DELL EMC

HYBRID DATA MANAGEMENT ARCHITECTURES FOR LIFE SCIENCES

Leveraging Cloud Infrastructures to Maximize Efficiency in Life Sciences Workflows

ABSTRACT

This white paper shares insights and best practices to design and implement on-premise, Public Cloud, and hybrid architectures. These architectures mix file and object approaches to achieve an optimal balance between performance, archiving, and data governance and protection requirements.

December, 2016

Table of Contents

EXECUTIVE SUMMARY

The economic benefit of life sciences research is huge across multiple sectors. With increasingly affordable gene sequencing, it is much cheaper to generate larger volumes of raw data at much higher rates. But storing all this raw and derived data quickly, reliably, and cheaply for the very long term is challenging. Additionally, to generate valuable insights, it is imperative to quickly prepare, analyze and interpret the rapidly growing data volumes while keeping pace with instruments' output speed.

To address these challenges, Research IT teams must implement the right balance of compute and storage that conforms to the budget, data governance and other organizational policies. For this, these teams are actively assessing various file and object storage solutions.

This white paper shares insights and best practices to design and implement on-premise public, private, and hybrid cloud architectures. These architectures mix file and object approaches to achieve an optimal balance between cost, performance, data governance, and protection requirements. Several life sciences institutions are deploying these hybrid architectures based on file and object offerings from Dell EMC.

OVERCOMING DATA CHALLENGES TO REALIZE LIFE SCIENCES VALUE

The rate of progress in genomics technology is astounding. Rapidly declining gene sequencing costs, advances in recording technology, and affordable clustered compute solutions that process ever larger datasets is transforming life sciences research. Today, a human genome can be sequenced within a day[1] and for about $1000, a task that took 13 years and $2.7 billion to accomplish during the Human Genome Project.[2]

By 2025, the economic impact of next-generation sequencing (NGS) and related technologies could be between $700 billion to $1.6 trillion a year[3]. The bulk of this value results from the delivery of better healthcare through personalized and translational medicine. NGS enables earlier disease detection, better diagnoses, discovery of new drugs, and more personalized therapies.

But the rate of growth of genomics data continues to explode. For instance, the Illumina HiSeq X Ten System—designed for population-scale whole genome sequencing (WGS)—can process over 18,000 samples per year at full utilization. Each HiSeq X Ten System generates up to 1.8 terabytes (TB) per run. When the HiSeq X Ten System operates at scale, it can generate as much as 2 petabytes (PB) of persistent data in one year.

As the cost of sequencing instruments become more affordable, smaller institutions are increasingly deploying them. Even larger existing research organizations are purchasing more instruments. This only compounds the growth of distributed raw data. Raw data must be consolidated, aligned and packaged; making storage requirements even greater. Unfortunately, storage costs are not declining as fast as sequencing costs.[4] Estimates are that in 2025, 2 to 40 Exabytes (EB) will be required just for the human genomes.[5]

But beyond storage systems acquisition costs, life sciences research organizations are realizing that the operating costs (including downtime and productivity loss) of managing, securing, tracking, and cleansing these exploding volumes of data are growing even more.

KEY CONSIDERATIONS TO MANAGE THE DATA DELUGE

Data volumes and access patterns intensify and vary widely as the use of life sciences data becomes more prevalent in time-critical clinical/diagnostic analyses. What took days to analyze in a pure research context must now be done reliably in hours, even as larger number of projects and files must be tracked. Research IT must collaborate closely with end users to agree on investment decisions, data management policies and infrastructure deployment choices. This helps optimize the total Costs and Benefits for the entire data lifecycle.

Costs

A range of key direct and indirect costs should be carefully considered:

- IT: Capital for servers, storage, networks, software purchase, etc. or corresponding Public Cloud service provider charges. Today, many IT organizations' budgets are narrowly focused on these direct costs; ignoring indirect costs that may hamper innovation and productivity.

- Operational: Labor (salaries for end-users and IT staff), energy, IT hardware and software maintenance, software license, etc.
- Other: deployment and training, downtime, bandwidth, productivity loss, lack of adequate compliance, security and data protection, etc.

Benefits

The benefits from an IT investment for life sciences come in several categories:

- Strategic: better patient outcomes, improved reputation for the institution, better stakeholder partnerships, leadership IP portfolio, ability to attract and retain top talent
- Research: more innovation, better collaboration, greater insights, improved quality
- Operational: faster time to results, greater throughput and more users supported, improved user productivity, better capacity planning
- Infrastructure: improved system management, administration, and provisioning, enhanced security, higher utilization, scalability, reduced downtime, access to robust proven technology and infrastructure management expertise

Storage Choices

Storage architectures (Private Cloud or Public Cloud) and Data Management policies have a significant impact on costs and benefits. Key influencing factors include:

- Data Location and Movement: As data volumes grow exponentially, the costs of moving the data in and out of a compute location becomes prohibitive. So, it is imperative to keep the data local to the best compute location. This minimizes data motion and reduces access overheads especially when reusing the same data.
- Applications and Analysis Pipelines: The raw instrument data is consolidated, aligned, and packaged by multiple automated applications and analysis pipelines working in tandem with end users; typically increasing the active data size by a factor of two to three.
- Performance: As more data is generated and stored, more data must be processed on the hundreds of compute cores in a cluster. This means that the storage systems must perform and be able to feed these cores to keep the pipelines operating in full-throttle.
- Active and Archive Data: Once data passes into the archive tier as part of a repository, it is important to quickly access data and metadata when needed, regardless of where it is and which operating system is requesting the file. The frequency and type of access determines specific data management policies.
- Data Security, Privacy and Protection: Key considerations include: How secure and private is the data? Is data stored in a redundant manner to ensure recoverability? How much control does the user have over remote storage? These are especially critical in clinical or commercial settings.
- Data Transfer Speed (Public Cloud): How to move large datasets to and from the Public Cloud efficiently, especially frequently changing data that must be ingested from a range of sources? The limited Wide Area Network (WAN) bandwidth to and from the Public Cloud makes the data transfer process manual, cumbersome and expensive. This significantly impacts the productivity of scientists who often collaborate across multiple locations. So it is critical to minimize bulk data transfers.
- Compliance (Public Cloud): Most life sciences data are subject to compliance/regulatory constraints. This is cumbersome and expensive on the Public Cloud. So, having the data onsite avoids many of these issues.
- Staff: What are the ingrained institutional processes and procedures? Can the staff easily adapt to new technologies and/or processes? Do they have the requisite skills?

But to achieve this economically, it requires implementing fast file storage for the smaller-sized (10s of TB) active data; combined with lower-performance and more cost-effective storage for the larger-sized (10s of PB or more) archive data. Leading life sciences organizations[6, 7] are precisely doing this; implementing hybrid storage systems consisting of very fast scale-out Network Attached Storage (NAS) for smaller-sized active data and more economical Object Storage for the larger archive data.

BENEFITS OF HYBRID STORAGE: COMBINING FILE AND OBJECT STORAGE

A hybrid cloud is a composition of two or more clouds (private-public, private-private, or public-public) integrated in a way that enables data and application portability. Cloud-to-cloud integration is across a single workload as well as between workloads. While the entire hybrid storage solution can be deployed onsite, cloud computing (particularly hybrid cloud) is also being used for compute and/or storage.

Both scale-out NAS and object storage can scale storage and are complementary that each offer unique strengths. Scale-out NAS typically has a single highly scalable file system across a cluster with a single name space. This simplifies adding new capacity as the

number of shared files grows, while delivering excellent performance for local files. But performance could suffer in large cloud-scale environments particularly those that cross geographical boundaries. Dell EMC Isilon is the leading scale-out NAS system that consolidates files onto a single, shared pool of storage with its OneFS operating system. It delivers flexibility, extensive multiprotocol support and ultra-high performance.

Object storage can scale almost infinitely and provides a simple way to manage storage across multiple geographies with rich (and user-definable) metadata. However, deploying object storage often requires changes to the application code; and for smaller-sized (10s of TB) active data, performance could suffer.

Metadata can include content, retention, data protection, security and other organization-specific information. This metadata can be used to quickly search and track data as well as automate the management of stored objects throughout the lifecycle for a much broader range of applications. These are very valuable features for extremely large cloud-scale (10s of PB or more) data. Dell EMC Elastic Cloud Storage (ECS) is reliable, cost-effective cloud-scale object storage with geo-replication protection.

Many life sciences organizations are discovering that a combination of products works best for different needs within the same environment as data moves from active to near-line archive to deep archive to permanent archive tiers. They are deploying a hybrid storage architecture which combines scale-out NAS with Object Storage.

DELL EMC HYBRID STORAGE SOLUTION OPTIONS FOR LIFE SCIENCES

Dell EMC provides highly optimized storage solutions for both file and object storage.

Active Data: Dell EMC Isilon S or X Series provide a large, high-performance shared file system to concurrently gather and process the predominantly streamed and sequential incoming instrument data. This data is typically actively shared by on-site researchers. So it is important to have file-locking capabilities and the ability to manage permissions and read/write access of very large files across both Windows and Linux Operating Systems. Dell EMC Isilon supports this over the course of multiple back-to-back sequencing runs through concurrent read/write operations. To accommodate the high throughput of data, a single file system is layered across multiple Isilon nodes, allowing each to serve the same namespace to a wide number of connected clients.

Archive File Data: Dell EMC Isilon NL Series provides cost-effective, highly scalable nearline storage and data-at-rest encryption (DARE) with self-encrypting drive (SED) options to meet rigorous compliance and security needs. The Dell EMC Isilon HD Series is a highly efficient and resilient scalable storage platform that scales over 68 PB in a single file system. It provides a very cost-effective deep archiving solution with robust data protection and security options.

Within OneFS, Smartpools provides the ability for customers to automatically move data between storage pools based entirely on rules they select. This automated data management functionality is highly useful when managing performance pools and archive pools within the same cluster.

Archive Object Data: In order to benefit from all its attractive capabilities, organizations interested in using an object store archive, can use Dell EMC Elastic Cloud Storage (ECS) on commodity infrastructure. ECS delivers a complete storage platform – as an appliance or just software – with multi-tenancy, metering, self-service provisioning, etc. The ECS Object Service is compatible with Amazon S3 and OpenStack Swift.

Dell EMC ECS features include:

- Universal protocol support in a single platform with support for block, object, and Hadoop File System (HDFS)
- Single management view across multiple types of infrastructures
- Multi-site, active-active architecture with a single global namespace enabling the management of a geographically distributed environment as a single logical resource using metadata-driven policies to distribute and protect content, and
- Multi-tenancy support, detailed metering, and an intuitive self-service portal, as well as billing integration.

Metadata Management: Dell EMC is also contributing to iRODS—integrated Rule-Oriented Data System—a widely deployed open-source system for managing large volumes of data that requires extendable metadata. iRODS virtualizes data storage resources, and data can be tracked, cataloged, operated upon in limited fashion without end users dealing with the complexities of the underlying file systems. It provides many of the attractive attributes that object storage systems provide and is compatible with Dell EMC Isilon and other systems.

EVALUATING INFRASTRUCTURE DEPLOYMENT OPTIONS

Life sciences organizations have several alternatives to deploy the generalized hybrid storage architecture. Each choice, including several that leverage the cloud for computing and/or storage, has several advantages and limitations. Key choices include:

All Private Cloud: The entire storage and compute infrastructure is all onsite. This is well understood. Existing institutions often have optimized their workflows and methods for sharing data for onsite use. It is also easy to de-duplicate copies of data sets and recover data when the data owner is no longer at the institution. Unlike on certain Public Cloud storage systems, assured deletion of data is not a concern.

All Public Cloud: In recent years, primarily driven by lower usage costs (for compute and storage) per user, there is a growing interest in using public clouds for life sciences workloads. Key benefits of public clouds for some life sciences clients and workloads include:

- Unified, location-independent platform for data and computation
- Pay-per-use and available even to small labs
- Affordable infrastructure costs for individual end-users
- Linear scaling with parallel execution for some workloads
- Ability to quickly scale to large number of compute resources when needed
- Basic infrastructure deployment and management complexities handled by cloud provider
- Up-to-date base technology platform for processing and storage
- Better collaboration between scientists with a centralized computing environment.

But there are many challenges with running all life sciences workloads entirely on Public Clouds. Some limitations that magnify with the number of Public Cloud users and/or data volumes include Data Security and Protection, Data Transfer Speed, and Compliance.

Hybrid Cloud: These approaches have the potential to offer a better cloud solution for a broad spectrum of life sciences workloads at a lower Total Cost of Ownership (TCO) with data protection, security and regulatory compliance. Key scenarios include:

- Compute on the Public Cloud with onsite archive: Many analytics algorithms benefit from the fast and scalable compute on the cloud. Onsite archive from scale-out NAS systems such as the Dell EMC Isilon NL or HD Series.
- Compute on the Public Cloud with all storage onsite: If all storage is object and delivered with ECS; organizations benefit from extreme scale and very low costs. For ultra-high performance, Dell EMC Isilon S or X Series can be added for active file management.
- Compute on the Public Cloud with storage deployed both onsite and in the Public Cloud: For ultra-high performance, Dell EMC Isilon S or X Series can be added for active file management. For near-line archive, depending on the type of storage needed the Dell EMC Isilon NL or HD Series or ECS can be used. For deep archive, Public Clouds can also be used.

CONCLUSIONS AND RECOMMENDATIONS

Life science technologies like NGS and Imaging can have a huge economic impact. But storing and analyzing all this raw and derived data quickly, reliably, and cheaply for the long term is challenging.

Many life sciences research organizations are implementing a hybrid approach with high-performance file and cost-effective object storage to overcome these performance and capacity challenges—bringing together the best attributes of file and object storage.

Dell EMC's hybrid data management solutions for life sciences research consists of:

- File Storage: Dell EMC Isilon scale-out NAS based on OneFS – a networked file system – provides the requisite high-performance, throughput and onsite capacity for active data (Isilon S and X Series) and archive data (Isilon NL and HD Series).
- Object Storage: Dell EMC Elastic Cloud Storage (ECS) is a reliable, cost-effective cloud-scale object storage platform with geo-replication protection. With active-active read/write support with strong consistency, Dell EMC ECS provides efficient near line access and performance for small and large objects before going to permanent archive.

These Dell EMC storage solutions can be fully implemented onsite and/or on a hybrid cloud.

REFERENCES

[1] http://www.prnewswire.com/news-releases/study-whole-genome-sequencing-technology-enables-26-hour-diagnosis-of-critically-ill-newborns-nearly-halving-previous-record-for-speed-300151018.html
[2] http://www.veritasgenetics.com/documents/VG-PGP-Announcement-Final.pdf
[3] McKinsey Global Institute,"Disruptive technologies: Advances that will transform life, business, and the global economy", May 2013.
[4] http://www.genome.gov/sequencingcosts/
[5] Zachary D. Stephens, et. al., "Big Data: Astronomical or Genomical?" PLOS Biology, 2015.
[6] Chris Dwan, "Research Computing @Broad", Bio-IT World Expo, April 2015.
[7] Dirk Petersen, "Economy File Project", Fred Hutchinson Cancer Research Center, Bio-IT World Expo, April 2014.