Processed 1.5 Millon Documents with Aadhaar Masking Solution



Client: Choice Finserv, India

A Leading Financial Service Provider

Introduction

Choice Finserv, a leading financial services provider, faced the challenge of complying with Reserve Bank of India (RBI) guidelines for masking Aadhaar details within 1.5 million documents stored on Amazon S3. These documents included loan files, credit reports, and bank statements. The RBI mandates that only the last four digits of the 12-digit Aadhaar number be visible, and the QR code must be obscured to protect sensitive customer information.

The Challenges

This project presented several key challenges:

- Massive Data Volume: Processing 1.5 million documents efficiently required a highly scalable and performant solution.
- Diverse Document Types: The dataset included various file formats (PDFs, images, text) requiring robust file classification to accurately identify and mask Aadhaar details without impacting other sensitive information.
- Security Constraints: The processing environment had limited internet access, necessitating an offline-friendly approach with minimal reliance on external resources.
- Accuracy Requirements: Precise masking was crucial to ensure compliance. False positives (incorrectly masking non-Aadhaar data) could degrade document quality, while false negatives (missing actual Aadhaar data) would result in non-compliance.
- S3 Download Efficiency: Downloading a large volume of data from S3 efficiently was critical for project timelines.

The Approach

- The solution was deployed on a high-performance server (16 cores, 32 GB RAM) with an Ubuntu operating system
- AWS CLI with read-only credentials was used to efficiently download documents from S3.
- Documents were processed in parallel using Python multiprocessing, with the AI model analyzing each document to identify and mask Aadhaar details.
- Rigorous quality checks were implemented to ensure accurate masking and validate the integrity of the processed documents.
- Masked documents were stored in a secure location, ensuring compliance with data security and privacy regulations.

The Solution

- Trained an AI model to accurately detect 12-digit Aadhaar numbers within diverse document types.
- Intelligently categorized documents, processing only those containing Aadhaar information.
- Effectively masked the first eight digits of the Aadhaar number and obscured the QR code while preserving document integrity.
- Packaged the solution into a Docker container for offline deployment and consistent performance.
- Utilized multi-threaded downloads from S3 and optimized batch processing to accelerate document processing.

TECH STACK

PYTHON

The Outcome

- The project was successfully completed within one month.
- The solution achieved a 95% accuracy rate in masking Aadhaar details.
- The solution ensured full compliance with RBI guidelines by effectively masking sensitive information and maintaining data integrity.
- The optimized solution efficiently processed the large volume of documents within the stipulated timeframe.

The Learnings

- Careful consideration of server resources (CPU, RAM) was crucial for optimal performance.
- Packaging all necessary libraries and dependencies within the Docker image simplified deployment and ensured consistent performance in the offline environment.
- Utilizing multi-threaded downloads significantly reduced the time required to retrieve documents from S3.
- Continuous refinement of the AI model was necessary to address edge cases such as low-quality images or unusual Aadhaar formats.

The Conclusion

This project successfully demonstrated the effectiveness of an AI-powered approach to address complex data compliance challenges. By combining advanced AI techniques with a robust and scalable infrastructure, we delivered a solution that met the stringent requirements of the RBI while maintaining data integrity and security. This successful implementation provides a valuable framework for addressing similar compliance challenges across the financial services industry.

Visit us at: www.habilelabs.io Contact us at: +91-8769611996

© Copyright @2025 HabileLabs Pvt. Ltd. All Rights Reserved.