PESGB EAGE

ML P17

An Automated Information Retrieval Platform For Unstructured Well Data Utilizing Smart Machine Learning Algorithms Within A Hybrid Cloud Container

N.M. Hernandez[1], P.J. Lucañas[1], J.C. Graciosa[1], C. Mamador[1], L. Caezar[1], I. Panganiban[1], C. Yu[1], K.G. Maver[2]*, M.G. Maver[2]
[1] Iraya Energies, [2] KGM geoconsulting

Summary

There is a large amount of historic and valuable well information available stored either on paper and more recently as digital documents and reports in the oil and gas industry especially by national data management systems and oil companies. These technical documents contain valuable information from disciplines like geoscience and engineering and are in general stored in a unstructured format. To extract and utilize all this well data, a machine learning-enabled platform, consisting of a carefully selected sequence of algorithms, has been developed as a hybrid cloud container that automatically reads and understands the technical documents with little human supervision. The user can upload raw data to the platform, which are stored on a private local server. The machine learning algorithms are activated and implement the necessary processing and workflows. Structured data is generated as output, which are pushed through to a search engine that is accessible to the user in the cloud. The aim of the platform is to ease the identification of important parts of the technical documents, automatically extract relevant information and visualize it for the user, so they can easily do further analysis, share it with colleagues or agnostically port it to other platforms as input.

Introduction

There is a large amount of historic well information available stored either on paper and more recently as digital documents and reports in the oil and gas industry. These technical documents contain valuable information from diverse disciplines such as geology, geophysics, petrophysics, reservoir engineering, drilling and other subject matters and are in general stored in a unstructured format.

Especially national data management systems and oil companies hosts these large amounts of very valuable historical well data, which contain information such as reservoir metadata, images, texts, and processed information, such as lithology, geology, shows, drilling risks etc. Due to the large volume, vintage variety, and non-standard formats, extraction of valuable information, which can be used as input for further work, is an arduous task as the manual nature of data mining is very time-consuming.

To extract and utilize all this well data, a machine learning-enabled platform has been developed as a hybrid cloud container that automatically reads and understands hundreds or thousands of technical documents with little human supervision. The aim of the platform is to ease the identification of important parts of the technical documents, automatically extract relevant information and visualize it for the user, so they can easily do further analysis, share it with colleagues or agnostically port it to other platforms as input.

Methodology

The platform utilizes a hybrid data service architecture, which leverages the 2-tier strength of both cloud and private servers. The hybrid architecture serves to:

- Enhance the platform's security and data privacy by storing raw data locally
- Increase data shareability in real-time by utilizing a cloud solution
- Reduce data redundancy and increase data integrity among users
- Provide a pragmatic solution to optimize data storage costs

The user can upload raw data to the platform, which are stored on a private local server. The machine learning algorithms are activated and implement the necessary processing and workflows. Structured data is generated as output, which are pushed through to a search engine that is accessible to the user in the cloud (Figure 1).

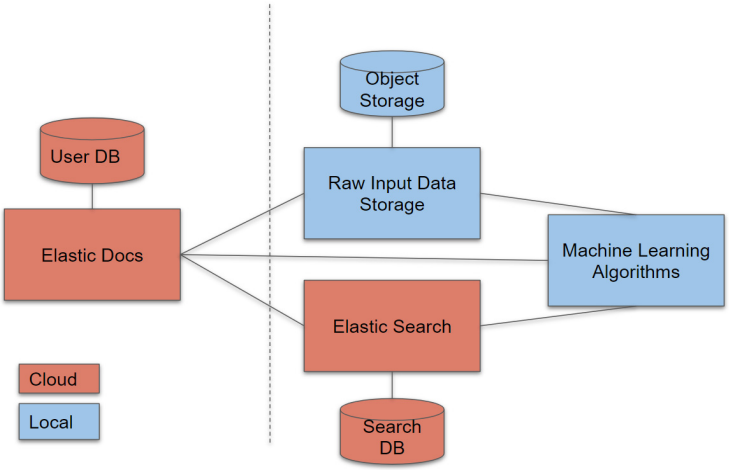

Figure 1 *The hybrid architecture of platform (ElasticDocs) utilizing the 2-tier strength of local and cloud sever applications for data security, integrity and shareability. Carefully selected machine learning sequence for automated text and image analysis include: optical character recognition, deep convolutional neural network and image clustering.*

The platform capitalizes on the machine learning algorithms that automatically process the unstructured data into a condensed format in which only pre-selected information are stored. The machine learning algorithms employs a unique sequence of separate steps, which are set-up to mimic the human experience of processing unstructured documents.

Workflow

A Norwegian dataset consisting of 400 well reports (58,000 pages) and an Australian well database consisting of 6,000 pages have been used as training data for generating structured data.

For the unstructured data the first machine learning step is the digitization and conversion of .pdf or .docx file formats into an editable format. This conversion uses Optical Character Recognition (OCR), where the machine identifies each character in the image.

After the documents are digitized important information has to be identified. This metadata extraction and tagging utilizes Natural Language Processing (NLP) to tokenize each digitized text and identify terms of significant value. Named Entity Recognition (NER) is then performed to create a model to extract the metadata like well name, basin, permit, operator, well classification, latitude, longitude, spudding date, kelly bushing etc.

For the images extracted by the digitization process, a modified VGG-16 neural network is used to automatically classify tables, charts, stratigraphic chart images, maps, seismic, core samples and scanning electron microscope images within each document (Simonyan *et al*., 2014)

For the visualization of the images an at-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is used to quantify the similarity of each image, which has been developed to visualize high-dimensional datasets and reveal clustering within the datasets (van der Maaten *et al*., 2008).

The output from the machine learning sequence is then exposed to the users through the platform to ease the work of identifying important information and perform analysis in a more efficient way. The extracted information is outputted in an agnostic format, which can be efficiently loaded to other platforms or used as is, i.e. X, Y or Latitude/Longitude, formation tops in csv or excel format, digitized maps as shapefile for loading into GIS software.

Discussion and conclusion

Wells provide key information about the subsurface in oil and gas exploration and production but at a substantial cost. As this valuable information associated with a well is often stored as unstructured data, it is difficult to do further analysis or apply additional artificial intelligence processing to the well database to enable geoscientists to gain new insights and extract new relationships.

The carefully selected sequence of machine learning algorithms in the workflow deals with these large unstructured datasets, is housed within a hybrid cloud platform to automatically extract relevant information within technical documents and convert the unstructured data into a shareable structured dataset.

References

Simonyan, K. & Zisserman, A., 2014: Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint* *arXiv:1409.1556*

van der Maaten, L.J.P. & Hinton G.E., 2008]: Visualizing High-Dimensional Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2576-2605.