TOP TIPS FOR SEISMIC DATA STORAGE AND ORGANISATION

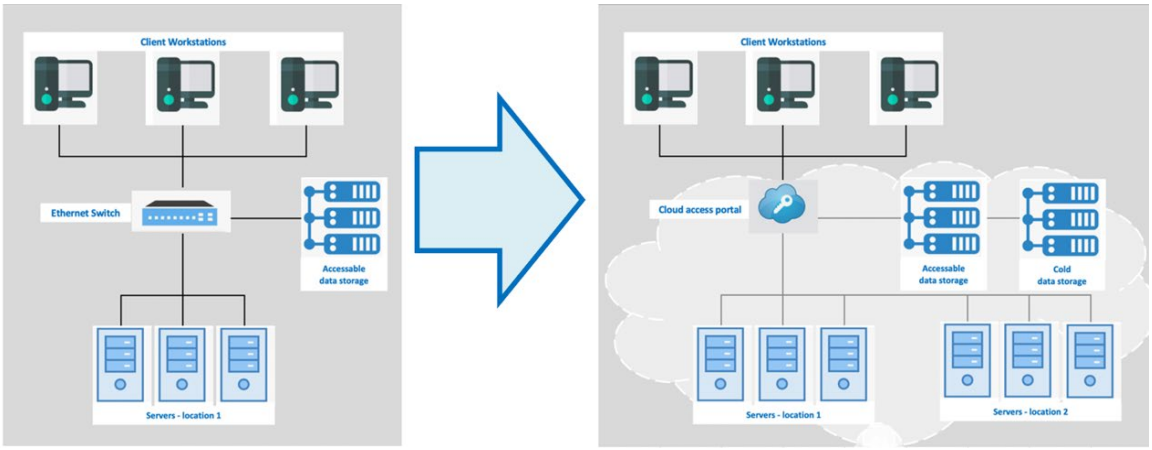
PART 2: CLOUD-BASED SEISMIC DATA STORAGE AND MANAGEMENT

This is the second in a short series of articles in which we are highlighting the key considerations for seismic data storage and organisation. **Part 1** focussed on the storage and management of seismic data stored on physical tape and disk media. In part 2 we focus on the functionality of cloud storage for seismic data.

Cloud versus Standard IT Infrastructure

Reliable access to seismic data requires sufficient computing and storage resources to prevent a negative impact on performance, which is challenging when access and usage requirements can vary greatly. In physical IT systems, it can be difficult to accurately forecast demand and scale resources accordingly, leading to significant system downtime and cost.

However, utilising virtual infrastructure in the cloud addresses these challenges and offers infinitely flexible scalability without associated downtime, as the cloud environment can effectively remove the entire physical IT infrastructure associated with seismic data management. Data hosted on virtual servers can be viewed, manipulated, and shared via web-based portals. Applications can be run on virtual machines in the cloud (so-called 'serverless computing') which increases speed and removes the requirement for an organisation to build and maintain its own IT infrastructure.



Simplification to illustrate the essential differences between Network-Attached Storage (NAS) architecture (left) versus cloud-based architecture (right)

Cloud-based seismic data storage

Seismic data is digitised and ingested into the cloud host provider's servers where it is stored in tiers depending on the level of access required; from "hot" with immediate access (at higher prices) to "cold" archived data (at much lower prices).

In contrast to a physical IT system, the cloud handles key factors of data security, disaster recovery and varying usage demand for seismic data in the following ways:

Data security is a shared responsibility between client and host; in general:

- **The cloud host** is responsible for security **of** the cloud (i.e. the server & storage infrastructure, networking, access management, availability of resources); and
- **The client** is responsible for who has access to their data and for its integrity and security **in** the cloud (i.e. the infrastructure the client builds within the cloud platform).

Disaster recovery is achieved by replicating data across multiple servers, possibly in different geographical areas, removing the requirement for costly regular data back-ups.

Varying usage demands can be managed by moving data from "cold" to "hot" tier for analysis. Cloud-based storage should reduce latency significantly to the of the order of hours, rather than the days it would take in a physical IT structure. Whether data is "live" or "archive" is simply a function of the tier in which data is stored and then becomes a factor in cost management.

Replicating data across several servers provides the ultimate flexibility to assign resources to meet demand, rather than only for peak usage. Scaling is simplified without the delays inherent in physical IT systems.

Flexibility of scaling is a key factor in cloud cost:

√ **Manual** – operator managed - has potential for human error to incur unnecessary cost
√ **Scheduled** – by pre-set peak usage times that minimises requirement for active monitoring
√ **Autoscaling** – triggered by pre-defined parameters to optimise IT resources and cost efficiency

Comparison of how key factors are handled between physical and cloud-based platforms		
Factor	**Physical storage**	**Cloud storage**
Scalability	Increase physical compute infrastructure – slow, costly, incurs downtime	Infinite & flexible - CPU/Memory/ I/O Auto-scaling options – fast Host on multiple servers
Disaster recovery	Archive backup dataset is stored off-site	Data replication across several servers
		Data replication across several locations
Security	Fire wall protection Secure log-in to access data	Data encryption Multi-tiered security process options
	Live data – Data owner's responsibility Fire wall protection	Shared responsibility Data encryption

The infinite scalability of serverless computing offers compelling reasons to use cloud-based services which can match or exceed a physical system, even accounting for data upload and download, depending on the tasks required. By way of example, we investigated indicative times taken to create a 1GB data file using both local and cloud-based IT infrastructure. In our simple test, the file could be created on a lap-top or modest office based computer in around 3 to 7s. On a cloud based Azure server, writing to a hosted fileserver, similar times were achieved.

However, the main bottle neck was in writing the file from a local workstation to the cloud, which is mainly governed by the internet upload speed. In tests where the upload was 10Mbps (a typical domestic provision) it took around 20 minutes to write the file, however this reduced to 10-15 seconds via a 1Gbps upload connection. This test, along with Moveout's cloud experience to date, suggests

cloud computing is limited by the user's connection speed, especially regarding upload times, which can vary considerably, but are only really acceptable with a 1 Gbs upload connection."

Further benefits from serverless computing include:

- Task automation / machine learning - speeds up time-intensive processes, e.g. interpretation
- Removal of manual tasks - eliminates repetitive processes, e.g. data sorting, backups
- Collaboration is more easily facilitated, whether in the next office, or across the world
- Reduced cycle times – with demand-driven scaling flexibility options
- Increased information management efficiency – via AI and analytics applications
- Reduced operating costs – with efficiency management processes and pay-per-use charging

Summary

The pros and cons of storing data in the cloud		
	Advantage	**Disadvantage**
Cloud Storage	Avoids holding/storing physical data	You no longer hold a physical data asset
	Improved data access & reduced latency	You are reliant on the cloud host
	Broader & faster analytic applications	Dependant on internet connection
	Removes data duplication requirement	Requires risk mitigation for insolvency of host [1]
	Inbuilt disaster recovery	Cloud outage risk [1]
	Access to serverless computing	Costs may be difficult to ascertain [2]
	Very flexible scalability (autoscaling)	
	Improves operational efficiency	
	Reduces/eliminates IT hardware requirement	
	Unit storge costs likely to decrease with time	
	Applications & options to reduce cost	
	Data sharing is easy and fast - customers > data	

[1] These risks would be significantly reduced by hosting data with a market leading cloud host provider
[2] Whilst offering huge potential for cost efficiency, the range of data accessibility, management and serverless computing application options available in the cloud necessitates careful planning to fully understand and manage costs

Conclusion

Storing seismic data in the cloud promises to remove the requirement to hold or store seismic data on physical media and offers significant advantages in terms of data access and potential for collaboration. The additional benefit of unlimited serverless computing can significantly speed up the application interpretation tools and analytics, and provide easy access to machine learning applications, so long as local connection upload and download speeds are sufficient.

Whilst tiered storage promises to keep storage costs low, it can be difficult to gain clear foresight on costs. Our next article in this series will explore this further. Some seismic data owners may have concerns about the reliability of the cloud host, connectivity and data security not being entirely within their control. These concerns may, however, be mitigated by considering a hybrid solution of storing post-stack data in the cloud and archiving pre-stack seismic data to magnetic tape media, held in a physical storage facility.

Whether stored in the cloud or on tape, it is recommended that seismic data undergo thorough specialist QC and verification ahead of either uploading to the cloud or transcribing onto modern physical media. This will ensure that the data is compliant & contiguous, that duplications are removed, and that it is ready for immediate use.

******

About Moveout Data

Moveout is a specialist in the management of geoscience data, with a mission to enable our customers to fully realise the value of their information asset by delivering optimised, application-ready data, whenever and wherever they need it. Moveout was appointed by the Oil & Gas Authority in 2020 to support the development of the new National Data Repository.

moveout

Better data at your fingertips