IDfy Presents

FRAUD BUSTERS

Think you've got what it takes? Prove it in this ultimate hackathon!

Privacy Sherlock - Efficiently Uncover Sensitive Data

Problem Statement:

Develop a robust and efficient data discovery tool capable of identifying and classifying Personally Identifiable Information (PII) within diverse data repositories, including relational databases, cloud storage services (e.g., Google Cloud Storage, Amazon S3) and file systems. The tool should accurately determine the presence and type of PII in each data point and subsequently assess the associated risk level for the entire database or object.

Key Objectives:

- Comprehensive PII Identification: Accurately detect a wide range of PII types, including but not limited to Aadhaar numbers, PAN numbers, dates of birth, emails, names, Social Security numbers, driving license numbers, medical reports, and credit card information.

- Accurate Classification: To enable targeted risk assessment, categorize identified PII into specific types (e.g., financial, medical, and personal).
- Risk Assessment: Quantify the potential risk associated with the discovered PII, considering factors such as data sensitivity, regulatory compliance requirements, and potential consequences of a data breach.

Requirements

Functional Requirements:

- Data Ingestion: Support the ingestion of data from various sources, including databases (e.g., MySQL, PostgreSQL) and cloud storage services (e.g., GCS, S3).
- PII Detection: Employ advanced techniques (e.g., regular expressions, machine learning) to identify PII patterns within unstructured and structured data accurately.
- PII Classification: Categorize identified PII into specific types based on predefined criteria or machine learning models.
- Risk Assessment: Calculate a risk score for each database or object based on factors such as PII sensitivity, data volume, and regulatory compliance requirements.
- Visualization: Provide clear and informative visualizations to represent PII distribution, risk levels, and potential vulnerabilities.

Technology References:

PII Identification Techniques:

- Regular expressions
- Machine learning algorithms (e.g., decision trees, random forests, neural networks)
- Natural language processing techniques

Data Privacy Regulations:

- General Data Protection Regulation (GDPR)
- California Consumer Privacy Act (CCPA)
- HIPAA (Health Insurance Portability and Accountability Act)

Risk Assessment Frameworks:

- FAIR (Factor Analysis of Information Risk)

Data Discovery Tools:

- Apache NiFi
- Apache Airflow

Machine Learning Libraries:

- TensorFlow
- PyTorch
- Scikit-learn

Judging Criteria:

- The solution should be easy to use and have an intuitive UI. A nice clean working demo is mandatory
- How accurate is the PII identification & classification logic
- At least 1 SQL Database & 1 Cloud Storage Integration should be implemented
- Code quality, extensibility
- Well-defined Service Design Documentation - Sequence Diagram, HLD, API Contracts.