LLNL Unified Storage Namespace

A metadata clearinghouse with novel use cases for the HPC center

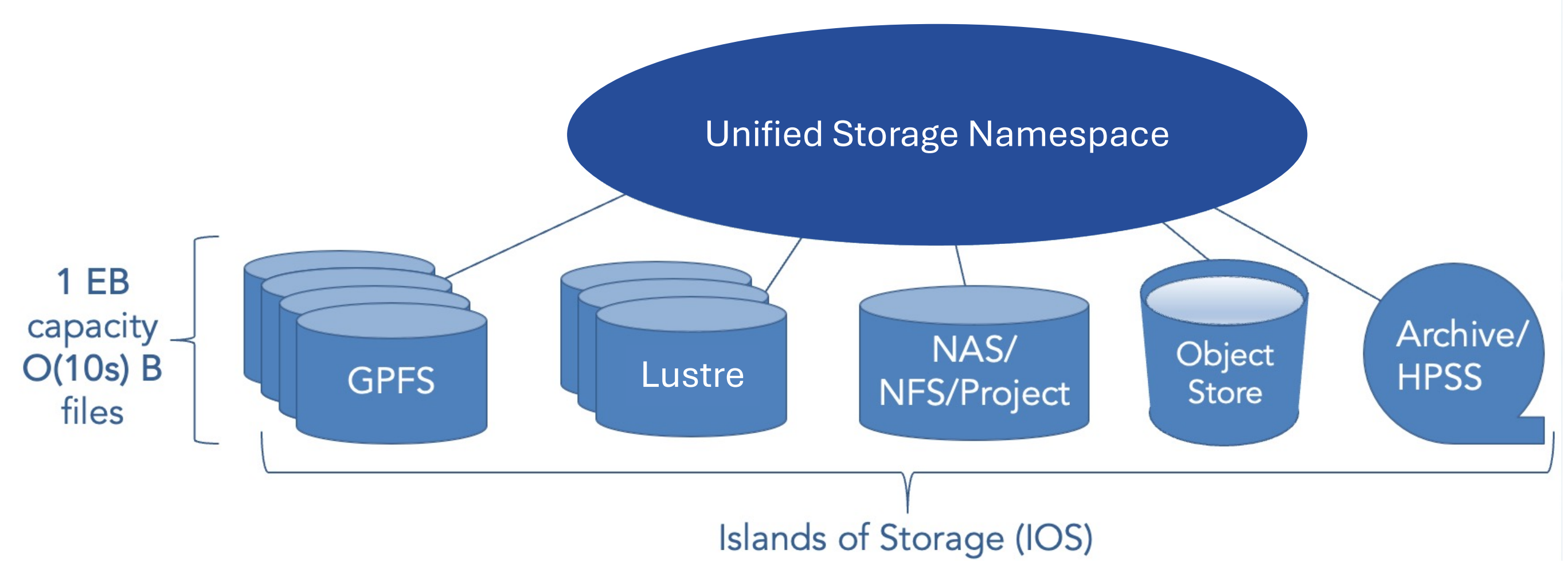
Herb Wartens (LLNL), T. Heer (LLNL), O. Faaland (LLNL)

A Unified Storage Namespace (USN) has been deployed across both the unclassified and classified production environments at LLNL. The USN is a decade-long vision to provide a single point of access to metadata for all files and directories contained within LC's massively scaled out islands of storage.

Originally launched as a **3-year ISCP** project to address the challenge of managing tens of billions of files in constant flux, USN received **$527k in staff funding** from FY21 to FY23. The result of that investment was a steady-state operational USN for tens of billions of objects for the first time.

Initially, USN ran on secondhand hardware, but after successful implementation, dedicated hardware for both environments was funded by the program.

What Is a Unified Storage Namespace and Why Do We Need It?



USN Contains

- File/object names
- Source IOS
- Attributes(stat())
- Tags
- Key-values pairs

USN Does Not Contain

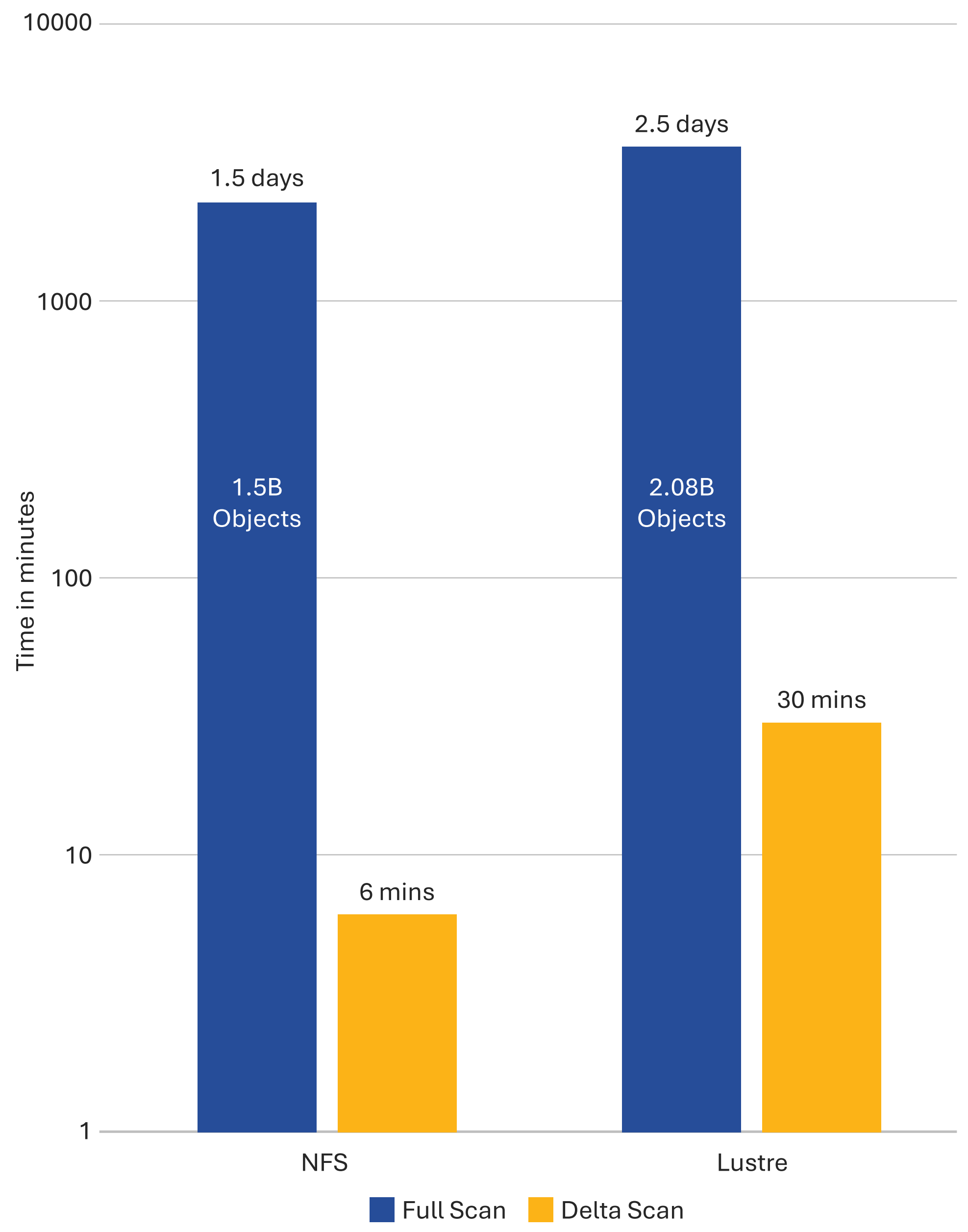
- Data*

- Database that stores metadata for all shared filesystems in our HPC center
- Helps us understand how data is distributed across filesystems
- Useful for capacity planning, understanding data aging, finding inefficiencies across domains of expertise
- Grants the ability for data discovery from a single database
- Allows for data movement based on traits

Initial Capability Deployment (Phase One)

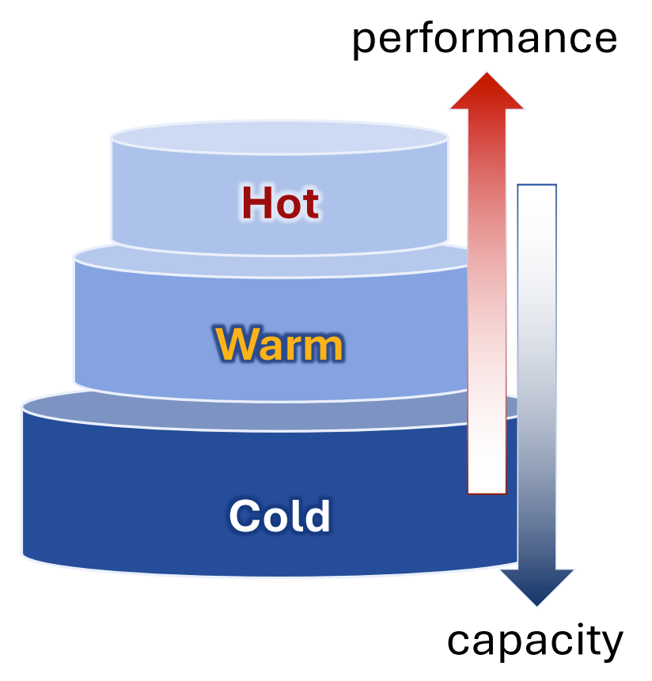
- First-of-a-kind capability in LC. Prior to USN, we would have to run time-consuming commands on each filesystem to find what we were looking for. This often required an expert for each filesystem type. Now, we can execute complex queries in minutes without requiring extensive domain expertise.
- USN is eventually consistent; real-time precision is not required.
- Prior attempts at a USN could not handle the workload scale. Lustre filesystems regularly see 300,000 events/second for 20 consecutive hours.
- Developed delta interfaces for large capacity namespaces to keep up with the filesystem changes (Lustre Changelogs, HPSS Rumble API, GPFS Policy Engine, etc.) and was successful due to ongoing collaboration with Starfish
- Initial deployment is for LC administrator use.
- Unclassified GA deployment in FY24 and classified deployment in FY25.
- LLNL showed that using ZFS for database workloads was viable. Consequently, the Starfish LLC began recommending ZFS for their industry engagements.

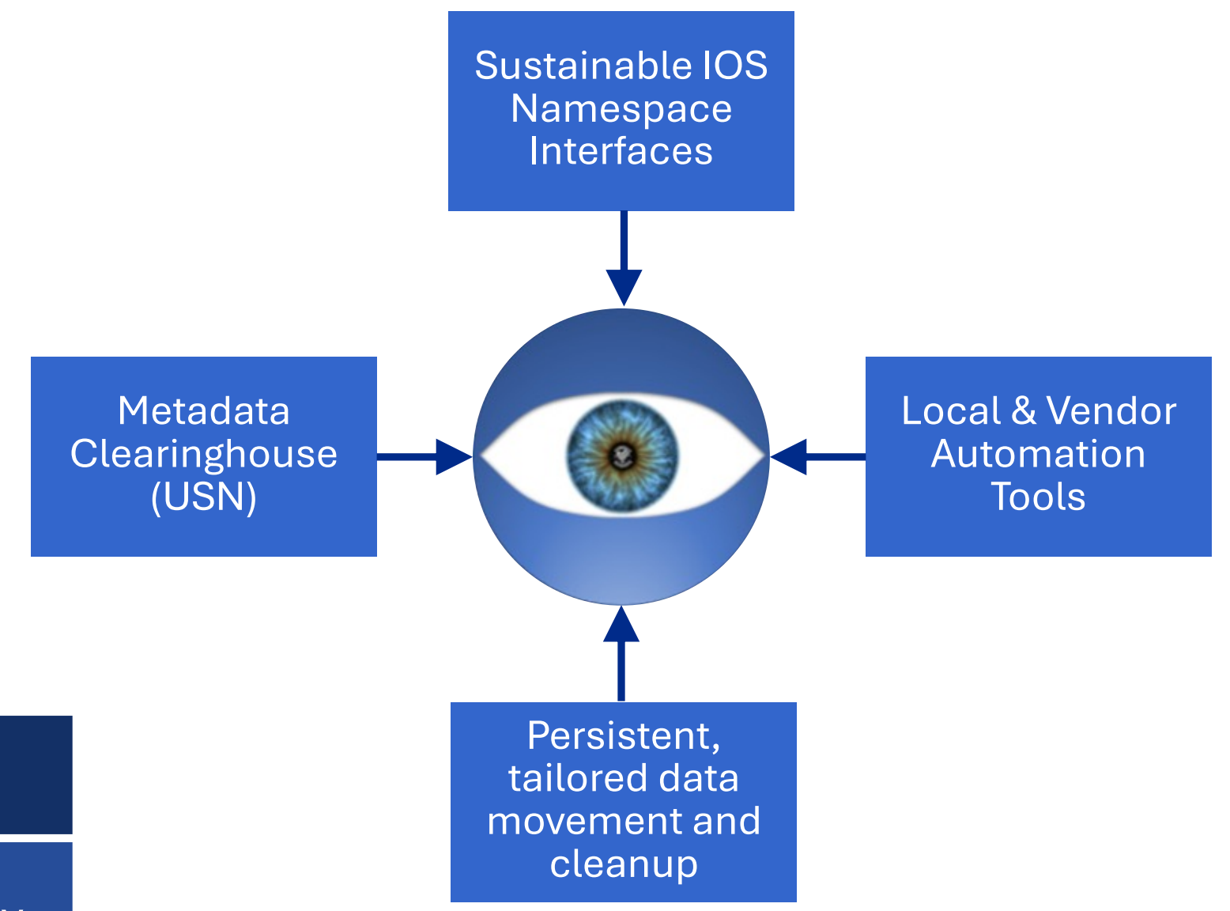
	Physical Size	Object Count
OCF NFS	6.19 PiB	5.14B
OCF Lustre	27.61 PiB	6.11B
OCF HPSS	62.76 PiB	2.38B
OCF GPFS	17.85 PiB	1.14B
SCF NFS	454.51 TiB	517.28M
SCF Lustre	22.18 PiB	3.84B
SCF HPSS	85.90 PiB	1.91B
SCF GPFS	N/A	N/A
Total	222.93 PiB	21.04B



Advanced Capability Deployment (Phase Two)

- Data discovery and rich metadata tagging.
- MPIFileUtils automatically moves data between parallel filesystems to archival storage. Changes already made in upstream MPIFileUtils with HPSSFS-FUSE support.
- Tackling security challenges, allowing access for unprivileged LC users.




Sustainable IOS Namespace Interfaces
Metadata Clearinghouse (USN)
Local & Vendor Automation Tools
Persistent, tailored data movement and cleanup

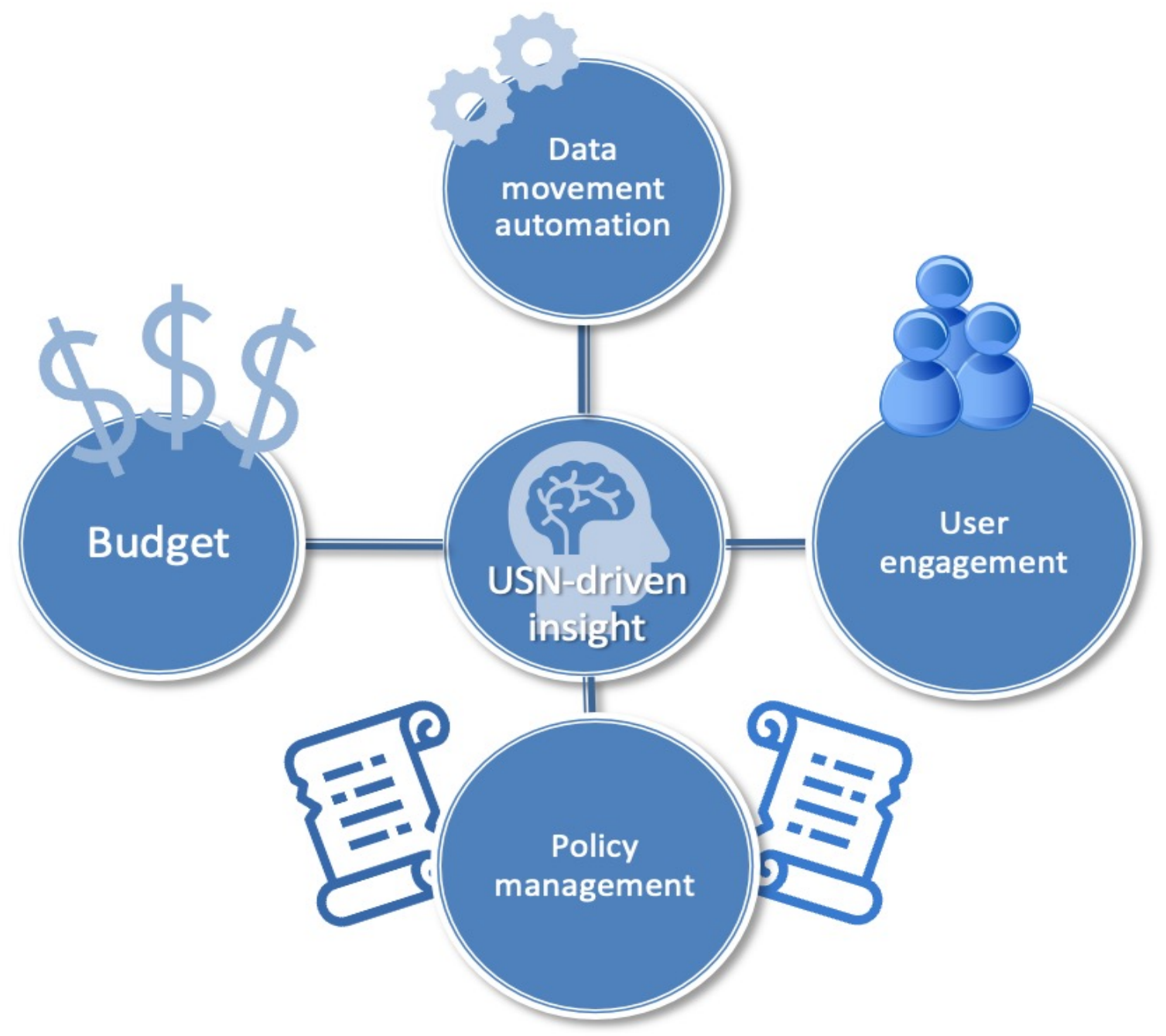
Notional Data Lifetimes

Tier	Type	Data Availability
Hot	Cluster-local parallel file systems, node-local storage TTFB=0	~1 day – 6 mo
Warm	Center-wide PFS, Project NFS, Object Storage, Campaign Store TTFB=0	~6 mo – 3 yrs
Cold	Archive TTFB=45 seconds	Forever

Starfish jobs engine will kickstart the process of moving data between tiers.

Conclusions

- Advanced data stewardship capability needs are here now and represent a service that HPC center is best suited to provide.
- LLNL built and improved delta interfaces, allowing us to scan at scale. Close collaboration with Starfish LLC was required.
- Advanced capability challenges will require ongoing work and deep expertise.



Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-POST-2004489

NNSA National Nuclear Security Administration