NSF CSSI PI Meeting, Seattle, WA, Feb. 13-14, 2020

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

PI: Ivan Rodero, Co-PIs: Vasant Honavar, Jenni Evans, Grace Agnew, James von Oehsen
Institutions: Rutgers University and Pennsylvania State University

Award #: 1640834

VDC VIRTUAL DATA COLLABORATORY

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

RUTGERS Discovery Informatics Institute · PennState · NJEdge · KINBER A Commonwealth of Collaboration

Project Overview and Goals

Motivation:

- Explore robust, configurable, extensible, data and computational infrastructure to support collaborative, reproducible, and data-intensive science

Goals:

- Seamless access to data & tools for researchers, engineers, and entrepreneurs
- Train the next generation of scientists in leveraging data, cyberinfrastructure, and tools to address research problems

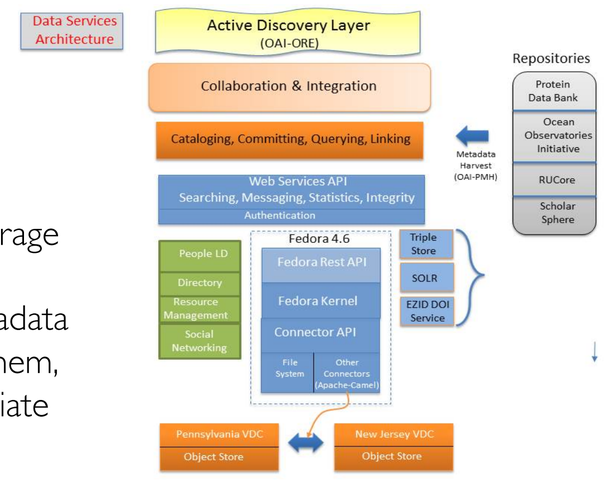
Key Components:

- Scalable data-intensive computing platform
- Data services to support research workflows
- Regional science data DMZ network

Data Services Layer – FAIR Data

Enabling Interdisciplinary Research and Context:

- Based on Samvera
- Designed for intuitive organization and discovery of research products
- Integrates tightly with other big data frameworks, storage and computing infrastructures
- Record, store and search provenance and other metadata
- Products are linked to the researcher who created them, to the tools that analyzed them, and to any intermediate products (analyses, visualizations, etc.)



- An integration framework to enable VDC to interoperate with other large data repositories
- APIs enable VDC users to discover resources outside the VDC, using powerful search and browsing capabilities of external repository, but leverages resources within the VDC

VDC System and Data DMZ Layer

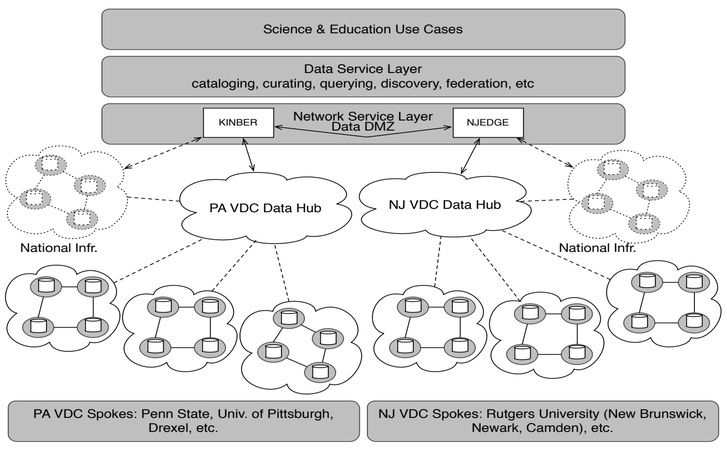
Computing Platform:

- Support for large-scale multi-site workflows
- Big Data frameworks, including in-memory and streaming platforms

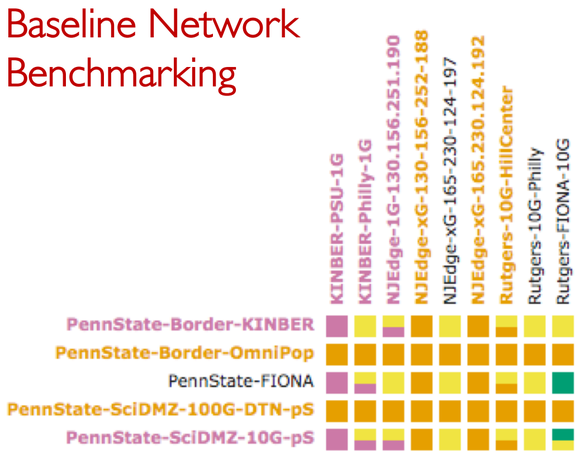
Data DMZ Components:

- Data DMZ backbone: direct 10 GE connections between Rutgers, NJEdge and KINBER
- Data HUBs connect directly to the Data DMZ or through their regional network
- Data Spokes connect through the regional network
- (FIONA) Data Transfer Modes deployed at Data HUBs and regional networks

VDC Architecture



Baseline Network Benchmarking



Education and Outreach Activities

Goals:

- Provide tools for using large-scale data in the classroom
- Work with scientists to translate their research into educational products that can be used by K-16 students
- Unique VDC educational programs focused on interdisciplinary aspect of research, connectivity to external repositories, collaboration
- Sensitize students and early career researchers to issues important for data management (e.g., curation, reproducibility)
- Tools, resources, and learning modules available for broad implementation at other locations via the VDC website
- Raise awareness of VDC resources among high school teachers, faculty and researchers

Application Case Studies – Demonstrative data-intensive, interdisciplinary and collaborative research

Tsunami Early Warning

Description:

- Increase precision and delay for Tsunami warning by analyzing multiple sources of data simultaneously, in collaboration with UNAVCO.

Data Sources:

- High-precision GPS stations (UNAVCO)
- Bottom pressure (OOI)
- Underwater seismometers (IRIS)

Stream Analytics:

- VDC spokes deployed close to the data sources compute local analytics.
- When a GPS station triggers SLA/LTA, the VDC Data Services are queried to find the nearest stations on all networks and get an estimate arrival times to more distant stations
- Additional instrument can also be queried dynamically to compute radial distance and determine the hypocenter of the earthquake

Monitor:

- The VDC computing infrastructure hosts the central rule-based algorithm responsible for automatically issuing warning based on observations and models.
- Live graph shows displacement of earthquake over all the stations to improve.

Structural Bioinformatics

Description:

- Collaborative assembly, integration, and analyses of several data sets of protein-nucleic acid complexes derived from the Protein Data Bank (PDB)
- Shared data and computational infrastructure, complete with computational workflows

Examples:

- Characterization of conformational changes in proteins upon binding to DNA
- Computational prediction of protein-DNA and protein-RNA complexes

Expected Outcomes:

- Curated datasets, assigned DOIs, versioned, indexed, and shared to support intentional revisions to data and analyses tools.
- Digital artifacts linked to the work products using the VDC's Data Services Layer