PRIVACY SHERLOCK - EFFICIENTLY UNCOVER SENSITIVE DATA

A robust tool for identifying and classifying PII in diverse data repositories

Team Name - kumawat.7
Member - Mohit Kumawat
+917378242131

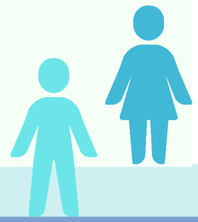
PROBLEM STATEMENT & KEY OBJECTIVES

Problem Statement:

Develop a tool that detects Personally Identifiable Information (PII) across diverse data repositories (databases, cloud storage, file systems) and assesses the associated risk.

Key Challenges:

- Variety of Data Formats: Supporting both structured (CSV, DataFrames) and semi-structured data (JSON).
- Accuracy in PII Detection: Identifying multiple PII types like names, phone numbers, and emails.
- Scalability: Handling bulk data across different formats.
- Anonymization: Replacing detected PII with anonymized values while preserving data integrity.



SOLUTION OVERVIEW

Key Features:

- PII Detection: Implemented using Presidio to analyze DataFrames and JSON objects for sensitive information.
  - Supported PII Types: Names, phone numbers, email addresses, and other common identifiers.
- Data Format Flexibility:
  - Processed data in Pandas DataFrames, converting it into a dictionary format for easy analysis.
  - Supported semi-structured data in JSON format, enabling nested PII detection.
- Batch Anonymization: Anonymized PII across entire datasets (CSV, JSON), ensuring that personal information is securely replaced.
- Skipping Keys for Flexibility: Provided the ability to ignore certain keys in JSON or DataFrames during analysis to focus on specific columns/fields.

Structured and Semi-Structured Data → **Batch Processing** → **Customizable Flexibility** → **Accurate PII Identification**

FUTURE IMPROVEMENTS:

- **Integration with Databases**: Apart from mysql, integration with non tabular data structures for flexible data ingestion and processing.

- **Cloud Storage Integration**: Adding capabilities for processing data from cloud storage services like AWS S3 and GCS.

- **Enhanced Risk Assessment**: Implementing a comprehensive risk scoring mechanism to quantify the sensitivity and impact of detected PII.