SOLUTION BRIEF

StreamSets

BENEFITS

- **Faster migration to cloud with less overhead on data engineering resources**
- **Easily bring data from multiple disparate sources using a drag-and-drop interface**
- **Better management of data quality and performance for cloud data lakes with Delta Lake**
- **Change Data Capture (CDC) capability from several data sources in to Delta Lake**
- **Decreased risk of disruptions for Hadoop migrations with quicker time-to-value on on-prem to cloud initiatives**
- **Continuous monitoring of data pipelines to lower support cost and optimize ETL pipelines**

StreamSets + databricks

StreamSets and Databricks's joint solution accelerates the delivery of critical data engineering projects (ETL) and democratizes the Apache Spark and Delta Lake experience.

Companies want to accelerate digital transformation projects with reliable data lakes that deliver on their analytics and data science plans. However, high data volumes from disparate sources make it difficult to update a data lake while maintaining data quality and availability.

To solve this, Databricks and StreamSets have partnered to accelerate value to cloud analytics by automating ingest and data transformation tasks. The joint solution brings rapid pipeline design and testing to cloud data processing. StreamSets applies a DataOps mentality for rapid development on Databricks Unified Data Analytics Platform and Delta Lake, helping organizations extend the power of Apache Spark to the entire data team for modern analytics and data science.

StreamSets provides a drag-and-drop interface to design, manage and test data pipelines for cloud data processing. Together, this partnership brings the power of Databricks and Delta Lake to a wider audience.

About Delta Lake

Delta Lake is a storage layer that sits on top of your existing data lake file storage, such AWS S3, Azure Data Lake Storage, or HDFS. Delta Lake makes it possible to unify batch and streaming data from disparate sources and analyze it at data warehouse speeds. It supports transactional insertions, deletions, upserts and queries. It provides ACID compliance, which means that any writes are always complete and failed jobs are fully backed out.Delta Lake also stores a transaction log to keep track of all the commits made to provide expanded capabilities like ACID transactions, data versioning, and audit history. To access the data, you can use the open Spark APIs, any of the different connectors, or a Parquet reader to read the files directly. Multiple data pipelines can read and write data concurrently to a data lake. ACID Transactions ensure data integrity with serializability, the strongest level of isolation.

About Transformer

StreamSets Transformer is an execution engine within the StreamSets DataOps platform that allows any developer to create data processing pipelines that execute on Spark. Rather than writing code, developers can utilize Transformer's simple-to-use drag-and-drop UI to develop data pipelines. Transformer's heavily instrumented pipelines provide heightened visibility into the execution of Spark applications. 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 (without having to parse through complex Spark logs). Transformer pipelines run in batch or streaming mode, processing data between a myriad of sources and destinations.

SOLUTION BRIEF

StreamSets

COMMON USE CASES

ETL
Create Apache Spark ETL pipelines using a drag-and-drop UI.

Migrate Data for Cloud Analytics or ML
Move data easily to the cloud and perform self-service analytics and machine learning.

Change Data Capture to Delta Lake
Save money by replicating batch data on the Spark cluster.

IoT Analytics
Capture high-velocity sensor and time-series data in real-time for predictive analytics.

Easy Ingestion for Spark Analytics
Leverage 100's of pre-built connections for enterprise data systems.

ABOUT STREAMSETS

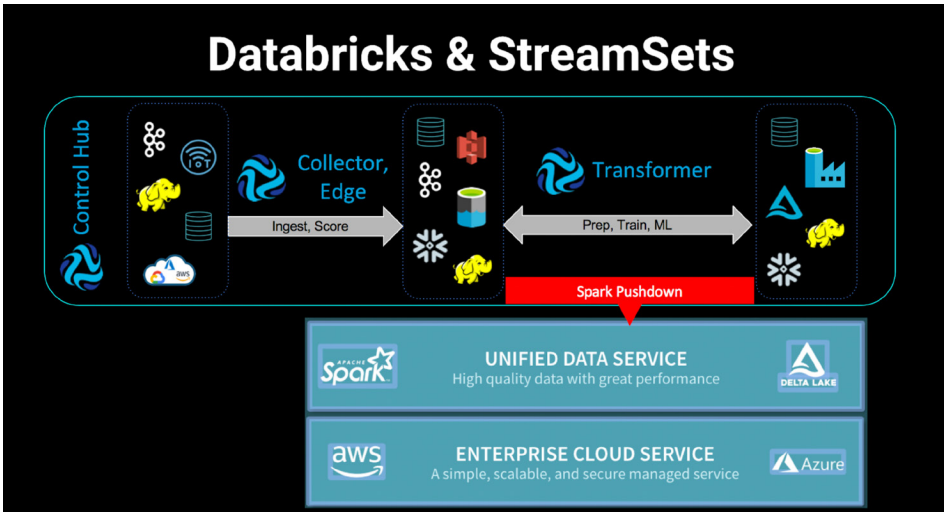
StreamSets built the industry's first multi-cloud DataOps platform for modern data integration, helping enterprises to continuously flow big, streaming, and traditional data to their data science and data analytics applications. The platform uniquely handles data drift, those frequent and unexpected 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.

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 streamsets.com

LEARN MORE

Visit us at:
www.streamsets.com
www.databricks.com



Features

Rapid Development on Apache Spark

- Empower data teams with rapid, no-code development using Apache Spark and Delta Lake for modern analytics and data science at scale.
- Superior visualization for monitoring and metrics to optimize and manage spark jobs.
- Single pane of glass for all Spark workloads.
- Easily design Spark workloads regardless of the data complexity.
- Allow non-Spark data engineers to utilize Spark without having to learn to code.

DataOps for Analytics and Data Science Projects

- Self-service ingestion and ETL to allow data engineers and data scientists to advance projects faster.
- Simplified architecture that handles batch and streaming workloads.
- Native integration to Delta Lake's real-time and batch data management.
- Support for tensorflow, MlLib, SparkSQL.
- Build, test, and deploy in one system.

Change Data Capture (CDC)

- Out of the box CDC capability for popular relational data sources (such as MySQL, MS SQL Server, Oracle and more).
- Read the binary log of any relational system to capture changes without any performance or load impact from CDC pipelines.
- Implemented Delta's MERGE functionality allows users to reconcile changes from CDC sources to Delta tables with a simple visual pipeline automatically.

Easy Migration to Delta Lake

- Move existing Hadoop and other on-premise workloads and maintain a hybrid environment until you are ready to cut over.
- Delta Lake supports transactional insertions, deletions, upserts, and queries—making your data lake reliable and highly performant.
- Data from Hive tables can be streamed directly into Delta tables, using a no code visual approach.
- Auto-scaling and on-demand cloud compute better align costs to actual analytics workloads and outcomes over On-Prem monolithic clusters.

Together, this partnership brings the power of StreamSets and Databricks to a wider audience.