Flexibility in the Cloud Elevates Innovation and Research Impact at The Wharton School

The Wharton School of the University of Pennsylvania is the number one ranked business school in the United States and the world's oldest collegiate business school. Wharton fosters its rich legacy of ground-breaking innovation and collaboration by supporting its award-winning faculty and graduate students as they pursue cutting-edge research in fields such as AI and Analytics, Behavioral Economics, Finance, Leadership, and Public Policy.

With over 20 research centers, 225 faculty members, and more than 150 graduate students, there is significant pressure to utilize a data infrastructure that is highly performant and flexible enough to support such a breadth of research and intensive-compute workloads.

> In the increasingly fast-paced world of academic publication, reducing the time that it takes to run workloads can lead to measurably shorter time to insights and the ability to not only meet deadlines, but exceed them and be the first to publish ground-breaking research in a field.

Challenges

- Complicated data management in the cloud
- NFS-based I/O not performing at scale
- Inflexible and insufficient resource allocation

Solution

- The WEKA Data Platform in the cloud provides an extendable file system that offers the scale and economics of cloud object stores and the speed of SSD and dedicated networking links for hot and recent data.

Benefits

- Automation and policy-driven management simplifies administration of the HPC3 cluster
- Increased throughput to every compute node allows significantly better time to solution
- Realizing the promise of elastic cloud computing allows us to grow and run as many jobs as our researchers require

Central to its mission of supporting innovation is the institution's High-Performance Computing Cluster Cloud (HPC3), a cloud-based cluster available to faculty members, doctoral students, and their associated research assistants and coauthors. Led by Gavin Burris, research computing expert and senior IT leader at Wharton, HPC3 supports a wide range of research activities from statistical analysis to generative AI projects. In addition to data management, the HPC3 team provides comprehensive support, from setting up the computational environment to installing necessary software.

- HPC3 runs in AWS and is architected to ensure that research teams can run their analytics workloads at scale and efficiently handle large datasets and computationally intensive tasks.
- HPC3 is built with flexibility to allow researchers to grow their computing resources as their projects grow.

Cloud-Induced Growing Pains

Over the last several years, research projects at Wharton have grown in size and complexity. This led to challenges for the existing data infrastructure that struggled with efficiency and scalability. When the IT department was faced with a cloud mandate, they moved almost all their workloads to the cloud where managing data storage and performance became a struggle.

As data management became more challenging, Burris began to search for a solution that would streamline Wharton's data processes in the cloud. Many vendors proposed cloud storage solutions that didn't offer any command line options or the scriptable admin interface that Gavin's team relied on. Gavin went on to clarify that "Their systems were essentially local hardware setups, modified to function in the cloud but lacking the depth and granularity of a **truly cloud-native solution.**"

After extensively evaluating various file systems, Burris identified WEKA as the ideal solution because it met his needs for a cloud-native data platform. As he explains, "Unlike retrofitted solutions, WEKA is **purpose-built** for the cloud. It is an **integrated product** that combines a wide range of storage mechanisms and the cloud. It provides an **extendable file system** that offers the scale and economics of cloud object stores and the speed of SSD and dedicated networking links for hot and recent data."

Dynamic Data Management and Storage Provisioning

Wharton relies on WEKA automatically managing data placement to ensure that frequently accessed data resides on fast SSD storage while less-used data is moved to more economical object storage. Burris notes that automation is key, "allowing us to optimize our storage based on workload and ensuring efficient resource use." "We don't want to have valuable SSD resources sitting idle. With WEKA, we've made it so that after 16 hours, a more transient user's file system is almost nonexistent on the SSD," he explains. This ensures that the most active, demanding users can fully utilize available SSD resources to meet their critical research deadlines.

"My deployment is a little bit unique in that I create a file system per user," Burris says. "We have hundreds and hundreds of file systems. WEKA's thin provisioning allows us to assign and provision SSD for all users efficiently. If a user is inactive, all those resources are freed up for the rock star who's cranking."

WEKA's flexibility also makes storage highly scalable. "We've created giant volumes and file systems for individual users," Burris says. "And we've created dedicated job queues tied to their own backing and compute, effectively giving users their own EC2 instances."

Wharton leverages WEKA's easy configuration and management to their advantage. "I can go to a webpage at WEKA, pick a cluster size, amount of SSD, and the total file system size I need," Burris explains. "Then I click launch, and it just comes up." This straightforward process contrasts sharply with previous systems, which required manual setup and did not fully utilize cloud capabilities. For example, "WEKA has a strong command line capability and a great web UI for spot-checking things and going in and doing admin work by hand with the web interface," he says.

WEKA's automation and policy-driven management also simplify the administration of the HPC3 cluster. "With WEKA, everything works together as one responsive, tunable cloud-based file system," he says. "You don't have to think about what's underneath it." This setup eliminates the need for procedural code to manage data movement, making the system user-friendly and efficient.

Increased Cluster Elasticity Optimizes Resource Distribution

Wharton can adjust their HPC3 cluster resources flexibly based on demand thanks to WEKA's elasticity. The research schedule often fluctuates, with heavier loads during the summer when faculty and students have fewer classroom responsibilities. Burris explains, "When the demand is high, we can tap research budgets and scale up. And then when the demand is low, we scale back to only what we need." This dynamic adjustment prevents the institution from paying for idle capacity, further optimizing financial efficiency.

"With WEKA, we can expand input/output capabilities to ensure that even the most data-intensive research projects are supported efficiently," Burris says.

Overcoming I/O Bottlenecks Clears the Way for Research Breakthroughs

One common challenge in high-performance computing is I/O bottlenecks. Burris reflected on the pre-WEKA era, noting, "We had been using an NFS-based solution on the last version of the cluster. That was fine, but the I/O just wasn't there at scale, leading to bottlenecks and limited scalability. With WEKA, every compute node becomes a member of the storage cluster, each with dedicated links, enhancing throughput and scalability."

Unlike NFS, which had a single network link limitation, WEKA leverages distributed multiple link data transfers from backing stores, allowing parallel data loading and retrieval across multiple nodes. "WEKA can parallelize the loading and exiting of data into the SSD cluster," Burris notes, which is crucial for supporting Wharton's compute-intensive research projects.

> “Going from the old NFS solution to the new WEKA on the new HPC3 has been such an upgrade, Burris says. The amount of throughput we get to every compute node allows significantly better time to solution and helps researchers meet their publication deadlines.

Underneath, data in the WEKA environment resides in Amazon S3. The WEKA software automatically manages storage tiering to ensure that hot, recent data is kept on SSDs across 10 Amazon EC2 I3en instances that support the WEKA storage cluster. This setup keeps current data fast and accessible, leveraging the aggregate performance of multiple servers to achieve high I/O throughput.

Burris has configured an Amazon Machine Image (AMI) with a thin client so researchers can join the cluster quickly and efficiently. "The system comes up with multiple internet/network interfaces, and WEKA subsumes one or two of the processors and an entire network interface dedicated just to I/O," Burris explains.

> “ HPC tends to turn CPU problems into I/O problems. WEKA turned those I/O-bound problems back into CPU-bound problems. That means we can focus on throwing more compute and GPUs to support the code, explains Burris.

Effortless Data Management Lets Researchers Focus on Research

Each Principal Investigator (PI) has a dedicated file system, including a home directory and a subdirectory for their team. Despite the complex backend, the user experience is simple. Researchers interact with their directories like standard file systems on their local machines. "We export a network file share via SMB so that they can mount a home directory on their Mac or Windows PC," Burris explains. This integration makes it easier for researchers to access and manage their data without needing to understand the underlying infrastructure.

Streamlining Data Infrastructure Provides Ease of Management for Data Architects

WEKA significantly lightened the administrative workload at HPC3 which has allowed the team to focus on essential tasks, all while providing sophisticated configuration options when needed.

> “With WEKA, I no longer require a full-time employee dedicated to file system hardware administration,” Burris says. “The ability to manage this cluster with reduced overhead is a significant relief.

The process is streamlined through automation, allowing for rapid response to changing demands. “We script that so it can be one command line to create a user, add a new volume and get them up and running. Or we can go into the WEKA web interface and allocate an obscene amount of SSD to account for their 10 million files,” Burris elaborates. The result is that researchers can efficiently meet their deadlines without the institution bearing the cost of over-provisioning.

WEKA also supports granular security control through Access Control Lists (ACLs), which are set within the file system. “This is important for compliance with data usage agreements, ensuring that sensitive data is appropriately protected,” Burris says.

Burris underscored the strength of WEKA’s support team. “The support team is great. I can’t say enough good things about how good they are. Whether it is upgrades or bugs, WEKA proactively contacts us, schedules time, and sits with us virtually to get whatever is needed done. I’ve learned so much from the support team about how to tune, adapt, and deploy my storage cluster to fit our needs.”

Unlocking the Full Potential of Wharton’s Research Capacity

Feedback from faculty and doctoral students has been overwhelmingly positive. “They’re just wowed at how much better and more consistently their jobs are running,” Burris shared, adding that **“the improved consistency and performance has allowed researchers to tackle larger and more complex problems,** handling millions of files, which WEKA efficiently supports.”

Burris says, “Now when a researcher worries that with the base compute allocation it will take six months or a year to chew through their parameter sweep, we can say, if you have the budget, we have the tools to make it happen. That’s the power WEKA gives us.”