baffle

BAFFLE DATA PROTECTION FOR AI

The Safest Way to Use Private Data with Generative AI (GenAI)

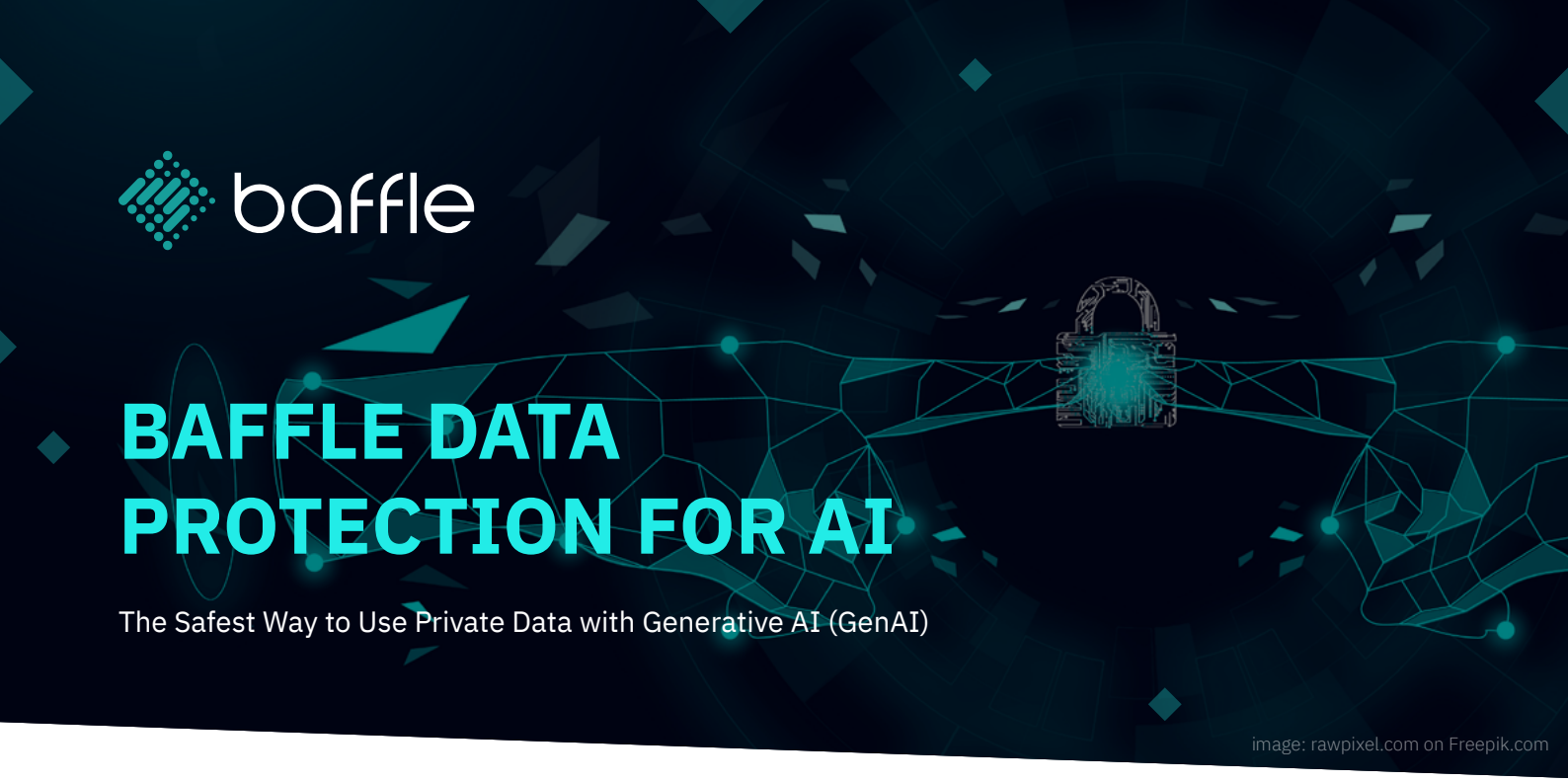


image: rawpixel.com on Freepik.com

GenAI is driving companies to accelerate their plans to leverage private data for insights into their business. Baffle enables AI teams to ensure their GenAI private data sets are secure and compliant with privacy regulations.

Key Benefits

- **Secure:** Regulated data is anonymized on ingest and remains protected in the AI system
- **Easy:** No application changes required to protect data cryptographically everywhere it is used
- **Control:** Role-based access and tenant key controls ensures that only authorized users have access to your data

Use Cases for Augmenting GenAI with Private Data

Large Language Models (LLM) are trained on publicly available data. For enterprise use cases they are missing the context necessary to generate responses that are meaningful to business users. Here are a couple of examples of use cases where private data is necessary:

Example 1: Customer Support

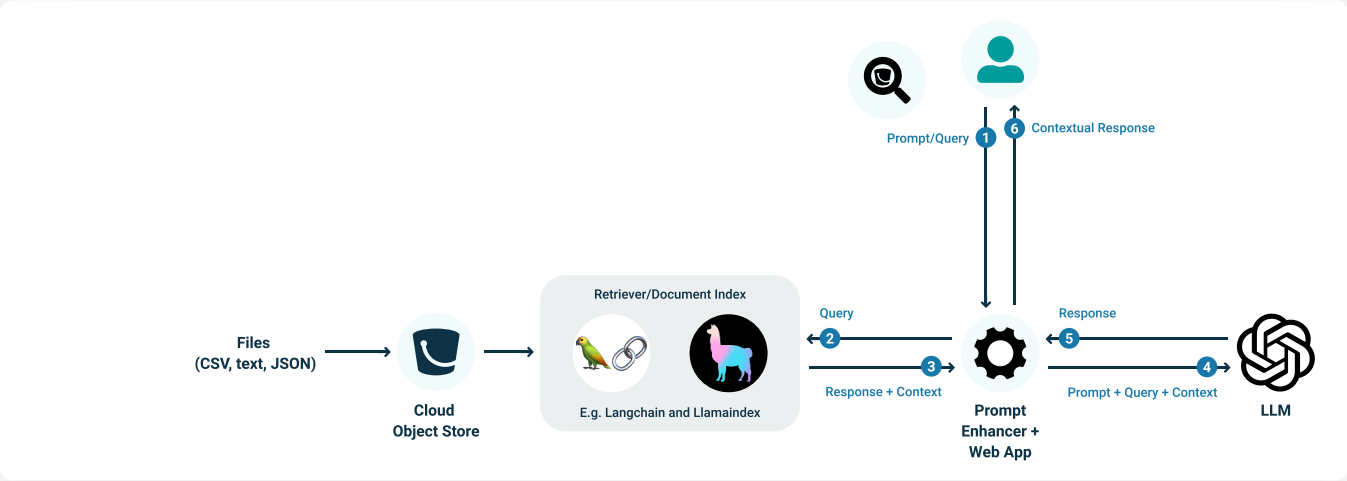
A company wants a GenAI chatbot that can help their support reps diagnose and solve their customers' issues. The LLM would need to leverage all of the information in support cases in the company's support database, including structured, semi-structured, and unstructured data, to provide accurate responses. However, the company wants the names of the customers and other identifying information to be protected.

Example 2: Insurance Agents

A company wants a GenAI chatbot that would guide less experienced agents in recommending policies for customers based on best practices of existing policies for current customers. The data would come from their policyholder database which is primarily structured, with some semi-structured and unstructured fields. Identifying information about policyholders should be protected, either encrypted or masked. architectures such as microservices or serverless PaaS.

Retrieval Augmented Generation (RAG)

To enable organizations to provide LLMs with context from their private data, Retrieval Augmented Generation (RAG) is a mechanism that provides AI models the ability to retrieve facts from an external knowledge base after a user prompt and to use that information to augment its training data as it responds. The augmented context is only used for generating responses and is not ingested into the public LLM model. The external data can be structured or unstructured, and stored in a cloud data store.

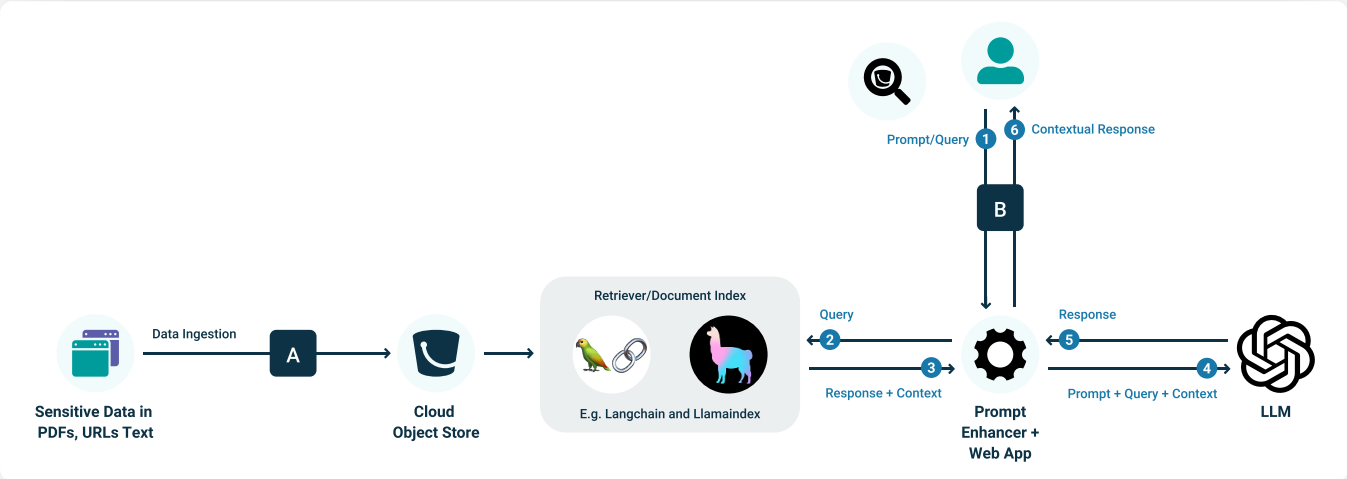


As illustrated above, a special Prompt Enhancer (PE) is the interface between the user and the GenAI system.

0. Private data is collected in a object store and then ingested in a Retriever/Document index
1. The enterprise user enters their request (a “prompt”) into the Prompt Enhancer (PE) app.
2. The PE forwards the query to their private data index which is generated from their private content, which potentially contains sensitive data.
3. The PE receives the Context back from the private index.
4. The PE sends the original user Prompt plus the Context to the public LLM.
5. The public LLM processes the information using the Context (and its own trained data set), and responds back to the PE. The Context is not stored or used to train the public LLM.
6. The PE responds back to the user.

Data Security Issues with RAG

When using a RAG to feed private data into a LLM, there are critical data exposure issues to address.



Data Anonymization (Point A):

Any sensitive data that is in the private data set can be inadvertently exposed in responses to user prompts. Specific challenges include:

Unauthorized users can see private data in clear text

- Infrastructure admins and other privileged users can see private data in clear text, even though they haven't been explicitly authorized to see it

Meeting Compliance requirements, e.g. right to be forgotten

- GDPR requires companies to be able to delete the data of any individual, so a mechanism is need to identify a person's private data and the ability to remove it from the data set

Loss of control once data moves downstream

- Once the data is extracted from the source (consumed by other applications or replicated to other repositories), you can no longer control the exposure of PII and sensitive data

Data Access Control (Point B):

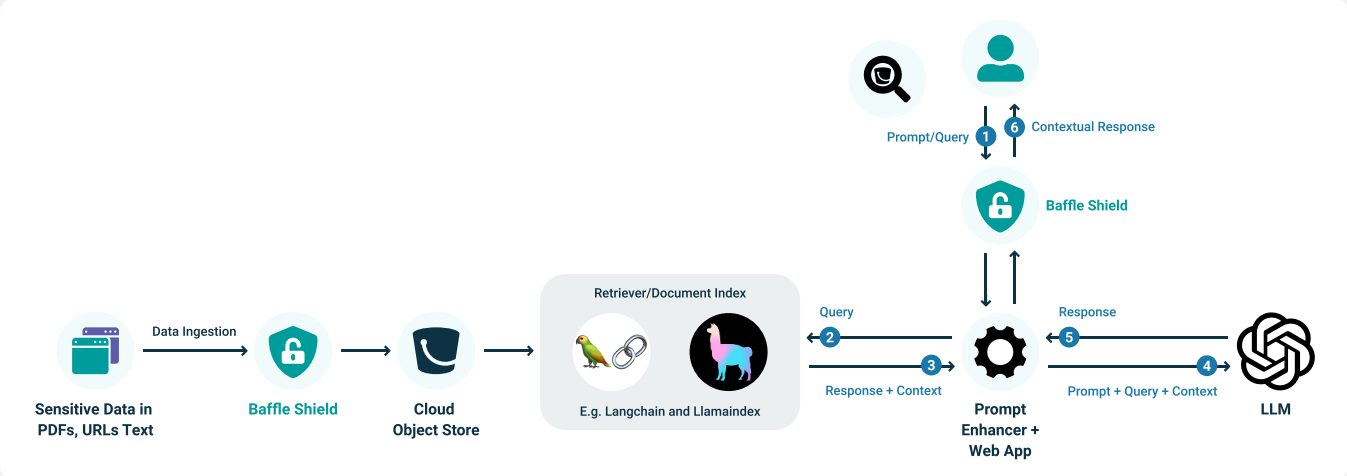
Access to sensitive data should be granted based on user role and corporate policies. Prompts should not reveal sensitive information in any form to unauthorized users. Specifically:

- Data science team and/or any user of the private GenAI service effectively have full access to sensitive data through the prompt/response interface unless security filters are in place
- Even with state-of-the-art security filters, users are readily able to bypass these filters with "adversarial prompting" with new techniques for bypassing filters shared online

Baffle Protects Sensitive Data In Retrieval Augmented Generation Systems

Baffle Data Protection for AI uses a unique proxy based approach to seamlessly secure sensitive data before it is ingested by AI applications. Sensitive data is encrypted or masked with standard AES algorithms. When this data is used in a GenAI service, whether public or private, any sensitive data values are anonymized, so data leakage can't happen.

Organizations can enable GenAI-RAG projects to use protected regulated data while meeting compliance requirements.



Baffle provides data security policy enforcement at two critical points, seamlessly alleviating the challenges of feeding private data into LLMs via RAG.

A. Policy-Based Data Anonymization

Sitting in-line between the source content and the cloud data stores, the Baffle data proxy can anonymize data on-the-fly. Data can be encrypted, masked, or tokenized. Field-level encryption provides maximum flexibility and utility by anonymizing only the data that is sensitive.

Baffle's no-code approach means data can be protected at the source without the need for burdensome application coding and use of SDKs, and the anonymized data can be persisted in any other part of the GenAI system without risk of exposure.

B. Role-Based Access Control

As the responses are returned from the Prompt Enhancer, Baffle can automatically detect which fragments of the data have been previously anonymized. Combining knowledge about who the user is, their role, and previously defined access policies, Baffle can then decrypt the data on the fly and present clear-text (or appropriately masked data) to the user.

Baffle can enforce data security policies to ensure that protected data is decrypted for users with the appropriate access rights. Centrally defined policies make it easy to determine which users can see what data and confirm that compliance requirements for proper controls are met.

Privacy, Security, and Flexibility

Baffle Data Protection for AI allows organizations exploring using GenAI with RAG to balance security with ease of use. Organizations can confidently use private data with LLMs while meeting privacy regulations as well as business needs. The unique architecture offered by Baffle provides easy deployment, flexibility and high scalability for deploying data protection across the enterprise.

Learn More

Get personalized insights and recommendations from Baffle's GenAI data security experts. Schedule a consultation here.

info@baffle.io
https://baffle.io
3979 Freedom Circle, Suite 970
Santa Clara, CA 95054