genomeweb

Jackson Laboratory Modernizes its Research Data Storage Infrastructure

Sep 30, 2021 | Neil Versel

Premium *Save for later*

CHICAGO – The Jackson Laboratory is nearing the end of a multiyear project to upgrade and reconfigure its data storage to support rapidly growing research activities. To increase efficiency of storage management and help with grant budgeting, the institution is now slowly phasing in a plan to assign storage-related costs to each research project.

According to Shane Sanders, Jax's senior manager of cyberinfrastructure, the lab has spent between $5 million and $10 million reorganizing and restructuring data storage since he joined in early 2016.

Sanders, who was elevated to his current position in June, is in charge of all the computing infrastructure for research activities. The business and operational side of Jax has its own computing resources and manager, but that was not always the case.

Sanders said that the high-performance computing cluster at Jax was not well managed prior to 2017. For his first eight or nine months at Jax, Sanders spent most of his time observing the computing environment and gathering information from research users.

"All of our eggs [were] effectively in the same basket," with the genomic sequencing lab at the Jackson Laboratory for Genomic Medicine in Farmington, Connecticut, using the same storage as more traditional researchers at Jax headquarters in Bar Harbor, Maine. Sequencing workflows often failed, sometimes with irreplaceable samples.

Jax thus designed systems for core facilities and established quotas and queues for storage computing systems, and standardized data intake from scientific instruments. "We went from one large basket full of eggs to many smaller baskets," Sanders said last week during a virtual presentation to last week's Bio-IT World conference in Boston.

He said that Jax has seen tremendous growth in the last decade and a half, in part due to the proliferation of next-generation sequencing and the 2014 opening of the institution's genomic research center on the campus of the University of Connecticut Health Center.

The old model of *ad hoc* provisioning of computing resources on the space-constrained Bar Harbor campus — which largely conducts more traditional mouse-based research — was no longer viable, Sanders said.

When the COVID-19 pandemic struck in March 2020, Jax had the computing infrastructure to handle the abrupt shift to remote working. The organization had brought its newest HPC cluster online in December 2019, and it had not seen strong uptake until the pandemic hit, Sanders said.

The modernization has not been easy. "There is also a significant cultural inertia to change in the computational biology community," noted Sanders, who has a Ph.D. in computational biology. "The situation is very much like rebuilding a World War I-style biplane into a modern jetliner while in flight."

Early in his tenure, Sanders worked out an asset-management plan with the finance department at Jax, aligning hardware purchases to the organization's five-year depreciation schedule. "This got us out of the cycle of piecemeal capital purchases and away from the often painful price negotiations with hardware vendors for

purchasing support for our patchwork pool of hardware assets."

Every compute node now has a direct path to the storage infrastructure.

Significantly, Sanders also classified several tiers of storage at Jax, labeled 0 through 3, with 0 and 1 being in an HPC environment.

The 450 terabytes of storage classified as Tier 0 is transient, meant for high input-output activities. Jax is in the process of upgrading this disc-based storage to an array of 12 Dell EMC Isilon servers linked to 600 TB of Dell EMC PowerScale F600 flash storage. Files are deleted 10 days after they are created.

"A lot of the bioinformatics tools generate hundreds of intermediate files that people don't look at or only look at briefly," Sanders explained. "[Tier 0] is really to provide some flexible space for people to do their computation."

Tier 1, for persistent storage, has a capacity of 7.25 petabytes and is entirely housed in Farmington. This, Sanders said, is for holding active research data over a longer term.

Tier 2, with 5.36 PB of non-HPC research storage, is split between Bar Harbor and Farmington. It is available to research scientists when they do not need a high-throughput environment.

This tier is kind of a bridge between persistent and archival storage. "We do allow the researchers to mount it to things like their laptops or various virtual machines or instrumentation computers, because it's more of a file share-grade storage in terms of the performance we're seeing out of it," Sanders said.

It is suitable for generating data and documents for publication, he added.

Jax also has a research data archival platform dubbed Tier 3. The organization this year completed a migration to a system featuring 20 Dell EMC Isilon A2000 flash storage nodes at Bar Harbor and another 20 at Farmington, offering a total usable capacity of 2.85 PB, replacing disk storage and tape backups.

The IT department compresses large datasets when it saves data here. Users are asked not to submit requests for Tier 3 access if they expect to need the information again within six months. Eventually, Sanders wants to migrate "colder" data to a cloud platform.

Sanders and the research IT team found that the legacy HPC cluster frequently got overloaded and saw a drop-off in file system performance when platform demand exceeded 85 percent of capacity. "Users had no incentive to manage their storage effectively," he said. He said that he regularly found text-based log files as large as 2 TB that had not been accessed in two years or more. "There was no incentive to archive, compress, or delete older data, and some even used the expensive storage as a backup for their vacation pictures and home movies."

That situation has spawned a "chargeback" program that the research IT group at Jax is now implementing so researchers understand the true time and financial costs to the institution for using the various storage tiers.

"Your typical wet-lab biologist has never been trained on application profiling and may not have ever used an HPC cluster," Sanders said. "This can lead to significant overestimates in the resources they ask for in their computational work."

With the chargeback program, Sanders wants to "provide a more efficient and effective planning mechanism around our HPC data storage," which he believes will improve forecasting and budgeting. "The goal here is not to provide a cost recapture mechanism," Sanders said.

Labs are given monthly storage quotas in tiers 1 and 2 and are sent added charges if they exceed those limits. IT fully subsidizes tiers 0 and 3. The program is being phased in slowly so labs can learn to incorporate the costs into grant applications, with charges not hitting 100 percent of costs until 2025, though Jax fully subsidizes the archive tier and has promised to continue doing so indefinitely.

"If somebody comes tomorrow and says, 'I need an extra 500 terabytes of storage,' that probably is going to trigger some financial discussions," Sanders said. "If somebody comes tomorrow and says, 'I need an extra 5 TB

of space for the next two months to finish this project out,' we've probably got that in reserve."

Currently, the plan is essentially in a pilot stage to educate researchers about the amount and types of storage they are consuming. This "showbacks" activity has been in place for about two years now, and Sanders said that the organization is ready to start phasing in the actual chargebacks program in July 2022.

"There's a little bit more incentive for people to go in and clean up and work with us to get things archived and move to the right tier," Sanders said. He noted that the "showbacks" program has reduced the number of requests for additional storage from faculty members who would otherwise let terabytes of data from since-departed postdoctoral researchers sit unused in a low storage tier.

"We can work with those researchers and help them identify … the stuff [they] want to keep, and we can archive everything else," Sanders said. "As it phases in, we're working with our grant writing and our sponsored programs office to help make sure those costs are accommodated," Sanders said.

"While I am a research scientist and I can probably sit down and help identify the files that are valuable and what aren't, most of my team aren't," Sanders added. "They really need a partner from somebody in the labs [who has] done this."

Filed Under + Informatics + Jackson Lab + North America + storage + research center + HPC + computational biology + Next-Generation Discovery Workflows + Editor's Pick