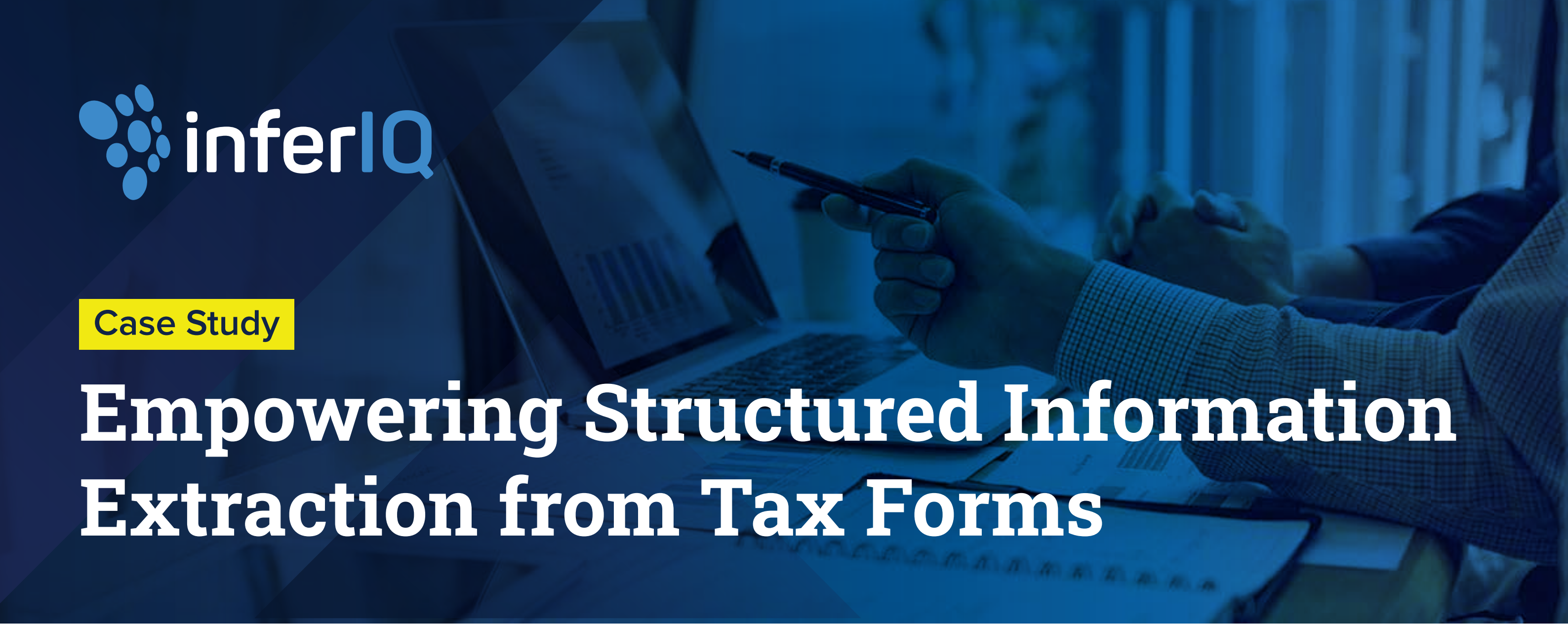


Case Study

Empowering Structured Information Extraction from Tax Forms

According to Everest Group, the Intelligent Document Processing (IDP) market reached $250-350 million in 2018 and is expected to grow at a compound annual growth rate of 70-80% through 2020. Approaching US $1.1 billion this year, this market continues to see growth potential. As unstructured document processing is critical in the financial industry, organizations are challenged with implementing an efficient way to extract data into a structured format. Key documents processed in this solution are standard US tax return forms. Idexcel built and applied an InferIQ solution to make the extraction of data from these critical tax forms possible, while also achieving a shorter processing time to perform the task.

CHALLENGE

We are now in an era of fast-growing data, where many data resources available are in unstructured or semi-structured formats. These come in multiple formats like text, images, audio, video, blogs, websites, etc. There are also diverse sources like social media, financial, legal, and medical domains. 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 of data, in combination with its complexity, presents major challenges that organizations need to adapt to in order to remain competitive. In response to this need, the popularity of machine learning techniques and the process for information extraction from unstructured documents has increased tremendously. Information Extraction (IE) is the automated retrieval of specific information related to a particular topic from unstructured or semi-structured documents. Extracted structured information can be used for further data analysis to guide stakeholders in key business decision-making.

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 extraction of the key-value pairs from form fields and tables from digital .pdf or scanned .pdf documents. Since there were scanned .pdf documents and image files in our dataset, the solution required a supporting tool to perform Intelligent OCR (Optical Character Recognition).

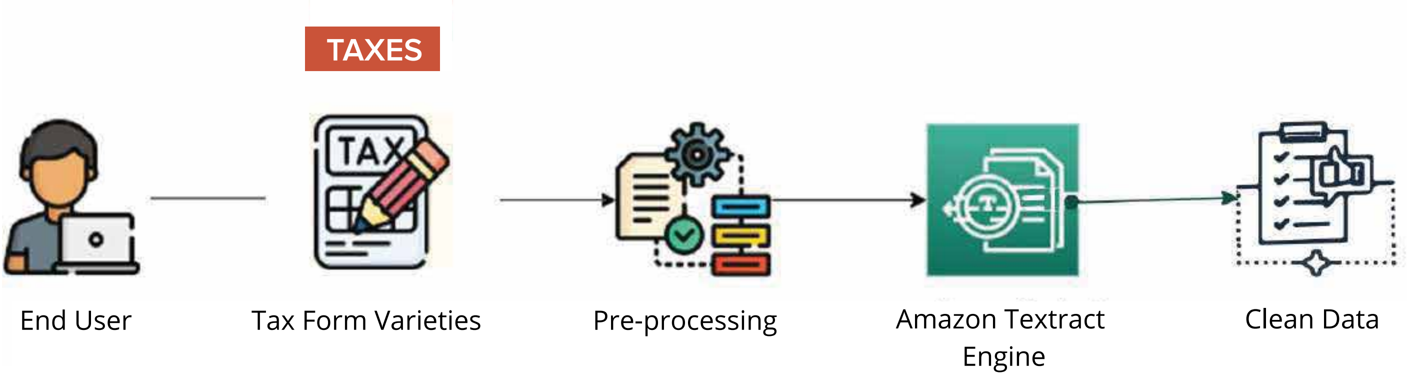
SOLUTION

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 improve image document clarity.
- Extraction of Key-value pairs from Form Fields.
- Extraction of Content from Tables.
- Extraction of Check-box Fields.

The extracted data was finally exported to a .csv which serves as an input in downstream data analytic workflows. An overview of the procedure is given below:



BENEFITS

Faster Processing Time: A Processing time of about 10 minutes was saved per file using the InferIQ extraction solution.

Improved Accuracy & Efficiency: This solution extracts contents of form fields and table fields from documents efficiently with reduced error. It also supports the most common document formats - viz. digital and scanned .pdf.

Improved Document Output: With this approach, structured output in .csv format and other suitable formats can be generated. It also produces data rich .csv as an input in downstream data analytic workflows.

Easy Integration: This solution can be integrated into any loan management or financial decision-making system, that runs on a different technology stack via the microservice architecture.

OUR AWS COMPETENCIES

aws
PARTNER
Advanced Tier
Services

- Public Sector
- Solution Provider
- DevOps Services Competency
- Financial Services Competency
- Migration Services Competency

Contact us

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

inferIQ

Find out how InferIQ solutions can help your business. Contact us today!
✉ info@inferiq.ai I 🌐 www.inferiq.ai