Virtual Data Collaboratory (VDC): A Regional Cyberinfrastructure for Collaborative Data Intensive Science

Rutgers University, Pennsylvania State University, City University of New York, NJEdge, KINBER

Project Overview, Goals, and Milestones

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

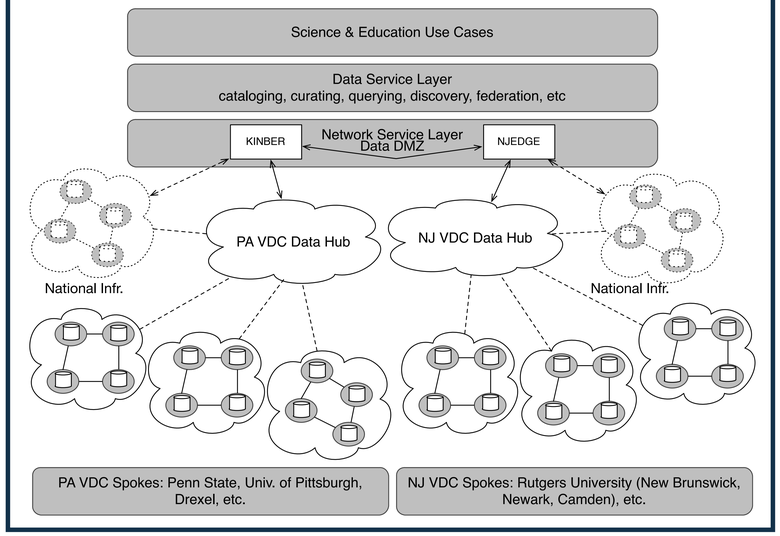
Key Components:

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

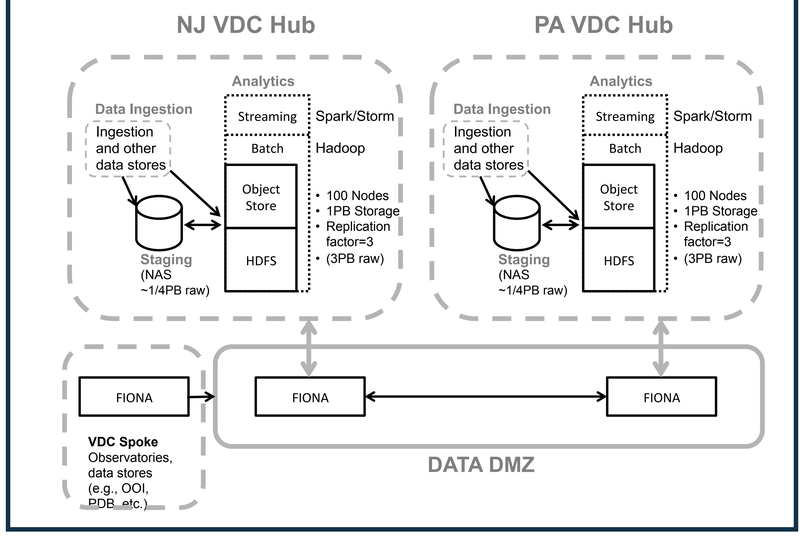
Milestones:

- Phase I (1.5 Years): Acquisition, deployment, and testing of computing infrastructure and initial configuration of the Data DMZ
- Phase II (1.5 Years): Commissioning of science DMZ; deploy network and data infrastructure at spoke campuses
- Phase III (1 Year): Transition data to VDC and intensify outreach within the two hubs, spokes, and the region, using experiences and feedback to improve.

VDC Architecture



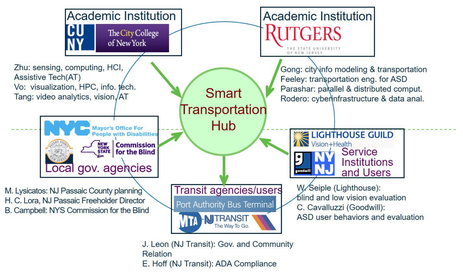
Computing Platform



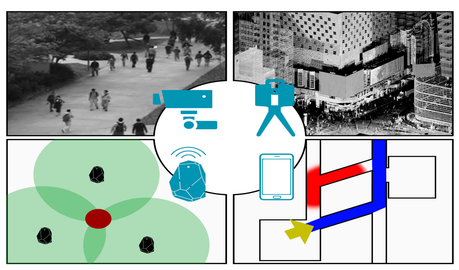
Application Case Studies

Smart and Accessible Transportation Hub to All

In U.S., the visually impaired population has reached 6.6 million people and expected to double by 2030, and ASD is the fastest-growing developmental disorder affecting 1 in every 45 people. With smart city technologies, many people will be able to travel with greater confidence, while others may be able to travel independently for the first time.



Crowd Analysis | Digital Facility Model



Mobile Localization | Personalized guidance

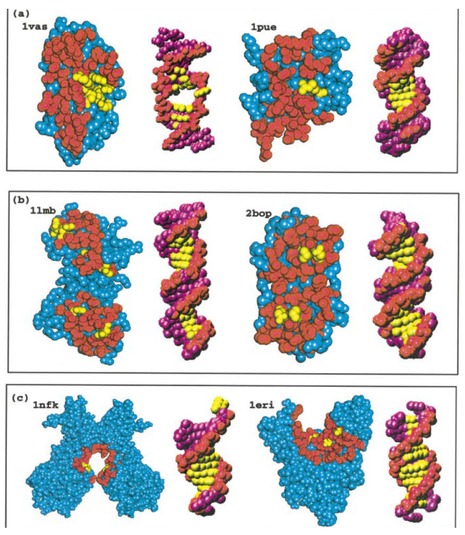
(1) Internet of Things: 3D semantic modeling, wireless sensor network, surveillance cams, and active navigation while requiring minimal infrastructure changes
(2) Analytics: Deep-learning, image-to-model registration, localization, sensor placement optimization, and path planning
(3) Services: Personalized & adaptive travel guidance for multiple disadvantaged groups

Studies of Protein Nucleic Acid Interactions

Protein-DNA and protein-RNA interactions play a central role in cellular processes. We aim to develop novel computational methods to automatically characterize and predict protein-nucleic acid interactions, interfaces, and complexes.

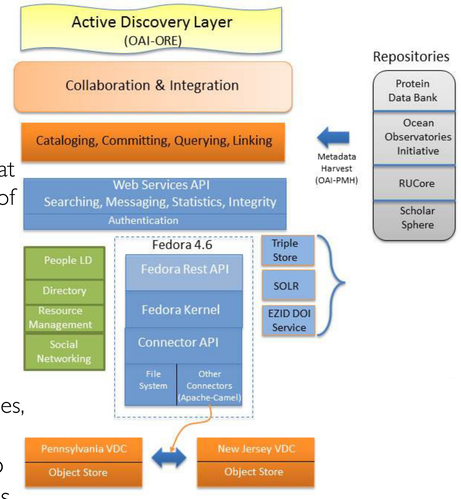
We will develop workflows to:

- Select appropriate complexes from the PDB
- Incorporate various annotations and have them reviewed by experts
- Compute a variety of characteristics of bound and unbound structures
- Use state-of-the-art machine learning methods to cluster or classify interfaces
- Visualize clusters and their annotations
- Share data sets and results



Data Services Layer

- How the researcher experiences the VDC
  - Design for intuitive use. How the researcher expects to organize and discover research products
  - Integrates tightly with other components so that the researcher can leverage all the capabilities of the VDC
- Enabling Interdisciplinary Research and Context
  - Use linked data to show context. Research products are linked to the researcher who created them, to the tools that analyzed them, and to any intermediate work 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



Data DMZ Layer

- How the researcher connects to the VDC
  - A high-bandwidth high-performance network infrastructure connecting the VDC Data Hubs and Spokes
  - Supports data import/export services with necessary qualities of services to enable efficient and transparent access to data and compute regardless of a scientist's location
- Data DMZ Components
  - Data DMZ backbone consists of 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