Virtual Data Collaboratory (VDC)

A Regional Cyberinfrastructure for Collaborative Data Intensive Science

Manish Parashar, Rutgers Univ., Vasant Honavar, Penn State Univ. and the VDC Team

Motivation: Scientific progress across disciplines is increasingly enabled by our ability to examine natural phenomena through the computational lens (e.g., using algorithmic or information processing abstractions of the underlying processes) and our ability to acquire, share, integrate, steward, and analyze disparate types of data. However, realizing the full potential of data to accelerate science calls for robust, configurable, extensible, data and computational infrastructure to support collaborative, reproducible, data-intensive science by teams of researchers across institutional as well as disciplinary boundaries.

Overview: The research team will design Virtual Data Collaboratory (VDC), a federated infrastructure that integrates the state of the art data-intensive computing platforms, storage, and networking, with an innovative data services layer across three geographically distributed Rutgers University (RU) campuses in New Jersey (Camden, Newark, New Brunswick), multiple campuses in Pennsylvania (Pennsylvania State University (PSU), Drexel University, Temple University, University of Pittsburgh) and beyond (City University of New York (CUNY)), coupled by a high-speed network managed by New Jersey's Regional Education and Research Network (NJEdge) and KINBER. VDC will build on and integrate with existing national/international and regional data repositories (including NSF funded repositories such as the Ocean Observatories Initiative (OOI) and the Protein Data Bank (PDB)), and leverage local/regional/national ACI investments, such as the NSF funded Pacific Research Platform (PRP), Big Data Regional Hubs, XSEDE, OSG, Campus Cyberinfrastructure projects, among others.

Challenges: The primary challenge to date has been the coalescing of a diverse, interdisciplinary, geographically distributed team of researchers and data infrastructure experts around a shared vision for the design, deployment, and use of the VDC. To address this challenge, the team met face-to-face for a day-long project kickoff meeting at Rutgers on October 31. During the meeting, the overall goals of the project, VDC architecture and implementation, data services, and the use cases were discussed and an action plan and timeline were developed. Leads for each of the major thrusts were identified, along with team members from the participating campuses. Action items and coordination mechanisms were established. The teams for each of the main thrusts have been meeting regularly and making good progress. The team plans to have its next face-to-face meeting during Spring 2017. A second challenge has to do with the hiring of the right personnel with the necessary cross-disciplinary skills. The team is collectively addressing this challenge.

Future Directions: The project is on track for deploying the VDC hardware at Rutgers and Penn State in about 6 months. Central to the VDC are three infrastructural innovations: (i) regional science DMZ (Data DMZ) that provides the data import/export services and necessary services to enable efficient and transparent access to data and computing capabilities regardless of location, (ii) expandable and scalable architecture for federated data-centric infrastructure and (iii) data services layer that supports data linking, search, and sharing; mechanisms to attach DOIs, archive data, data and compute-intensive research workflows for collaborative research. The team will leverage the VDC infrastructure for research-based graduate and undergraduate education and training in the Data Sciences at the participating institutions. The team has agreed on the design of the software stack. The data services to be supported are being prioritized based on the needs of the use cases. The teams focusing on the use cases are making good progress. For example, the collaboration between Helen Berman, a Rutgers structural biologist, founder of the Nucleic Acid Database and former director of the Protein Data Bank, and with Vasant Honavar, the principal investigator of the project at Penn State is well underway. They have begun to assemble curated data sets of protein-DNA and RNA complexes and interfaces and have developed plans for computational workflows characterizing and predicting protein-DNA and RNA interfaces, interactions, and complexes using machine learning and other computational tools. This will not only help develop and evaluate the VDC infrastructure, but the results will advance the understanding of protein-nucleic acids interactions.