STARFISH

Case Study | Higher Education

Harvard FAS and ColdFront

Harvard FAS Research Computing Uses Starfish with ColdFront to Eliminate Petabytes of Legacy Data and Generate $1.5M in Chargeback from Hundreds of Labs

The Challenge

Achieve sustainability and manage data storage capacity growth in one of the world's largest grant-funded research facilities

The Solution

Starfish combined with ColdFront provides up-to-date capacity consumption metrics with full auditability for grant chargeback while providing users with tools for ROT cleanup and archiving

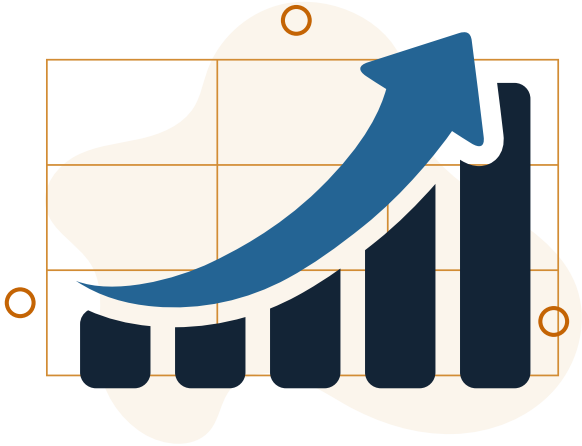
The Results

- Provides clear, easy visibility into data that can be archived or deleted, potentially enabling the deletion of ~20PB of data
- Eliminates need for continually adding one-off storage servers
- Simplifies grant-related project tracking, chargebacks, and audits
- Provides insights into where data volumes are growing for better planning
- Improves billing accuracy and revenue streams, generating $500K in revenue from chargebacks in year one and $1.5M in year two
- Provides an interface for end-users to identify data for disposition and archiving

Introduction

The Research Computing organization of the Harvard Faculty of Arts and Sciences (FAS-RC) is one of the largest research computing environments at Harvard. FAS-RC services more than 40 academic departments in both liberal arts and other scientific disciplines. This includes more than 600 research labs and nearly 4,000 users.

FAS-RC's primary high performance computing (HPC) cluster, called Cannon, is made up of 1,800 compute nodes with over 80,000 cores. Cannon runs more than 45 million jobs each year and provides more than 60PB of storage, consisting of more than 10 billion files.



The Challenge

Runaway storage demands, granular auditability requirements.

Harvard's researchers are often one of the first to embrace new methods of discovery and thus have a never-ending hunger for storage capacity. Given continuous growth and increasing complexity of the storage environment, FAS-RC had lost the ability to account for storage consumption with the level of granularity required to charge it back to the research programs.

Usage tracking is especially important because government granting agencies provide significant funding to Harvard FAS and have specific rules for charge-back. For example, in 2022 alone the National Institutes of Health (NIH) and National Science Foundation (NSF) awarded Harvard FAS researchers more than $78 million and $53 million respectively.[1,2] To recover costs for storage capacity consumption, FAS-RC needed a way to enumerate files belonging to specific grants along with the ability to audit file details on command.

> **Our data volumes grow by 20% year after year. Since continued growth at this rate would quickly become unsustainable, we started considering how to help users be smarter about what data they store and where—or even if they really need it. Ultimately, our goal is to see overall year-over-year storage demands flatline.**
>
> — Raminder Singh, Associate Director Data Science & Research Facilitation, Harvard FAS RC

The Solution

Combining Starfish with ColdFront

The FAS RC team had already adopted ColdFront, an open-source allocation management system developed by the Center for Computational Research (CCR) at the University at Buffalo. The CCR developed ColdFront to provide faculty members with greater control when managing student user groups, including monitoring active/inactive accounts and gathering information on the research being conducted on lab systems through mandatory annual project updates.

The FAS RC team saw an opportunity to use ColdFront to automate storage resource management and billing consumption back to individual grants. At the same time, Singh explains that his team needed to overcome some hurdles with ColdFront: "We have such a diverse storage infrastructure, and we realized that each storage device reports end-user usage differently. To ingest usage information into ColdFront, we would have needed to build custom integrations for each storage device and access huge file systems."

> **Without Starfish, querying individual user data would require very heavy file system operations. With our Starfish database, there is no penalty for collecting and updating user information daily.**
>
> — Raminder Singh

FAS-RC turned to Starfish, an innovative platform for managing the large, complex file stores used by R&D, because of other installations at Harvard. Harvard Medical School started using Starfish in 2013 and was one of the first production customers for Starfish. Harvard Libraries also uses Starfish as a bespoke data protection solution for its sprawling digital asset management system. The FAS-RC team saw an opportunity to complement Starfish's data classification and capacity optimization capabilities with ColdFront's allocation and monitoring capabilities, but instead of creating one-off integrations, the team developed a ColdFront plug-in for Starfish. "The plug-in enables our team to aggregate metadata from multiple file system silos into a database to deliver an up-to-date, unified view of key project metrics," Singh says. The metadata in the Starfish database supports querying and reporting on individual user resource consumption and ingestion into ColdFront to track per-project usage and facilitate chargeback. "Starfish has enabled us to track individual usage and map it to grants and projects without the need for accessing a file system," adds Singh.

FAS RC began ingesting data from Starfish into ColdFront in early 2022. Initially, the team piloted the solution with several labs and quickly started seeing promising results.

STARFISH

Starfish is an unstructured data management solution that enables organizations to analyze and manage file systems and object stores and to move data based on those discoveries. Specific capabilities include:

- Metadata and content classification – allows users and applications to associate metadata with files and directories
- Reporting and analytics – provides industry-leading reporting and file system analytics
- Data protection and preservation – supports multiple data protection strategies to give users complete control over files
- Tiered storage and capacity optimization – enables users to participate in storage management
- Archive and recovery – provides comprehensive capabilities for deciding what to archive and recover
- Workflow automation – combines metadata with batch processing, making it simple to automate pipelines and workflows involving file processing and data movement

ColdFront is an open-source resource allocation management system for administration, reporting, and measuring the scientific impact of HPC resources. It is used to chargeback computing resources to grant-funded programs and enables organizations to:

- Collect project, grant, publication, and other research output data from researchers
- Define custom attributes on resources and allocations
- Email notifications for expiring/renewing access to resources
- Integrate with third-party systems for automation and access control

"When we developed ColdFront, rather than focusing on a rigid set of requirements, we aimed to create a flexible framework that allowed extensions to be tied to it. We've seen people integrate things like authentication systems and storage quotas, but Harvard's implementation using something as advanced as Starfish is truly impressive."

—Dori Sajdak, Senior Systems Administrator, University at Buffalo

"Giving users clear visibility into what data they had and how they were spending their money helped them get serious about what they wanted to keep and what could be deleted... Based on what we've seen so far, we might be able to delete upwards of 20PB of existing data.

— Raminder Singh

The Results

Sustainable Storage

The Starfish and Coldfront solution has been a storage-management game changer for FAS RC.

"Giving users clear visibility into what data they had and how they were spending their money helped them get serious about what they wanted to keep and what could be deleted. Within a few weeks of rollout, we were able to help users delete nearly a petabyte of data," observes Singh. "It's eye-opening because users can easily see the actual dollar values specific data is costing their labs and then decide if they should move it to a lower tier of storage or delete it."

On the technical side, the solution works nearly seamlessly in the background with minimal overhead on systems. "With Starfish, you rely on metadata and query a database rather than accessing the actual files. That's critical because our environment supports billions of files," says Singh.

Today, FAS RC scans every storage solution it purchases with Starfish and automates allocation of space through the Storage Service Center. ColdFront is also integrated with the FAS RC billing system, which automatically calculates charges by project, department, and storage system each month and sends bills to designated account holders. With the help of this efficient, automated billing, FAS RC has seen significant year-over-year revenue growth. It generated $500K in the first year, growing to $1.5M in the second. Revenue in the third year is projected to reach upwards of $2.5M as more labs are added to the system. (By the end of 2022, FAS RC expanded the use of the Starfish and ColdFront solution from 10 labs to more than 230, and it is continuing the rollout on a phased schedule.) In addition to improving billing accuracy and processing, Starfish has simplified grant-related project tracking because it keeps a running history of file system details, enabling users to see exactly what files were charged to what grant at any point in time.

> **One of the big problems we've solved with this effort is eliminating the need to continually buy one-off storage servers**
>
> — Raminder Singh

Starfish also gives the FAS RC clear insights into why data volumes are growing so it can both better plan for and manage future expansion. "Although data volumes in workloads will continue to grow, if we more consistently and quickly move legacy data to our tape system and purge data that's not being used, we will have a much more sustainable model. Starfish makes it easy to identify candidate files for archive and to automate data movement and recovery," notes Singh. In addition to providing a compelling ROI, Starfish helps users appreciate the value of data and practice better overall storage hygiene.

With the cost savings from purged data and more accurate chargeback, the Starfish solution pays for itself and has given FAS RC more opportunities for educating users about best practices. "Currently, we are looking to hire someone to help us continually identify data that is not being actively used and to train our user groups to be more efficient with data management… We also want to provide tools for users that help them better align their usage with funding agency mandates for data management and Starfish will be integral to those efforts," Singh explains.

ColdFront Plug-In for Starfish

Harvard FAS RC has generously released an open-source version of its ColdFront plugin for pulling usage data from Starfish.

You can download it here:
https://github.com/fasrc/sftocf

> **What Harvard did was great. They didn't hack up the code, they made a plug-in. And then they made that plug-in available to other people, which has been fantastic.**
>
> — Dori Sajdak

About Starfish Storage

Starfish is a unique software application for managing unstructured data at very large scale. Starfish combines a file system metadata catalog with a parallelized data mover and batch processor. You make discoveries and reports using the catalog. You move data fast and furiously and take other actions using the batch processor. The software is agnostic to storage vendors and works great with HPC file systems, enterprise NAS, object stores, and archives. Starfish allows your users to participate in storage management. Use cases include ROT cleanup, duplicate detection, backup, archiving, reporting, chargeback, content classification, and much more.

STARFISH

Sources

[1]NIH Awards by Location & Organization, NIH RePORT Harvard University, 2022.

[2]NSF Award Summary: by Top Institutions, Top50 Institutions FY 2022, Harvard University.