Sensitive Data Leak Detector for Cloud Storage

GUVI + HCL
Project 2

Student: Mayank Bhushan

Problem Statement

Problem Statement:

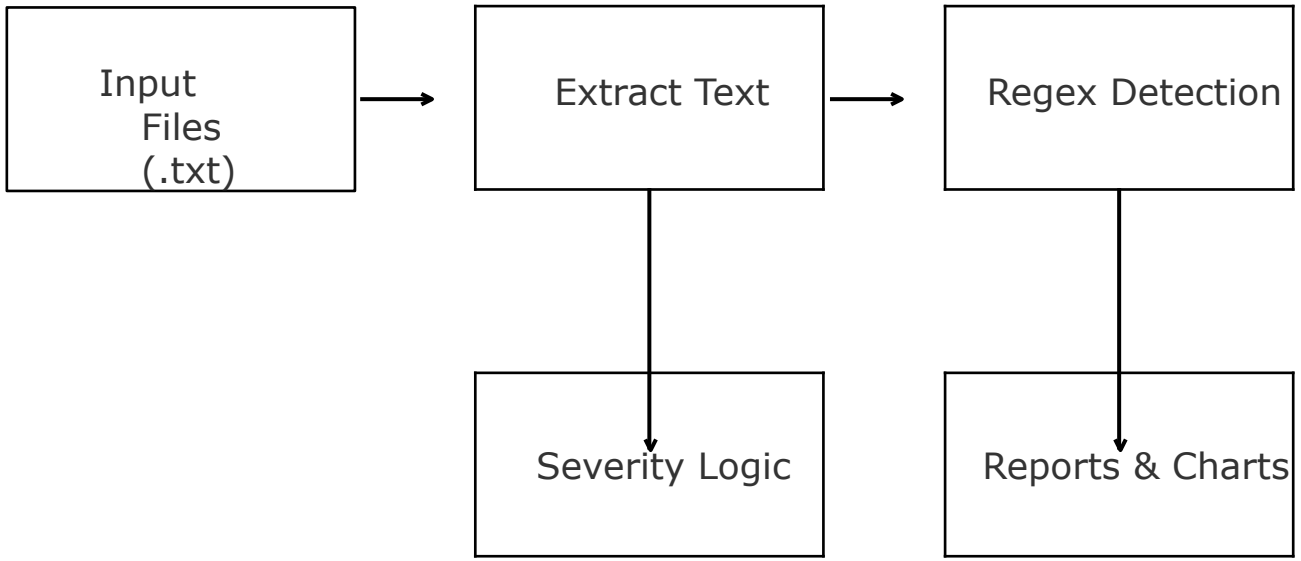
Sensitive data like Aadhaar, PAN, emails, and passwords can leak from cloud storage. Manual review is slow and error-prone. We need an automated tool to detect and classify risk.

Proposed Solution

Proposed Solution:

A Python-based scanner that extracts text from files, runs regex-based detection for Aadhaar/PAN/emails/password phrases, and classifies severity (Critical/Medium/Low). Generates reports and charts.

Architecture Overview



Technology Stack

- Python 3
- Flask
- Pandas (reporting)
- Matplotlib (charts, presentation)
- Optional: pypdf2, python-docx, openpyxl (export)
- Regex (re)

Im p l e m e n ta

- Pat ter n s: Aa dh aa r ,

t i o n

P A N , e mail, password phrases

≥3 emails) > Low > None

- Severity rules: Critical (Aadhaar/PAN) > Medium (password phrase or
- Outputs: results.csv, results.json, severity_chart.png, results_table.png
- Easy to extend to PDFs/DOCX and a Flask UI

file	severity	aadhaar_hits	pan_hits	email_count	password_phrase
ector/data/sample_clean.txt	None	0	0	0	False
ctor/data/sample_leak_1.txt	Critical	1	1	1	True
ctor/data/sample_leak_2.txt	Medium	0	0	3	True

Future Scope

- Add NLP/ML for smarter detection and fewer false positives
- Integrate with S3/Azure/GCS for real-time scans
- Build a web dashboard with risk trends and alerts

Conclusion

An automated detector for sensitive data in cloud files that is simple, fast, and extensible. It produces clear severity reports and visuals to support proactive data security.