Data agility: Innovative approaches to subsurface data management

Jose Chapela[1]* explores the challenges associated with data management and reviews solutions for overcoming them.

Abstract

Data management has become a critical component of oil and gas exploration. The industry has spent years collecting, storing, analysing and interpretating subsurface data. This data is invaluable for making informed decisions on where to explore for hydrocarbons, but properly managing subsurface data presents a multitude of challenges, each of which impacts on the quality, accessibility, and utility of this crucial resource. In this article, we explore the challenges associated with data management and review solutions for overcoming them.

Introduction

Dealing with data in the upstream oil and gas sector has always presented distinct and complex hurdles. This industry handles a vast array of data, each marked by its own unique formatting intricacies. Consider file types like SEG-D, SEG-Y, ACSII, UKOOA, Multibeam, LAS, GeoTiff, and more, encompassing seismic data, well logs, horizons, interpretations, and various other data categories. Furthermore, a significant portion of these data formats include spatial components that demand meticulous handling to ensure accurate geolocation.

This article will primarily address techniques for effectively managing your seismic data, whether it's in SEG-Y, SEG-D, or even older formats like SEG-A, SEG-B, SEG-C, or SEG-X.

As you contemplate the next phase of data management, it's crucial to acknowledge that the industry is currently shifting from its historical focus solely on hydrocarbons as the primary energy source to integrating renewables like wind and solar energy. This transition will bring entirely new challenges to our existing data storage and retrieval systems.

If today's data management already poses formidable challenges, one can only imagine the obstacles that tomorrow's data managers will encounter. When considering the future of data management, here are the key questions you should ponder:

1. How can you devise a storage strategy that is both manageable and cost-effective, considering the exponential surge in data volumes, primarily seismic, witnessed over the past decade?
2. How can you amass sufficient metadata pertaining to your data, making it not only usable for your data management team but also for your processing, interpretation, exploration, machine learning, and artificial intelligence teams? Does this metadata support eventually align with the Open Subsurface Data Universe (OSDU) standards?
3. How do you facilitate the efficient transfer of data, not only within your internal teams but also with external partners and collaborators?
4. How can you organise and cleanse your data, gathered over decades, to ensure your machine learning and artificial intelligence initiatives are poised for success?
5. How can you design the next generation of adaptable data management solutions to effectively address these challenges while remaining flexible enough to accommodate future data sets?

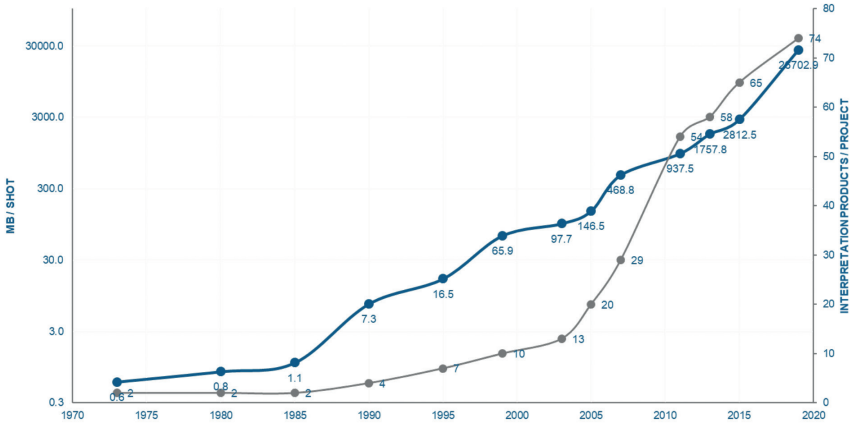

Figure 1 Size per shot on left axis and number of products on right axis.

[1] TGS

* Corresponding author, E-mail: Jose.Chapela@tgs.com

DOI: 10.3997/1365-2397.fb2023103

Addressing storage — the elephant in the room

Primarily, we must address the limitations of contemporary seismic data storage practices. In our industry seismic data is conventionally stored and shared in the SEG-Y format, a standard introduced by the Society of Exploration Geophysicists (SEG) in 1975. The inception of this standard occurred during a time when the predominant storage medium for seismic data was nine-track tapes. Consequently, the SEG-Y file format is tailored for sequential access, lacking optimisation for random input/output (I/O) operations. To illustrate, if you need to extract the last five inlines, this happens quickly but the process becomes significantly more resource-intensive, particularly when extracting or viewing crosslines. The complexity escalates when numerous nodes are attempting to work with a single file for processing or machine learning, causing I/O bottlenecks. Many organisations turn to parallel file systems to mitigate these challenges. Those parallel file systems, in turn, add expense and complexity to an already complex ecosystem.

Secondly, the ever-increasing volume of data poses a formidable challenge. Seismic data management grapples with the massive data sets generated at an unprecedented speed. Modern seismic surveys employ advanced sensors and technologies that yield vast datasets. The resulting datasets also generate a greater number of products at an increased density compared to previous projects. See Figure 1, which illustrates the size increase per shot and the increase in the number of products delivered to TGS data management per project from 1975 to 2020.

The high volume and rapid velocity of seismic data place considerable strain on storage systems and overpower traditional data management infrastructures. Handling such extensive datasets necessitates substantial computational resources, resulting in elevated expenses and resource intensity. For instance, within a decade, traditional streamer surveys have grown fivefold in size. Presently, a large 3D streamer survey demands more than 2 petabytes (PB) of storage for a single copy. For more data-intensive Ocean Bottom Node (OBN) surveys, this figure balloons to over 6 PB of storage for a single copy. Best practice suggests storing a 'production' copy and a second copy for disaster recovery for every piece of managed data. According to the aforementioned surveys, this amounts to 16 PB of stored data!

To address these issues, we have implemented a file format created by TGS and subsequently released to the open-source community, known as Multidimensional Input/Output (MDIO). A comprehensive explanation of the benefits of MDIO can be found in the October issue of *First Break* in the article 'Integrating Energy Datasets: The MDIO Format' (Sansal et al., 2023) or at www.mdio.dev

In the context of our data management requirements, this format has enabled us to achieve an average of 41% disk space savings compared to storing the data in the conventional SEG-Y format. Furthermore, as the data is now in a format that permits fast concurrent and random access, we can develop routines to efficiently and swiftly subset the data in a scalable automated fashion. Additionally, it enabled us to store the data in a format that remains readily accessible by our processing nodes and machine learning initiatives while at rest.

Utilising the hierarchical storage management (HSM) system embedded in cloud object storage, we can achieve further cost reductions by setting up policies to move data that is not actively being accessed by applications to lower-tier storage subsystems. The policies are configured so that the data is always immediately accessible and the HSM takes care of the file movement between tiers automatically in the background. With a significant decrease in storage requirements for the production copy of the data, we are free to explore the best possible path forward for storing disaster recovery copies of our data.

Conversion to MDIO (ingestion)

Now that you have established a data storage method, the next phase involves transitioning (ingesting) your datasets into this new format to harness various enhancements. When converting from SEG-Y to MDIO, this process offers an excellent opportunity to capture crucial metadata and address any discrepancies within your dataset.

Drawing upon four decades of experience in geological and geophysical data management, TGS is currently developing an automated quality control (QC) application. This application will swiftly extract vital information from each seismic product. The primary aim of this QC tool is to provide the team with an overview of the product's condition and to identify any issues that need attention before ingestion into MDIO. It will also extract essential information to facilitate the conversion to MDIO. For instance, it will attempt to determine the storage locations for CDP X and CDP Y values. Furthermore, it will extract key data from the EBCDIC header and geographically position the data on a map. This functionality enables data management personnel to efficiently review the data, extract pertinent information, and then proceed with the ingestion into MDIO or send the file for updates and corrections.

TRACE HEADER INFORMATION:

HEADER VALUE	BYTES	TYPE	HEADER VALUE	BYTES	TYPE
FOLD	33-34	2I	CMP Y-COORDINATE	205-208	4I
CDP ELEVATION	41-44	4I	TRACK(INLINE)	213-216	4I
SURFACE IN TIME	155-156	2I	BIN(CROSSLINE)	217-220	4I
CMP X-COORDINATE	201-204	4I			

Figure 2.1 Trace Header locations extracted from EBCDIC.

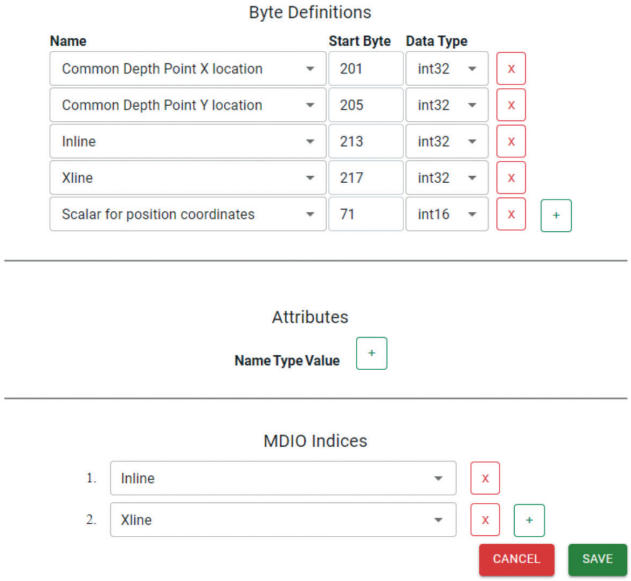

Figure 2.2 Byte definition screen.

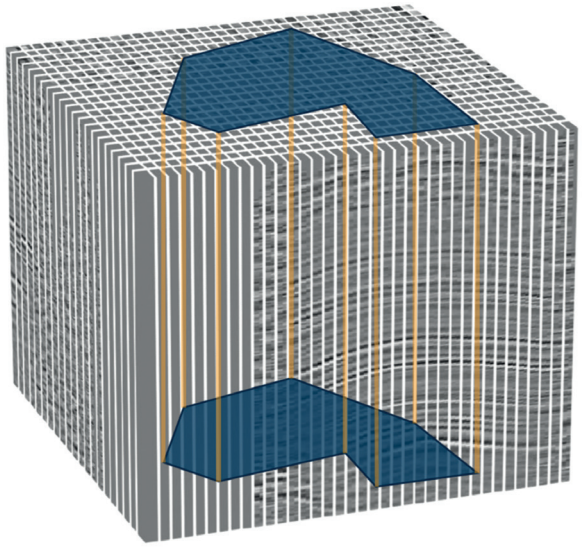

Figure 3 Section of seismic in MDIO format.

It is commonly known that adherence to the SEG-Y standard is often lax, making it vital to store crucial metadata alongside the file to ensure accessibility for consuming applications. For instance, when TGS was ingesting SEG-Y data for a US land survey, certain trace headers did not conform to the SEG-Y rev 1.02 specifications for locations such as 181-184 for CDP X and 185-188 for CDP Y. As per best practice, trace headers that deviated from the standard locations were specified in the EBCDIC header. As depicted in Figure 2.1, it is evident that within the EBCDIC header, CDP X was stored in positions 201-204, and CDP Y was stored in positions 205-208.

Within MDIO, we have the capability to configure these parameters, ensuring that any future applications accessing this dataset are aware of the key trace header locations through our byte definition screen. See Figure 2.2.

In this case CDP X, CDP Y, Iline, Xline, and the scalar for position coordinates, were noted and added to the metadata of the file.

If there is a need to define additional trace header locations, this can also be accomplished through the attributes section. To enhance data viewing and sub-setting efficiency, we have employed MDIO to create indexes for inlines, crosslines, and time slices. This allows for the quickest retrieval of our data in all three dimensions. Figure 3 illustrates the concept well. The blue-shaded polygon is our area of interest. The application will simultaneously access only the blocks of data necessary to create the data cut, returning a result set in a fraction of the time of traditional data cutting methodologies.

In addition to defining byte structures and indices, a standardised nomenclature was introduced, facilitating automated data tracking, versioning, and retrieval. This naming convention can be cross-referenced with a relational database containing extensive project and product information and metadata. The synergy between the data stored within the file and the metadata housed in the relational database provides a comprehensive view, serving the needs of internal data management applications and allowing for the retrieval of the data necessary to comply with Open Subsurface Data Universe (OSDU) Data Platform standards.

Once in the MDIO format, we can easily output the data back into SEG-Y for delivery to our external clients. Since it was cleaned and key metadata was captured, we are assured that the data being delivered to our clients is in the best condition possible. We can also output JSON files with key attributes included to ease data loading for our clients.

MDIO also allows for the storage of non-seismic data sets as well. It can effectively store any multi-dimensional data set, so it is also employed on wind data sets that are now being added to the data library.

Data movement

Ever since we decided to transition our complete data library to the cloud, we have harnessed the data movement technology developed by the leading public cloud providers: Microsoft Azure, Google GCP, and Amazon AWS. These providers have invested substantial resources to ensure efficient data transfers, whether it is between clouds or from the cloud to on-premises systems. Consequently, their data movement applications are finely tuned to deliver the fastest and most seamless data transfers. For instance, with a 10 Gb/s internet connection, transfer speeds of 75 terabytes per day (24 hours) were consistently achieved. This translates to an impressive utilisation rate of 69% of the link speed, while sharing the link with the rest of the organisation's normal internet traffic. What is noteworthy is that these data-moving utilities eliminate the need for us to concern ourselves with complex tasks such as multi-threading the copies, bandwidth throttling, encryption, checksums, and more, as these aspects are already integrated into the utilities themselves.

To provide context, TGS undertook the transfer of a massive single dataset, moving a remarkable 1.5 petabytes from our facility to a client's Azure cloud. The transfer, theoretically achievable within 20 to 25 days at speeds ranging from 70 to 75 terabytes per day, experienced some delays in practice. This was primarily due to the client's need to adapt and optimise their data intake procedures to handle the substantial daily influx of data. Once the client successfully adjusted their data intake routines to accommodate these large daily transfers, the process proceeded seamlessly, with approximately 70 terabytes of data delivered within each 24-hour period.

In a traditional delivery model, 1.5 PB delivery would have taken approximately 45-60 days before we would have been ready to ship, depending on the availability of resources for data copy procedures. It also would have required approximately 188 IBM 3592JD tapes or 150 12 TB USBs hard drives. Once the tapes or USBs arrived at our client's datacentre, they would have had to read that information from the media onto their systems before it was accessible to their internal teams. With the cloud transfer, the data was available to the client in a significantly reduced time frame.

Data management as a service

Standardising our data sets has opened up new possibilities. With data now described in a universally consumable manner and metadata accompanying each product, seismic data seamlessly integrates into the larger TGS data lake without any adverse impact on performance through API calls. For instance, having full certainty about position information enables us to query the data lake for all data falling within a project's boundaries, yielding

results such as acquisition reports, processing reports, OB logs, navigation files, contracts, well data, and more.

Furthermore, in our MDIO format, ordering data for internal processing or interpretation projects no longer necessitates involvement from data management personnel. Users can effortlessly request the entire project or a specific subset, whether a single product or all products. The ordering process is streamlined, user-friendly, and entirely automated. Once an order is submitted, the system promptly allocates the essential cloud resources for data extraction and subsequent quality control. Upon completion, the requester receives an email containing the location of the requested data and tools for downloading it to a location of their choosing.

TGS has recently introduced Data Verse, a Data Management as a Service offering that enables customers to take advantage of the storage savings, performance, ease-of-use, AI/ML readiness, and scalability available from TGS' data management systems.

Conclusion

Now is the opportune moment to contemplate your data management strategy. Does your strategy not only provide solutions for the four questions presented but also accommodate any unique concerns specific to your role as a data manager? While you explore the intricacies of your current data management solution, search for a system that possesses the flexibility to address present challenges and the agility to adapt to the ever-evolving demands of our industry in the future. Consider how your organisation plans to handle multi-client data and the management of your proprietary datasets.

Don't shy away from dealing with the elephant in the room, the SEG-Y format, and the large datasets we store in SEG-Y.

Acknowledgements

We would like to thank Dr Eugenio Maria Toraldo Serra, Vidar Mikalsen, Kevin Nicholls, Scott Tompkins, Karen Malick, Lisa Sanford, Daniel Phelps, Joao Gusso, David Garner, Gary Stafford, and the rest of the TGS Data Management team for the countless hours spent not only in design discussions but also executing on the vision on top of their normal duties.

I would also like to thank Sathiya Namasivayam, Charles Nguyen, Mohammad Nuruzzaman, Altay Sansal, Alejandro Valenciano, Inah Arthasarnprasit, and the rest of the TGS Data and Analytics team for its support, and unending patience during the design and creation of these solutions for the next generation data management ecosystem. Without this talented team of men and women, this would still be a vision and not a reality.

Special thanks goes to Jan Schoolmeester, and Laura Arti for supporting the vision and providing much-needed guidance.

References

Sansal, A., Lasscock, B. and Valenciano, A. [2023]. 'Integrating Energy Datasets: The MDIO format', *The Leading Edge*, **41**, 69-75.

SEG Technical Standards Committee, [2002]. *SEG Y rev 1 Data Exchange format.*