SOLUTION BRIEF

StreamSets

STREAMSETS TRANSFORMER

StreamSets Transformer is an execution engine within the StreamSets DataOps platform that delivers next-generation ETL at impressive scale.

Batch and Streaming

Reduce operational overhead, optimize data ecosystem spend, and better utilize data teams.

Execute Apache Spark Anywhere

Execute Apache Spark on hosted services in the public cloud, on prem, or in virtualized workloads.

Scale to Meet Your Needs

Use StreamSets Transformer to scale your analytics and machine learning.

No Hand Coding

StreamSets Transformer brings the power and scale of Apache Spark to every developer.

Continuous Data, Continuous Monitoring

StreamSets Transformer handles batch and streaming semantics in the same platform.

Progressive Error Handling

Transformer offers progressive error handling with the ability to build, preview, and debug—all before pushing Apache Spark applications live.

StreamSets for ETL

Overview

ETL—shorthand for the extract, transform, and load triumvirate of critical data processing capabilities—has been a fixture in the data management ecosystem since the 1970s. Data storage and processing has gained in complexity over the last few decades, meaning that ETL tools and approaches have been forced to evolve, as well. Today, ETL is necessary for vital enterprise operations such as cloud migration, machine learning/AI, Internet of Things (IoT) data integration, data warehousing, and business intelligence.

Challenges

Enterprises in all industries increasingly rely on insights gleaned from data analysis, which requires leveraging cloud-based, streaming, real-time, structured, and unstructured data. As a result, they also rely on the unglamorous but entirely necessary ETL processes. But all the different ways that data is generated, transmitted, and stored put increasing pressure on ETL systems that struggle to keep pace with both the ever-expanding variety and volumes of data.

An effective modern ETL solution must:

- Be logical and intuitive enough for any data professional to use.
- Handle data from any source, in any format.
- Be designed for extensibility.

Organizations are hampered in their ETL efforts because:

- While Apache Spark has significantly reduced ETL processing windows, using its sophisticated frameworks demands equally sophisticated—and expensive—data processing expertise.
- ETL tools designed for structured batch processing can't handle workloads of unstructured or streaming data.
- The list of possible data sources and formats continues to grow, as does the pressure to process and analyze the data more and more rapidly.

Data is the lifeblood of organizations, and enterprise success and even survival depends on how well and how fast they can organize, convert, and analyze that data. At the same time, both the sources and target workloads continue to grow in complexity. Whereas historical ETL tools operated in relatively well-defined data storage and data processing environments, today's environments are wilder and more free-form. The same set of data might be needed, immediately, by multiple systems, all requiring different data formats.

Modern ETL tools must be built to deal with the realities of these complex environments. It's no longer acceptable to rely on different ETL tools for different data environments, or to restrict any ETL tool's use to a few highly skilled, highly experienced, and highly paid data experts. For organizations to thrive now and in the future, next-generation ETL tools must be designed for any data situation and for use by any member of the data team.

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

Solution

StreamSets enables next-generation ETL through StreamSets Transformer. The product provides enterprises with the flexibility to create ETL pipelines for both batch and streaming data as well as clear visibility into their data processing operation and performance.

Easy enough for any data professional to use. Instead of depending on a couple of superstar Apache Spark experts who have mastered hand coding, enterprises can use StreamSets to extend next-generation ETL capabilities to their entire data team. Developers can use the drag-and-drop StreamSets UI to create pipelines for ETL that execute on Apache Spark as well as stream processing and machine learning operations.

Handles data, in any format. The data that organizations need to make the best business decisions arrive from multiple sources, in multiple forms. By not requiring separate ETL tools for each data format, StreamSets decreases organizations' operational overhead of supporting multiple tools.

Designed for extreme extensibility. StreamSets not only makes ETL work with any data source, format, or environment, it also is extensible for use by any member of the data team. StreamSets provides higher-order transformation primitives for ETL developers, SparkSQL for analysts, PySpark for data scientists, and custom Java/Scala processors for Apache Spark developers.

StreamSets Benefits

StreamSets enables organizations to:

- Take full advantage of Apache Spark, no matter your level of coding expertise.
- Stop worrying about whether processing is batch or streaming, if it's structured or unstructured, or what format it's in.
- Work equally well with data in hosted services in the public cloud or on premises.
- Reduce the operational overhead of operating Apache Spark in production.
- Significantly speed the creation of ETL pipelines that are accurate, secure, and reliable.
- Diagnose and debug ETL pipeline issues quickly and easily.
- Migrate more of their mission-critical workloads to the cloud.
- Incorporate machine learning, AI, IoT, and other advanced data-reliant operations into their businesses.

Closing

The ability to process data for analytics—rapidly, accurately, and securely—has become a core competitive capability for enterprises across industries and geographies. ETL is a crucial process for enabling analytics, but ETL tools have not kept pace with today's fast-multiplying data sources and formats.

Find out more about how StreamSets can deliver next-generation ETL to both streamline and democratize your data analytics processes. Contact a StreamSets representative today.