Data Science
and Analytics
Drive Life Sciences
Innovations:

Is Your IT Infrastructure Ready?

Underwritten by

WEKA.IO
Radically Simple Storage

Introduction

Thanks to advances in next-generation sequencing and imaging lab equipment, researchers in life sciences have an opportunity to explore new drug candidates, better understand the root causes of diseases, identify new biomarkers associated with specific diseases, and personalize treatments based on a patient's genetics.

As more data becomes available, researchers find they also have new analytic methods to explore that data. And the variety of techniques is rapidly changing. Today, there is growing use of predictive analytics, neural nets, machine learning, and other technologies that just were not considered practical to use a year or two ago. This is made possible due to many factors. For example, the newest generation processors give each node more raw processing power, GPU use is on the rise to accelerate computations, and cluster technology such as Hadoop makes distributed computing available to the masses.

In some ways, life scientists have a perfect storm for conducting revolutionary research in areas ranging from new drug discovery to precision medicine. They have vast volumes of data to work with, more powerful analytics to explore that data, and access to high-performance compute cluster technology to run their models and analytic routines.

New research highlights the potential of modern analytics applied in the life sciences. Examples include:

- A machine learning system that accelerates drug discovery by predicting whether a candidate drug molecule would bind to a target protein by predicting with 99 percent accuracy
- Medical image analysis that assists doctors diagnosing patient symptoms (often outperforming the doctors) and helps determine the best treatment option
- A system that combines predictive analytics with machine learning to eliminate common drug discovery pipeline problems by predicting molecule toxicity and reactivity without the need for expensive trials

For more details about these applications, **click here**

However, not all organizations can use modern analytics tools to their full potential. Due to the volumes of data involved when conducting innovative life sciences research today, most IT infrastructures cannot support the analytic workloads nor deliver results in a timely manner. Simply put, in many organizations the IT infrastructure is an obstacle to achieving success.

> **"Simply put, in many organizations the IT infrastructure is an obstacle to achieving success."**

Life Sciences Big Data Analytics Challenges

Life sciences lab equipment is producing enormous amounts of data. Old IT infrastructures are quickly overwhelmed with this volume of data and introduce significant bottlenecks in analytic workflows.

To put the issues into perspective, consider the amount of data routinely being generated in labs today.

Next-generation sequencers produce terabytes (TBs) of genomics data per experiment. A lab might typically have several of these devices running simultaneously. They would generate petabytes (PBs) of data a month that needs fast analysis to speed new drug discovery or better understand a biological entity.

In some labs, other equipment is starting to produce even more data. Light sheet microscopes can generate up to 25 TBs of data a week. Medical research centers routinely use X-rays, CT scans, and MRIs that produce increasingly higher resolution images. Those images need fast analysis and interpretation to diagnose a patient's problem and prescribe a treatment or course of action.

For quick analysis, data from these devices must be staged on the highest performance storage (SSDs and all-flash arrays) to keep expensive HPC resources fully utilized. Given the multi-disciplinary nature of life sciences research, data is often reused and shared (in collaborative research efforts) by different researchers at different times. And as such, must be easily accessible throughout its lifetime. In many cases, where government agencies fund the work, shared access is mandated by the funding grants. In most contracts, data must be retained and made available for six to seven years.

Additionally, genomic and bioinformatics Big Data analysis workloads vary significantly from lab to lab and project to project. Thus, any HPC / storage solution must be able to support a wide variety of demanding I/O and throughput requirements. In many cases, there is a need for very low latency, especially if using GPU servers that are quite costly. Labs do not want them sitting idle.

Trying to accommodate these conditions is the grand challenge in the life sciences. In fact, "storage and data management are the two biggest headaches," said Ari Berman, vice president and general manager of life sciences computing consultancy BioTeam, in **an interview with HPCwire's Editor John Russell.**

What's Needed?

Life sciences researchers and organizations want to focus on innovative research and do not want to get bogged down with arduous IT infrastructure management chores.

They need a storage solution that removes the performance bottleneck, scales with growing demands, works with existing infrastructure, and is easy to use. A significant challenge is that typical workloads must work with datasets that have very different file characteristics, as well as wide-ranging I/O, throughput, and read/write requirements.

Some examples of common data types and workloads include:

- Large binary and text files: Many labs work with large flat text files of DNA or protein sequences, such as BLAST formatted databases. And those studying genomics with work with Fastq and SAM files. Running a sample against one of these databases or comparing properties requires streaming the entire file to nodes for analysis. A suitable storage solution would need low latency and high sustained throughput.
- Many tiny files in a single directory: Output from mass spectrometers often consists of many thousands of tiny files (a few kilobytes each) sent to a single directory. A storage solution must be able to handle many reads taking into account both the data and metadata associated with each file.
- Many files in a complex file and folder hierarchy: Output from some of the commonly used next-gen sequencers can produce multiple nested folders of data. A storage solution with an easy to use global filesystem can help make locating and managing data over time much easier and faster, saving precious time and money.

Algorithms that access these different types of datasets create different stresses on the infrastructure. For example, when performing a BLAST search, many bioinformatic analysis routines compare a query file against the entire database. This requires ingesting database as the whole resulting in long sequential reads of a big file. This action is performed once for the execution of the algorithm. So, storage and networking solutions must be capable of sustained high throughput to make the analysis job run in the fastest time.

In contrast, other work such as molecular modeling or structure prediction often requires working with many small algorithms and workflows that create highly variable random I/O access patterns and read/write requests.

As the use of AI and deep learning in the life sciences grows, a properly selected storage solution becomes even more critical. Typically, organizations use specialized servers (usually making use of GPUs to accelerate workloads) that can cost $100,000 or more per node for their AI efforts. A ten-node system means an organization has one million dollars of compute power available to run its workloads. A low-performance storage solution will not be able to make full use of this capacity. If the nodes sit idle 20 percent of the time, the organization has effectively lost $20,000 per node or $200,000 for the entire system. This does not consider the indirect costs of lost researcher and scientist productivity and delayed results.

Further complicating matters, many AI, deep learning, and other life sciences analytics applications must be able to scale to meet future compute demands. Traditionally, the fastest way to crunch data was to use direct-attached storage. But as compute requirements grow, a better alternative might be to use a shared file system that provides comparable performance to keep an organization's high-priced HPC servers running at peak utilization.

Simply put, life sciences research involves some of the most complex analytics workloads being run today. Each organization has unique needs concerning compute and storage performance, scalability, and accessibility. For instance, file-based storage is ideal for high performance compute clusters used during the analysis phase of the workflow. For long-term storage and sharing of valuable research data, a cloud-scale solution based on object storage is a more efficient and cost-effective approach.

What's needed is a storage solution that offers the best of shared scale-out file-based storage and parallelism coupled with the performance and low latency of flash memory. This will provide the flexibility, economics, and performance needed for advanced research.

Selecting the Right Technology Partner

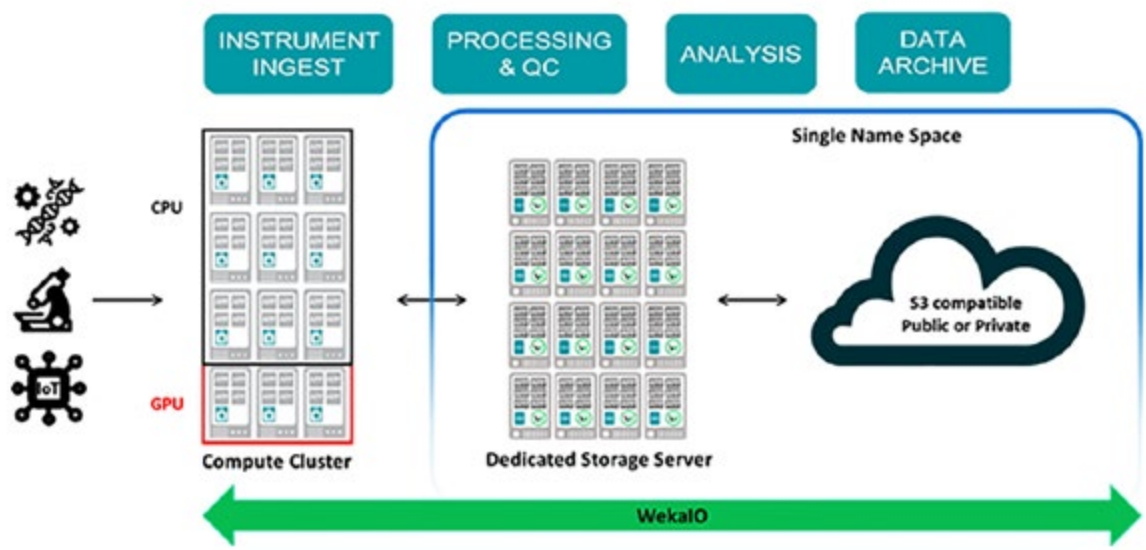
WekaIO is a technology partner that brings years of expertise in life sciences Big Data analytics to help organizations overcome IT infrastructure obstacles. It has developed a flexible architecture that supports the entire workflow.

The combination of WekaIO Matrix™ and any S3 compatible object storage system make an ideal two-tier storage solution, delivering the performance, scalability, and data resiliency critical to accelerating discovery and protecting valuable research results.

Performance starts with the file system. MatrixFS is a distributed, parallel, POSIX-compliant file system that runs on your compute cluster or as a dedicated storage layer and uses off-the-shelf SSDs, significantly improving storage system performance. With active data on flash-based storage inside the server and part of a global namespace, access is near instantaneous. Valuable data is protected using patented data protection and distribution algorithms that allow the system to sustain up to four simultaneous node or SSD failures.

Researchers can easily share massive data sets using any S3-compatible object storage solution from leading vendors such as AWS for public cloud, or Scality, Western Digital, Cloudian or Cleversafe for private cloud. With object storage, data is protected and always available with up to 17 nines durability, including site-level fault tolerance in a multi-site configuration. Robust data integrity checks occur automatically and transparently protecting data from silent data corruption known as bit-rot. Each object can tolerate up to 1000 bit-errors without the loss of data, which can eliminate the risks, costs, and media management activities associated with tape-based archives.

A SINGLE-STORAGE SOLUTION FOR THE ENTIRE WORKFLOW



Such a tiered solution is ideal for accelerating discovery and improving patient outcomes. **Benefits that can be realized using WekaIO include:**

Accelerate drug discovery with dynamic performance and capacity scaling: WekaIO's radically simple storage platform, Matrix, is ideal for the large data sets and complex workloads found in life sciences—large and small files, random and sequential access, structured and unstructured data. Independently and dynamically scale up or scale down performance and capacity without costly and disruptive forklift upgrades. Automatic load balancing provides consistent, sub-millisecond data access, so researchers are more productive and spend less time waiting for results.

Improve patient outcomes with low latency, fast access to data: Improved patient care means the right person must have the right information in the right place and at the right time. Matrix was designed for flash technology and provides a 10x improvement in throughput, and 8x improvement in IOPS at less than half the cost of traditional network attached storage. Its software-only architecture runs on any x86 based server with any type of solid-state disk technology (SAS, SATA, NVMe) for true hardware independence.

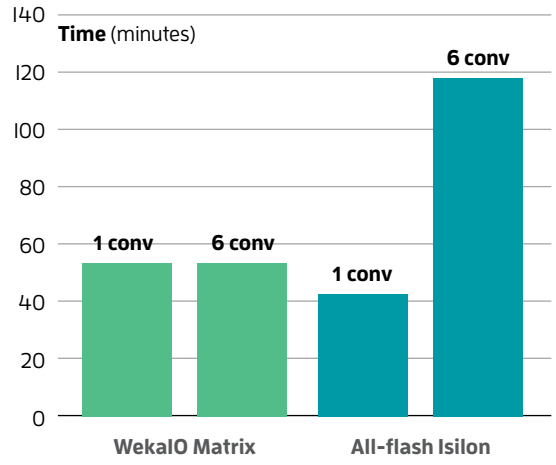
Increase collaboration with integrated cloud connectivity: Researchers must have the ability to quickly and easily share their results with colleagues and partner organizations. Matrix features a POSIX compliant, parallel, and distributed file system that allows shared file access. Integrated tiering to any S3 compatible object store enables seamless connectivity to the cloud without special software. All files remain part of the global namespace and appear local to users and applications, and multi-protocol support provides anywhere, anytime data access. No changes are required to host applications. Fluidly expand compute resources with a flexible platform that adapts to your needs: IT environments and application requirements are continually changing. Matrix provides a single platform that quickly and easily adapts to those changes. The solution supports multiple deployment models, including hyper-converged or as a dedicated storage server. It can be run on bare metal, a virtual platform, or directly in the Amazon cloud, and the Snap-to-S3 feature enables on-prem deployments to burst to the cloud to scale resource on demand. An intuitive graphical interface provides point-and-click ease of management so that a single administrator can manage petabytes of storage.

Benefits Abound Using the Right Storage Solution

Genomic research workloads are characterized by millions of small files and hundreds of very large files. Storage systems must be able to handle both workloads equally well. In clinical settings, lives depend on rapid sequencing analysis. This real-world application test was conducted by a leading genomics and cancer research center. The test involved converting BCL files from an Illumina sequencer to FASTQ files for use in the organization's bioinformatics pipeline. This particular step in the pipeline is a very I/O intensive operation.

Matrix was run on a six-node cluster; the benchmark system was an all-flash NAS system. Matrix completed a single file conversion in approximately the same time (5 minutes longer). However, when six file conversion jobs were run, Matrix completed all six conversions in exactly the same amount of time as it took for a single run. The benchmark system took over twice as long as Matrix to complete all six conversions.

BCL2FASTQ CONVERSION TEST



In real-world situations, research organizations generally run many of these jobs simultaneously. Matrix handles larger workloads than purpose-built systems at a fraction of the cost, providing far greater value.

To see how WekaIO can help accelerate your research efforts, visit https://www.weka.io/ or sign up for a 30-day free trial https://start.weka.io/

Machine Learning goes Mainstream

Life sciences and healthcare organizations are beginning to use machine learning in innovative ways. Application areas include:

DRUG DISCOVERY:

Using just a few training references, a machine learning system predicts whether a candidate drug molecule would bind to a target protein by predicting with 99 percent accuracy. With this approach, drug discovery could be significantly accelerated. **For more details, visit: https://tinyurl.com/wekaio-drugdiscovery**

HEALTHCARE:

An image analysis system evaluates how a patient's aortic valve narrows and blocks blood flow to the rest of the body. It is one of the most challenging conditions for cardiologists to diagnose. The application looks at medical images and determines if the problem is due to a tumor, an infection, or just an anatomical quirk. It then helps the cardiologist figure out which patients need follow-up care for aortic stenosis. **For more details, visit: https://tinyurl.com/wewkaio-medicalimaging**

BIOLOGICAL RESEARCH:

One area of interest is to predict molecule toxicity and reactivity. These activities are often a considerable burden on the drug discovery pipeline or even drug repurposing. AI and machine learning potentially can speed work in this area. **For more details, visit: https://tinyurl.com/wekaio-bioresearch**