EUNIS 2019:
Object storage backup for research and academic data: Discover how the University of Lausanne (UNIL) went beyond NDMP

Speaker 1: Hervé Collard – VP Marketing Atempo
Atempo – 23 Avenue Carnot, 91300 Massy, France – herve.collard@atempo.com

Speaker 2: Pascal Potier – VP Pre-sales & Professional Services
Atempo – 23 Avenue Carnot, 91300 Massy, France – pascal.potier@atempo.com

Keywords
Backup, Data movement, Unstructured Data, ACLs, File Systems

1. SUMMARY

The University of Lausanne (UNIL) in Switzerland is a higher education and research institution composed of seven faculties where approximately 15,300 students and 3,000 researchers work and study. UNIL is focused on several academic disciplines, especially Medicine, Life Sciences, Geosciences, Environment, Business, Humanities and Social Sciences.

UNIL has several petabytes of data stored in three data centers on campus. In 2015, the IT teams concluded that the current backup solution was unable to respect the available backup windows because of increasing data volumes and the growing number of files. Without backup, the only data recovery route would be from multi-site replications. This was deemed to be insufficient to safeguard data sets and data security in the event of a cyberattack.

More generally, unstructured data plays a rapidly growing role in the mix of generated data and consequently storage of data encapsulated in, for example, a structured database format is becoming less preponderant. Finally, Swiss law requires university research to have specific retention periods; this long-term aspect of data protection would be key to the success of the UNIL backup project.

2. STATEMENT OF THE PROBLEM AND BACKGROUND

The volumes of unstructured data are set to grow from 33 zettabytes in 2018 to 175 zettabytes or 175 billion terabytes by 2025 (source IDC). The accessibility and usage of unstructured data for regulatory, analytic and decision-making purposes is driving the need to search and scrutinize this data. What was once cold data stored on tape will be used more and more for analytics, machine learning and business intelligence.

Legacy archiving methods typically fall short when compared to cloud computing and AI applications where extracting value from data is built into storage processing. Traditional data management is moving towards automation and Delivery as-a-Service. In order to reduce costs and relieve IT management overheads, organizations need powerful scale-out unstructured data management solutions for the long-term which ensure data is readily available.

Traditional backup technologies such as Network Data Management Protocol (NDMP) max out at approximately 100TB of data or 100 million files. With too much data and too many files, NDMP does not guarantee disaster recovery but rather forces many organizations to rely on data replication resulting in little or no backup history. Cyberattacks also propagate to replicated data sets. Limitations in terms of IO parallelization or filesystem scanning capabilities mean that NDMP backup technology and legacy backup software has truly served their time.

These traditional approaches are giving rise to many challenges that can be solved by new technologies for storage and data synchronization for petabyte-scale, scale-out NAS and parallel storages. In other words, there will be solutions to continue to protect data even when volumes exceed the scope of standard protection.

3. HOW UNIL FACES THIS STATEMENT OF THE PROBLEM

UNIL's case perfectly illustrates this with a large volume of data generated by the faculty of biology and medicine – for example DNA sequencing tests or the results of microscope samples which generate tens of millions of relatively small files each year. The necessity to respect specific file access security constituted an additional layer of complexity to any future project. The growing threat of ransomware and other cryptoviruses meant UNIL no longer felt sufficient safeguards were in place to ensure data security. In the event of file and folder lockdown, the damage to a primary NAS can quickly spread via replication to secondary NAS. Lastly, UNIL needed identical protection for a mix of file systems (CIFS and NFS).

UNIL's former storage backup relied on NDMP (Network Data Management Protocol). A single full backup could take several days. Reading a file tree structure when the tree contains over approximately 100 million objects and/or over 100 TB of data equates, in short, to NDMP hitting its realistic operational ceiling.

4. PROPOSAL

The solution chosen by UNIL is Atempo's Miria Backup for large NAS. Atempo's solution does not use NDMP but an incremental forever approach where only new, changed or deleted files are part of the backup process. The source storage is EMC Isilon and the target is S3-compatible EMC ECS (Elastic Cloud Storage). The data is moved from the source storage via a pool of dedicated physical data movers managing NFS and CIFS file systems. The internal Atempo database manages file data, metadata and ACLs.

Miria manages data and network load balancing on source storages, data movement machines and target storages. The architecture is fully scalable with the chosen storage to match internal data growth requirements and ACLs (Access Control Lists) are entirely respected in the event of a restoration, even if the restore is to a different environment. Restoration can be granular (file level), directory of volume-level or even complete recovery in the event of catastrophic storage loss.

The standard data protection paradigm which requires the backup solution to perform a “file tree walk” to isolate new, modified or deleted files has been broken. The new approach dispenses with file tree walks and deploys FastScan technology developed by Atempo.

Atempo’s FastScan rapidly collects and processes the list of new, changed or deleted files minimizing the load on the source storage and avoiding full scan. FastScan technology both enables and improves the incremental forever backup process and ensures any restarts are never from scratch. After the first full backup and subsequent synchronizations, incremental forever backups take over and there is no pressure on the available backup window.

5. CONCLUSION

What seemed to be a simple statement of intent: “back up petascale file volumes” was, in reality, a complex project requiring an extensive period of upstream specifications and pre-production testing during the proof of concept phase with Atempo. The reality is that there are few solution vendors - including the storage vendors themselves- capable of managing the movement of several hundred terabytes or multiple petabytes plus millions or even billions of files. Putting an end to NDMP and integrating a valid backup using FastScan to respect the backup window was key.

The implications go beyond the primary need to protect the priceless data resources of UNIL. The ability, in the future, to move or migrate data to new storage locations for analysis, cleaning, archiving or other tasks will be crucial to this type of organization. Machine Learning and Artificial Intelligence will add to these requirements. For example, transferring volumes to high tier storage for analysis and processing before returning this data to lower-tier storage.

UNIL’s academic staff and students can today benefit from this assurance of long-term data protection. The scale-out object storage currently in place and the balance of data mover servers to handle the backup transfers and data and metadata flows will ensure lasting protection even in the event or primary storage loss.

6. SPEAKERS' BIOGRAPHIES

Hervé Collard – VP Marketing Atempo

With more than thirty years of experience in the world of storage, backup and archiving software solutions, including 20 at ATEMPO, Hervé Collard is recognized as an expert in the field of data lifecycle.

Pascal Potier – VP Pre-sales & Professional Services

Recognized as a Technical Expert with more than 20 years of experience in the management of Pre-Sales and Professional Services activities, Pascal Potier has extensive knowledge in the field of data protection as well as in the design of efficient technical and scalable infrastructures.