DATA SHEET

StreamSets

BENEFITS

Build data drift-resistant pipelines in hours, not weeks, to greatly reduce development costs.

Connect to leading big data systems, including Apache Hadoop™, Apache Kafka and Apache Spark.

Deploy flexibly on-edge, on-cluster or in the cloud.

COMMON USE CASES

Apache Kafka™ Enablement

Build real-time streaming pipelines with drag-and-drop connectors for Kafka and many other sources and destinations.

Cloud Migration

Transport data to, from and across multiple clouds with connectors for leading providers.

Apache Hadoop™ Ingest

Continuous ingestion of new and traditional data into Hadoop and the surrounding ecosystem.

Search Enablement

Populate search engines with data from any source and route to multiple search indices based on data values.

StreamSets Data Collector™

Open source software for building dataflows quickly and easily, spanning on-premises, multi-cloud and edge infrastructure.

Big data opens new horizons for business, but also creates a challenge in the form of unpredictable changes known as data drift. The complexity of modern data sources and platforms ensure that data structure, semantics and infrastructure will change unexpectedly, as will your business requirements for data movement.

In this dynamic environment, custom coding or rigid integrations built using schema-centric ETL tools is problematic. Data drift, plus sprawling data architectures and the urgency of real-time analytics, put new pressures on data movement that traditional integration solutions can't handle. Hand-coded pipelines take too long to develop and break frequently, putting the performance of data-driven applications at risk.

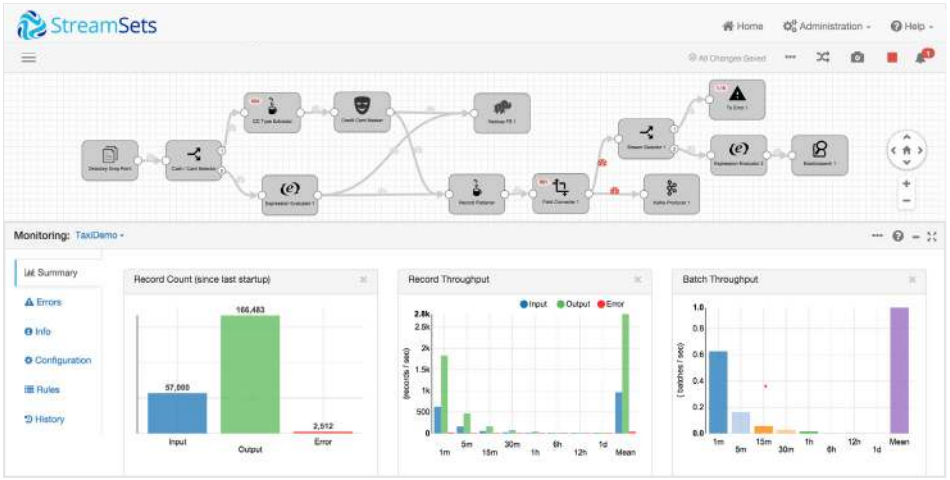
StreamSets Data Collector (SDC) solves these problems by simplifying the process of building, executing and operating modern dataflows.

SDC's graphical user interface lets you efficiently build batch and streaming data flows with minimal schema specification, connecting myriad sources to multiple big data solutions with built-in transformations for data normalization and cleansing.

As data architectures become increasingly complex, you need the ability to deploy anywhere. SDC enables you run robust pipelines wherever you need to, be it on premises, in the cloud or – using the new StreamSets Data Collector Edge – on constrained endpoint devices.

SDC supports common streaming data platforms, including popular tools such as Apache Kafka™, Amazon Kinesis and Apache Spark™, so you can accelerate the time to build real-time solutions.

Built for continuous operations in the face of change, SDC enables new sources to be added in minutes. Dataflow stages are logically isolated to enable zero-downtime upgrades of underlying systems. Unexpected data drift that impacts schema or semantics is handled automatically to maintain dataflow reliability and data quality.



StreamSets Data Collector: An open source IDE for drift-resistant any-to-any batch and streaming dataflows.

FEATURES

Build adaptable pipelines with minimal coding and maximum flexibility.

- Easy to use GUI for building pipelines without hand coding.
- Dozens of transformations built in, plus advanced scripting and triggered execution of external code such as Apache Spark.
- Rapid troubleshooting using data injection, snapshot and replay functionality.

Operate continuously in the face of constant change

- Zero downtime for upgrades and migration of underlying infrastructure (e.g. HDFS, Kafka).
- Direct integration with big data governance tools including Cloudera Navigator™ and Apache Atlas™.
- Manage complex dataflow topologies with StreamSets Dataflow Performance Manager (DPM).

Execute pipelines wherever you need to

- Flexible deployment in your cluster, across multiple clouds, or on edge nodes and devices.
- Deploy and scale efficiently using Kubernetes.
- 100% in-memory operation for high throughput and low latency.

Monitor pipeline performance and data quality

- Pipelines automatically detect and adapt to data drift as schema and semantics evolve.
- Customizable runtime metrics on dataflow operations and data fidelity.
- Real-time early warning of anomalies and outliers via data introspection, sampling, threshold rules and alerts.

The StreamSets Data Operations Platform is designed to simplify the entire dataflow lifecycle, including how to build, execute and operate enterprise dataflows 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.

ABOUT

StreamSets is headquartered in San Francisco. Our mission is to help enterprises harness their data in motion. StreamSets software is in use at hundreds of organizations and we're backed by top-tier Silicon Valley venture capital firms, including Accel Partners, Battery Ventures, Ignition Partners, and New Enterprise Associates.

LEARN MORE

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

www.streamsets.com