

SSI11

A new way of handling unstructured data in the age of digitalization

F. Baillard¹, K.G. Maver¹, N.M. Hernandez¹

¹ Iraya Energies

Summary

Years of field development and production have generated huge amount of data produced by different G&G disciplines. Any operators will therefore be facing during the field lifetime common challenges such as 1. An exponential amount of data created 2. A variety of document type and format being generated 3. A retention of knowledge from senior personal leaving the company.

A new way of handling unstructured data in the age of digitalization

Kim Gunn Maver* (Iraya Energies), Nina Marie Hernandez (Iraya Energies), Francois Baillard (Iraya Energies),

Introduction

Years of field development and production have generated huge amount of data produced by different G&G disciplines. Any operators will therefore be facing during the field lifetime common challenges such as 1. An exponential amount of data created 2. A variety of document type and format being generated 3. A retention of knowledge from senior personal leaving the company.

These challenges leave the geoscientist drowned into the data, he/she are tasked to structure the data, curate it and then create meaningful insights. He will then ultimately share this to his/her colleagues into a new piece of data that will likely be unstructured which by definition will difficult to share within the organization.

In this paper we will be looking into a native cloud platform providing a seamless ingestion of complex geoscience unstructured data by applying Machine Learning pipelines.

Unstructured data platform

The unstructured data is provided in an unstructured manner including final well reports, completion reports, drilling reports, seismic processing and acquisition reports to mention huge and diverse range of disciplines and format. The files are commonly in .pdf, .docx, .xlsx, pptx, .las and .segv formats.

All the data is ingested through a consecutive pipelines and workflows using machine learning techniques such as NLP or deep CNN to provide the user with a structured set of data including metadata, index of documents, search through text and image corpus capabilities (Hernandez et al., 2019). Figure 1 summarizes the applied workflow for each document.

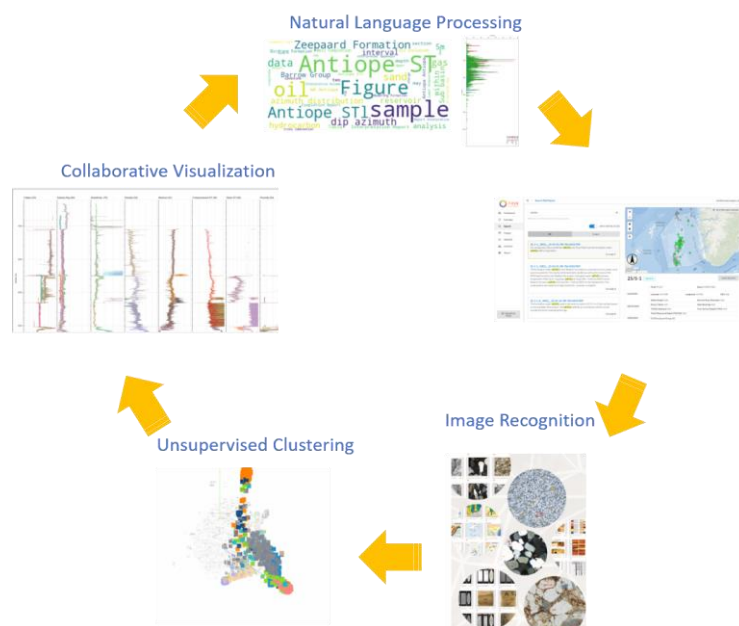


Figure 1 Workflow for data crushing with ElasticDocs

The final data can then be interrogated through the platform functionalities.

Time saving process

In comparison to traditional geology databases populated by army of geoscientists. The platform as seen on Figure 2 combines a searchable structured database with Machine Learning techniques shortening significantly the time between receiving the files and deriving insights.



Figure 2 ElasticDocs combination of ML techniques and a structured database

The past and the future may reveal how much time could be saved through the full process by the use of new Machine Learning technologies.

Looking at the past history, recent case studies has revealed that a corpus covering 50 years of diverse G&G data generated by a team of 100 geoscientists for 1 year (equivalent of 5,000 man year) would be fully processed in a couple of months which can be translated by a production gain of more than 500 times (upper bound).

Similarly looking at the future, considering how much a given geoscientist would be taking to read and process this data, we could consider an experience geoscientist to take 10min per page to find and extract the information he is looking after. Therefore this would require the said gescientist to spend 10 years to cover the full corpus. In comparaision, the current automated process would crush the full amount of data in a couple of months which corresponds in this case to a production gain of 40 times (lower bound).



Figure 3 ElasticDocs production gain

Conclusion

Application of Machine Learning platform provides tools for geoscientist to finally curate and extract insights from the (unstructured) data in an efficient way. This allows a better use of his/her time, focusing on more complex processes and workflows; ultimately leading to a better and faster decision-making process. For an organization this will be translated into a production gain between 40-400 times compared to today's traditional techniques.

References

Hernandez N., Lucañas P., Graciosa J.C., Mamador C., Panganiban L. C. I. Automated Information Retrieval from Unstructured Documents Utilizing a Sequence of Smart Machine Learning Methods within a Hybrid Cloud Container. EAGE Workshop on Big Data and Machine Learning for E&P Efficiency 25 - 27 February 2019.