Sanitizing Clinical Data Using Big Data Frameworks

Objective

This project aims to develop an end-to-end, privacy-preserving clinical data sanitization pipeline using Apache Spark to ensure compliance with healthcare regulations (e.g., HIPAA, GDPR) while maintaining data utility for research and analytics. The pipeline will integrate multiple anonymization techniques, including k-anonymity and pseudonymization, to process structured (EHRs) data at scale.

Motivation

Healthcare datasets (e.g., MIMIC-III, eICU) contain sensitive patient information, including demographics, diagnoses, and treatment records. Traditional data sanitization techniques such as k-anonymity, l-diversity, and t-closeness are often inefficient when applied to large-scale datasets. This project will explore the application of big data frameworks, particularly Apache Spark, to implement these techniques efficiently on clinical datasets.

Project Overview

The project will involve the following steps:

1. Data Collection: Load raw clinical datasets (e.g., MIMIC-III CSV files) into Spark DataFrames. (https://synthea.mitre.org/downloads)
2. Data Preprocessing: The raw dataset will be preprocessed to handle missing values, normalize data, and prepare it for the anonymization algorithms.
3. Sanitization Algorithms Implementation: Classical sanitization methods such as k-anonymity will be implemented in a distributed environment using Apache Spark. These algorithms will be tailored to ensure that sensitive patient information is protected while maintaining the data's utility for research purposes.
4. Performance Optimization: The project will focus on optimizing the performance of these algorithms in a big data environment to improve computational efficiency.
5. Evaluation: Verify k-anonymity compliance and measure re-identification risk.

Expected Outcome

The expected outcome is a robust, scalable data sanitization framework that can handle large-scale clinical datasets efficiently, ensuring privacy preservation while maintaining the utility of the data for research. The implementation will provide a proof of concept that demonstrates the feasibility of using big data frameworks to enhance privacy protection in clinical data analysis.

Tools and Technologies

Apache Spark (PySpark), Python, Java, Ubuntu/Linux, Public clinical datasets (MIMIC-III, UCI Machine Learning Repository).

Conclusion

This project will address a critical challenge in the healthcare industry—protecting sensitive patient data while making it available for research and analysis. By leveraging big data frameworks, the project will provide scalable solutions that can be applied to large clinical datasets, contributing to the broader field of data privacy protection.