10.1071/PVv2016n185p36

Data trends

Guy Holmes
guy.holmes@katalystdm.com

Hadoop-de-do

There have been some real game changers that have hit the ground in the last few years. Big Data Analytics, Watson from IBM, IoT and Hadoop are among these. I hear that pockets of the industry are starting to adopt some of this technology, but it still feels like we are a little behind the times.

If there is one thing that the oil industry has it is a lot of data, and most of the technology that I have listed would be easy to adapt to our needs. As an industry we also tend to only use a small portion of our data, after we weed out what conventionally has been considered surplus to requirements – but nice to have. In fact, it has always appeared to me that in seismic in particular we start with very large files and then go about reducing the size of those files as quickly as possible in order to make the data more manageable. The decision to make the data more manageable - instead of better quality, or smaller - instead of more insightful, seems odd to me, especially now that tools like Hadoop have changed the imperative to reduce the size of datasets. What I think this means is that we have some great and ground breaking options available to us – especially if we take a step back and think about our new options.

So what is Hadoop?

Hadoop is an open source programing framework that allows users to process and store extremely large datasets (large files or large volumes of files) in any environment. Imagine having a 50 Tb file you need to process but your computer only has a 2 Tb hard disk. In conventional processing you would either get a bigger hard disk (not always cost effective or even possible), or process the file in small parts, one at a time until you are able to process all of the data and gather the smaller more manageable version to move forward.

In seismic we tend to process the data in parts and then go about distilling it down to smaller packages to make it more manageable. In my view we carry out the 'distilling' process more to make the dataset smaller and more manageable than because the data we are getting rid of is not useful. For me this means that we make scientific decisions further down the track, without the full benefit of all of the data that could have been used. With Hadoop a user can store and process the entire dataset and interact with the entire file all at once. No need to distil it at all. It means we can now work in broad generalities, while at the same time retaining full access to all of the data – even if we are not using it. If we see something interesting while working we can go back and change our sample size, or dig deeper from surrounding data to confirm results that previously would have meant going back and processing the data again from scratch, which typically would be cost prohibitive.

Hadoop is an open source programing framework that allows users to process and store extremely large datasets in any environment

Why is Hadoop important?

For me Hadoop is important in the oil and gas space because it allows deep diving into data, and the ability to change tack, refine, review, validate and, in the end, derive the best possible results and make the best possible decisions. It allows previously cost or time prohibitive transactions in data analysis and processing to be undertaken in an inexpensive and interactive way. You really don't need to make the types of decisions we now make up front in data processing – made only to get the data down to a manageable size to move the project forward. If you combine Hadoop and Cloud storage you get such a scalable and cost effective work environment that more data becomes a better way to work, and teams don't need to be pressured into reducing their footprint on the network. No one needs to be in a rush to pick the attributes they feel are important up front, and teams can go back to all available data to validate things as they move forward. Almost infinite storage, massive processing power and the ability to address huge files or datasets all at once has not been possible until the last few years in any cost effective way. I don't think that as an industry we have woken up to this fact yet.

Imagine a world….

Imagine a world where newly acquired exploration data starts off life in in the cloud. No tapes to read or transport, and no need to break up datasets into pieces that fit our storage media. Navigation, positional data and support materials stored permanently in a raw form in a state that is always available to users – never offline, never on storage media that you cannot read and, since it is all online at once, always in the most modern format available (it can be converted on the fly as standards change or as new datums are derived etc.).

Imagine that we find in our geophysical community a certain thing or indicator in data that if used on any dataset gave you confidence about some geophysical property that in turn gave you exploration certainty about targeting resources. Imagine then if you could apply that knowledge to all data, on all prospects, globally, in one place, at a very low cost with the click of a button.

Hadoop, cloud storage and processing, analytics, and big data make all of this possible. All we need to do now is make it happen.