Qumulo File Data Platform Solution Profile

2022-23 DCIG TOP 5

RISING VENDORS IN STORAGE FOR LIFE SCIENCES SOLUTIONS

By Mike Matchett

Rising Vendors in Storage for Life Sciences Solutions

Table of Contents

Rising Vendors in Storage for Life Sciences Solutions

SOLUTION

Qumulo File Data Platform

COMPANY

Qumulo, Inc.
1501 4th Avenue
Suite 1600
Seattle, WA 98101
+1 (206) 260-3588
Toll Free +1 (855) 478-6856
qumulo.com

DISTINGUISHING FEATURES OF QUMULO

- Simple global hybrid cloud
- Consistent workflow integration
- Real-time analytics

DISTINGUISHING FEATURES OF TOP 5 SOLUTIONS

- File performance
- Large capacities
- Broadening application support
- Resiliency at scale

SOLUTION FEATURES EVALUATED

- Deployment capabilities
- Data protection
- Product and performance management
- Technical support
- Licensing and pricing

Rising Vendors

DCIG recently conducted research into storage for life sciences. DCIG classifies vendors with revenues less than $200M as "rising vendors." The use cases are the same, but rising companies tend to be younger companies bringing newer technologies and disruptive solutions to market. Those organizations looking for modern approaches and potential competitive advantage may want to consider working with a rising vendor.

Storage for Life Sciences Challenges

Life Sciences organizations depend on some of the most compute and data-intensive applications in the world for primary research, on-line analysis, global collaboration and product development. Many of these applications are High Performance Computing (HPC) quality workloads that include genomics sequencing, molecular simulations, protein folding, AI/ML optimization, and intensive media processing. These applications can push even cutting edge IT data storage implementations to their performance and capacity limits.

And increasingly, upstream life sciences workloads feed critical data into downstream workflows that might for example manage real-time medical-grade production lines or oversee global distribution.

The stakes are high for IT in life sciences with competitive global pressures, the search for life-saving solutions, expensive data "source" equipment, elite research staff and ever-increasing data volumes and performance demands. For example, voracious genomics sequencing equipment can easily generate TB's of raw data in a few hours, overwhelming local legacy file storage often configured with the default operator workstation. This data must then be offloaded into downstream research-feeding storage while that equipment is idling, wasting opportunity, time and resources.

Other workloads present challenges too—molecular simulations running on scale-out HPC clusters impose HPC data consumption patterns, which require local high-speed parallel file IO of very large or very many files to many client nodes at once. Many life sciences workloads need to strongly leverage critical GPU resources, and maximizing the utilization of large numbers of GPUs may require specific GPU-related storage features. And on-line data capacity demands mount with each type of research conducted, each form of analysis required, and every AI/ML model built.

Traditionally, data requirements for these highly demanding applications have been served with parallel file system solutions. Parallel file systems can deliver massive IO volumes to data hungry applications (like those running in large HPC or HPC-like compute clusters), but historically implementing, operating and tuning massively parallel file systems for top performance at scale has been the province of dedicated PhD level academics.

Today, huge volumes of data are not only created and processed by primary life science applications, but that data then must be shared with multiple consumers and global collaborators. It is obviously expensive (and often prohibitively slow) to make and maintain many multiple copies of very large data. It can also be quite expensive to maintain large histories of data online for downstream research activities. It is traditional to archive HPC data sets into object storage for downstream use. However, this secondary

storage dataflow means that data lives in multiple locations and in different forms, adding friction and delay to sharing, complicating data access and preventing optimal data value extraction.

In addition, "academic" storage solutions have tended to be light on enterprise storage management features like data protection, security and backup/disaster recovery functionality. While primary HPC storage must first meet the HPC application requirements, it is increasingly the case that research data stays on-line for longer time periods to be recalled and leveraged on-demand and accessed by a wider set of applications. Maintaining this key data over time properly then becomes more critical to the ongoing overall success of the organization.

An effective storage solution will ingest and store data faster than source equipment can produce it and then deliver it as fast as the sum of consuming workloads calls for it.

Benefits of an Effective Storage for Life Sciences Solution

At the top level, scalable storage performance is critical. A life science organization should try to fully utilize their high-end research and laboratory equipment, HPC clusters and GPU-intensive analytical servers. The objective of life science storage then is to store and flood massive amounts of scientific data into all the expensive data pipelines serving the organization's goals. An effective and equally scaled storage solution will ingest and store data faster than source equipment can produce it and then deliver it as fast as the sum of consuming workloads calls for it.

Scalable storage will scale-out almost indefinitely, as needs and requirements grow or expand. The best solutions can offer a single "namespace" for all files even as the data storage under management grows into exabytes, eliminating multiple arrays, needless replication and extraneous data copies. Storage solutions that really support massive scale-out to thousands of devices also offer resilient designs and non-disruptive upgrade/repair features to avoid single points of failure and downtime.

Perhaps even more important than maximizing resource utilization is fully empowering scientific staff. The top storage solutions work transparently to keep data flowing to all users, in real-time on-demand, driving their applications and analytical workflows without lag or downtime. The best scalable storage solutions also present simple (and automated) data storage interfaces, freeing staff from onerous storage management concerns and enabling them to focus more on their productive research.

There is plenty of competition in the global race to life science insight and discovery, with massive opportunity for organizations that can most efficiently leverage resources, empower researchers and minimize IT risk and distraction. Life sciences storage can make a significant difference in organizational outcomes by delivering world-class performance, scaling readily to handle the largest of online data requirements and significantly increasing organizational efficiency.

Finally, we are seeing top end storage solutions increasingly supporting downstream and collaborative workflows through wider multiprotocol support, native storage tiering, inherent data protection, multi-tenancy and data security features. Overall, top life sciences storage solutions, even though often accelerated on high-end appliances or custom hardware, are becoming more cloud-like in utility, economics and management.

Distinguishing Features of Storage for Life Sciences

In addition to the broad capabilities mentioned above, all of the life sciences storage solutions evaluated in this report share some features that help distinguish them from the broader IT storage market.

File performance. First, these are not simply scaled up NAS solutions, but designed from the start for high-end file storage performance and capacity.

Large capacities. Scalable architectures are the norm, with the ability to add or expand on-line and tiered capacity without grossly affecting operations or performance.

Broadening application support. Life sciences research encompasses a wide variety of applications, usage, access, cost, risk and data management concerns. The storage solutions evaluated for this report have demonstrated significant utility in some slice of life sciences and threaten to increasingly consolidate storage with broadening application support.

Resiliency at scale. All of the life sciences storage we evaluated have features that address the resiliency and resulting availability of the solution at large scales of deployment.

Similarities Among the DCIG TOP 5 Rising Vendors in Storage for Life Sciences

In addition to the distinguishing features above that all of the evaluated storage solutions share, the selected DCIG Rising Vendors in Storage for Life Sciences solutions have the following traits in common:

File performance first. The top solutions need to be highly performant file storage solutions designed to serve HPC class performance requirements for high-volume data reads and writes.

Boundless scale-out. The top solutions all are capable of scaling-out to many PB's (or even EB's) of storage without impacting performance, often to hundreds or thousands of storage/server nodes.

Large and small files. High-speed ingest and performant read of large files to feed HPC-class workloads are key life science storage design points, but also increasingly these systems support billions (or trillions) of small files, random IO access patterns and low latency applications.

The DCIG TOP 5 rising vendor solutions also deliver the following product features:

Files integrated with objects. All of the top solutions provide unified object services or internally tier to cloud and/or cloud-like object storage.

Integration simplicity. They all provide modern storage consoles for management of storage at large, but also REST APIs for management integration at scale.

Aim to ease management burdens. These solutions are all designed to ease storage management burdens at scale especially those caused by siloed NAS deployments.

Rising Vendors in Storage for Life Sciences Solutions

Qumulo is especially suited for the capacity storage of petabytes of media files (images, video) while simultaneously servicing demanding video processing or research workloads with cloud-elastic or bursty IO patterns.

Qumulo

Upon completing its independent research and marketplace analysis, DCIG ranks Qumulo as a DCIG TOP 5 rising vendor in storage solutions for healthcare.

Qumulo is a cloud-native, "single-tier" global file system designed for efficient, extreme-scale unstructured data capacities and HPC-class file performance. Qumulo is especially suited for the capacity storage of petabytess of media files (images, video) while simultaneously servicing demanding video processing or research workloads with cloud-elastic or bursty IO patterns.

Qumulo applies native intelligent file-level analysis, predictive pre-fetch and caching best utilize assigned NVMe resources to maximize performance, while delivering linear scalability through auto-tiering across diverse hybrid cloud and active-archive storage capacities. Qumulo clusters can be deployed across all major public cloud providers, HPE, Dell, Pure, Fujitsu and other underlying storage arrays or are available through Qumulo pre-integrated hardware.

Three of the key features that earned Qumulo a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include

Simple global hybrid cloud. Qumulo delivers a single namespace with multiprotocol access (SMB, NFS, FTP, REST) across all assigned storage. No application migration or transformation is required to access files in any environment, including in the cloud, easing both storage consolidation and data collaboration tasks.

Consistent workflow integration. Rich API's from Qumulo enable the tight workflow integration of storage with complex data pipelines and critical application workflows found in key use cases like genomics, molecular simulations, image processing and PACS.

Real-time analytics. Qumulo provides instant data and usage visibility across billions of files. The resulting intelligence enables the "full utilization" of storage resources and powers the Qumulo platform's automatic predictive file-level pre-fetch and caching for high performance. ■