DATA SHEET

## BENEFITS

By using StreamSets for data initiatives, customers have been able to:

### Go Fast

- Rapid, continuous delivery of fresh data
- Reduced dependence on hard-to-find technical skill sets
- Greater cross-team collaboration and productivity
- Flexibility to adopt new platforms and technologies as needs change

### Deliver with confidence

- Decreased outages and breakages resulting from data drift
- Enforced data SLAs for performance, quality and security

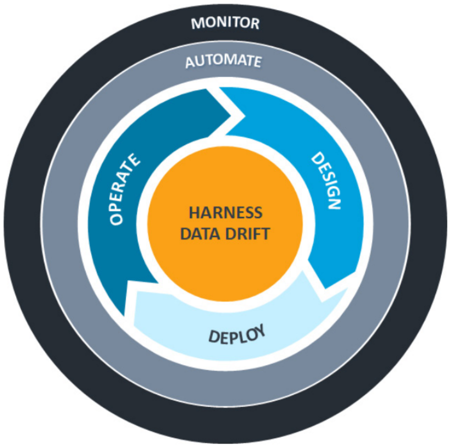
# StreamSets DataOps Platform

Delivering Continuous Data For Modern Analytics

## DataOps: Modernizing your Data Practice

Data is more pervasive and more valuable than ever, and analytics has modernized to harness the value of this data. How you deliver data to drive those analytics has to modernize, too. But that's no easy feat, as data is fragmented and proliferating in more places, with more and more of it outside of your direct control.

DataOps—based on the DevOps concept of continuous delivery—has emerged as the new approach to bring data integration and management into the modern era. A combination of processes, organization and enabling technology, DataOps has become the lynchpin to reimagining data integration and management in an always-on, always-changing world. By architecting your processes, technology and organization to deliver data continuously, you can go faster, with confidence.



The Practice of DataOps

## The StreamSets Approach

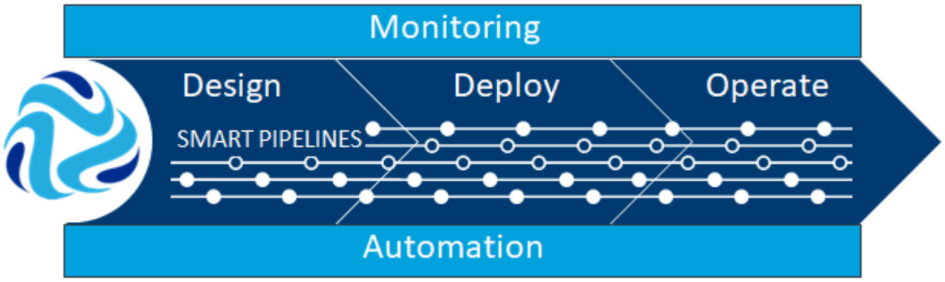
The StreamSets DataOps platform is a key technology foundation for a DataOps practice. The DataOps Platform is designed to simplify the entire dataflow lifecycle, including how to design, deploy, and operate enterprise data pipelines at scale. Developers can design batch and streaming pipelines with a minimum of code, while operators can aggregate dataflows into topologies for centralized provisioning and performance management. More important, the StreamSets DataOps Platform powers continuous data integration where smart data pipelines can be designed, deployed, operated, and adapted on an ongoing basis.

At the heart of StreamSets is the ability to harness data drift—those frequent and unexpected changes to data that break pipelines, delay projects and damage data integrity. Traditional approaches assume a static environment, and thus break when unexpected changes occur. StreamSets applies automation and monitoring, the cornerstones of DevOps, across the entire data integration lifecycle so you can deliver continuous data at the speed of need, without sacrificing confidence.

**Data Drift** *(noun)*
unexpected, unannounced and unending changes to data structure, infrastructure, and semantics.

## CORE CAPABILITIES

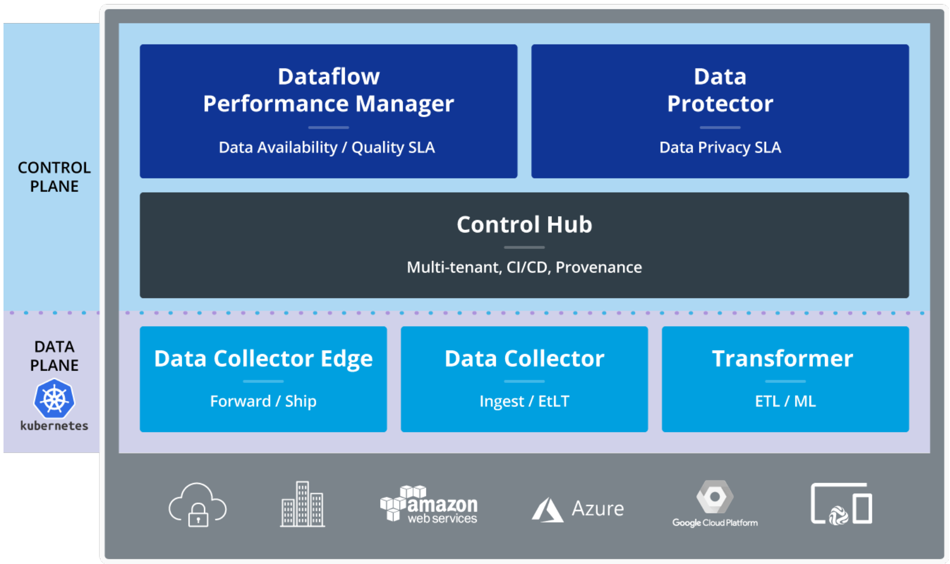
- Collaborative, visual pipeline design, deployment and operations
- Unified tools for both batch and streaming workloads
- Flexibility to deploy on-edge, on-premise or in the cloud
- Fully instrumented pipelines for real-time monitoring of data in motion
- Automatic data drift detection and remediation
- End-to-end visibility across all dataflows
- Measurement and enforcement of Data SLAs for availability, quality and privacy



**Smart data pipelines** are instrumented throughout, and are able to sense and respond to changes and issues on-the-fly, even as they are running. Where traditional data integrations break, smart data pipelines ensure continuous operations and let you harness change for business innovation.

### StreamSets DataOps Platform and Products

From the bottom up, the StreamSets DataOps Platform consists of several components which work seamlessly together to accelerate delivery, provide transparency and harness data drift.



## PRODUCTS AND KEY FEATURES

### StreamSets Data Collector

Open source execution engine which moves data between any source and any destination, performing transformations and push down analytics along the way.

- Leverage dozens of built-in processors or design your own.
- Trigger custom code when needed.
- Identify and handle personal data (PII) as it arrives.
- Minimal schema specification.
- Smart sensors detect and correct data drift detection automatically.

### StreamSets Control Hub

Hosted environment for collaborative design and deployment of dataflows. It includes a pipeline repository for reuse and manageability.

- Test pipelines wherever execution happens—on cloud, on edge or on premise.
- Controlled publish and version management, including rollback.
- Automated deployment and provisioning.
- Metadata collection and visibility across the architecture.
- Get real-time metrics for throughput, latency and error rates.

### StreamSets Data Collector Edge

Lightweight version of StreamSets Data Collector for a fit-for-purpose edge solution to support use cases such as IoT and cybersecurity.

- Data Collector Edge provides a fit-for-purpose edge solution to support initiatives such as IoT and cybersecurity.
- Less than 5MB footprint.
- Utilizes 1-2% of CPU.
- Apache licensed open source binary.

### StreamSets Dataflow Performance Manager

Management tool for enforcing Data SLAs for data availability and accuracy.

- Review historical view of dataflow metrics, at any point in the pipeline.
- View changes over time and across versions.
- Identify hotspots and performance issues.
- Configure SLAs for a range of requirements.

### StreamSets Transformer

Execution engine that allows any developer to create data processing pipelines that execute on Spark.

- Build, preview, debug, and execute on Spark using a UI.
- Brings the power and scale of Apache Spark to every developer.
- Easy-to-use interface and rich tools democratize the process of data transformation.
- Progressive error handling means the system finds exactly where and why errors occur, without needing to decipher complex log files.
- Execute on any Spark Cluster, on premise on Hadoop clusters or on cloud-hosted Spark Services.

### StreamSets Data Protector

Tool for discovering and securing sensitive data "in-flight" at run-time, ensuring pervasive, automatic data protection and compliance with data privacy regulations.

- Easily build PII monitoring into pipelines.
- Continuously scan structured and unstructured data.
- Obfuscate data using reversible and irreversible algorithms.
- Build rules by department, user type and data type.

## USE CASES

Organizations large and small have used StreamSets to support their modern data and analytics use cases, including:

**Data lake ingestion**

**Data warehouse modernization**

**Migration to cloud data platforms**

**Real-time event streaming**

**IoT and edge device integration**

**Machine learning**

RingCentral®

RingCentral, Inc. is an award-winning global provider of communications and collaboration solutions. RingCentral built a data lake creating a 360 view of conference calls. With StreamSets, RingCentral was able to build data pipelines that gracefully handled data drift and traffic bursts to ensure that business users had timely access to analysis-ready data. By having relevant and reliable data immediately available, RingCentral was able to address call quality issues and detect fraud in real time, as well as understand product usage to optimize the customer experience.

gsk

GlaxoSmithKline, a global pharmaceutical leader, transformed its R&D data and analytics infrastructure to shorten the drug discovery and development timeline. GSK uses StreamSets to operate data pipelines sourcing from millions of data elements, bridging operational silos such as genomics, clinical data and scientific research. With StreamSets, GSK is able to provide self-service data access and exploratory data science to over 8,700 scientists, analysts, and domain experts without the need for IT involvement.

## 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.

## LEARN MORE

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

www.streamsets.com