Automating the Digitization of Handwritten Forms

A union with thousands of members in the United States was confronted with a challenge. The union would solicit handwritten responses on printed forms from their members during various in-person meetings. They needed a way to digitize the forms and inject responses into their online spreadsheets without tedious and time-consuming manual transcription.

Union management approached MTech seeking assistance with automating the digitization process. The process needed to ensure careful authentication for two sets of users each of which had distinct access authorizations to perform different tasks.

The first set of union members would be authorized to merely submit the handwritten forms without being able to retrieve any other member's forms; this would ensure privacy of personal identifiable information (PII).

The second set of of union members, consisting of union management, would be authorized to view the original forms, and their corresponding digital versions (digital twin).

After a review of the initial specifications, MTech evaluated multiple architectures with different capabilities, development costs, and operational costs. After union management selected an architecture, MTech developed a project plan involving multiple milestones that would incrementally deliver ever-increasing functionality. The incremental approach would permit union management to evaluate the progress of the new system, and evaluate whether to proceed to the next milestone without risking payment for the full system up-front.

Requirements and Architecture

User Interface and User Authentication

Union management deliberated on the type of user interface that they would expose to the two sets of users. The users were various locations with different devices and different technical capabilities. After careful consideration, union management decided to present the automation system as a web application so that the users could easily access from any modern web browser on any device.

Furthermore, as the two sets of users required different functionality, they would each require distinct web applications.

The front-end of the web application is delivered through AWS CloudFront to provide a low-latency response for the users via its rapid content delivery network. The first line of defense makes use of CloudFront's geolocation filtering to limit access from selected locations where union members were permitted to connect from. The next line of defense makes use of AWS Cognito providing secure, frictionless user identity and access management (IAM).

Uploading of Scanned Forms

The web application for the first set of users permits them to upload scanned forms (i.e. digital scans or digital photographs) to a secure and ephemeral AWS S3 bucket coordinated by a serverless AWS Lambda function.

Securing access to the uploaded documents is paramount: download access from the AWS S3 bucket is technically impossible by design of the architecture and by design of the access restrictions in AWS which are configured to deny by default.

Automation and the User Experience

Each upload of the document triggers, via AWS CloudWatch, its own independent and parallel automation process handled by an AWS Step Functions state machine. As each independent process could take several minutes, it is desirable to provide the user with a responsive feedback regarding the automation's progress. That is accomplished by the automation process posting status updates to AWS DynamoDB which are subsequently relayed to the user's web application via a separate AWS Lambda function.

Text Extraction and Synthesis

AWS Textract is quite capable of extracting text content including handwritten text. Nevertheless, there are several deficiencies that require post-extraction enhancements. The first deficiency pertains to the raw handwritten text which is sometimes illegible. A second deficiency pertains to whether the handwritten text is assigned by AWS Textract to the appropriate question. A third deficiency pertains to whether the handwritten text can be validated against expected information (e.g. name, location, etc.)

Addressing those challenges requires extensive auto-corrections, re-interpretations, and synthesis with other information already know by union management in order to make more meaningful interpretations of the responses. Upon completion, the "digital twin" of the form is recorded within AWS DynamoDB table along with accuracy confidence levels as to each piece of interpreted text.

Integration with Online Spreadsheets

Each "digital twin" response is subsequently directed according to the confidence-level of the text extraction accuracy. Those responses with high-confidence accuracy are copied over to an online spreadsheet to integrate with other business processes used by union management. In contrast, those responses with lower-confidence are copied over to another spreadsheet to be scrutinized and corrected by union management.

Data Sovereignty and Cost

All of the scanned documents and processing are maintained within the union management's AWS account. Access to the data is restricted to union management whose access is permitted only by proper authentication with AWS Cognito.

Furthermore, as the various system components are configured to be serverless and consume minimal storage, the system incurs practically zero cost while not in use.

Deployment into AWS

Testing the various versions of the the system without affecting the current version in production requires a method of provisioning new and updated infrastructure reliably and quickly. Towards that end, the entire system is defined in AWS CDK which is a framework for defining cloud infrastructure in code and provisioning it through AWS CloudFormation. It further provides the ability to audit the evolution of the infrastructure over time.

Outcome

After internal testing, the automation system was activated and deployed to the field. The benefits were immediate and notable.

- Union members were able to easily upload responses from anywhere in the field.
- The responses were
  - automatically corrected and cross-referenced,
  - integrated with subsequent business processes, and
  - available to union management within minutes.

Across the thousands of responses uploaded from the field, the digitization system has proven to be fast, reliable, increased productivity, all while incurring near zero-costs while idle.

If you'd like help with your next project, reach out to discuss.

Get it done right
Get MTech