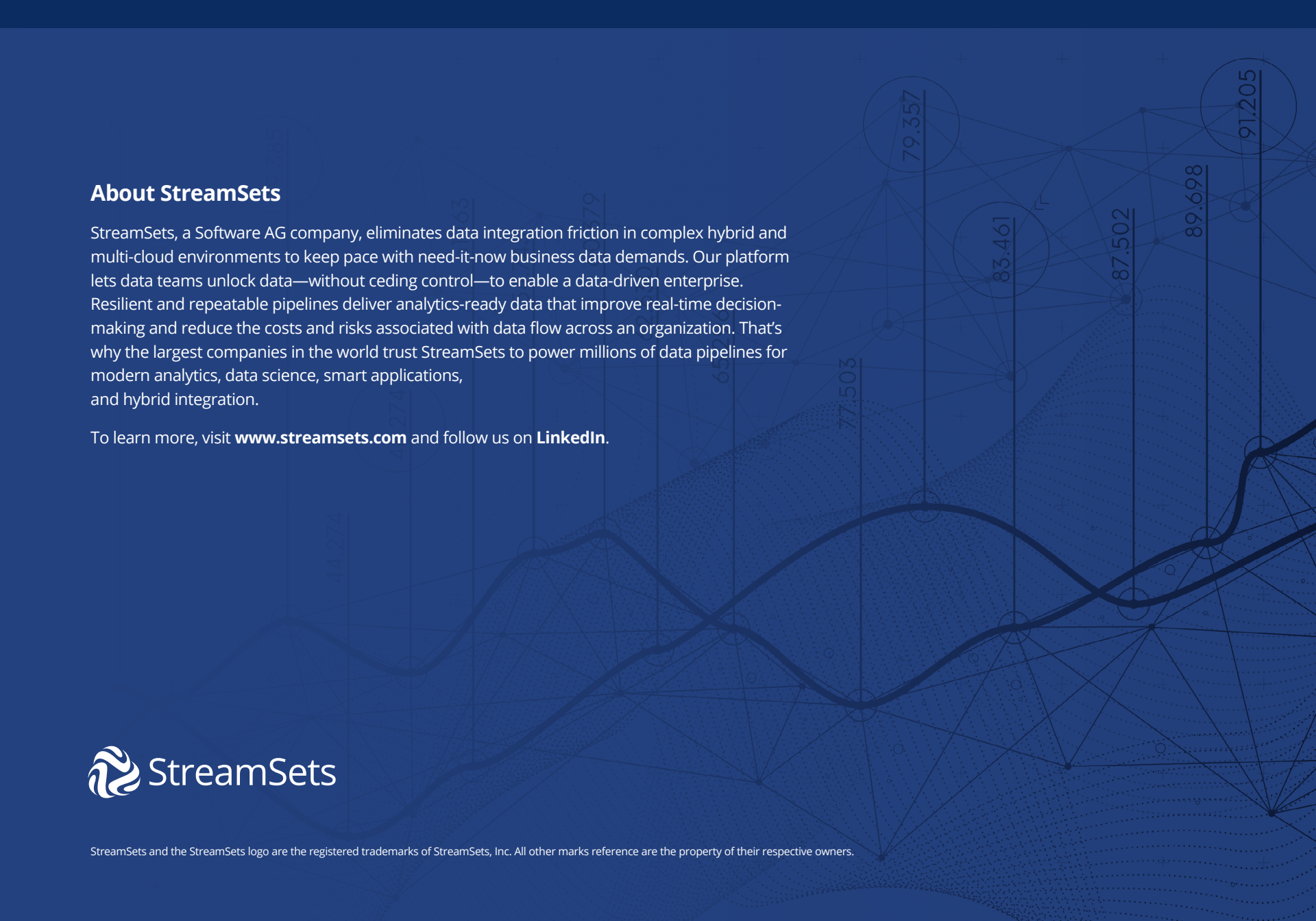
Take the guesswork out of modern data integration with resilient and repeatable data pipelines. Deliver continuous data to every part of your business with StreamSets.

Visit StreamSets.com **Request a Demo**

About StreamSets

StreamSets, a Software AG company, eliminates data integration friction in complex hybrid and multi-cloud environments to keep pace with need-it-now business data demands. Our platform lets data teams unlock data—without ceding control—to enable a data-driven enterprise. Resilient and repeatable pipelines deliver analytics-ready data that improve real-time decision-making and reduce the costs and risks associated with data flow across an organization. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, data science, smart applications,
and hybrid integration.

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



StreamSets