Securing Apache Kafka: How to Find the Right Strategy

Secure your data to ensure compliance and reduce the risk of breaches

Overview

Enterprises worldwide see the need to access and stream huge amounts of data to generate new digital services, business insights and analytics – essentially, to disrupt and innovate. However, the data landscape has changed dramatically. While it was relatively easy to handle classic data sets, such as orders, inventories and transactions – today we see massive growth in valuable data types, such as sensor data from IoT devices, clicks, likes or searches.

Customer information streamed in real-time is necessary to create a holistic view of customer behaviour in order to feed analytics and even run machine learning and predictive analytics. Kafka solves this problem. Apache Kafka is a distributed, partitioning, and replicating service that can be used for any form of "data stream". It's been making an enormous impact as many organizations from SMBs to large enterprises have started to use this system to organize their data streams.

While Kafka has many advantages in terms of reliability, scalability and performance, it also requires strong data protection and security. Not only is it a single access point to read data streams, it is also the perfect place to implement data-centric security, which protects the data at the earliest possible point, before it is distributed to various other systems where it may be difficult to keep track of.

Why secure Apache Kafka?

Given the fact that Kafka is used to to organize your data streams,
there is often sensitive data passing through Kafka which needs to be secured.
This could be PII, PANs, SSNs, health care records or any other sensitive value.

Main reasons to secure Apache Kafka:

- Ensure and maintain compliance & reduce compliance scope –
  keep consumer systems out of compliance scope
- Protect sensitive data in the confluent platform (Kafka) environment –
  reduce risk of data breaches
- Reduce risk of distributing sensitive data to unprotected confluent platform
  consumers
- Enable secure analysis of sensitive data & secure elastic search

Industry examples

Retail	Enable secure and compliant customer insights & big data environments Secure processing & analytics of sensitive customer data & payment transaction processing
Financial Services & Insurance	Secure & compliant payment transaction processing Improve insight into customer behaviour & optimize risk-management culture by enabling secure analytics of sensitive data Perform secure & compliant fraud detection analytics on sensitive data Secure & compliant data-driven Omni-channel sales, service and customer engagement
Healthcare	Secure omics, clinical, financial and operational data to enable compliant decision-support analytics
Public Sector	Secure and compliant processing & storage of personal data of citizens such as social security numbers

How are you doing it today?

An out-of-the-box Kafka platform setup allows any user or application to write and read any messages in any topic. In the Kafka platform data is plaintext by default. As soon as a cluster starts to handle critical and confidential information, you need to implement security.
But classic protection mechanisms come with disadvantages.

Encryption of data in motion using SSL / TLS:
Data encryption between producers and the confluent platform as well as between consumers and the confluent platform

- Data needs to be decrypted to be useful
- Complex key management
- Increased CPU usage to encrypt and decrypt
- Encryption is only in motion, data still un-encrypted on broker's disk
- Complicates the process of adding new consumers to topics

Encryption at the data layer

- Complex key management
- Security – when a consumer is compromised all of the data can be accessed
- In most cases data format is not preserved

Data protection only at the endpoints/at consumer level
(e.g. VLE – data is protected on Databases via volume level encryption)

- Stream unprotected
- Data needs to be decrypted for usage

Other examples of how to control access and Authentication are protection of the confluent platform network – but not the data itself:

- ACL's (Access Control lists)
- Authentication for communication between brokers and ZooKeeper
- Authorization of read/write operations

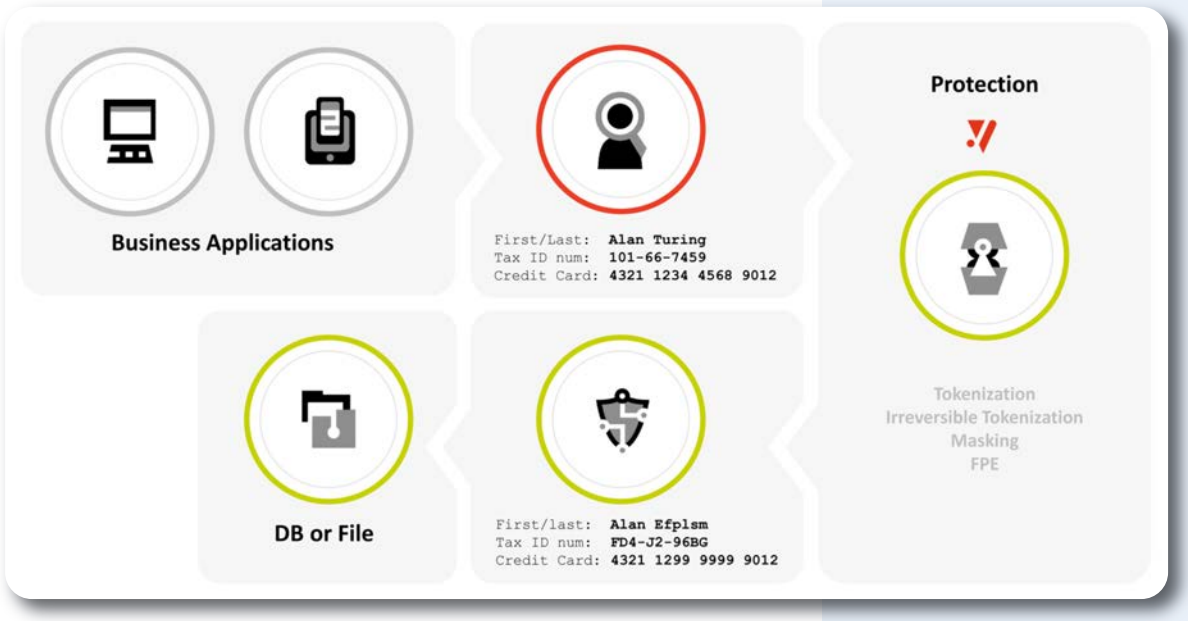
Data-Centric Security

A security solution needs to support Kafka's benefits in terms of scalability, performance, and reliability. Protection mechanisms such as tokenization address the shortcomings of classic security solutions and are essential components of a data-centric strategy. Tokenization protects sensitive data while preserving its original format, giving it referential integrity and resulting in a dataset that is the same size and utility as the original. The de-sensitized data has the identical statistical distribution as the original data, to ensure that all the characteristics and properties of the dataset are preserved. This eliminates the dilemma of having to choose between either security or analytics because data scientists are able to perform analytics and produce reports on the protected dataset.

With comforte's data protection security solution, it is possible to protect sensitive data while retaining its utility.

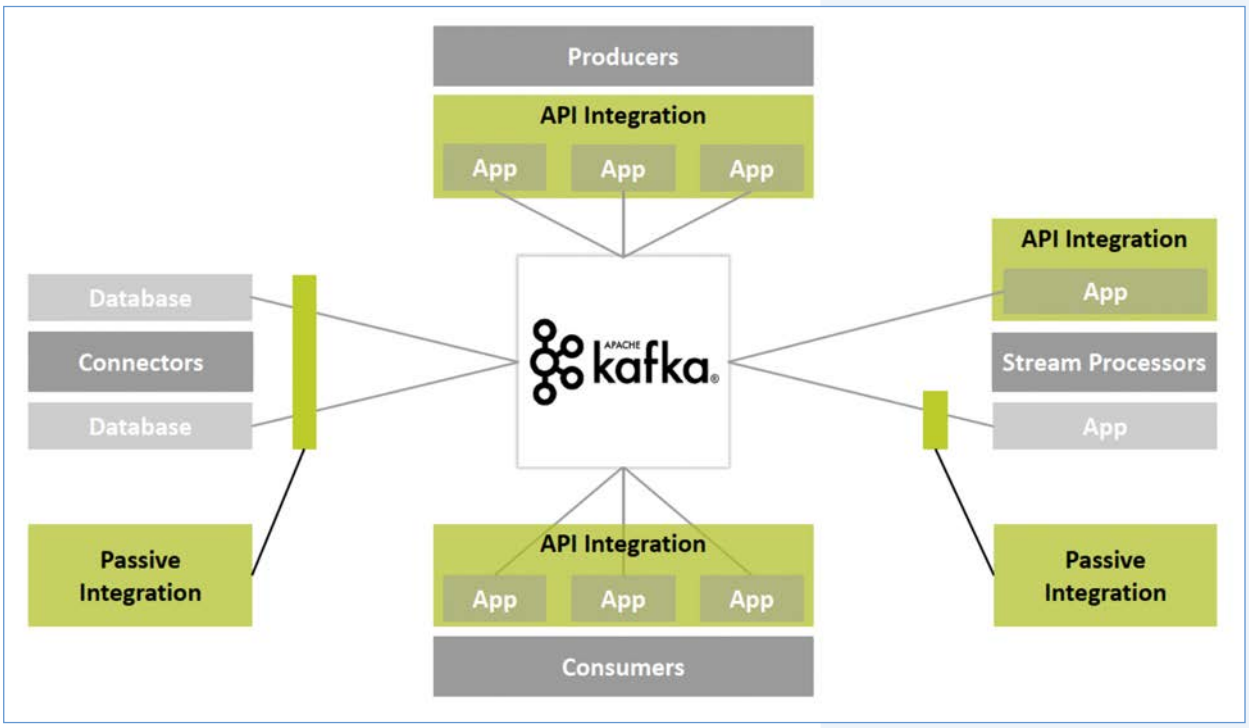
Data Protection Mechanisms	How They Work
Tokenization	Tokenization replaces the original data with a randomly generated, unique placeholder. There is no mathematical relationship between the token and the original data, consequently hackers cannot reverse-engineer it.
Format Preserving Encryption (FPE)	Similar to tokenization and unlike classic encryption, format-preserving encryption (FPE) encrypts the data in such a way that it maintains the same format as the original data, resulting in minimal, often no application modifications.
Masking	Data masking anonymizes sensitive data by creating a structurally similar but inauthentic version of the data. Unlike tokenization and FPE, masking is permanent; that is, it's impossible to reverse it to obtain the original values.

Tokenization removes confidential data from internal systems and big data environments by replacing it with randomly generated data of no exploitable value to cybercriminals.



Implementation

As a general rule with data-centric protection, the sensitive data should be protected as closely as possible to the point of ingestion and then only be revealed at the minimum number of places desirable throughout the enterprise. This strategy ensures that the attack surface and thus risk is as small as possible. Therefore, with Kafka as the enterprise's "central nervous system", data should always be stored in its protected form, and only be revealed when necessary.



Overview diagram: Integration options for data-centric security with the Kafka platform

The protection mechanisms provided by comforte can be easily integrated into Apache Kafka and the platforms based upon it. In the context of Apache Kafka, the relevant points of integration of data protection with SecurDPS are the four Kafka core APIs:

- The Producer API allows an application to publish a stream of records to one or more Kafka topics.
- The Consumer API allows an application to subscribe to one or more topics and process the stream of records produced to them.
- The Streams API allows an application to act as a stream processor, consuming an input stream from one or more topics and producing an output stream to one or more output topics, effectively transforming the input streams to output streams.
- The Connector API allows the building and running of reusable producers or consumers that connect Kafka topics to existing applications or data systems. For example, a connector to a relational database might capture every change to a table.

Implementation

The Application Integration for comforte's data protection solution SecurDPS into Apache Kafka can be done either "passively", i.e. using Transparent Integration, via Translation and Processor modules provided as part of SecurDPS, or, alternatively, via direct integration using comforte's SmartAPI.

For Kafka Producers and Consumers, SecurDPS integration can be performed using the SmartAPI. For Kafka Streams on the other hand, SecurDPS provides a dedicated "passive"/transparent Kafka Streams Integration module out-of-the-box to make integration as easy as possible. This transparent integration option does of course not preclude the alternative to use the SmartAPI for integration of SecurDPS Enterprise into Kafka Streams.

While API Integration using the SmartAPI could in theory also be used for home-grown Kafka Connect modules, the key value of Kafka Connect results from the fact that there is already a large set of official and ready-to-use Kafka Connectors available. To integrate data protection into Kafka Connectors, SecurDPS provides a dedicated transparent integration for Kafka Connect, allowing integration of data protection without the need to change anything in the actual Kafka Connector.

Every new consumer accessing a protected topic is out of scope. This is protection where it needs to be – at the core of your business.

With comforte's end-to-end data protection you can protect the whole stream, independent of your Apacke Kafka cluster. This can be accomplished with no impact on scalability, performance, or reliability of your system and enables you to:

- Replicate data centers without any additional work (no key management): simply replicate protected data – works with active/active and active/passive
- Enable secure multicloud integration: run your streaming data service on the cloud platform of your choice
- Enable secure data streaming between on-premises data centers and public clouds

Benefits

By adopting a data-centric security strategy, enterprises can:

- **Protect sensitive information within big data analytics environments, without impacting the ability to use the data in existing applications and systems**
- **Comply with regulatory mandates, without prohibiting or restricting access to certain datasets containing sensitive information**
- **Prevent costly and reputation-damaging data breaches**

Why comforte?

With more than 25 years of experience in data protection on truly mission-critical systems, comforte is the perfect partner for organizations who want to protect their most valuable asset: data.
comforte's Data Protection Suite, SecurDPS, has been built from the ground up to best address data security in a world that is driven by digital business innovations, empowered customers and continuous technology disruptions.
We are here to enable your success by providing expertise, an innovative technology suite and local support. To learn more, talk to your comforte representative today and visit www.comforte.com.

Find out more

Let's schedule a discovery call to qualify your needs in more detail.
To find out more, talk to your comforte representative today and visit us online at www.comforte.com/