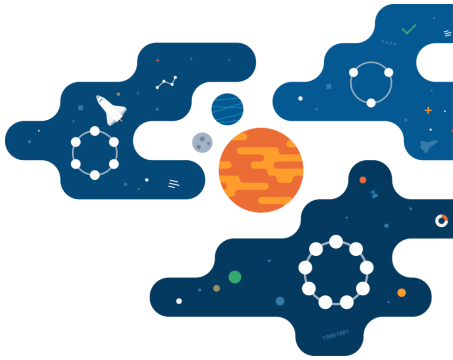
WHITEPAPER



Simplify GDPR Compliance with Apache Pulsar

The General Data Protection Regulation, or GDPR, is a regulation of the European Union that governs the rights of users with respect to the collection, processing and storage of their personal data.

The GDPR applies to personal data for any citizen of the European Union regardless of whether or not the organization collecting that data has a presence in the EU or not. Because the penalties for violating GDPR regulations can be severe, it is imperative that organizations comply with these regulations.

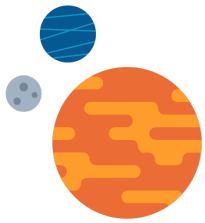
In this white paper we will look at specific aspects of GDPR and offer insights into the capabilities within Apache Pulsar that help you to comply with GDPR. We will examine design patterns that can be taken into consideration as your organization is formulating its approach to GDPR compliance.

Disclaimer: This whitepaper is intended to provide you with an overview and understanding of Apache Pulsar's capabilities for addressing GDPR compliance requirements. The information contained herein does not provide any guarantees that by following the guidelines in this paper that you will achieve compliance with GDPR regulations. It is provided for informational purposes only and does not constitute legal advice.

DataStax

01

Design for Privacy



Personal data privacy begins with you and ***your system design.*** No platform replaces the need for you to build privacy into the design of your software systems.

Therefore, while it is important to know the capabilities of Apache Pulsar, let's first discuss some of the requirements that we will need to take into account as we are building solutions that involve Apache Pulsar.

Right to be Forgotten

Under the GDPR, a user has the right to have their personal data permanently deleted by a data controller. Under GDPR, personal data is defined as:

> *...any information relating to an identified or identifiable natural person (data subject).*

As you can see from the definition, an important aspect of this requirement is how data can be related back to an identifiable person. Later in the paper, we will discuss implications of Pulsar's approach to storing messages and how those pertain to supporting Right to be Forgotten requirements. We'll also cover techniques to support this requirement by focusing on the connection that allows message data to be tied back to an identifiable individual and how you can take full advantage of Pulsar's capabilities while still complying with GDPR.

Right to Data Portability

Under GDPR, users have the right to transfer their personal information to another data controller, over a secure network, in a common machine-readable format. This implies that we must have a way to retrieve personal information for each identifiable user and provide a way to prepare and package that data in a form that allows us to transmit that data to the user upon request.

Later in this paper we will discuss design techniques that leverage capabilities in Apache Pulsar to make this process more straightforward to implement and less computationally intensive.

Lawful Basis and Consent

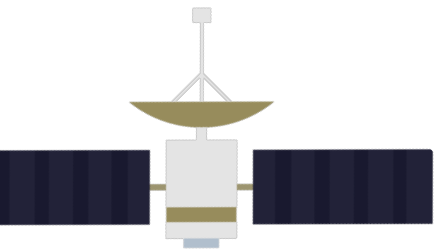
The GDPR requires that businesses have a lawful basis for collecting and processing data. There are two primary areas that we will consider in this white paper.

The first is when you have a situation where you have multiple requirements for retaining personal data for users. A common situation that can arise is that your business may be legally required to retain some data for an extended period of time while others you wish to purge as part of your privacy design. We'll talk about capabilities that Pulsar provides that allow you to handle this situation with simple configurations.

The second is where your user revokes their consent to have their personal data used as part of the processing that they originally agreed to.

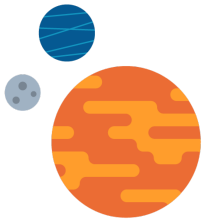
Data Protection

GDPR requires that businesses take the necessary steps to protect personal data. While we'll touch on security and safeguards that Pulsar offers in this paper, we will not go into depth on security best practices as this has been covered in depth previously in the DataStax whitepaper on security best practices for Apache Pulsar.



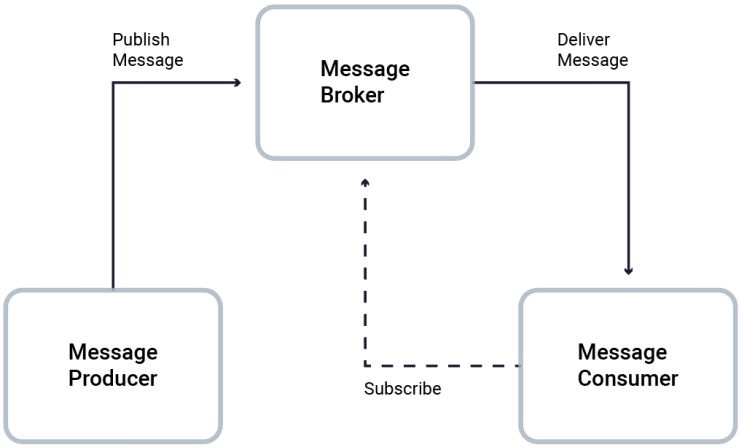
02

Understanding Pulsar's Architecture



Apache Pulsar is a messaging and event streaming platform that can be used to exchange and process data between multiple systems.

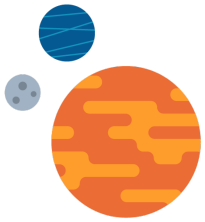
In a messaging system such as Pulsar, there are two primary roles that a system can play: producer or consumer. As the names suggest, a producer creates a message and publishes it to Pulsar. Consumers are interested in reading messages that get published, so they tell Pulsar which messages they want to subscribe to and when a new message is published, Pulsar will deliver the message to the consumer. The component within Pulsar that handles these interactions between the producers and consumers is called a message broker. The general pattern is shown here:



When we examine the internal workings of Pulsar, we will discover that the way messages are stored is with a data structure known as a distributed, append-only log. From a GDPR perspective, the worrisome words here are "append-only". After all, if the only operation you can perform on a log is to append to it (that is, add messages to the end of the log), then how can you comply with Right to be Forgotten requirements which require that personal information be deleted?

03

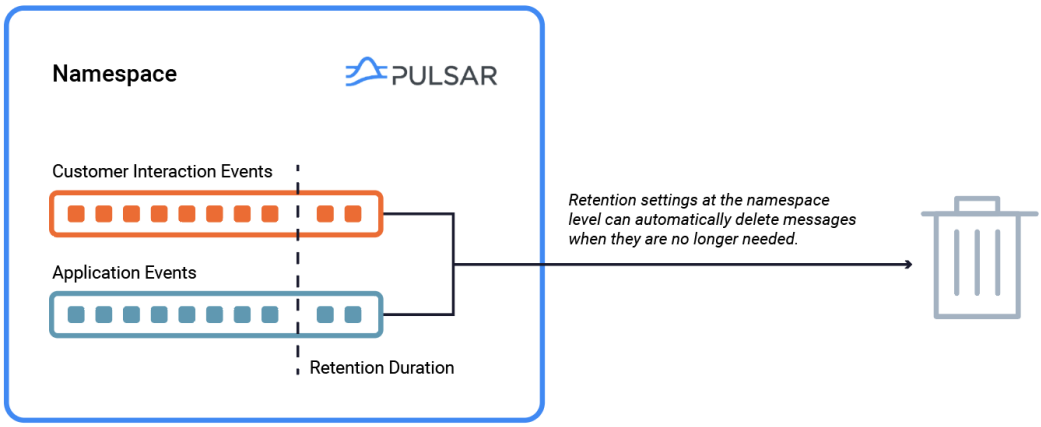
Right to be Forgotten



There are essentially two ways for you to comply with Right to be Forgotten requirements using Apache Pulsar: message deletion and pseudo-anonymization.

As a general privacy design approach, it is a best practice to collect personal data for the shortest amount of time possible to perform the processing that the data was collected for. This is further reinforced by ***GDPR mandates*** as well. For instance, if you visit a restaurant with a long wait to get a table, you might give the host your name and phone number so they can text you when your table is ready. The host may plug that information into a mobile device he or she is using and when your table is ready an automated system may deliver a text message to your phone. Once you are seated, and assuming there was no process for you to give consent for the restaurant to use your data in other ways, the restaurant has no reason to retain your personal information. As a best practice, the restaurant should design their system to purge your data in a timely manner once it's fulfilled its reason for collecting that data, namely to notify you when your table is ready. If your personal data is deleted after you have been notified, the restaurant drastically reduces the chances that a breach would result in the loss of your personal data.

Apache Pulsar allows you to configure data retention behavior out of the box. For situations where there is no need to store data long term for processing, you can simply have Apache Pulsar purge any messages from the system. This is the most simple approach to handling Right to be Forgotten requirements.



The retention capabilities in Pulsar allow you to specify a certain time duration that Pulsar will retain messages for or you can configure Pulsar to retain messages only until they have been acknowledged (processed) by all consumers who are subscribed to a particular topic. By default, Pulsar will not delete unacknowledged messages, so to ensure that messages are purged you will also need to consider setting a TTL on your messages so that Pulsar will remove unacknowledged messages as part of your overall retention strategy.

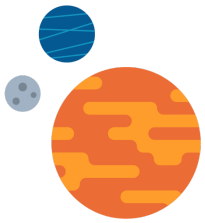
Because GDPR applies only to data that can be connected back to an identifiable person, a common technique used when designing GDPR compliant systems is pseudo-anonymization. Pseudo-anonymization typically involves using a token or identifier that can be used to associate a set of data with the user that data was collected from. Should a user invoke their right to be forgotten, you can simply remove the token associated with the user, effectively anonymizing the data and removing any way to associate the data with an identifiable person.

This can be a highly desirable approach when event stream and message data in Pulsar offers benefits from a data science or analytical perspective. This approach allows us to continue to extract value from this data in a way that facilitates compliance with these important regulatory requirements.

While Pulsar can play an important role in this design, ultimately it is one component of many that work together to form a GDPR compliant solution. We'll look more closely at an end to end solution in the design patterns section of this paper.

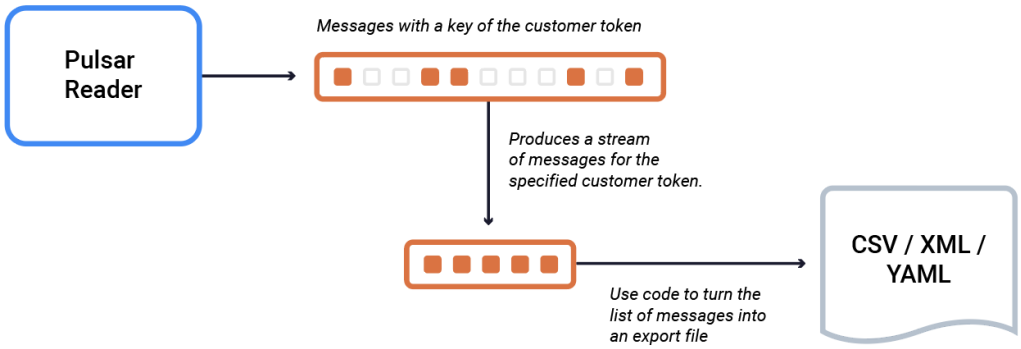
04

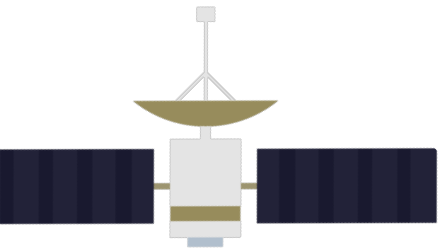
Right to Data Portability



To satisfy GDPR requirements, you must be able to export personal data and make it available to a user. This of course means that we must have a way to clearly identify the data that is associated with a user.

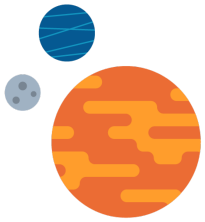
When working with personal data, the easiest way to ensure that you can identify messages that contain personal data for a particular user is to specify the user's identifier as the message key when the message is published. Using Pulsar's Reader interface, you can create a Reader and specify the user's identifier as the message key that you're interested in. You can then use the Reader instance to replay each message associated with the user and write code to export that message data into a machine readable format such as a comma separated file, YAML or XML.





05

Legal Basis



While we normally think about personal data as needing user consent to collect and process, that is only one legal basis that GDPR identifies for collecting personal data.

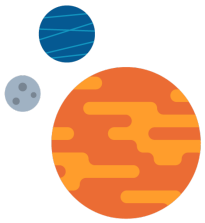
The list specified in Chapter 2, Article 6 of the GDPR regulations lay these out as follows:

- The data subject has given consent to the processing of his or her personal data for one or more specific purposes;
- Processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
- Processing is necessary for compliance with a legal obligation to which the controller is subject;
- Processing is necessary in order to protect the vital interests of the data subject or of another natural person;
- Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
- Processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

While our starting point for privacy-driven design is to retain personal data for as short of a time as possible, there will be situations under which you are permitted, sometimes even legally obligated, to retain messages for longer than would be needed otherwise. For situations where there is a legal basis to retain messages, Pulsar can again be configured with message retention policies that specify the duration for which messages should be retained. One approach to meeting this requirement is covered in the design patterns section below.

06

Data Protection



GDPR requires that organizations take appropriate steps to ensure that the personal data they collect is protected.

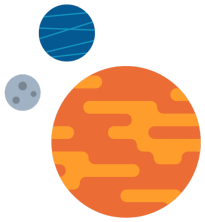
This generally entails the use of industry best practices and standards to ensure the security of personal data from a number of different perspectives.

These include network security, encryption, authentication and authorization as well as internal controls put in place by the data collector to ensure proper data handling.

Apache Pulsar provides support for standard security capabilities used across the technology industry. These include support for TLS to ensure that message publishing and delivery occurs over a secure network channel. They also include end to end encryption capabilities of message data to ensure that unencrypted message payloads are only available to the publisher and consumer and no intermediaries (assuming appropriate cryptographic key access controls are in place). Additionally Pulsar provides standard mechanisms for handling authentication and authorization which allows you to configure the precise level of access a client should have when interacting with Pulsar.

07

Other Important GDPR Considerations



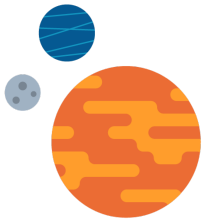
It is important to note that Pulsar must operate within a broader context of compliance within your organization.

There are a number of critical aspects of compliance which are outside the direct scope of Apache Pulsar, but which must be addressed through controls, documentation and other means to ensure full compliance with these regulations. They include, but are not limited to:

- **Communication** - before you publish messages that contain personal data to Apache Pulsar you must ensure that your use case has been appropriately described and communicated to the user in an appropriate form such as a privacy policy or in your terms of use.
- **Consent** - before you publish messages that contain personal data to Apache Pulsar, when required you should ensure that you have obtained the necessary consent from the user to collect and process their data (the process for handling consent revocation is similar to handling Right to be Forgotten, see the Design Example section for more information).
- **Age Requirements** - before you collect personal data you must ensure that you have taken the necessary steps to comply with regulations that govern the collection and processing of personal data for children.
- **Breach Notification** - while Pulsar provides security capabilities and access control to prevent breaches, should there be a breach (leaked credentials, etc.) Pulsar's logging capabilities can help give forensic clues as to the nature of the breach, but ultimately the notification requirements must be addressed through appropriate controls within your organization.

08

Design Example for GDPR Compliance



Let's imagine that you work for an organization that has a series of brick and mortar stores that sell specialty products.

Your products are very popular with your customer base and you have provided them with a web and mobile app that allows them to interact with your business through digital channels as well as in physical locations.

An executive business sponsor wants to fund a new project to improve personalized offer creations for customers. The project scope includes the following features:

- Allow users to find the nearest store using their current location on the mobile app.
- Use beacons within the store to determine if a user visits the store and at what time.
- Deliver a personalized discount offer (coupon) to the user's mobile app with a unique QR code.
- Add a QR code reader to the point of sale in store locations to allow users to claim the offer.

The business sponsor expects the organization's data science team to use the data for customer location and offer redemption to improve the algorithm which determines the best offer to give the customer over time.

Start with Consent

Before we implement anything in Apache Pulsar, we should first start by ensuring that we have taken the necessary steps to get consent from the user. Given this example, this would likely be achieved through multiple steps. You might decide that anonymous users can use the store locator on your mobile app. In that case, a simple device level consent to provide you with their location is sufficient.

However, you may decide that in order to claim offers you will require a user to register for an account. As part of this registration, you will decide what data to collect from the user and likely make the user aware of the location of your privacy policy and terms of service which define the data that you will collect and for what purposes that data will be used. Along with explicit consent from the mobile platform to share their location with your app, your data compliance team has assured you that you have complied with GDPR requirements for obtaining consent.

Data Retention

As part of your solution design, you consult with the data science team and they inform you that they want to track four pieces of information about user activity in this use case:

1. Where was the user when they searched for a store and when did they search?
2. What stores did the user visit and when did they visit?
3. What offers were presented to the user?
4. What offers did the user redeem in store and when?

They explain that they will use this data in aggregate to build an ML model that will help them decide what the best offer is to present to each user. They will look at a particular user's activity history when deciding what offer to present, but they also want to have a composite view of activity to help them present more relevant offers to users with no history.

They explain that they often create new models and replay historical activity as part of improving their models and want to retain user activity for an indefinite period of time.

Solution Design – A Mistake to Avoid

Based on these requirements, you start to think about how to best design this solution. As an initial design (and to simplify this example) you decide to use a single topic in Pulsar called `user_activity`.

You ask as junior engineer to design the structure that this message should take and they come up with something like this:

```
...
{
  "key": "user@gmail.com",
  "payload": {
    "eventType": "search",
    "location": "39.562868, -104.878789",
    "timestamp": "1629577309588"
  }
}
```

Looking at this from a GDPR perspective, you immediately notice that the message key contains personal data that could allow someone to easily tie this location data back to an identifiable user. At this point you consider whether this is a problem or not. You realize that you could use Pulsar's security mechanisms to prevent new clients from consuming this topic to use the data for purposes other than what it was intended. You also realize that you could use Pulsar's message retention policies to eventually delete this data which is helpful.

However, you know that your data science team specifically requested that this data be retained indefinitely to help them refine their ML models over time. If you use this message structure, how would you handle a user who decides to invoke their right to be forgotten? How would you handle a user who revokes consent and whose personal data should no longer be used by the data science team? After all, Pulsar's data storage approach does not allow us to delete individual messages for a given user.

Solution Design – Improved Approach

In the original message design, you realize that the key challenge from a GDPR perspective is that you have tightly coupled personal data that can be used to identify a specific user with the message data pertaining to location and a timestamp.

Rather than use the user's email as the message key, you decide to assign every user in your system a unique identifier. You take a mental note that even having a unique identifier may have some challenges and tell yourself to investigate ways to rotate these identifiers for an added layer of data protection. (Note: this technique is outside the scope of this paper, but it is a common approach you'll find with tech companies that are privy to large amounts of personal data, a simple Google search will lead you to detailed write ups from several companies on their approaches).

Rather than use the user's email as the message key, you can use a unique pseudo-anonymous token that is associated with that user as the message key:

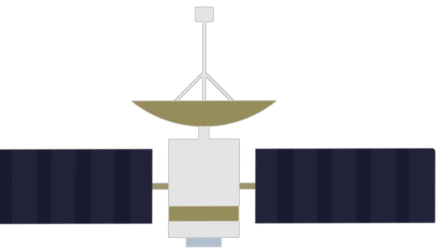
```
...
{
  "key": "E342C11D-CA07-40E2-ACEB-8B44D7F3C673",
  "payload": {
    "eventType": "search",
    "location": "39.562868, -104.878789",
    "timestamp": "1629577309588"
  }
}
```

When a user signs into the mobile app, the response for a successful login can deliver the user's pseudo-anonymized identifier which can then be passed along with subsequent requests rather than personal data.

With this solution in place we now have simple and easy to understand mechanisms we can rely on to comply with common GDPR scenarios:

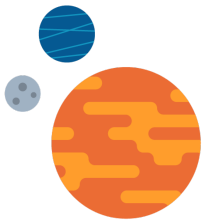
- **Right to be Forgotten** - if a user wishes to exercise this right, we can simply delete the user's pseudo-anonymized identifier and from that point forward we can no longer associate this data with an identifiable person.
- **Right to Data Portability** - if a user wishes to export their data, we can lookup that user's pseudo-anonymized identifier and use Pulsar's Reader API to retrieve a history of their activity using the message key. We can then write code to export this as a CSV, YAML or XML file based on our requirements.
- **Revocation of Consent** - if a user revokes consent, we can disallow lookups to discovery their unique identifier which prevents anyone using this data in a way that allows it to be connected back to the user; depending on the implementation you can create a design that would allow users to toggle consent as well, however that approach is outside the scope of this paper.

Using this approach, we are also able to retain this event stream data indefinitely. The data science team is able to correlate an event stream for a particular user. As long as the user continues to provide consent, we can even use that event stream to deliver more relevant offers and promotions to that user when they use the mobile app, all with an approach that lets us comply with GDPR.



09

Design Pattern—Legal Basis for Extended Retention

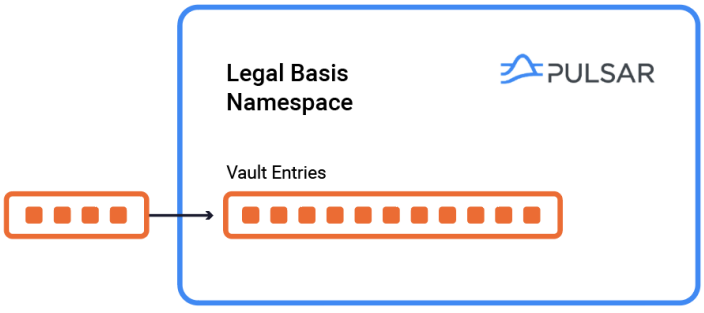


Let's say that you work at a bank and in your country there is a legal requirement that all entries into the bank vault must be recorded and kept for a period of seven years.

Since your bank's policy is to only allow employees into the vault, these vault entry records will be tied back to a specific employee. In addition to vault entry details you have a considerable amount of data about that employee that you have collected as part of the normal ongoing employment relationship.

Of course, employees may decide to leave the company and like anyone else, an employee's data protection rights are applicable under GDPR like anyone else. Therefore, the employee could request that their personal data be forgotten or revoke consent from the company to use their data. However, you have a legal basis to retain entries to the vault for a period of 7 years.

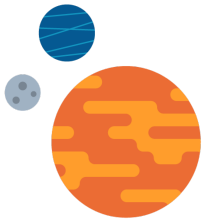
In this case, you can use Pulsar's message retention setting to retain messages appropriately. We can choose to isolate any topics which are exempt from Right to be Forgotten requests into a separate namespace which has a retention policy of 7 years in our example:



As opposed to our previous example where we opted to anonymize the message data using an identifier, in this case we may choose to include personal data directly in the message. In this approach, we can maintain the level of auditing required to comply with our local laws, but still have the option of decoupling other event data from an identifiable person for other uses which are not exempt from Right to be Forgotten requests.

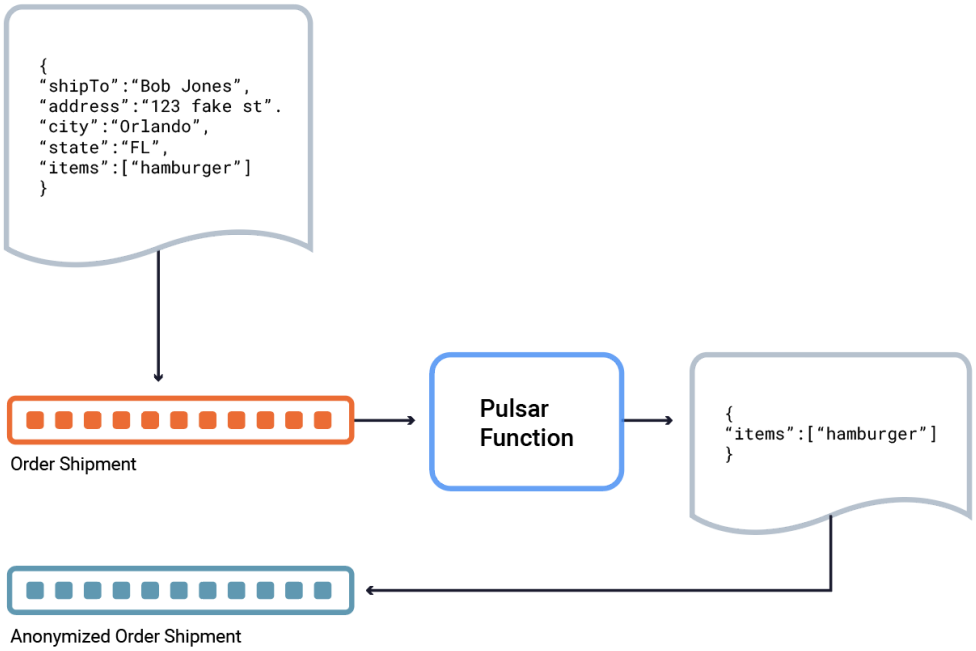
10

Design Pattern—Anonymizer Functions



Sometimes you may have messages that need to contain personal data that you do not wish to retain as a matter of privacy-driven design. However, there may be some information in these messages that would be beneficial to retain from a data science or analytics perspective.

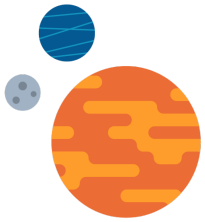
In this case, Pulsar Functions provide you with a convenient method to accomplish this:



Using Pulsar functions, you can use very short lived retention on the message that contains personal data. You can then use a Pulsar function to modify the message payload to remove any personal data from the payload before republishing the message to an anonymized event stream. This gives you considerable flexibility to build streaming solutions in a way that ensures that you maintain privacy and protections for your users while still extracting the value contained within your organization's many event streams.

11

Conclusion



GDPR provides users with strong legal protections for how organizations can use their personal data. Like many regulations, GDPR defines a series of intricate rules and definitions that define your requirements for safeguarding and handling data to achieve compliance. Apache Pulsar has many capabilities that make it an effective choice to use within a GDPR compliant environment, but no product or platform can guarantee compliance on its own. You must work with your organization's data protection officer to build a strategy for compliance that ensures your users' personal data is protected and used in compliance with all applicable regulatory requirements.

Feedback / Questions / Contact

If you have feedback or questions about the content of this whitepaper, we'd love to hear from you! You can reach the DataStax team who is focused on Apache Pulsar by emailing ***pulsar-team@datastax.com.***

© 2021 DataStax, All Rights Reserved. DataStax, Titan, and TitanDB are registered trademarks of DataStax, Inc. and its subsidiaries in the United States and/or other countries.

Apache, Apache Cassandra, and Cassandra are either registered trademarks or trademarks of the Apache Software Foundation or its subsidiaries in Canada, the United States, and/or other countries.