High Performance Data Facility Project: Status and Plans

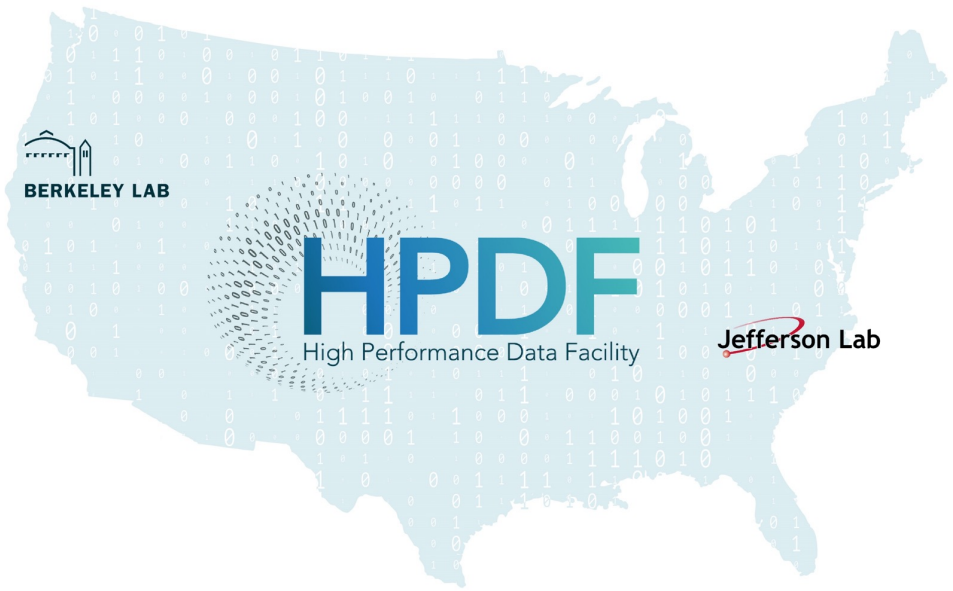
Representing the HPDF project team:
Graham Heyes, Technical Director

July 2024

Jefferson Lab
U.S. DEPARTMENT OF ENERGY | Office of Science
BERKELEY LAB

Innovation and Stewardship Through Partnership

Our mission: To enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools



HPDF is a DOE 413 Project announced in Oct. 2023

- Project team formed spring 2024
- Undertaking outreach with all SC programs & forming initial technical design

The HPDF project team leverages the strengths and complementarity of both labs:

- Decades of experience with scientific missions and user communities
- A shared understanding of resilient, distributed infrastructure that supports the data life cycle
- A shared commitment to the IRI initiative and ASCR ecosystem

The HPDF will be a first-of-its-kind SC user facility:

- A distributed operations model will be essential to long-term success and required performance levels
- Project structure is integrated with JLab and LBNL staff

Meeting the Greatest Needs

The DOE envisions a revolutionary ecosystem – the Integrated Research Infrastructure – to deliver seamless, secure interoperability across National Laboratory facilities.

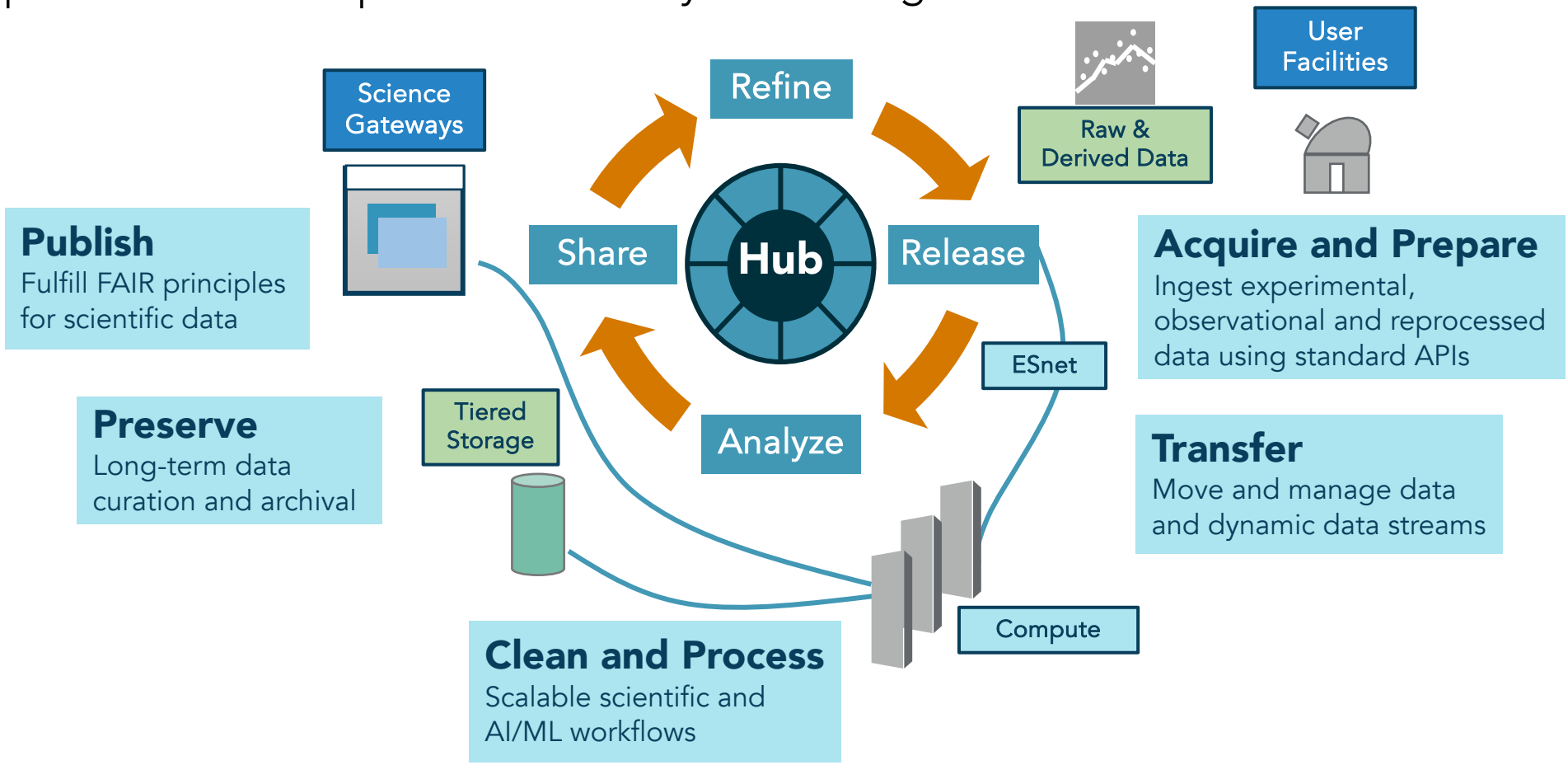
The 2023 IRI Architecture Blueprint Activity identified three broad science patterns that demand research infrastructure interoperability:

- Time-sensitive patterns
- Data-integration-intensive patterns
- Long-term campaign patterns

HPDF specifically will enable analysis, preservation, and accessibility of data produced by SC facilities through data focused resources and advancing data stewardship

HPDF will Support Data Lifecycle Management

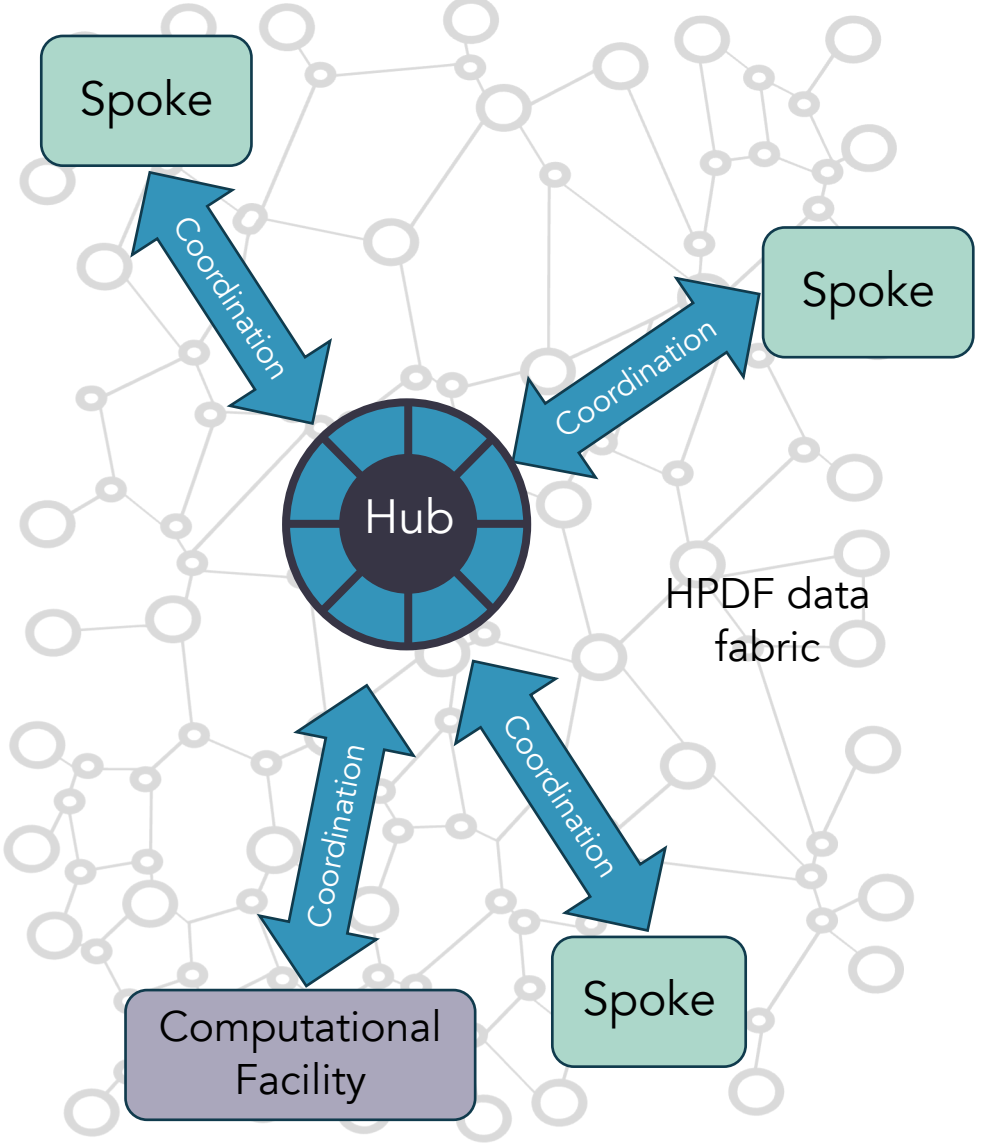
Data science requires curated and annotated data that adheres to FAIR principles, and data reuse will be a metric for HPDF. Office of Scientific and Technical Information (OSTI) services will complement HPDF to provide full life cycle coverage.



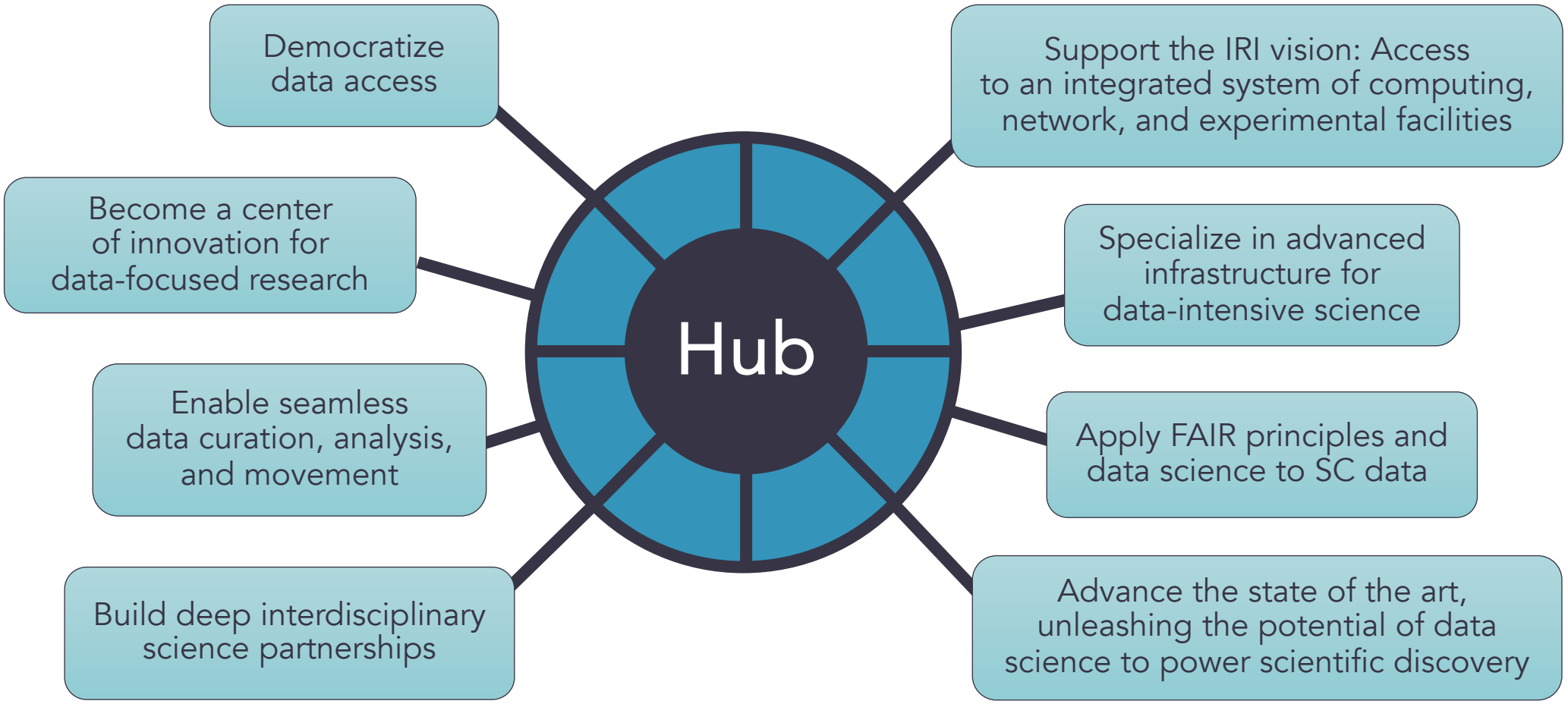
HPDF — A Distributed Facility

Concept: HPDF is a distributed facility with a hub and spoke architecture.

- **Hub.** Data-centric infrastructure with high availability and performance, as well as geographically and operationally resilient active-active failover.
- **Spokes.** Distributed data-centric infrastructure to enhance HPDF access and support for science users and integrate distributed computing or storage resources.
- **Integration and Services.** Orchestration hardware, software, and services for data movement, storage and retrieval, and science workflow automation. These will use a mesh data fabric building on ESnet6 capabilities.



HPDF Hub Will Address Key Strategic Goals and Capability Gaps



Technical Design — Core Capabilities

Hub Computing and Data Infrastructure

- High uptime
- Experiment-friendly availability
- Data-driven agility
- Support for new technologies
- Data storage, management, and interoperability
- Data preservation

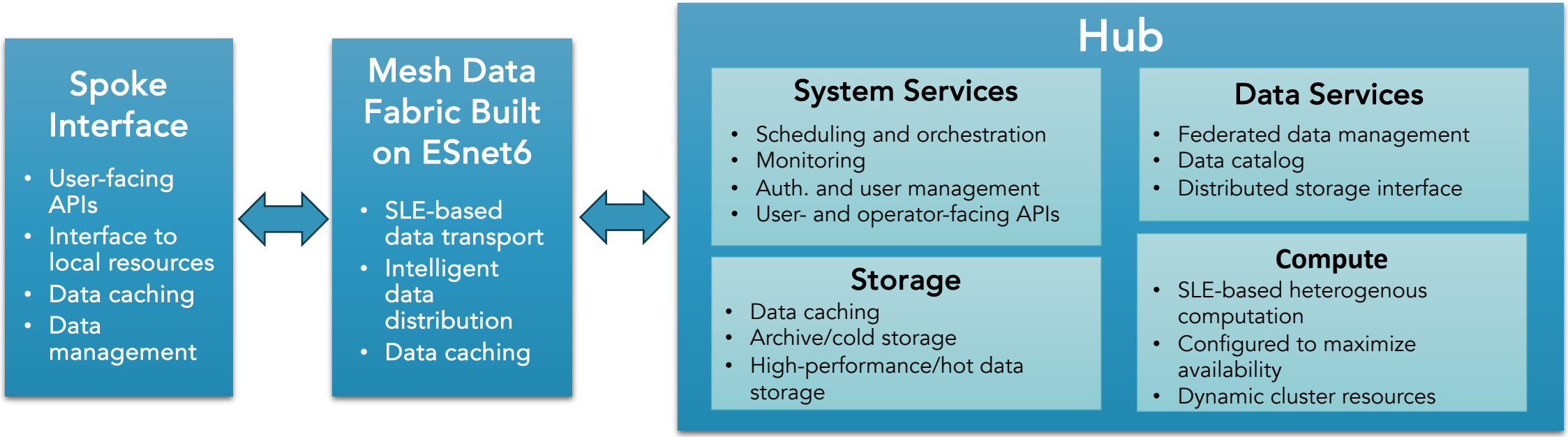
Distributed Spoke Infrastructure

- User support
- Scientific application tailoring
- Hardware resources that mirror, supplement, or complement Hub resources
- Low-latency or high-bandwidth coupling of HPDF services to edge compute

Data-centric Orchestration of Hardware, Software, and Services

- High availability
- High-performance mesh data transport fabric
- Secure data paths
- Monitoring
- Orchestration

High-Level HPDF Technical Concept



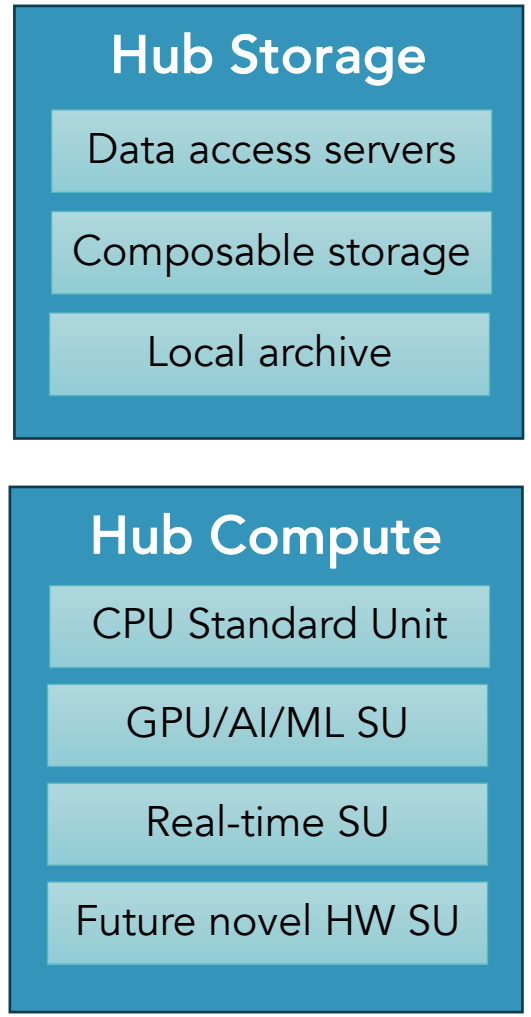
Design methodology, qualification, and approach:

- Pilot and phased delivery, enable early development, fine tune design
- Use of proven technologies to ensure a reliable, robust platform
- Hardware distributed and replicated at both sites to improve reliability and geographic diversity
- Modular heterogeneous approach to support a broad range of analysis

Approach to delivery and modularity allows composition adjustment during the design phase

The HPDF Hub: Unique Hardware Capabilities

- Combines high availability, flexibility, and support of time-critical workflows
- Composable storage will be configured to limit the need to modify existing code
- A local archive will be available along with a federated data catalog of data archived elsewhere
- The data processing design is based on the concept of "standard units," hardware elements following well-defined architectures targeting specific use cases
  - Batch jobs, AI/ML intensive, streaming, real-time, and dynamic reconfiguration
  - A mix of CPU/GPU flavors to run existing optimized code
- The Hub will incorporate a range of standard units in a mix that meets the science needs yet can evolve over time
- This is not a one-size-fits-all approach; it allows tailoring to needs and lowers the barrier to HPDF use



Software & Data Services Strategy: Innovation & Stewardship

Identify

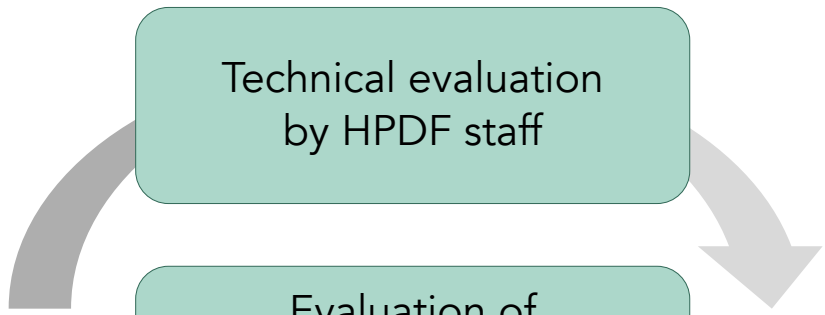
- Identify data services and software as part of DOE SC user facilities and projects
- Identify data services and software as part of DOE ASCR facilities and IRI
- Interaction with software stewardship organizations
- Interaction with ASCR research programs and Exascale Computing Project

Evaluate

- Technical evaluation by HPDF staff
- Evaluation of sustainability, usability, etc. by UX team
- Work closely with DOE PMs and advisory committees

Integrate

- Work closely with the community on identified software, ensuring it is usable and operational
- Identify stewardship path
- Integration and development efforts
- Advanced data technologies



Innovate & Partner

Coordination and collaboration with the governance for IRI and ASCR scientific software

User Experience Engagement: Core to Our Strategy & Plan

User research gives us a process to verify/validate our "intuition about what the user needs" (hypothesis) and convert it into action

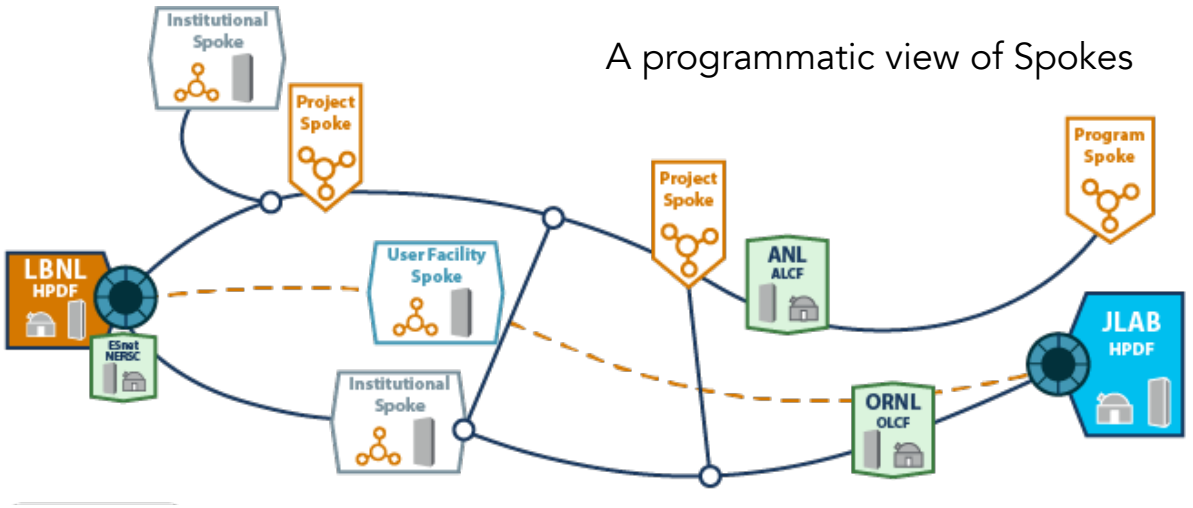
Deep partnership model to serve user needs, mature data stewardship across SC, and develop a workforce

User support and data stewards will provide critical help to HPDF users to leverage resources effectively and efficiently, allowing us to address computation and data needs early

"Tale of caution; you get one shot at making things accessible. Simple, easy (Google-esque) interfaces. You get one shot at it, or you'll lose your PIs. Interface simplicity."

– Dan Jacobsen, ORNL/UT

Spoke Architecture — Seamless Service



Selected Facets of Spoke Design

Community Structure
How strongly governed or united a community is around a set of policies or goals for its data products.

Organizational Structure
How the organization or institution is designed to support their user community's full data life cycle.

Size
Spoke size will be shaped by the confluence of anticipated user base, data volume and velocity, and resources (staff, compute)

Funding Model(s)
How the spoke is funded, for what lifespan, and how end users are supported (sub-awards, allocations, etc.) to leverage its resources.

Data/Compute Resources
The types and extent of technical functionality a spoke supports for its user community.

Exemplar Spoke Model Concepts

Spoke A – Provides edge services (hardware & software) close to experiment or instrument and interfaces with the Hub through APIs

Spoke B – Provides minimal computation and/or data services at the Spoke, primarily leverages computational or data services at Hub.

Spoke C – Co-designs software services with Hub and supports user community's use

Spoke D – Co-designs & supports software services with a particular Scientific Community

Science Community – Co-designs additional software services within community & supports users

Spoke E – Relies on Hub resources for preserving and publishing data while supporting & educating users

Summary

Next Steps

- Working toward CD-1: Conceptual technical design and scope and alternative analyses
  - Includes design of Hub and initial Spokes
  - Converting requirements into technical capabilities
  - Identifying:
    - Existing technical solutions
    - Gaps that need R&D
    - Partnerships
- Implementation stages:
  - Testbed
  - Early Access System
  - Early Access System with a beta spoke
  - A fully resilient Hub
  - The final system with a resilient Hub and early spokes

User & Community Engagement

Science engagements & partnerships are critical to success of HPDF

We are in the early conceptual design and community engagement (i.e., listening) phase.

We are eager to continue learning about science community's needs, challenges, science drivers, and so on.

Pointers to tools, reports, and so on are always helpful resources for our team.

Engagement opportunities:

- ✓ 6-way Light Sources meeting (Jan, in-person)
- ✓ IRI Management Council (April, virtual)
- ✓ FES PI meetings (June, in-person)
- HPDF/IRI workshop (July, in-person)
- Small-group interviews with groups identified through initial HPDF workshop (summer/fall, virtual)
- Supercomputing '24 (November, in-person)

Q&A

https://hpdf.science

https://linkedin.com/company/doe-hpdf

https://www.youtube.com/@doe-HPDF

Share thoughts & questions or request to be added to our mailing list via our form.

Answers will be provided via the website within a few weeks.