DEEPIQ

Geospatial Analytics at Scale

How to Perform Advanced Geospatial Analytics at Scale

Background

Geospatial analytics is the collection, preparation, manipulation, and reporting of geographic data and imagery. Data can come a variety of sources, including sensors, GPS, seismic, well logs, maps, and satellite imagery. Geospatial data analytics began in the 1960s in Canada to catalogue natural resources[i]. It has since grown to a multi-million-dollar industry. In industry, the most common use case is to improve efficiencies in field operations, including exploration. Though geospatial analytics can have an extraordinary impact on E&P, mining, healthcare, retail and other verticals, it has remained an esoteric field with many point solutions. Building industry leading models on large geospatial datasets is now a trivial task using DeepIQ DataStudio.

In this whitepaper, we first explain the challenges in geospatial analytics. Then, we explain how DeepIQ addresses these challenges leveraging your cloud providers scalability and customizability.

Challenges in Geospatial Analytics

In building machine learning workflows, the three main challenges that users encounter in geospatial analytics are explained. In later sections, we elaborate how we address them using the power of your cloud provider.

Data Formats

The datasets needed for complex geospatial analytics are themselves unique and not supported by standard software. Well logs are available both in Raster and LAS formats. Maps can come in multiple proprietary formats like:

- SHP Files – a vector format specific to GIS software.
- Esri JSON – A standard for encoding spatial data
- GeoJSON – A standard for encoding geographic data structures
- TIFF – Raster images enriched with GIS relevant metadata

Ingestion of this software and converting them to a structure that is amenable to machine learning is complex, error prone and messy.

Data Cleansing

Your geospatial data defined by well logs define 3D spaces where the measurements are sparse and unevenly distributed. These data sets and maps occur at multiple resolutions and can have multiple data quality issues. Now, what is the best way to handle missing data? What is the right convolution filter to use to down sample your geospatial data to the same resolution? It depends. Is it better for your machine learning model to preserve sharp edges between neighbouring points? Is it better to smooth out noise? Both are competing demands and require completely different approaches. The right approach is to iterate with multiple variants of data cleansing operations, along with your model's hyper parameter search, to determine what combination generates the best results.

Geospatial Modelling

Your data has been ingested into your data lake in a clean and normalized manner. It is important to retrieve the right data required for analytics. You might need to figure out the intersections or containment relationships between polygons in your dataset. Your data lake may not support geospatial indexing like creating a R-index. So, you might need data models that allows for fast retrieval using database structures.

Structured Learning

Structured Learning encompasses a set of machine learning techniques that learn and predict on structured objects, rather than scalar discrete or real value[ii]. That is, for structured learning, the outputs generated by the machine learning model have relationships with each other. A simple example is parts of speech tagging where models are trained to recognize the parts of speech of each word in sentence. Clearly, the part of speech of a word has a strong influence on the probability of the part of speech of its preceding and succeeding words.

Structured learning has been widely used for such problems in the fields of natural language processing, computer vision and time series analysis. Recent breakthroughs in deep learning have made structured learning successful in these fields. One limitation of structured learning is the learning models require large data sets to train a model, because the relationships between outputs need to be learnt in addition to the relationship between observed and output variables. For example, a popular dataset used to train models for object segmentation which is a popular structured learning problem has more than 300,000 images.

However, consider the problem of generating predictions for geospatial data sets that provide attribute information (the characters of object or phenomena) and the location information (coordinates or regions on earth where they occur).

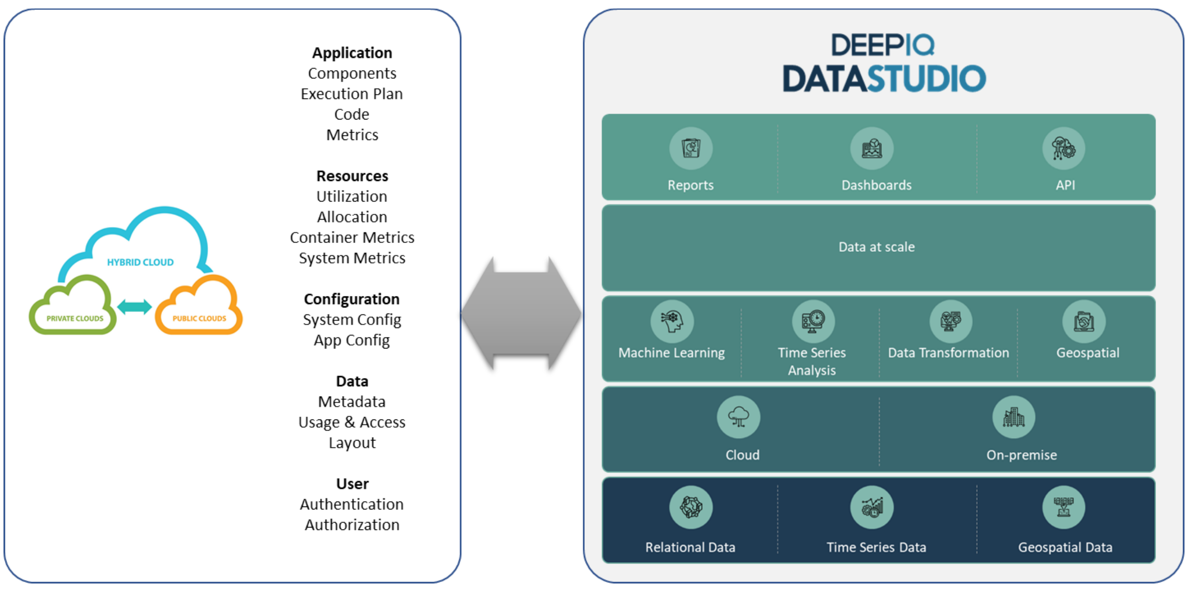
Modelling the relationships between outputs on a geospatial dataset has severe limitations. Geospatial datasets typically have significantly smaller amount of data to train a model. For example, consider a basin with ten thousand oil and gas wells where we are trying to predict the depth at which a particular formation occurs using geoscientific measurements like well logs. Clearly, structure plays a strong role in these predictions-that is, the depth at which a formation is present for neighbouring wells has a strong predictive value in estimating the formation depth at a particular well location.

In this example, we have only 10,000 data points that can be used to learn structure. Contrast this sparse dataset with an object segmentation problem where we have 10,000 images of 500 x 500 resolution to train model. For this problem, we have 2.5 billion pixels to learn how outputs from neighbouring pixels are related to each other.

The DeepIQ DataStudio

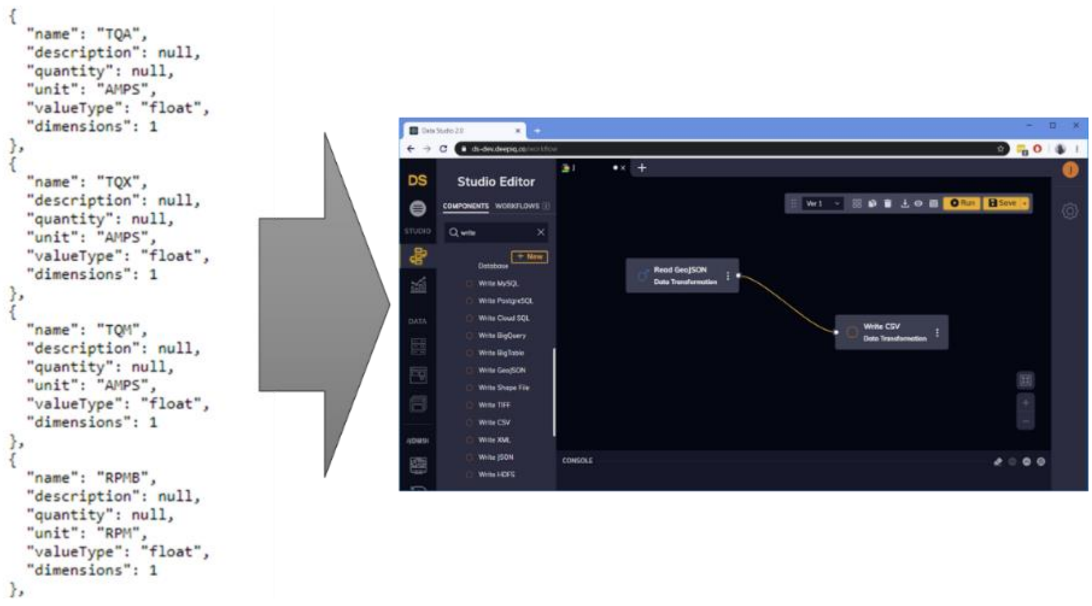
DeepIQ's DataStudio was conceived with the concept of Self-Service Analytics for the Industrial World. The software was designed so that process engineers, data scientists, and petro-physicists have an easy to use tool for complex data problems.

DeepIQ's DataStudio requires an execution platform that will run Apache Spark. The execution platform can be either cloud based, on premises, or a hybrid. If you do not have an existing execution platform, the DeepIQ team will be happy to make a recommendation.



Data Formats

DataStudio has prebuild components for ingesting a wide variety of geospatial datasets and converting them into structured datasets amenable for analytics. At the click of a button, you will convert all of your maps into a compact excel file or a cleanly structured table in your favourite database. DataStudio has native connectors for the industry's most common formats including LAS files, SHP files, Esri JSON, GeoJSON, and TIFF. The user simply drags a component on the canvas and then sets the properties. In Figure 1, DataStudio is ingesting a standard GeoJSON file, and storing the results in a comma separated format.



```
{
   "name": "TQA",
   "description": null,
   "quantity": null,
   "unit": "AMPS",
   "valueType": "float",
   "dimensions": 1
},
{
   "name": "TQX",
   "description": null,
   "quantity": null,
   "unit": "AMPS",
   "valueType": "float",
   "dimensions": 1
},
{
   "name": "TQM",
   "description": null,
   "quantity": null,
   "unit": "AMPS",
   "valueType": "float",
   "dimensions": 1
},
{
   "name": "RPMB",
   "description": null,
   "quantity": null,
   "unit": "RPM",
   "valueType": "float",
   "dimensions": 1
},
```

Figure 1: Ingestion of a GeoJSON File

Data Cleansing

DataStudio has multiple data cleansing operations built for geospatial datasets including:

a) Missing Data Replacement
b) Resolution Conversions
c) 2D Convolution algorithms for noise corrections

The user simply drags the cleansing components into the workflow and configures the component in properties. The workflows can be saved and re-run as more data becomes available.

Geospatial Modelling

Consider the problem of estimating of formation tops using well log data (Figure 2). Using our proprietary deep learning algorithm with just gamma logs, we have been able to achieve more than 97% accuracy on an open source dataset on both test and train datasets (Figure 2).

Figure 2 shows a sample well log where we are trying to estimate two tops-MNKT and OPCH. The machine learning algorithm identified MNKT top to be at 5,255 feet but is unable to detect OPCH because some of the data is missing. In this approach, we have not used structured learning.

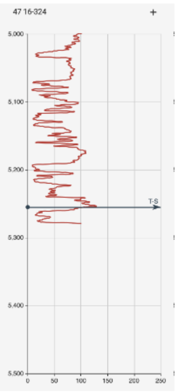


Figure 2: Sample Well Log

Let us look at the neighbourhood wells and how their data looks like as shown in Figure 3. Clearly, the neighbours have better data and better predictions. Standard machine learning algorithm treats each well where geoscientific properties need to be estimated as being independent of others. However, wells that are close to each other have a greater likelihood of having formation tops at similar locations and algorithm quality can be improved by modelling the geospatial relationships between neighbouring points within the model.

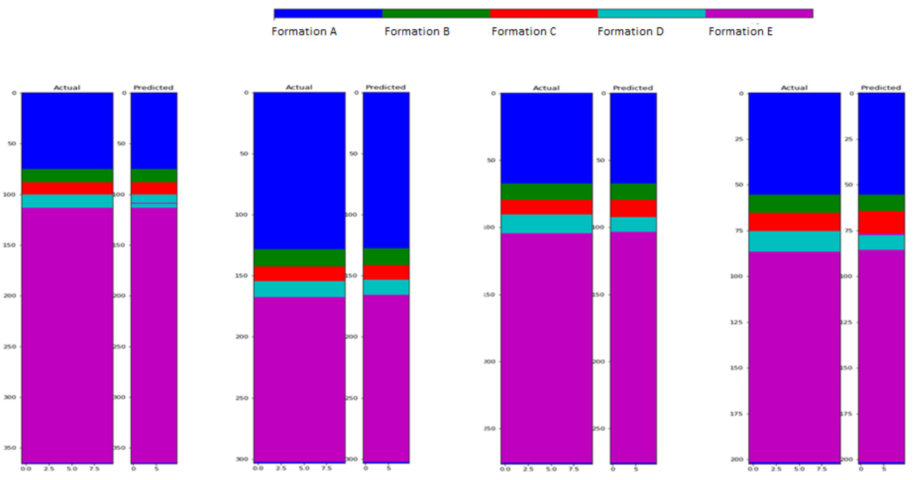


Figure 3: Sample Results from Formation Identification Algorithm

For structured learning, the outputs generated by the machine learning model have relationships with each other. Structured learning has been widely used for such problems in other fields. One limitation of structured learning is the learning models require large data sets to train a model because the relationships between outputs need to be learnt in addition to the relationship between observed and output variables. DeepIQ is able to overcome these limitations with their proprietary models.

Let us use structured learning to see if the model can estimate OPCH top for the second well here. The results are shown in Figure 4.

The second well has an estimation for OPCH even though the data is missing. Incorporating neighbourhoods in our prediction really helped in improving the quality of our results.

We had a geologist verify the ground truth using auxiliary data sources and we were off only by two feet, even on sparse datasets!

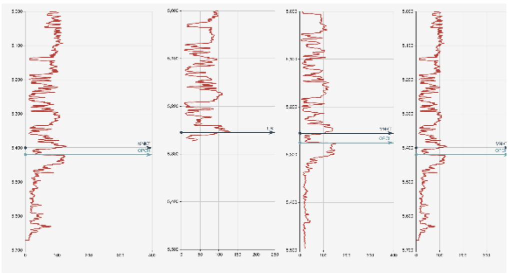


Figure 4: Neighbouring Wells

Conclusion

DeepIQ's structured Machine Learning capability captures the complex interactions between geospatial entities even when the training data is sparse. This unique capability can be a gamechanger for your predictive capability on geospatial datasets by coherently modelling interactions between neighbourhoods.

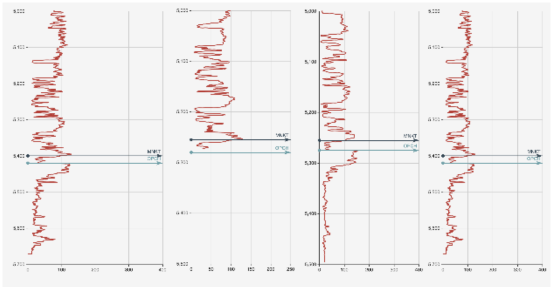


Figure 5: Formation tops prediction for a sample well with missing data using structured learning

[i] https://www.omnisci.com/technical-glossary/geospatial-analytics
[ii] Neural Information Processing Systems Foundation, Predicting Structured Data, MIT Press, 2007.