

High Performance Data Facility (HPDF) Project: Status and Plan

October 23, 2024
Lawrence Berkeley
National Laboratory (LBNL)



The concepts in this talk are works in progress. Feedback is valued and greatly appreciated

We are eager to discuss possible alignments, opportunities, & gaps

We would like to talk further at your meetings, organizations, etc.

Innovation Through Partnership

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

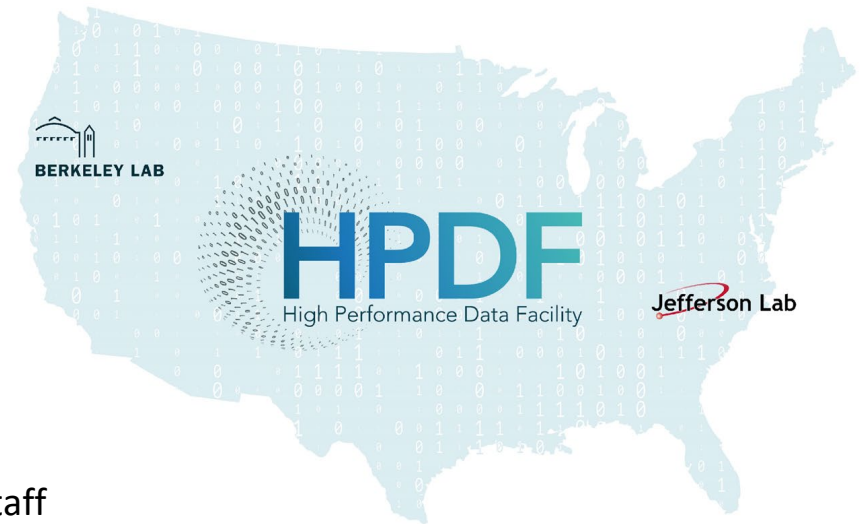
- Decades of experience with scientific missions and user communities
- A shared understanding of resilient, distributed infrastructure that supports the data life cycle
- A shared commitment to the IRI initiative and ASCR ecosystem

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

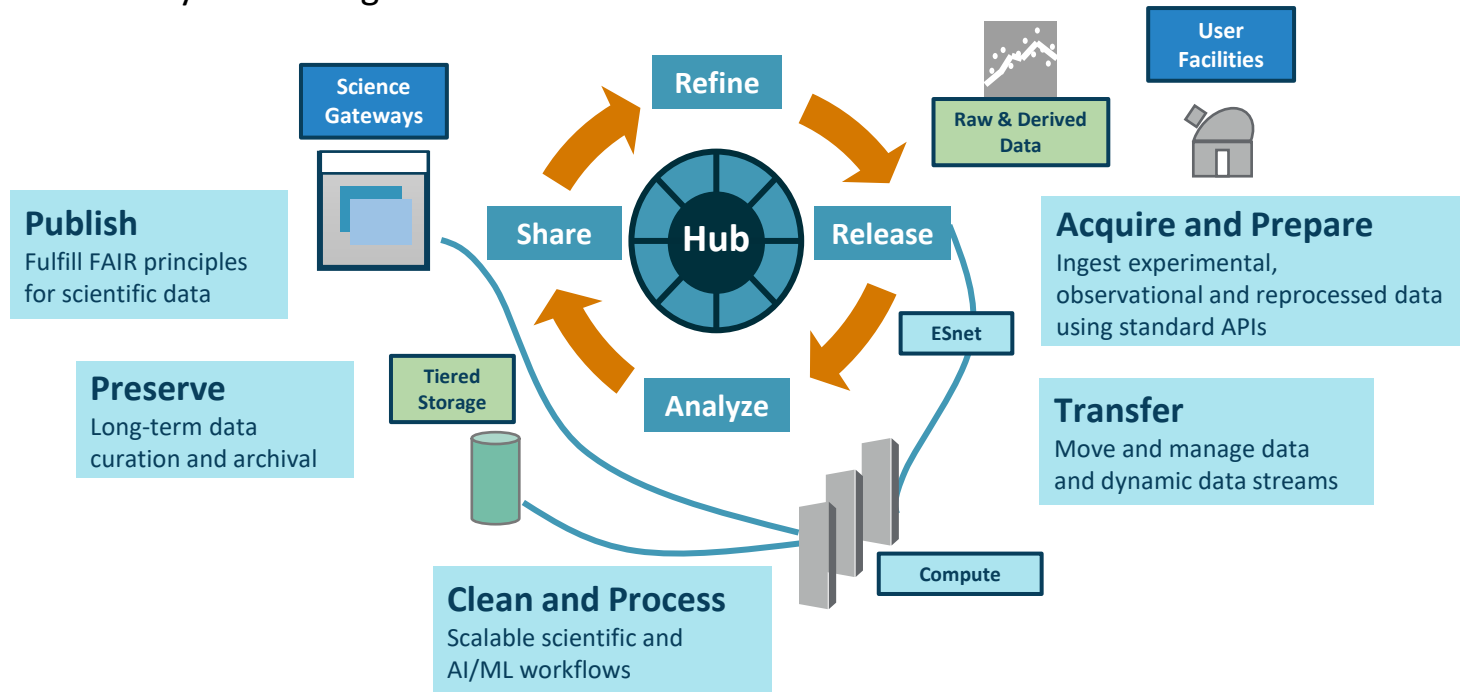
HPDF is a DOE 413 Project announced in Oct. 2023

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












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 Will Address SC Priority IRI Science Patterns

Drivers	IRI Patterns
Supporting data curation, repositories, and archives	 
Supporting data processing and analysis pipelines	  
Data federation, sharing, and collaboration	 
Real-time streaming and processing	

 Time-Sensitive

 Data Integration-intensive

 Long-Term Campaign



These patterns are seen widely in the larger community, in other parts of DOE, and outside

Hub and Spokes model allows for effective and efficient integration

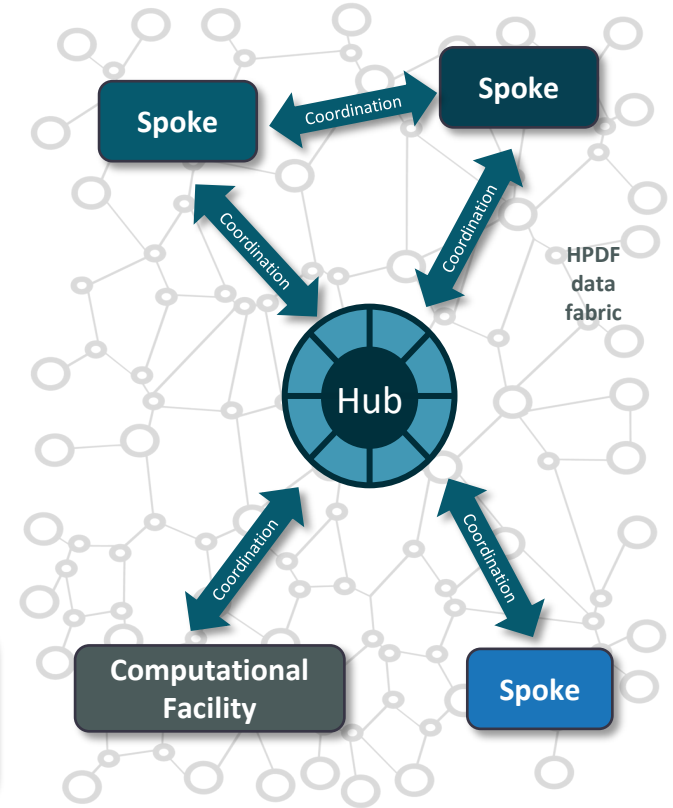
Hub-and-Spoke infrastructure design and management framework

- Will allow the scaling of resources and services
- Support various strategic priorities, partnership contexts, and scientific communities

Hub — Provides core infrastructure, services, and generalized expertise to support the diverse communities across SC

Spokes — Resources will be tailored & localized for their data lifecycle needs depending on the scope identified.

Impact: Together the Hub and Spokes will enable and accelerate multi-disciplinary novel science discovery through support for all IRI patterns across the entire data lifecycle.



Findings from ASCR IRI/HPDF Coordination Kickoff meeting [July 2024]

Vast majority of participants from the SC Program Offices (BER, BES, FES, HEP, IP, NP) expressed an urgent AND immediate need for HPDF capabilities.

Urgent Capabilities

- On-demand customized environment
- Support for high reliability and availability for data lifecycle
- Storage for critical data sets – replicated, distributed data storage

Software and Services

- Data Catalogs to enable metadata-rich query and search
- Data Portals to enable hosting, sharing, exploration, and access
- Data governance and policy
- Flexible APIs (aligned with the IRI software ecosystem)

Spokes

- Supporting and sustaining data repositories and portals
- Supporting automated experimentation and AI workloads
- Enabling real-time data streaming, processing, and analysis from experiments

Critically, HPDF must provide data lifecycle support to DOE user facilities and projects through: highly reliable and highly available resources; software and tools to effectively and efficiently manage, process, and share data; data stewardship in the DOE community.

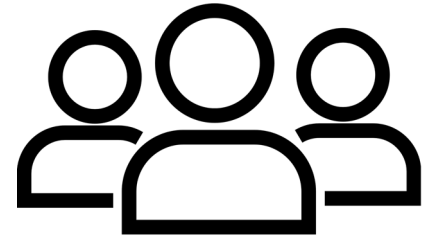
HPDF Project Next Steps

HPDF Project is focused on meeting **CD-1 milestones**, continuing and expanding community engagement, and developing partnerships in the coming year.

- Explore and develop HPDF's role in the practical implementation of data governance and policy, which is evolving in the DOE SC community.
- Develop a software-focused approach for early integration that leverages existing initiatives and community products
- Define the models and integration for partnerships and interfaces between the Hub and Spokes
- Develop a phased deployment approach that can serve the urgent needs of the community
- Develop the vision for the future HPDF facility, including operational and allocation models
- Lead the development and execution of the storage and data management roadmap for the ASCR Ecosystem in coordination with other ASCR facilities

2024/2025 Deliverables

- Priorities/Timeline/Roadmap
- Vision for Operational and Allocation models
- Workflow Archetypes
- Refine conceptual design priorities (Hardware, Software & Services, Spoke models)
- Set up Advisory groups
- Explore data policy and governance



Q&A

LRamakrishnan@lbl.gov

 <https://hpdf.science>



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.

