



VDC Data Services Integration with NSDF discussion (Use case scenario)

10 Feb 2022



CIF21 DIBBs: EI: Element: The Virtual Data Collaboratory: a Regional Cyberinfrastructure for Collaborative Data Intense Science (VDC)



Award #: 1640834



RUTGERS
Discovery Informatics Institute



PennState



THE UNIVERSITY
OF UTAH®

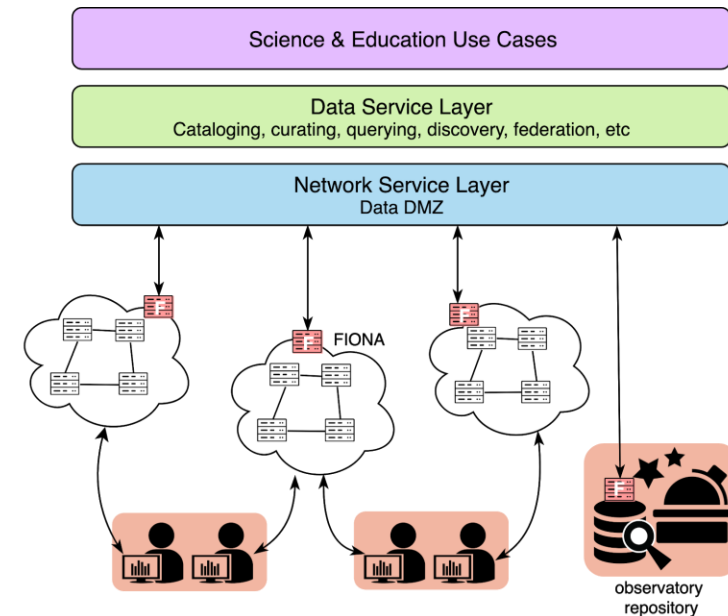
“A federated data cyberinfrastructure for data-intensive, interdisciplinary and collaborative research.”

Key VDC features:

- A dedicated high-speed network, compute and storage **resources federated** over the participating institutions
- **Data discovery**: a set of Data Services including indexing, cataloging, sharing and metadata management
- An **Internet-scale execution platform** for containers, allowing to distribute complex distributed analytics close to the data source
- A network of **high-performance DTNs** equipped with fast storage for implementing smart data delivery strategies.

Impact:

- VDC connects people with data and compute
- Enhanced early warning using online data fusion from large facilities (data streaming)
- Structural bioinformatics

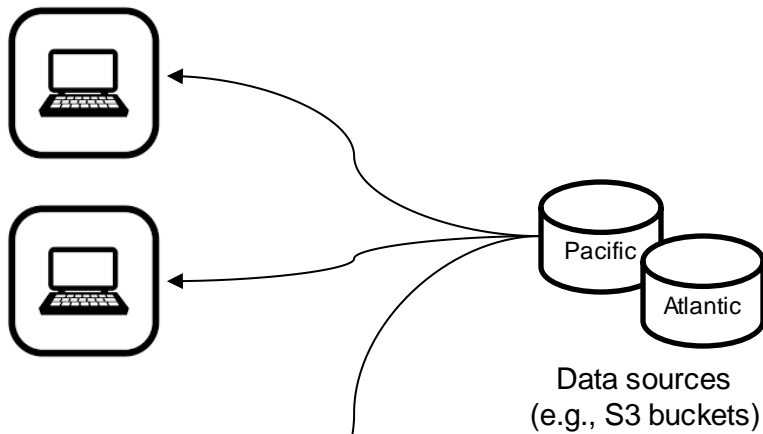


EVDC
VIRTUAL DATA
COLLABORATORY

Collaboratory – Use Case Scenario

Example: interactive training event

- Diverse data sources
- 50+ trainees



Potential issues:

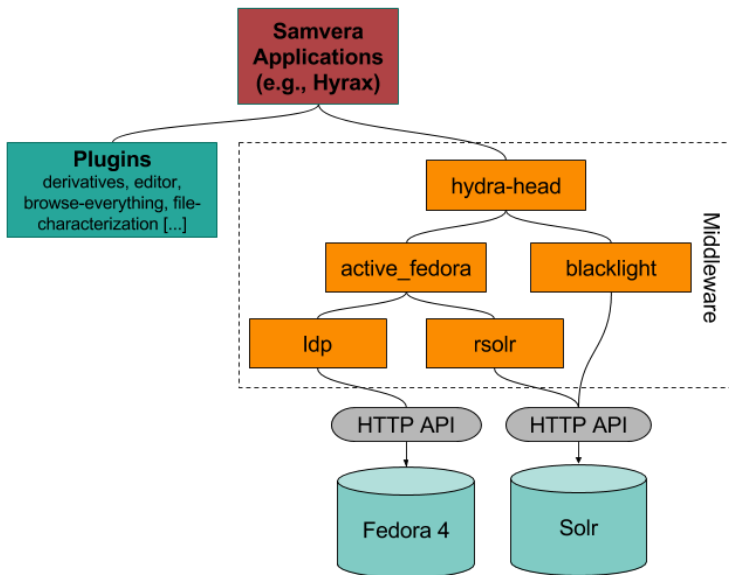
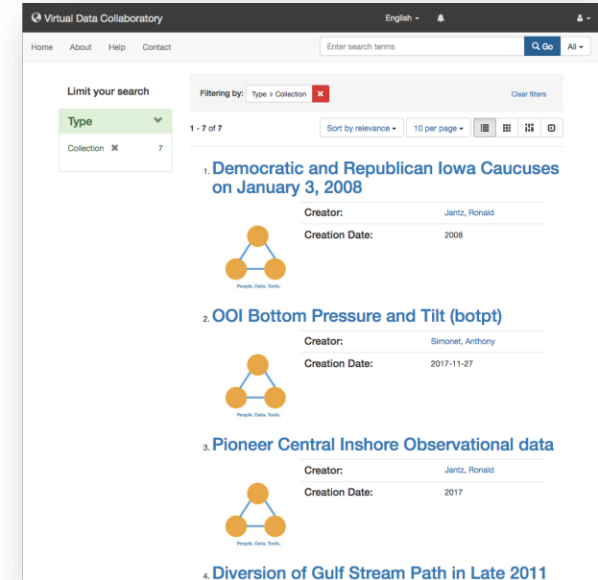
- Latency
- Bottlenecks
- Compute costs
- Potential egress costs
- Etc.

The screenshot shows the 'Pacific Sound' website interface. The top navigation bar includes links for Home, Installation, Models, Data Access, Blue Whales, Humpback Whales, Shipping Noise, Change Log, and License. A search bar and a user profile icon are also present. The main content area features a 'Welcome' message, a link to documentation, and a 'Quick start' button. Below this is a map of the North Pacific region, highlighting the Monterey Bay National Marine Sanctuary (MBNMS) and Monterey Bay. A circular inset shows a hydrophone array labeled 'MARS'. The map includes latitude and longitude coordinates and a 30 km scale bar.

<https://docs.mbari.org/pacific-sound/>
(Open data on AWS)

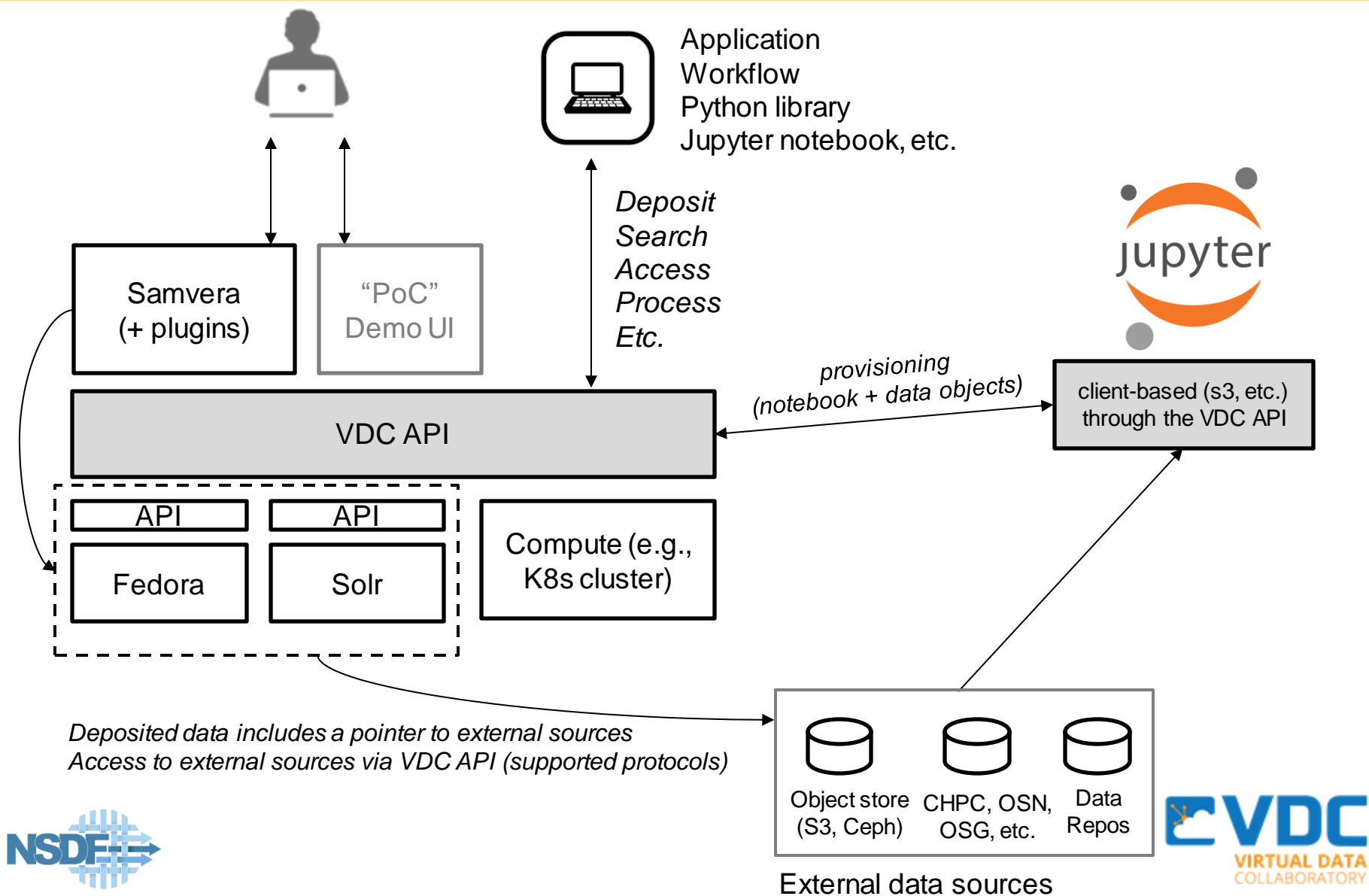
VDC Data Services

- Provide tools and services to work with large datasets
- Registering objects (collections/files/links)
- Searching, Discovering, Sharing
- Create DOI, Store/edit Metadata
- Deriving data, storing provenance
- Provides Globus endpoints for deposited data
- Multiple upload methods (direct, Dropbox, Google Drive, etc.)
- AAI integrated with CILogon

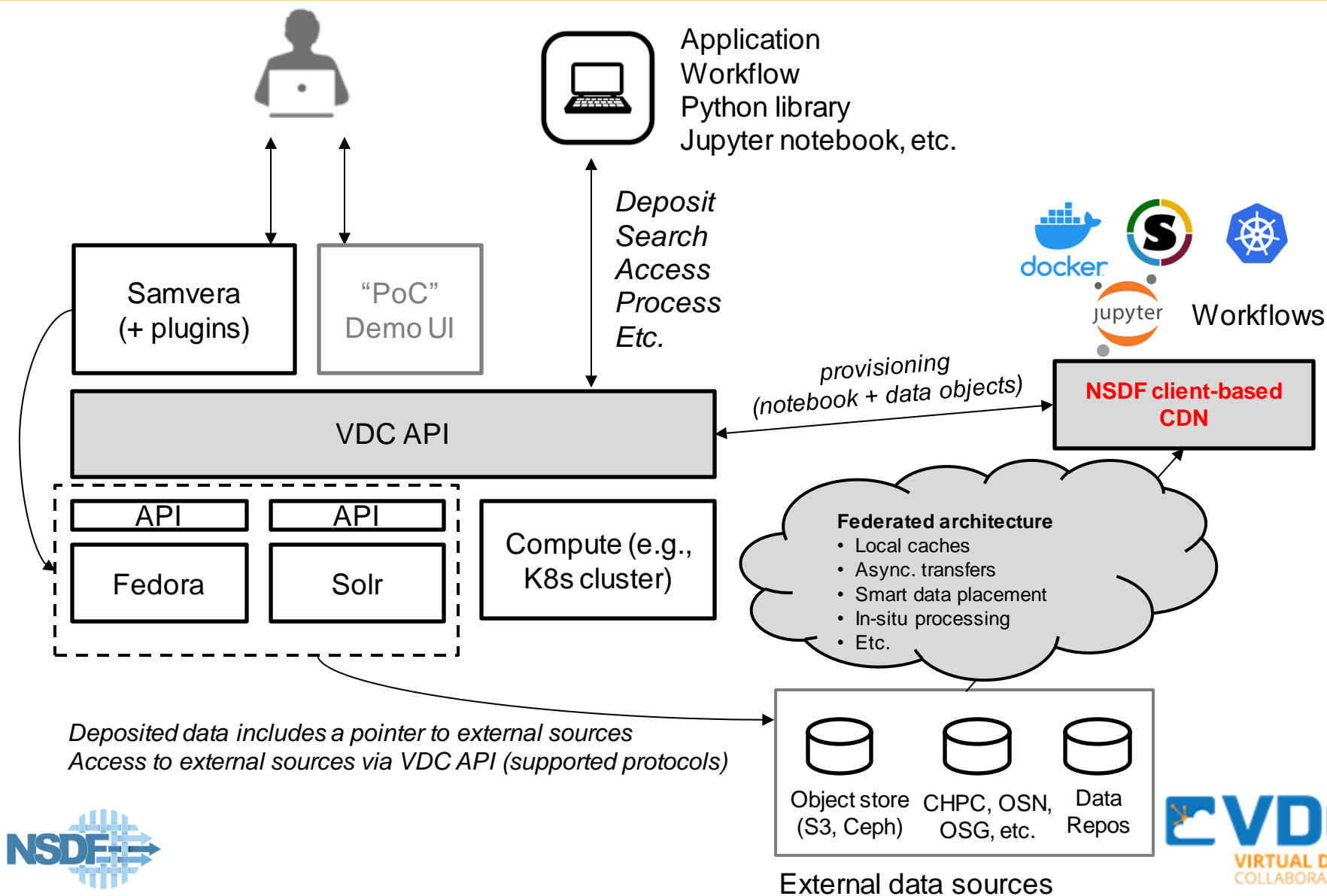


- FAIR Technology stack
- Based on Samvera (customized)
- Fedora provides content as linked data (RDF)
- Semantic support: data and metadata using any ontologies and vocabularies
- Advanced search: content indexed into Solr
- Advanced Query: can be easily indexed using triple store applications (e.g., Jena Fuseki) - SPARQL query language support

Today's Demonstration (1/2)



Today's Demonstration (2/2)



Proposed NSDF Architecture (1/2)

- User/community functionality requirements
- Metadata/data discovery and access
- Analytics, visualization, etc.
- Etc.



Users and Workflows

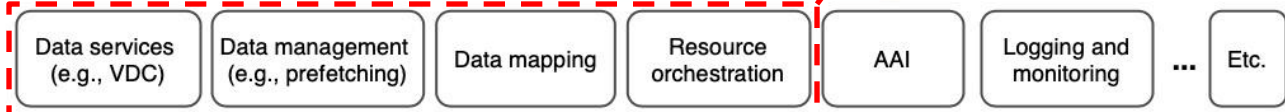
NSDF Interfaces
(API, clients, etc.)

- Technology requirements
- Federation (technical, non-technical)
- Integration
- Etc.

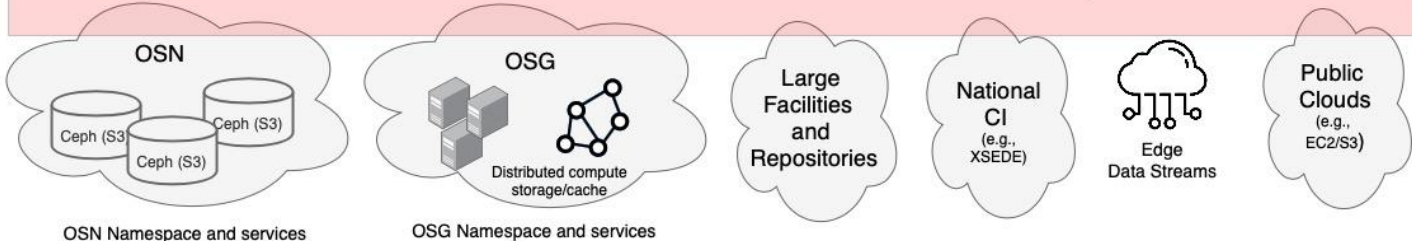
- Leverage core VDC components and APIs (e.g., data discovery/access)
- Building block for developing “smart” data management services
- Support for data/resource brokerage implementation

NSDF

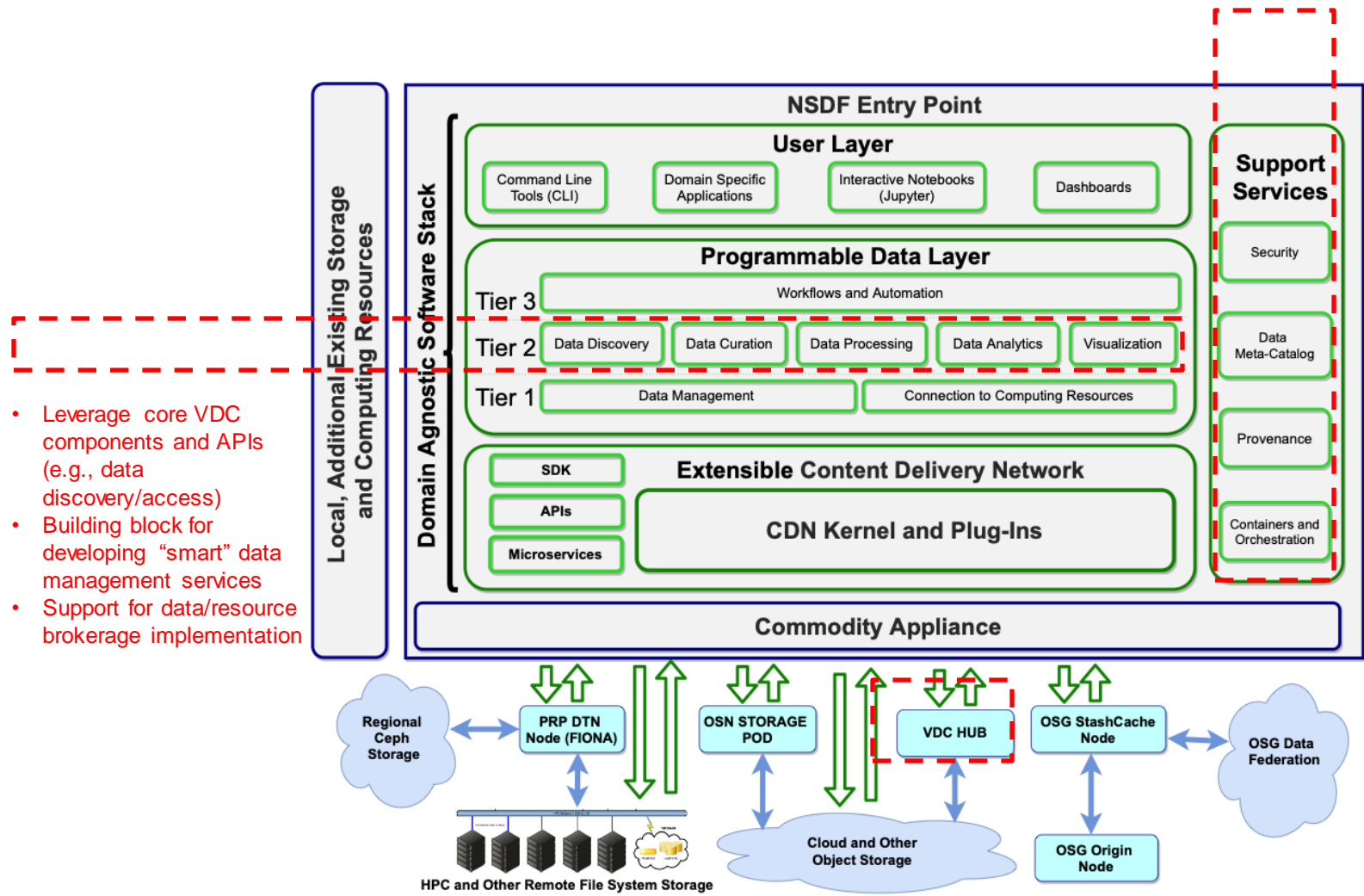
Core services



Federated Resources



Proposed NSDF Architecture (2/2)



- Building block for support services

- Leverage core VDC components and APIs (e.g., data discovery/access)
- Building block for developing "smart" data management services
- Support for data/resource brokerage implementation



Thank you!

ivan.rodero@utah.edu

