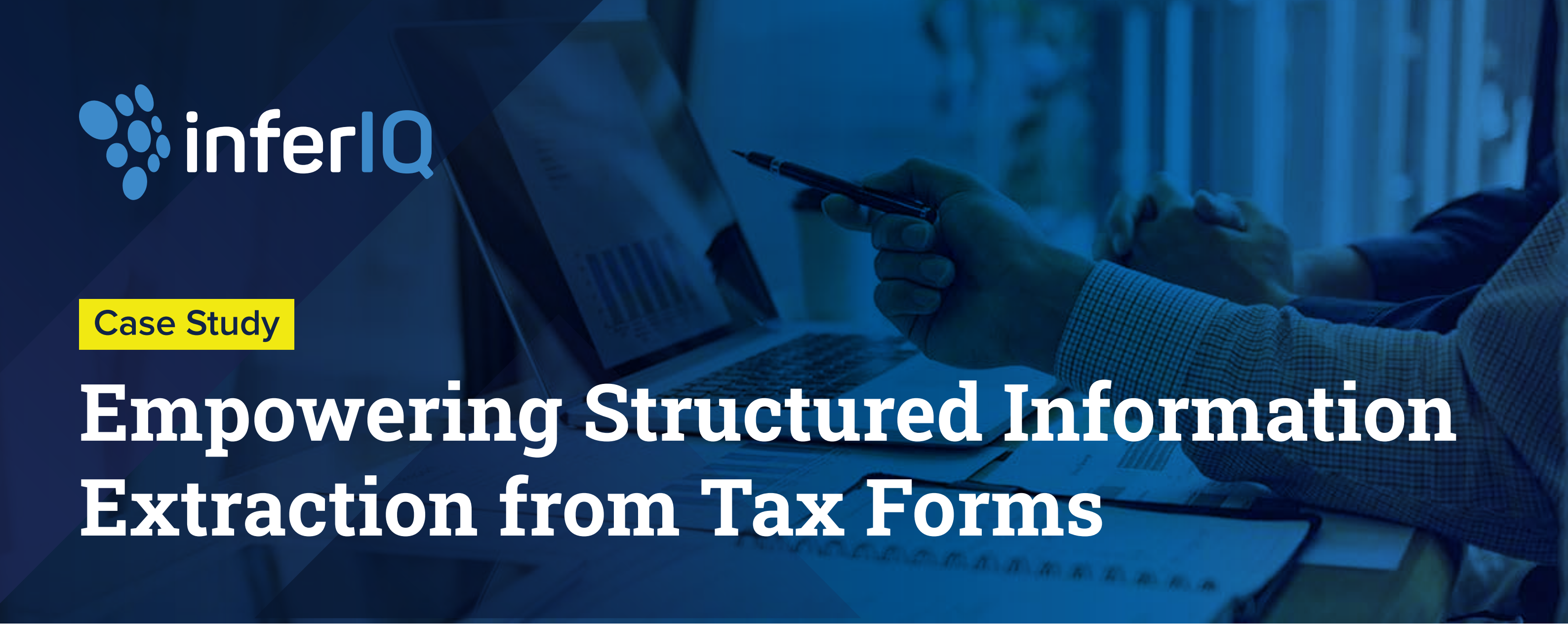


inferIQ

Case Study

Empowering Structured Information Extraction from Tax Forms

According to Grand View Research, the global Intelligent Document Processing (IDP) market was valued at USD 2.3 billion in 2024 and is projected to reach USD 12.35 billion by 2030, growing at a CAGR of 33.1% from 2025 to 2030. As unstructured document processing becomes increasingly critical in the financial industry, organizations are challenged to implement efficient methods to extract data into structured formats. A key document type processed in this solution is the standard US tax return form. Idexcel developed and applied a solution to enable data extraction from these critical tax forms, while also reducing the time required to perform the task.

CHALLENGE

We are now in an era of rapidly growing data, where many available data sources exist in unstructured or semi-structured formats. These come in multiple forms, such as text, images, audio, video, blogs, and websites. Additionally, they span diverse domains including social media, finance, legal, and healthcare. According to an article published in March 2019, the International Data Corporation (IDC) estimates that 80% of global data will be unstructured by 2025.

The sheer volume and complexity of this data present major challenges that organizations must overcome to stay competitive. In response to this need, the use of machine learning techniques and the demand for extracting information from unstructured documents have increased significantly. Information Extraction (IE) refers to the automated retrieval of specific data related to a particular topic from unstructured or semi-structured documents. Structured data extracted through IE can be used for further analysis to support key business decision-making by stakeholders.

Our client in the financial services industry approached us to architect a solution that would enable their team to extract data from standard US tax return forms. The specific data retrieval involved extracting key-value pairs from form fields and tables in digital or scanned PDF documents. Since the dataset included scanned PDFs and image files, the solution also required a tool to perform Intelligent OCR (Optical Character Recognition).

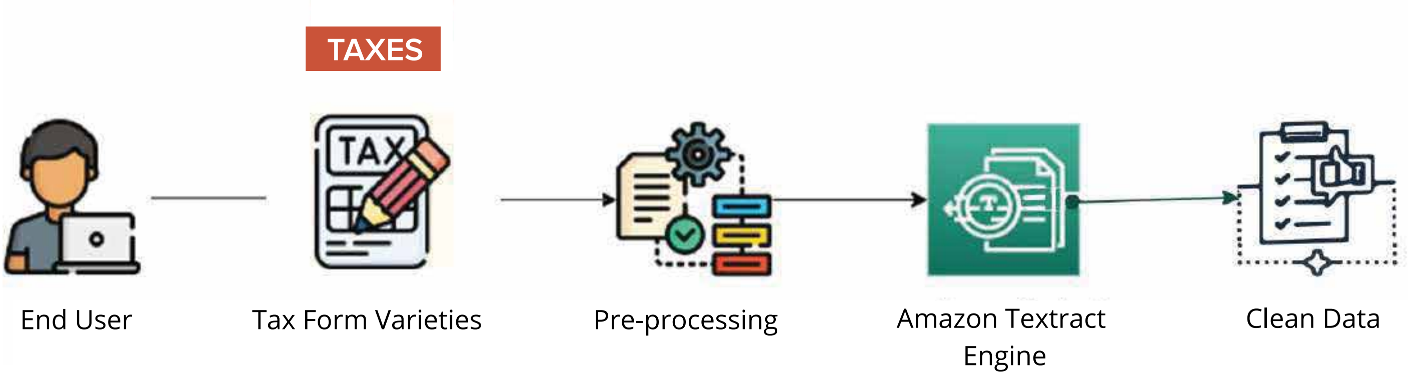
SOLUTION

After a thorough assessment of various tool options, we built a solution called InferIQ on top of AWS Textract.

Key Functionalities of Our InferIQ Solution:

- ✓ Pre-processing to enhance image document clarity.
- ✓ Extraction of Key-value pairs from Form Fields.
- ✓ Extraction of Content from Tables and Check-box Fields.

The extracted data was exported to a .csv file, which serves as input for downstream data analytics workflows. An overview of the procedure is given below:



BENEFITS

Faster Processing Time: Approximately 10 minutes were saved per file using the InferIQ extraction solution.

Improved Accuracy & Efficiency: The solution efficiently extracts content from form fields and table fields with reduced errors. It also supports commonly used document formats, including digital and scanned PDFs.

Improved Document Output: This approach enables structured output in .csv and other suitable formats. The output .csv files are rich in data and ideal for downstream data analytics workflows.

Easy Integration: The solution can be integrated into any loan management or financial decision-making system running on a different technology stack via microservice architecture.

OUR AWS COMPETENCIES

aws PARTNER

- DevOps Services Competency
- Financial Services Competency
- Migration and Modernization Services Competency

Contact us

Idexcel, Inc.
459 Herndon Parkway Suite 10, Herndon, VA 20170
Tel: 703-230-2600 | Email: info@inferiq.ai