DATA SHEET

StreamSets Transformer™

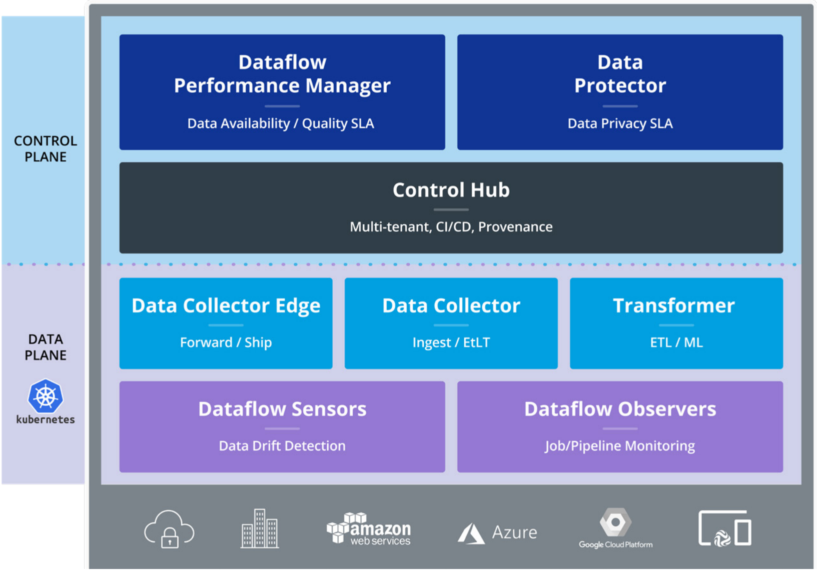
Harness the power of Apache Spark, minus the complexity.

Apache Spark brings the promise of stream processing, next-generation ETL, and machine learning at impressive scale. But for too many companies, taking advantage of powerful tools like Apache Spark has remained in the hands of those with mature data engineering skills and knowledge of big data systems. It's often difficult to recruit and retain talented Spark developers, making them a high priority hire in most organizations. Even for companies with mature Spark workloads in production, poor visibility into operation and performance leaves already taxed data teams with few tools for optimal operation.

StreamSets Transformer is an execution engine within the StreamSets DataOps platform that allows any developer, including those without Spark expertise, to create data processing pipelines that execute on Spark. Using a simple-to-use drag-and-drop UI, users can create pipelines for performing batch data processing, stream processing and machine learning operations—able to fully utilize the power of Apache Spark without a deep technical understanding of the platform. Conversely, advanced Spark developers are able to take advantage of Scala and PySpark language processors and maximize reuse and operationalization of their pipelines.

StreamSets Transformer enables users to solve their core business problems by abstracting away the complexity of operating the Spark cluster. Pipelines instrumented with StreamSets provide heightened visibility into the execution of Apache Spark. As a result, it's easy to see exactly how long every operation takes, know how much data gets transferred at every stage, and view any proactive and contextual error messages that appear if and when problems occur down to the row/column level.

Transformer can execute both batch or streaming operations, mixing and matching as required. Users never have to make batch, streaming, lambda, or kappa architectural decisions when designing pipelines and instead can focus on working with continuous data, when and where it's needed.



StreamSets DataOps Product Platform

BENEFITS

Execute batch and streaming in the same system.

Run Apache Spark anywhere: on prem, public or private cloud.

Powerful Apache Spark operation executed through a simple UI.

Bring Spark capabilities to the entire data team.

COMMON USE CASES

ETL

Create massive-scale ETL operations using a drag-and-drop UI.

High-Throughput Real-Time Streaming

Generate millions of rows per second and choose between micro-batch and batch.

Machine Learning

Train and execute ML models on massive data sets in a fraction of the time of non-distributed systems.

Self-Service Ingestion

Create self-service data transformations that power your advanced analytics.

FEATURES

Perform next-generation ETL and machine learning with no hand coding

- Brings the power and scale of Apache Spark to every developer.
- Easy-to-use interface and rich tools democratize the process of data transformation.

Achieve continuous data and continuous monitoring

- No need to master batch or streaming semantics because the system handles any of them, mixing and matching as required.
- Unparalleled visibility into Spark application execution.

Extend Spark capabilities to the entire data team

- The ETL developer can use higher-order transformation primitives.
- The Analyst can use SparkSQL, and the data scientist can use PySpark and Scala.
- The Spark developer can use custom Java/Scala processors.

Take advantage of rich data processing capabilities

- Progressive error handling means the system finds exactly where and why errors occur, without users needing to decipher complex log files.
- Build, preview, debug, and execute on Spark using the StreamSets Data Collector-style UI.
- Execute on any Spark Cluster, on prem on Hadoop clusters or on cloud hosted Spark Services.

The StreamSets Data Operations (DataOps) 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 without hand coding, while operators can aggregate dataflows into topologies for centralized provisioning and performance management.

ABOUT STREAMSETS

StreamSets built the industry's first multi-cloud DataOps platform for modern data processing and 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. StreamSets 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.

Founded in 2014, StreamSets is backed by top-tier Silicon Valley venture capital firms, including Battery Ventures, New Enterprise Associates (NEA), and Accel Partners. For more information, visit www.streamsets.com.

LEARN MORE

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

www.streamsets.com