CASE STUDY

Document Analytics for an Oil & Gas Service Company

CONTATA

OVERVIEW

The client is a part of an oil services company in the business of providing data related to oil wells. The data containing details of drilling and logging information yields is used by oil companies for the exploration and exploitation of new oil reservoirs.

CHALLENGES

Our client sought a solution that can be used to categorize documents, identify duplicates and extract actionable information. The process that the client used for compiling its data product was complex. Raw data was procured from several different sources – most of it in the form of paper files. The sheets relating to each well were first OCR-ed and then sent through a human process to remove duplicates and categorize the remaining ones together into one of many known types of filings. Finally, information was extracted from specific pages (such as forms) and captured into structured records representing the information contained in the documents. To ensure quality, several of the steps were repeated by independent processors and then checked to verify the match. All in all, the process was resource intensive, costly, and not immune to error.

INDUSTRY

- Oil & Gas

SERVICES

- Web Application
- Machine Learning
- Text Analytics
- Cloud Solutions
- Document Analytics

TECHNOLOGY

- Python
- NLTK
- Elasticsearch
- HTML
- AWS

SOLUTION

Contata deeply analyze and was engaged by the client to review the overall process for compiling the data and see what types of automation could be built to decrease cost and increase reliability. Based on an analysis of the process, Contata recommended several machine learning-based approaches to replace the human-based steps.

A full-text search database was first created to house, search and retrieve the OCR-ed documents. This allowed easy finding and deletion of duplicate documents. Multi-class categorization techniques were then used to categorize documents into different classes representing the types of filings then represented. Finally, contextual information extraction was used to retrieve structured information from forms. The whole system was made available for human review and QA through a web-based application interface.

BENEFITS

- The content processing engine successfully processed documents at high speed and with great accuracy.
- Manual processes were largely eliminated except for the final QA step, resulting in 90% savings in costs.
- The system accurately identified documents with 95% of the time, flagging the remaining 5% with low confidence and requiring manual intervention.
- Operational accuracy and flexibility achieved.

About Contata

Contata Solutions is a trusted leader in technology and digital innovation. Through our work in data engineering, data analytics, machine learning, marketing automation and app development, we deliver solutions that address complex problems in ways that are simple, insightful and impactful.

Our promise and value proposition to our customers is simple: we leverage our deep technical expertise and global presence to bring software products and data-driven decision capabilities to life.

Founded in 2000, Contata is a privately-held company headquartered in Minneapolis that serves clients globally from offices in the United States and India.