DATA SHEET

StreamSets

General Data Protection Regulation

Compliance Starts at Data Ingest

It is likely that 2018 will be known as "The Year of GDPR" as enterprises struggle to comply with its legal requirements or incur massive fines and negative publicity for having violated the regulations.

The General Data Protection Regulation (GDPR) is an EU law that impacts organisations that do business in EU countries or interact with EU citizens, which is to say nearly every major company on the planet.

The following organisations are subject to GDPR:

1. An organisation established in the EU
2. An organisation based outside the EU that either offers goods or services to EU citizens or monitors their behaviour

GDPR goes into full effect across all EU member states on May 25, 2018.

HOW STREAMSTS HELPS

- Watch for Incoming Personal Data
- Alert for New and Changed Data Fields
- Sanitise, Disguise or Discard
- Verify and Audit Actions on Personal Data
- Nothing Written to Disk

Know What You Collect and Where It Resides

The purpose of the General Data Protection Regulation is to give individuals the utmost control over personal data that is collected by organisations. Key provisions include the following citizen rights as relates to their data (from the United Kingdom's Information Commissioner's Office):

- Right to be informed
- Right of access
- Right to rectification
- Right to restrict processing
- Right to data portability
- Right to object
- Right to erasure (a.k.a."right to be forgotten")
- Rights in relation to automated decision making and profiling

In order to ensure these rights, companies must know every location where a user's personal data is currently being stored or processed. It is important to note that the GDPR definition of personal data — as described in Article 4 (1) — is straightforward but also broad, namely:

"Personal data" means any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.

Protection Requires an End-to-End View

In order to fulfill your obligations under GDPR, you first need to know — at all times — what sensitive data you have or are collecting and all the places where it is currently stored, from traditional databases to big data stores like Hadoop clusters, search platforms and even edge systems. Per the regulation, you must also have a mechanism to redact or remove data related to the identity of a specific individual "without undue delay", in response to a request from the individual or termination of their business relationship with you.

While a great deal of attention is being paid to mechanisms that scan and catalog data at rest, you can only have a complete solution by also inspecting and taking actions on the data as it is ingested. You need visibility and control over personal data from the point of ingestion all the way through to its use and being archived.

Manage Personal Data at First Contact

StreamSets has developed the industry's first data operations platform. It provides visibility and control over your data in motion from source to store. While you will certainly need to implement functionality for identifying and handling sensitive data at rest, you can make your life a lot easier, and reduce your GDPR surface area, by **using StreamSets to detect, sanitise, route and track personal data as it is ingested**. You can think of this as a multi-layered defense, where StreamSets protects data as it enters the organisation, complementing the work done by catalogs and other tools that examine your data stores.

Some of the ways StreamSets helps you deal with personal data at first contact include:

Watch for Incoming Personal Data

As you build pipelines that feed batch or streaming data sources into your data architecture, StreamSets scans the incoming data while it is in motion and automatically detects sensitive information. Today you can configure your own personal data patterns to detect; a library of the most common patterns is coming soon.

Alert for New and Changed Data Fields

Due to data drift, data structures can change at the source without notice, adding new fields or potentially modifying existing fields to serve new purposes. To help protect against a previously "clean" source evolving to contain personal data, StreamSets detects the occurrence of data structure changes and alerts you. You can then use StreamSets to route the new data to a well-protected temporary store for further attention, and reprocess the data once you have decided how it should be handled.

Sanitise, Disguise or Discard

Once sensitive data has been detected, you can choose how to handle it: by utilising built-in transformations that can mask, hash or encrypt the data, or perhaps by configuring a custom transformation that tokenizes the personal data before moving it to storage. StreamSets can also be configured to execute routing decisions such as sending personal data to a holding pen for inspection, restricting it to a specific security zone or simply discarding the data without storing it at all.

Verify and Audit Actions on Personal Data

StreamSets collects fine-grained lineage information for each data record that it processes, its origin, waypoints, destinations and any transformations that occurred. Lineage data is particularly useful if there are multiple data systems touching the data and can assist in a compliance audit by verifying that required masking, encryption or anonymisation is occurring *before* the data is stored.

Nothing Written to Disk Means One Less Worry

Whether running on edge or on your existing cluster, StreamSets operates 100% in memory which means there is no risk of personal data being written to disk as part of the data movement. To protect data while it's in motion, StreamSets fully supports the security protocols required by endpoints in the data movement pipeline.

LEARN MORE

To learn more about GDPR, visit www.eugdpr.org.
To learn more about StreamSets visit www.streamsets.com.
To download the open source StreamSets Data Collector, visit www.streamsets.com/opensource.