A brief about us

Technossus UK Limited is a part of Technossus Ltd, a Californian based 11 year old company firmly into Technology 4.0. It has a very strong cloud strategy and transformation capability along with a sound data science and automation team. Technossus's technology capability is very much used in Healthcare, BioScience and Financial Services domain where unstructured and cognitive pattern data in the from old artifacts or motion pictures can derive useful information for predictive analysis.

Thames Transformation Partners is a boutique consultancy firm whose talent base is a unique blend of practical, hands-on industry experience and consulting (advisory) expertise that strengthens client's business. It has a strong Finance and HR consultancy background guiding fortune 500 clients along with guiding them with their digital transformation needs.

Both **Technossus** and **Thames Transformation** team up together to conceptualise and bring this unique product of Intelligence OCR to the market.

Intelligent OCR Solution

The Problem

Optical Character Recognition (OCR) identifies text characters from digital assets such as pictures or PDF documents. OCR has been in existence since the 1970s and is used successfully in various industries to extract and convert the text found in basic documents.

However, the origin of text data and the format of documents are so diverse that most OCR tools cannot effectively read and extract text from these documents. Furthermore, handwriting forms a critical part of many documents. OCR solutions cannot process complex document layouts or recognize characters outside a predefined set of typefaces and fonts.

Business Application

The concept of Intelligent OCR solution came to light when we were given a challenge to help an NGO to carry out an investigation that related to examining the contents of 60 to 70 year old documents of Dioceses in various Church organizations. The NGO had to submit it findings within a short time frame as stipulated by the Court Of Law by the analysis of legal directories spanning from 1930 to 2020. This project was named **ZAP or Zero Abuse Project** as you shall see the mention of ZAP in the diagrams.

Many of these documents were so badly preserved that even human eyes could not read them. The first port of call was to identify existing solutions in the marketplace that could read and extract the text from the documents. **After reviewing the OCR tools provided by Amazon AWS (20% accuracy), Google (32% accuracy), and ABBYY (7%), it was clear that we needed to develop our solution.**

By utilizing computer vision and deep neural networks, we developed a solution that achieves a 99.5% accuracy in extracting and annotating text across the entire dataset—delivering a solution where caseworkers could simply enter a search phrase and receive a listing from multiple sources.

To enhance the user experience, we annotate the original document to show precisely where the results originated from and, by doing so, saved thousands of hours of manual research.

Our Solution

To solve this problem, we implemented an AI model capable of recognizing any font, typeface, or handwriting with 99.5% accuracy from only 1000 lines of examples. In most circumstances accurate AI models require a substantial amount of labeled or example data to produce accurate results. Our solution can produce this accuracy from only two pages of text.

Our document reading engine can process content such as tables and lists, recognize objects such as signatures and stamps and understand the flow of advanced content layouts.

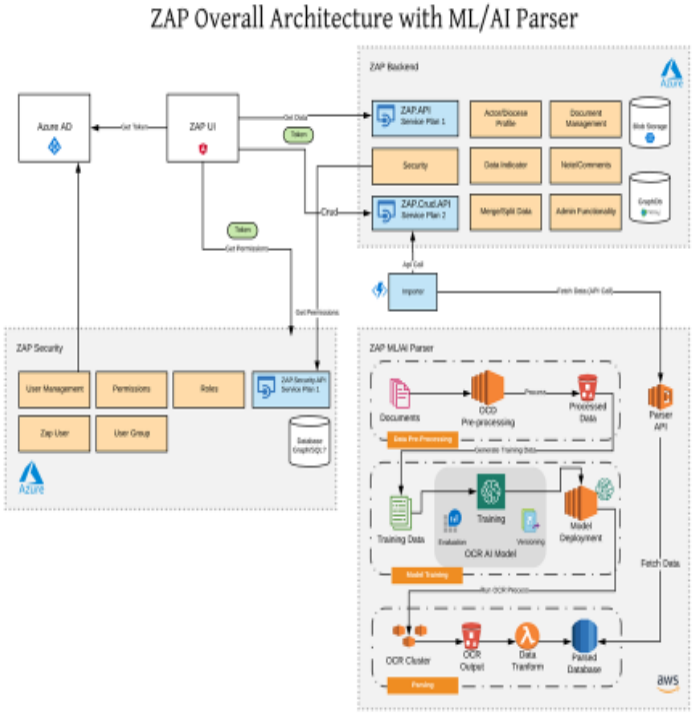
Finally, our solution can extract key elements from the document and customize these elements based on the underlying business domain. For example, it can extract all names, addresses, phone numbers, and company names for downstream processing.

Technology Brief

In the following section, we outline the core pieces of what comprises the technology as it relates or our Intelligent OCR Solution

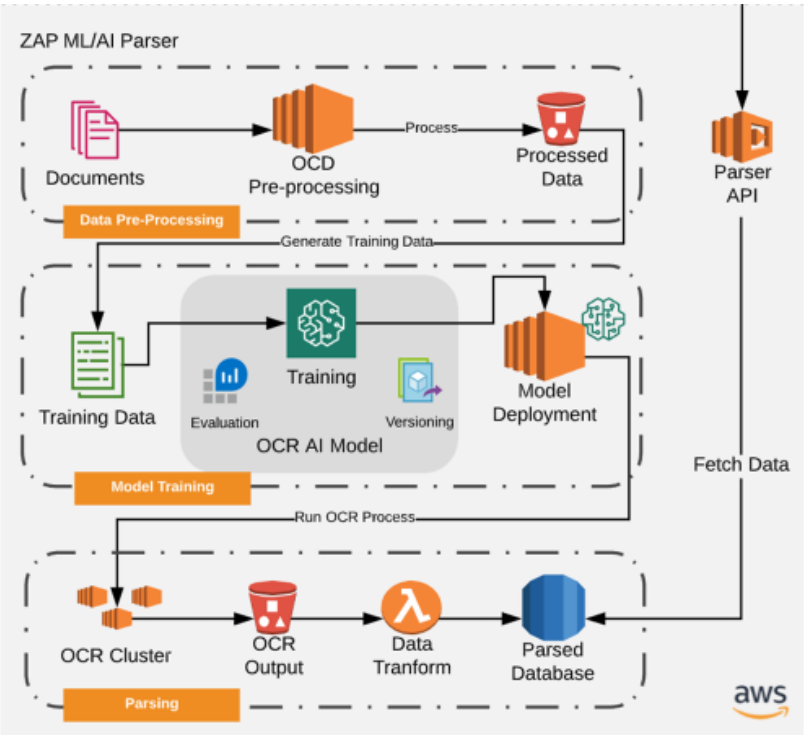
Application Architecture

- Azure AD for authentication
- Custom authorization
  - ZAP users stored in SQL
  - Roles not yet implemented
- 3 total APIs
  - Application API
  - Upsert API
  - Security API
- Neo4j graph database



Parsing and Insertion

- Replaced the regex-based parser
- Data patterns are *enormously* complex
- Insertion is done manually with CSVs
  - Insertion API took hours for one year's worth of data



Pre-Processing Pipeline

The data pre-processing pipeline allows for a resilient, scalable data ingestion process where we normalize and convert all data into high-resolution artifacts for downstream processing.

We also implement a UI interface allowing non-technical users to review and correct any results that performed below an accuracy threshold. This newly labeled data is then used to continually train the AI models for improved accuracy

Document Classification

The document classification engine utilizes advanced physics-derived topology analysis to identify content layout, document structure and object detection. The engine's sensitivity can easily be adjusted, but in most cases, it classifies a document and its content without any difficulty

Document Reading

The document reading engine is responsible for extracting structured text data from the document classification results. Our AI model can detect various text metadata such as the font, typeface, headers, footers, table headings, handwriting, signatures, and many more.

In addition, our AI model is initialized with a large pre-trained language model that is used to correct text that is either misspelled or where inherent noise exists between characters. This ability has resulted in high accuracy in almost impossible documents to read with the naked eye.

During the document reading process, the output contains the character level accuracy and coordinates on the document and any additional contextual metadata like tables and object types.

Document Understanding

The document understanding portion of the solution is responsible for extracting valuable components of the document text. Our AI model contains general entities such as Organization, Person, Numerical, Ordinal, Cardinal, and various other entity types agnostic to the specific business domain.

The model allows for training different entities specific to the business domain with an easy-to-use user interface requiring little technical fortitude. Similar to the document reading engine, the document understanding results appear in a rich user interface where non-technical stakeholders can easily label and correct the entities, which in turn is used to train and improve the model on a concurrent basis.