SECURITY WHITE PAPER

StreamSets: Control & Data Plane Separation

Ensuring Data Security & Privacy for StreamSets Customers

StreamSets
A SOFTWARE AG COMPANY

InfoSec

StreamSets
A SOFTWARE AG COMPANY
InfoSec

Executive Summary

StreamSets' services are used in constructing and managing data pipelines, both on premises and in the cloud. StreamSets' services allow customers to orchestrate their data pipelines, but do not involve StreamSets taking direct custody of any of the customers' pipeline data.

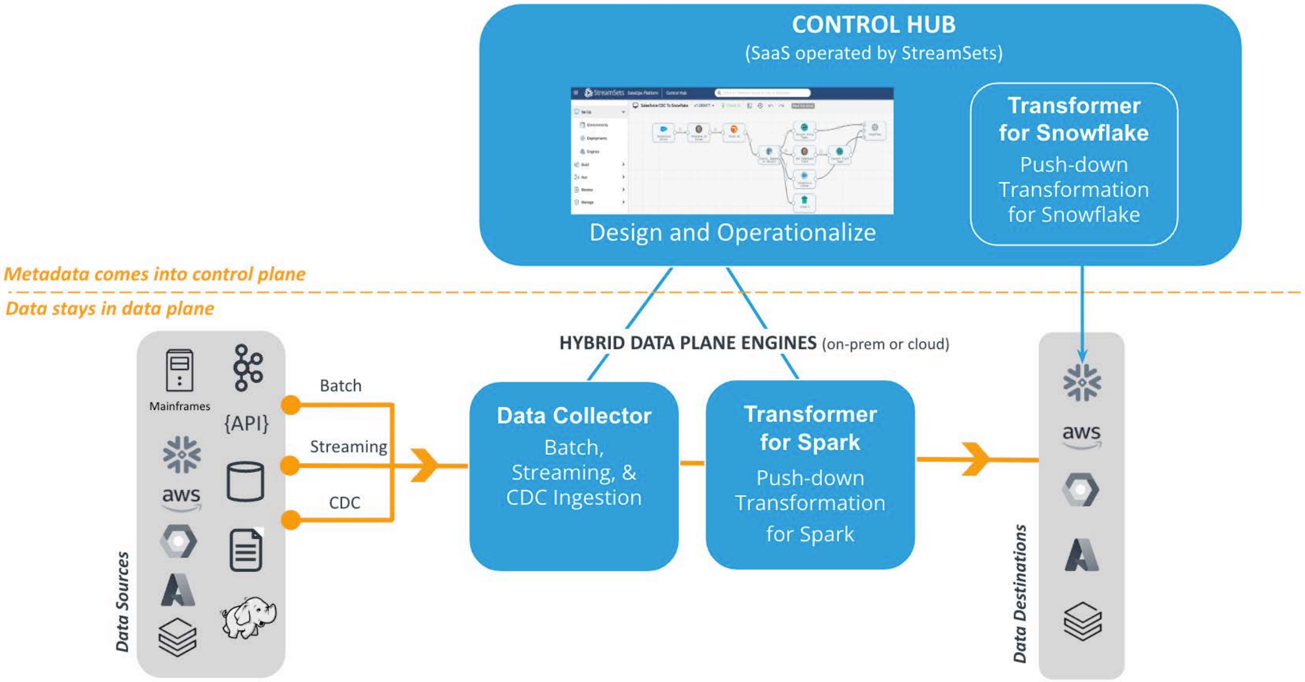
Questions on security and compliance issues around data security and privacy are regularly answered "not applicable" (N/A), because customer pipeline data are never exposed to StreamSets.

We can divide "data," in the StreamSets services, into three primary categories: configuration data; usage data; and pipeline data. The first is how customers are set up to use the services, and the second is generated by their use. The last, the actual data within the pipelines that customers create and manage using StreamSets' services, only resides within the customers' own resources, either on premises or in a cloud of the customer's choosing, and is never seen by StreamSets.

Date of Information: June 29, 2023

This white paper is one of a series produced by the StreamSets Information Security (InfoSec) Department as an aid in understanding the company, its services, and the security, compliance and privacy considerations in adopting and using them. All information should be accurate as of the date of publication, but corrections are welcomed, and white papers will be revised and reissued on a regular basis as required. Please contact InfoSec at security@streamsets.com with any questions or comments on this document, any of the other InfoSec white papers, or for general questions on StreamSets and its services.

Control Plane vs. Data Plane


CONTROL HUB
(SaaS operated by StreamSets)
Design and Operationalize
Transformer for Snowflake
Push-down Transformation for Snowflake
Metadata comes into control plane
Data stays in data plane
HYBRID DATA PLANE ENGINES (on-prem or cloud)
Data Sources
Mainframes
{API}
aws
Batch
Streaming
CDC
Data Collector
Batch, Streaming, & CDC Ingestion
Transformer for Spark
Push-down Transformation for Spark
Data Destinations
aws

StreamSets services take the form of a Control Hub, which is then used to configure and manage other components which collect, transform or deliver data. The Control Hub element can be delivered in several different forms:

- As a customer-premise service–software that the customer installs and runs directly;
- The original customer-premise service, but cloud-hosted and managed by StreamSets; and
- A Control Hub implementation designed for the cloud and managed by StreamSets.

The current implementations are commonly referred to as “3.x” and “3.x SaaS,” for the first two, and the StreamSets platform for the last, and our generally preferred solution is that last.

The other components of a StreamSets system, for a given customer, would be installed and run by the customer, and orchestrated by its choice of Control Hub implementation.

Customer Views into Data Pipelines

StreamSets’ services do include the ability for customers themselves to snapshot pipeline data to address configuration and debugging issues. In this case, the data are (temporarily) retained, solely within a data plane engine under the customer’s own management control; the data are not visible to anyone at StreamSets.

StreamSets employs web socket tunneling to allow for this visibility. Note that this feature is entirely under the control of the customer, disabled as a default, and for those customers that cannot allow their data to pass through firewalls, it is possible to implement this as a direct connection between the user’s browser and the data engines: Engine Communication.

For Transformer for Snowflake, a small data sample might be proxied through StreamSets Control Hub. This data sample is never stored or persisted and a customer can disable this feature completely.

What StreamSets Is Able to See

Configuration Data

Configuration data is data describing the customers assets, the pipelines developed by the customer, and administrative settings such as assigned user roles and permissions: what's required to turn this StreamSets instance into this customer's own implementation.

This is sensitive information–it describes the customer's business processes and choices, and identifies assets–and as such is maintained securely within the application. In the SaaS version of the application or the DataOps platform, this will be within a public cloud. Our approach to security within the cloud is described below.

If a customer has elected to host its own implementation of Control Hub, the customer will be responsible for doing so securely.

Usage Data

Metering

StreamSets collects metering data, for billing and other level-of-usage purposes. From a customer perspective, this would help in more easily sizing our services, e.g., to estimate costs in expanding usage.

Telemetry

StreamSets also collects telemetry data, in order to better support the customers' use of our services.

Such data can help to ensure we are developing and maintaining impactful features, e.g. if 70% of customers are using Snowflake or the field rename stages, we would prioritize enhancements or performance improvements for those first.

Telemetry data also allows us to improve overall stage/pipeline/service performance, efficiency, and service uptime.

And How We Secure What We Do

As noted above, the StreamSets DataOps platform is hosted within one of the major cloud service providers, with a backup instance hosted in a second data center. For additional backup and disaster recovery purposes, a third instance is hosted in an alternative cloud provider.

Data in the cloud is secured there (i.e., as data at rest) using strong encryption, with keys managed by StreamSets. Our current minimum requirements would be for AES-256 and RSA 2048-bit encryption. End user-communication with the Control Hub service is in the form of a secure web session (HTTPS) over port 443. Communications are encrypted with transport layer security (TLS)
1.2 encryption.

Communication between the StreamSets platform and a customer's pipeline management resources, such as Data Collectors, is also secured by default as HTTPS traffic.

About StreamSets

StreamSets, a Software AG company, eliminates data integration friction in complex hybrid and multi-cloud environments to keep pace with need-it-now business data demands. Our platform lets data teams unlock data—without ceding control—to enable a data-driven enterprise. Resilient and repeatable pipelines deliver analytics-ready data that improve real-time decision-making and reduce the costs and risks associated with data flow across an organization. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, smart applications, and hybrid integration.

To learn more, visit www.streamsets.com and follow us on LinkedIn.

StreamSets
A SOFTWARE AG COMPANY