Precision Care Challenge 2023

Phase 1 – Idea Submission

Explain The Problem You Are Attempting to Solve

Imagine a world where medical devices protect patient privacy automatically and Health information and log records can be shared instantly across the world to get analysis without being worried about the patient's identity getting exposed, an unbiased opinion towards decentralized ownership of data to build a Secure multiparty computation (SMPC). This is what we believe to achieve.

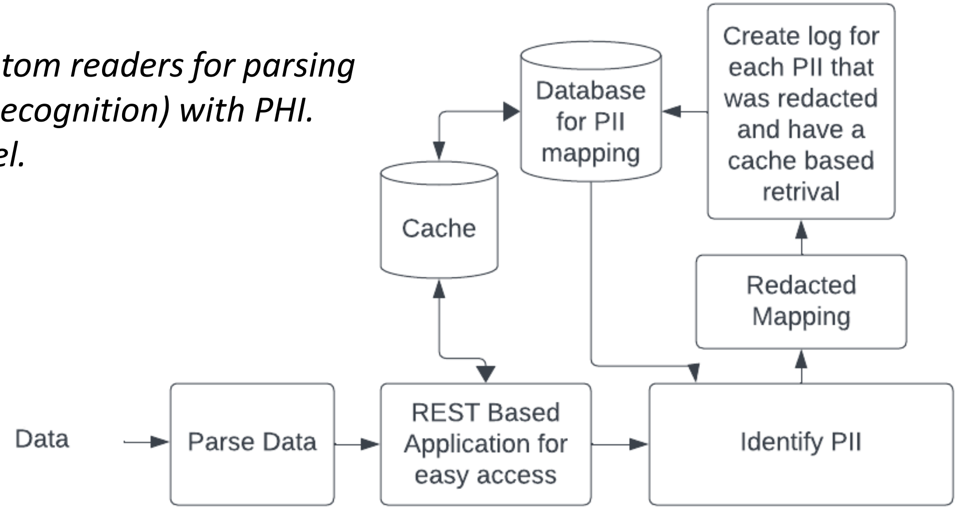
- *One of the biggest cybersecurity challenges is securing Personally Identifiable Information (PII) and Protected Health Information (PHI) on medical devices in on-premises, edge, and cloud environments.*
- *Adherence to legal statutes like HIPAA is vital to avert illicit access, revelation, or recording of confidential patient information.*
- *There are serious security risks associated with directly entering PHI/PII into system logs, which could result in fraud, identity theft, or other nefarious activity.*
- *Because healthcare equipment create large amounts of log data and require automated algorithms to read and analyse logs in real-time, detecting such direct logging is challenging.*
- *Ignoring this problem might have dire repercussions for healthcare organizations—legal ramifications, monetary losses, and harm to their brand.*
- *Trust between patients is crucial, and disclosures of private information can damage that trust and result in lost revenue.*
- *To limit the risks associated with the direct logging of PHI/PII and protect the security and integrity of patient data, a comprehensive strategy combining technical controls, compliance procedures, and continual monitoring and testing is required.*
- *By anonymizing the data itself, the suggested approach adds another level of security, making it difficult for unauthorised actors to identify patients connected to the data—even if they manage to access the logs. The suggested technique has several advantages, including better compliance with laws like HIPAA, increased security as anonymized data is less valuable to attackers, and a lower chance of data breaches because PII is removed from logs.*

Briefly Explain Your Idea/Solution

Imagine a world where medical devices protect patient privacy automatically. *The suggested method attempts to reduce the likelihood of unintentional disclosure from breaches or unauthorised access by anonymizing the data prior to logging it. This approach tackles the problem of protecting sensitive patient data (PII and PHI) on medical devices. Python is being used to build this application**.** Access control measures are the foundation of conventional systems, however they might not always be sufficient because authorised users could unintentionally reveal sensitive data. But putting the idea into practice comes with various opportunities to learn, notably the difficulty of creating and maintaining machine learning models for tasks like data anonymization and Named Entity Recognition (NER). In general, the suggested solution presents a viable method for protecting confidential information on medical devices; nevertheless, to guarantee efficacy and compliance, a thorough evaluation of its advantages and disadvantages must be conducted prior to deployment. We explore a novel idea: anonymizing Personally Identifiable Information (PII) directly on medical equipment before it's even logged. Thus, instantly making sharing and using the data for research. The problem statement can be broken down into simple problems,*

1. *One parse the data from logs, this can be structured and unstructured or custom. Build custom readers for parsing*
2. *Train and finetune necessary Language based ML models(BERT) to do NER(Named Entity Recognition) with PHI.*
3. *Identify the PII by first pre-processing Rule based recognition and then with a trained model.*
4. *Stack results from mixture of models and get the highest performance*
5. *Store the resulting identified PII information and redact the information*
6. *Anonymize the data using the identified PII and use a cache for faster access to rules*
7. *Automatic updation and retraining module for future scrapping of dataset and finetuning.*

YouTube Explanation link : *https://youtu.be/3D0TpPXrRAs*


Data
Parse Data
REST Based Application for easy access
Identify PII
Cache
Database for PII mapping
Redacted Mapping
Create log for each PII that was redacted and have a cache based retrival

Is Your Idea unique or is an improvisation of an existing solution?

Our Idea is Unique as it extensively uses all possible techniques that can be used for NER with fine tuning and stacking to get optimal results.

Explain Your Solution

Data Sources:

- *Unstructured and Structured Logs Data Reader: This reads unstructured data, which could be text files, logs, or other formats without a predefined structure.*
- *REST Based Application: This refers to a software application that uses REST architectural style for communication.*

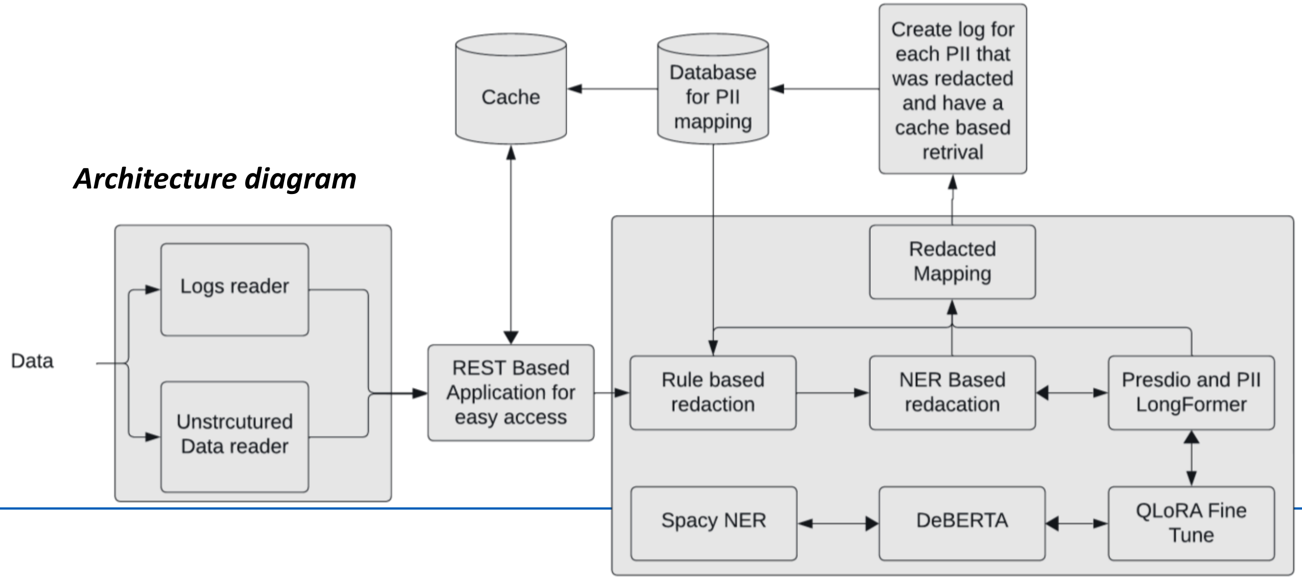
PII Identification:

- *Spacy NER: spaCy is a free, open-source library for natural language processing (NLP) that includes named entity recognition (NER) capabilities. It can be used to identify and classify named entities such as people's names, locations, organizations, monetary values, percentages, dates, times, quantities, ordinal numbers, and cardinal numbers.*
- *DEBERTA: This is a pre-trained transformer model for natural language understanding tasks, achieving state-of-the-art performance in many benchmarks. It can be fine-tuned for specific tasks like named entity recognition. Here the entity to recognise is PII and PHI information from unencrypted logs.*
- *QLORA Fine-Tune: QLoRA (QUery-Less Optimal Rationalization) is a technique for fine-tuning large language models (LLMs) to identify PII*
- *Rule-based Redaction: This refers to defining a set of rules to identify and redact PII data. The rules could be based on patterns or regular expressions.*
- *NER Based Redaction: This uses a named entity recognition (NER) model to identify PII entities in the text, and then redacts those entities.*

Other Techniques:

- *Presidio and PII Mapping: Presidio is an open-source library for information extraction and entity recognition, specifically designed to identify and classify PII data.*
- *Cache for PII Mapping: This stores frequently used PII mappings to improve efficiency. Furthermore, we can combine blockchain technology to make access of PII/PHI remain tamper-proof*

Architecture diagram



Any Additional Information / References

- *Here are the references that are used for creating the solution and mock architecture diagram*
- *https://towardsdatascience.com/custom-named-entity-recognition-with-bert-cf1fd4510804*
- *https://blog.px.dev/detect-pii/*
- *https://www.forbes.com/sites/forbestechcouncil/2023/07/18/the-future-of-personally-identifiable-information-and-health-data/?sh=d8a110124686*
- *https://github.com/dmoonat/Named-Entity-Recognition/blob/main/NER_with_spaCy.ipynb*
- *https://www.kaggle.com/code/emiz6413/rule-based-approach*
- *https://www.kaggle.com/code/awsaf49/pii-data-detection-kerasnlp-starter-notebook*
- *https://github.com/microsoft/presidio*