SafeLiShare

High-Risk Data Escrow and Confidential Controls

Secure AI/ML Model Inferences and Data Access in the Cloud

The case highlights a technology company specializing in transforming biomedical data into actionable insights and knowledge. Their primary business is the development and deployment of artificial intelligence (AI) and machine learning (ML) algorithms to analyze vast amounts of scientific and clinical data, such as medical literature, electronic health records, and genomic data. They offer their products and services to clients, including pharmaceutical companies, biotech firms, academic institutions, and healthcare providers, to help them accelerate drug discovery, clinical development, and patient care.

Privacy-Preserving Challenges

The Medical Academic Center's data contains privacy-protected health information (PHI). PHI includes any information that can be used to identify an individual and their health status, such as their name, address, date of birth, medical records, and test results.

Privacy concerns are an important consideration because the collection and sharing of PHI data must be done in compliance with privacy laws and regulations. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) sets standards for the protection of PHI data and regulates its use and disclosure by covered entities with a business associate agreement (BAA) in place, including healthcare providers, health plans, and healthcare clearinghouses.

To protect privacy, data collection efforts involve collecting only the minimum necessary PHI data needed to provide patient care. In addition, many digital tools, such as mobile apps or websites, collect and share information. To protect privacy, these tools typically use encryption and other security measures to protect the confidentiality and integrity of PHI data.

The Customer

The data escrow company consists of 150 people, with equal parts computer science and biotechnology expertise. The customer has access to high-risk data assets from a well-known research hospital (medical center) and has established relationships with several pharmaceutical companies. They are interested in establishing relationships with other medical centers in the US and abroad. The customer uses multiple cloud platforms, including Microsoft Azure, to store and run their SaaS. The customer is aware of Escrow's value proposition and solves their privacy issue through legal contracts with customers. As the intermediary of high-risk data assets, they understand they will face mounting difficulties as they get more customers and possibly start to mix datasets and provide models to clinical facilities.

Challenges

- Covered entities must ensure data custodianship and visibility that the chosen cloud provider has appropriate security controls in place to protect the data
- Need to keep up with appropriate access controls to prevent unauthorized access to PII or PHI data, both by their own staff and by the cloud provider
- Inaccurate data can cause a model to make incorrect predictions or decisions and affect the accuracy of the model when data is deliberately altered or anonymized
- Ensure appropriate data governance and data management policies per application identities throughout the data lifecycle

Overall, privacy protections are a critical consideration in healthcare data, and healthcare organizations and public health officials must take appropriate measures to ensure that PHI data is collected, used, and shared in compliance with applicable privacy laws and regulations.

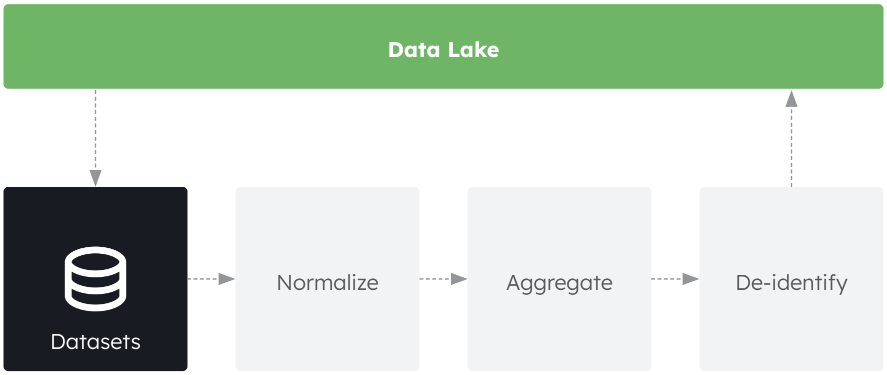
Data De-Identification Procedures

Since healthcare data contains PHI data intrinsically, the customer engages in active data de-identification procedures. Two methods are followed. In the first method, HIPAA regulations identify 18 data attributes, such as a patient's age, address, social security number, etc., which need to be anonymized. In certain kinds of healthcare datasets, such as tabular data, such privacy preserving transformations are relatively easy to implement.

However, many healthcare datasets are non-tabular and datasets may include doctor's notes, medical texts, reports, analysis and insights. In such cases, methods to remove PHI data are often quite complex. The customer in this case uses ML models that process textual and other non-tabular data to remove PHI. HIPAA regulations allow a second method called **expert determination**. In this method, an expert verifies that the probabilistic occurrence of PHI in the dataset is below a certain critical level.

Solutions

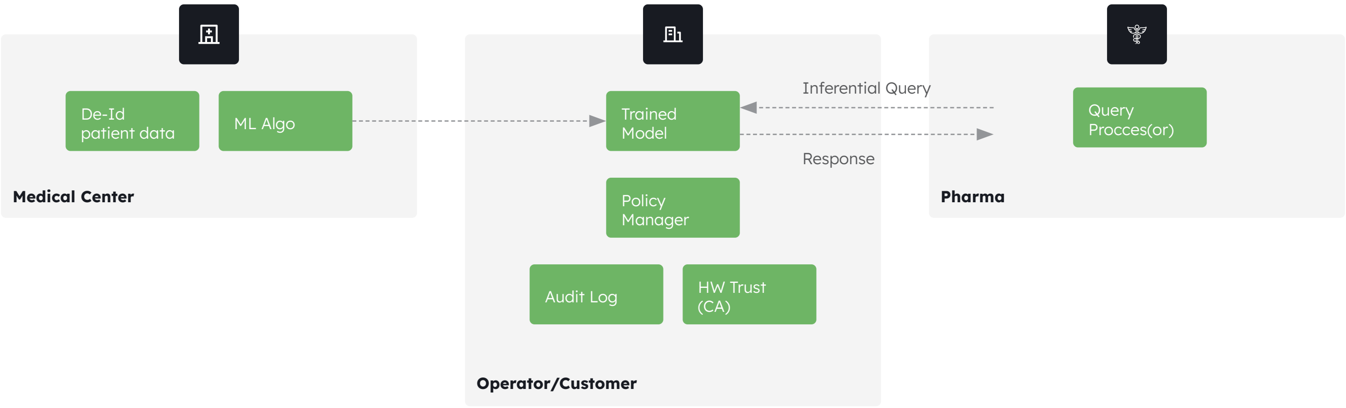
- Provision cloud-based secure enclaves for AI/ML or Federated ML for efficient and accurate data processing without de-identification
- Meet provisions of the Health Insurance Portability and Accountability Act (HIPAA) Privacy, Security, and Breach Notification Rules
- Avoid the need to move large volumes of sensitive data to external computing environments and reduce the risk of data breaches to unauthorized parties



The data de-identification process followed by the customer deals with PHI data and thus needs extra security measures to protect the data. In particular, the network domain in which the de-identification process is performed may be jointly administered by both the medical academic center personnel and the customer's representatives. Once the data has been de-identified, the anonymized data is provided to the customer in the latter's network domain.

Secure Enclave in the Cloud

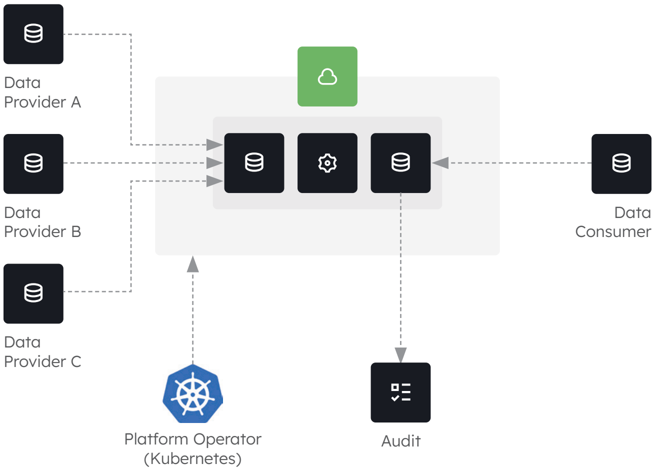
Confidential Clean Room processes created using Secure Enclave technology provide a highly secure computing environment which is isolated and protected from memory and network access controlled through PKI (Public-Private Key Infrastructure) provided by the Secure Enclave technology. These processes are also attested using hardware certificates at runtime, with all input and output being encrypted by default and decrypted only inside the clean room. Furthermore, Secure Enclave technologies are now offered by major technology companies such as Intel, AMD, and ARM, and can be utilized on cloud platforms such as AWS, Azure, and GCP. It is important to note that Secure Enclave technologies offered by different cloud providers differ in their trust guarantees and the operator of the Secure Enclave needs to be cognizant of the differences.



Future Considerations

Bring Compute to Data

The customer is considering providing the trained model to a third party practitioner such as a clinical facility. In this case, the clinical facility data does not need to travel to the customer premises. The data stays with the clinical facility where it can be used to query the model and derive inferences.

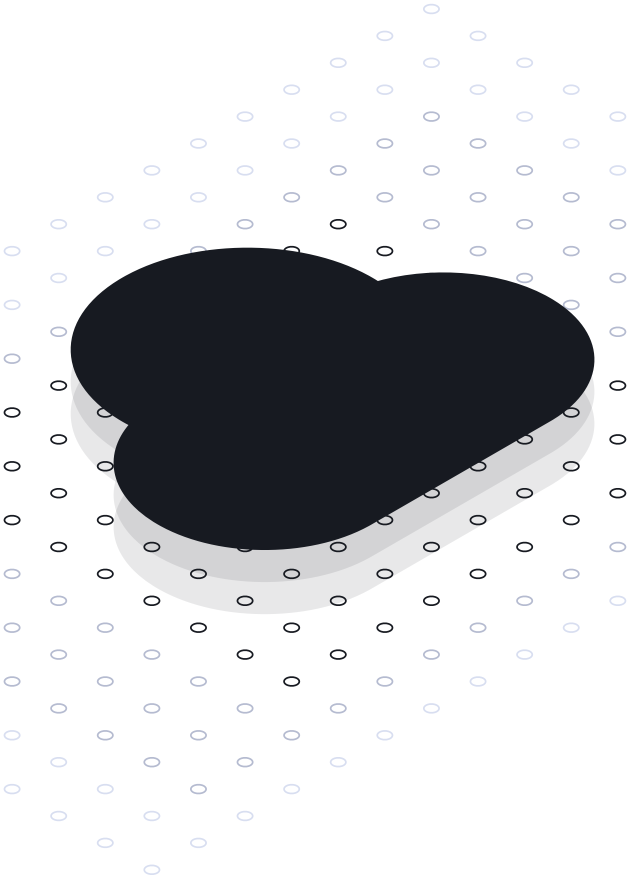


This process protects the data but does not provide protection to the model. It is possible for operatives or malicious party to duplicate or alter the provided model. To protect the model's IP and integrity, the clinical facility will use a Confidential Clean Room. The ensuing process may be tracked and audited with an immutable log provided by the clean room.

Secure Enclaves are a robust solution for secure machine learning, allowing data to be provided by multiple providers, third-party algorithms to be injected, decentralized control to be enforced through policies, and traceable, verifiable audit logs to be maintained.

Confidential Computing is Cloud Computing

The Federated "Confidential Clean Room" solution is a secure data-sharing approach that combines selective patient data from multiple providers in a secure cloud, generates models, and allows data and models to be only accessible inside secure enclaves for inferential queries. User identity is established via cryptographic keys and OAUTH, while logging all data access for auditing purposes. Model provenance and attestation are also maintained inside secure enclaves, enabling the secure sharing of models with external partners.



Summary

SafeLiShare's Confidential Clean Room solution enables federated machine learning and allows data-sharing deals to be made while addressing underlying issues of data ownership, and data and model privacy in secure enclaves. While larger datasets are crucial for convolutional neural networks (CNN) and non-convex models, blockchain technology solves governance issues but introduces performance issues and limits the kinds of models that can be learned. On the other hand, secure enclave technology offers a better solution by addressing governance, data, and model privacy issues without introducing performance concerns or limiting the models that can be learned. Although homomorphic encryption is a promising technology, it is not yet ready for prime time and very costly, making secure enclave technology the best solution for helping with data-sharing deals today.

If you or your covered entities are currently using de-identification or synthetic data to ensure your compliance, it's time to contact SafeLiShare on how you can use Confidential Clean Room technology to process real data with advanced encryption technology and process your AI/ML assets without performance impact. Visit **https://safelishare.com** or contact **info@safelishare.com**.

About SafeLiShare

In an era when data has become the product for many enterprises, and faces increased scrutiny due to tightening global regulations, SafeLiShare was founded with a vision to provide application specific access to data. All operations on data by applications are made visible, auditable, and trackable. Multiple governance policies can be enforced simultaneously. Powered by confidential computing technology, policies drive compliance and governance throughout an enterprises' multiple business functions, data domains, and even across to external business partners engendering a new class of business models based on ownership of data and applications.