🔐 Project Proposal Privacy Checker

An AI-Powered Data Privacy Compliance & Anonymization Suite

1. Project Title

Privacy Checker
A privacy compliance solution for detecting, anonymizing, and reporting personally identifiable information (PII) in structured datasets.

2. Prepared By

Group 19
INSA Summer Project 2025

3. Problem Statement

In Ethiopia and beyond, the mishandling of sensitive personal data presents major risks — from identity theft to regulatory non-compliance. Organizations frequently lack practical tools to:

- Detect and classify PII in raw datasets.
- Score compliance against privacy frameworks.
- Automatically anonymize or mask sensitive data.
- Export detailed reports for auditing and legal use.
- **Securely share sanitized datasets across local networks.**

This issue is especially relevant for government agencies, financial institutions, healthcare providers, and research institutions handling vast amounts of citizen data.

4. Project Objective

To develop a **modular, offline-capable, privacy compliance toolkit** designed to:

- Detect PII in CSV, Excel, and SQL database files.
- Check data compliance against customizable rules.
- Automatically anonymize data using safe transformation methods.
- Generate and export high-quality PDF and CSV reports.
- Visualize detected risks and compliance trends.
- **Support encrypted file sharing over local networks.**
- **Provide an administrative interface for system control and policy management.**

5. Target Users

- Cybersecurity analysts and compliance officers.
- Data privacy auditors and consultants.
- Government institutions (e.g., INSA, Ethio Telecom, NBE).
- Research centers, universities, NGOs.
- Software developers implementing secure data pipelines.

- **IT administrators managing data compliance tools in secure environments.**

6. Key Features (MVP Completed)

✅ Feature	Description
PII Detection Engine	Uses pattern-based matching to detect emails, phone numbers, IDs.
Customizable Rules	Accepts external `rules.json` for defining thresholds and allowed PII types.
Offline Compliance Scoring	Evaluates datasets based on loaded rules without internet dependency.
Anonymization Engine	Supports `mask` and `hash` techniques to protect sensitive fields.
Multi-format Input	Handles `.csv`, `.xlsx`, and `.db` formats.
PDF & CSV Reports	Generates audit-ready documentation of scans and scores.
PII Dashboard	Visual summaries via Plotly charts (bar + pie).
Unit Testing	Ensures detection accuracy and reliability.

7. Upcoming Enhancements

🔄 Feature	Description
Advanced Anonymization	Field-level control with preview and undo support.
Privacy Trend Dashboard	Tracks scan history and violation trends over time.
NLP-Based Detection (AI)	Detect names, addresses, locations using machine learning.
Scan History Viewer	View and export logs of previous scans with filters.
Custom Rule Editor	In-app rule management with editable UI.
Amharic Language Support	Tailored detection for Ethiopia's official language.
Smart Alerts	Real-time privacy alerts during scanning for high-risk indicators.
Encrypted Local File Sharing	Securely share anonymized datasets over LAN using built-in encryption.
Admin Interface	Central control panel for managing scans, rules, users, and file sharing.

8. Technical Stack

Layer	Technology
UI Framework	Streamlit
Data Handling	Pandas
Visualization	Plotly
Report Generation	FPDF, CSV
Storage	SQLite
Language	Python
Future NLP	SpaCy / Transformers
Testing & Logs	`unittest`, `logging`
Future File Sharing	Python Sockets + AES Encryption
Admin Interface	Streamlit + Session Management (future)

9. National Relevance for INSA

This project directly supports INSA's mission to protect digital infrastructure and citizen data in Ethiopia by providing:

- **Data compliance assurance** in air-gapped (offline) systems.
- **A foundation for local data protection enforcement** aligned with GDPR-like principles.
- **An extensible platform** that can evolve into a national data privacy framework.
- **Tools for secure data exchange within government intranets or LANs.**
- **Administrative control for institutional IT teams to manage compliance systems.**

10. Timeline Overview

Phase	Time Frame	Status
Research & Design	Week 1–2	✅ Complete
MVP Build	Week 3–4	✅ Complete
Feature Expansion	Week 5–6	🟡 In Progress
Final Testing & Documentation	Week 6	
Presentation & Delivery	Week 7	

11. Future Roadmap

- Integrate with **government systems** for automated auditing.
- Develop a **desktop GUI app** for wider public sector use.
- Enable **cloud deployment** with authentication and encrypted data handling.
- Create a **national open-source standard** for data privacy tooling.
- Collaborate with regulatory bodies to **align with Ethiopian data protection laws** under development.
- Implement **LAN-based encrypted file sharing modules** using Python and socket programming.
- Launch a **role-based admin dashboard** for managing compliance scans, audit logs, and sharing privileges.

12. Executive Summary

Privacy Checker, developed by **Group 19** during the INSA Summer Project 2025, is a powerful privacy compliance tool engineered for Ethiopia's evolving digital landscape. Designed for offline use and built on Python with Streamlit, it empowers cybersecurity teams, researchers, and policy makers to:

- Scan datasets for PII with configurable detection logic.
- Score compliance based on custom or predefined privacy rules.
- Anonymize sensitive information using safe, proven techniques.
- Export professional reports for documentation or legal audits.
- Visualize insights via interactive dashboards.
- **Securely distribute sanitized datasets across trusted local networks.**
- **Manage operations through a streamlined administrative interface.**

This tool lays a foundation for national-scale data protection efforts, with its modular design enabling future AI integration, Amharic-language support, and government-grade extensions.

With Privacy Checker, Ethiopia takes a confident step toward digital sovereignty, privacy assurance, and secure data governance.