978-1-5090-3525-0/16/$31.00 ©2016 IEEE

Cross-Institutional Research Cyberinfrastructure for Data Intensive Science

W.Christopher Lenhardt
Mike Conway
Erik Scott
Brian Blanton
Ashok Krishnamurthy
Renaissance Computing Institute (RENCI)
University of North Carolina, Chapel Hill

Mirsad Hadzikadic
University of North Carolina, Charlotte

Mladen Vouk
Alyson Wilson
North Carolina State University

***Abstract*—This paper describes a multi-institution effort to develop a "data science as a service" platform. This platform integrates advanced federated data management for small to large datasets, access to high performance computing, distributed computing and advanced networking. The goal is to develop a platform that is flexible and extensible while still supporting domain research and avoiding the walled garden problem. Some preliminary lessons learned and next steps will also be outlined.**

Keywords— distributed computing; risk; analytics; distributed data; open system architectures; data intensive computing; big data and distributed computing

I. INTRODUCTION

"Science as a Service" [1], "Backend as a Service" [2], science gateways and data commons are all labels for services that connect cyberinfrastructure (CI), data, and applications for end‐user scientists and researchers. However, this space is usually ignored in big data, data-driven science and analytics, and sustainable, reproducible, and multi‐ and interdisciplinary research discussions. Further, where integrated cyberinfrastructure exists, it is often stove-piped or a walled garden type of infrastructure hard-wired to support a specific research domain. This paper describes a collaborative project to develop such leading edge distributed data infrastructure which attempts also to address the walled garden problem for scientific research on risk and analytics. The project is part of a broader North Carolina Data Science and Analytics Initiative (NCDSA) lead by the University of North Carolina – Charlotte (UNCC) with partners at the Renaissance Computing Institute (RENCI) and North Carolina State University (NCSU).

The goal is to leverage advanced data science technologies and apply them to address challenges faced by contemporary scientists in data-rich and data-intensive scientific problem areas. Illustrative technological challenges scientists across domains face include access to distributed compute and data resources, an easy integration of applications and workflows, linking models, and coordination of distributed research teams. Data science challenges include increasing volumes, velocities and variety (and complexities) of data, issues related to understanding and assessing data quality (data veracity), data formatting, and lowering the burden associated with data curation and data publication. Unfortunately, researchers still spend a significant amounts of time finding, accessing, configuring, and connecting cyberinfrastructure resources, in addition to devoting significant time to transforming and managing data at the expense of conducting actual scientific research.

This paper describes our approach to developing a data science as a service platform. We utilize several open source cyberinfrastructure components, as well as some proprietary ones, to establish a flexible and extensible, cross-institutional research infrastructure. We will briefly describe a specific example of its use in subproject which leverages this architecture and incorporates additional leading edge technologies in new ways to address the challenges described above. In addition to the technical descriptions some preliminary lessons learned and next steps will also be outlined.

II. BACKGROUND

The NCDSA initiative has four main objectives: 1) create a distributed computational and data management architecture; 2) identify and incorporate relevant data; 3) curate new data developed in the context of the project; 4) identify and host relevant tools, including analytics software and policy management, for the researchers taking part in the initiative. The NCDSA topical focus is "big" data analytics and risk analysis for public and private stakeholders in North Carolina.

The umbrella cyberinfrastructure under development for NCDSA is called DSAi (pronounced "daisy"), Data Science and

The NCDSA effort is funded through a grant from the UNC system from funds allocated from the North Carolina General Assembly.

Analytics Initiative. DSAi is being collaboratively developed and implemented between the partners. DSAi incorporates distributed storage and computational resources. Each participating institution implements a common software stack and some specialized capabilities, then the resources are federated.

We highlight a particular use case for DSAi: the Risk Analytics Discovery Environment (RADE). RADE leverages the capabilities of DSAi and integrates additional data and analytical capabilities. The RADE project focuses on risk analytics for North Carolina. The goal for RADE is to provide a "[Data] Science as a Service" [1] capability connecting cyberinfrastructure and end-user scientists and researchers. The data and application requirements for RADE come from three domain science use cases. These use cases are quite diverse and cover a range of data velocity, volume, variety and veracity. This helps us develop a more robust data-science research infrastructure for NCDSA. The first use case applies natural language processing techniques (NLP) for automatic extraction of micro-level data from news sources (potentially high-velocity unstructured data). The second use case develops analytics for projecting future property values based on future hazards profiles (potentially high volume data). The third use case develops analytical methods for modeling the spread of non-native mosquito species and hosted viral disease.

III. Technical Aspects

A. DSAi

Data Sciences and Analytics Infrastructure (DSAi) is a three-campus federated storage system based on iRODS (the Integrated Rules-Oriented Data System, http://irods.org/, see Fig. 1). An iRODS deployment creates a data grid composed of a distributed system of storage resources and metadata servers. The DSAi architecture consists of a fully-independent installation at each campus with a unified filesystem tree and with authentication and authorization shared between institutions. Each institution continues to provide authentication and authorization of their own users in compliance with their own campus and departmental policies, while the data grid simultaneously allowing users to configure access to their data irrespective of the data storage infrastructure at each university. Data collections are thus abstracted from the physical storage system. The DSAi implementation includes 230Tb of usable space at RENCI, with similar amounts at the other institutions, and has access to computational resources such as the Hatteras cluster at RENCI (5168 cores), the SOPHI Hadoop environment at UNC-Charlotte, and the Virtual Computing Lab (VCL) cloud at NCSU.

The iRODS system implemented in DSAi also supports the execution of rules based on the iRODS rule language which defines policies and actions in the system. iRODS rules execute data management actions, such as file creation of a metadata update, which allows iRODS to actively manipulate the collection. Rules are already provided, for example, for tasks of interest to archivists, and the system supports the further addition of rules for any purpose – such as an integrated workflow for storm surge modeling research. Notably, rules may also access the large clusters at the three sites. This interconnection means that workflows may be controlled by data, rather than being limited conventionally to the workflows, and thus facilitates data-driven adaptive workflows.

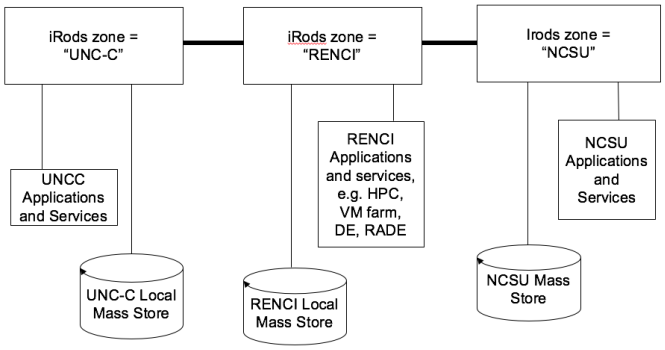

Fig. 1. DSAi Schematic

One of iRODS' most distinguishing features is the support for rich metadata in the storage system far beyond traditional filesystem support. Users can add, modify, and delete arbitrary metadata at will, and can access files based on the contents of the metadata. For instance, a file of Doppler current profiler data can have a metadata entry that contains the instrument's serial number. Users can then query the metadata server to return the files that were produced from that specific instrument. No longer is it necessary to encode metadata into unwieldy and long filenames and directory structures.

B. Discovery Environment (DE)

While much can be done with DSAi through leveraging the capabilities of iRODS, to develop a more complete data science as a service capability, the project is adding another set of tools in the form of the CyVerse (http://www.cyverse.org/) Discovery Environment. The Discovery Environment is a robust middleware architecture originally designed to provide cyberinfrastructure for grand challenge plant biology [3]. DE provides a platform to manage research data across the entire lifecycle, "from acquisition and analysis to publication and archiving" [4]. This lifecycle approach to cyberinfrastructure necessitates consideration of both data management and computation. The DE framework includes concepts of "tools" and "apps." DE allows researchers to register computational processes as Docker images, i.e. tools. These tools then appear inside the DE environment and can seamlessly stage data, schedule jobs, manage derived products, and notify users. DE is built from many open source cyberinfrastructure components, including the iRODS middleware platform. iRODS and the DE middleware provide a comprehensive architecture for full lifecycle research data management, including preservation, collaboration, computation, and discovery. From one coherent service architecture, researchers can manage and share both data and computational processes, as well as the products of workflows and analyses.

Fig. 2 describes the overall architecture of the DE stack. At the bottom, iRODS provides an abstraction over myriad storage technologies. It provides a global namespace and the ability to

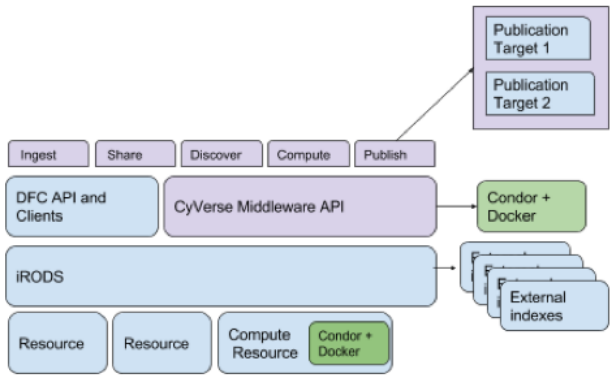

Fig. 2. Discovery Environment Schematic

federate data; it provides a rule engine that can enforce policies at each location; and it provides a metadata catalog. The DE middleware is a distributed, share-nothing architecture that communicates between components using a REST-ful API. This REST-ful API covers many necessary cyberinfrastructure functions for managing research data. The iRODS grid can signal all activities on data, metadata, and collections to registered indexers via AMQP. This allows different collections and parts of the grid, mediated by policy, to be projected into attached indexers, allowing data discovery through diverse technologies, including full text indexing and search, natural language processing, metadata extraction from structured data, and user curated metadata.

DE uses an "apps" abstraction to manage computation against data on the grid. A popular method is to allow researchers to define their own tools as Docker images, which are then wrapped as friendly apps. Condor is used to schedule these applications, and DE manages notifications, staging of data to and from a process, and gathering of provenance and other data to track how data sets were created, and allowing reproducible experiments. Current work at CyVerse has demonstrated discovery of the location of data at rest, such that the Docker image can be moved to the data. This is especially useful when large data sets are involved. A second, related functionality under development is the ability to configure plug-in publish mechanisms, so that research data may be published to various repositories and shared catalogs.

While the DE was built specifically to support iPlant and CyVerse researchers, the general patterns of research, and the infrastructure needs that support these patterns, are common across many domains of research [5]. The developers at CyVerse have been generalizing the DE infrastructure so it may be deployed outside of CyVerse as a middleware stack. Four groups are currently involved in this effort. First are the CyVerse developers, who manage the open-source DE repository on GitHub at https://github.com/angrygoat/DE. This GitHub repository contains both the DE codebase, as well as Docker images and Ansible playbooks that can stand up a configured DE instance. The second group is the DataNet Federation Consortium (DFC, http://datafed.org) who originally developed iRODS and many of the libraries used by DE. The third is the Odum Institute (http://www.odum.unc.edu/) at the University of North Carolina. Odum is interested in leveraging this stack to create a Virtual Institute for Social Science Research (VISSR) [6]. The fourth group is RENCI itself, participating in the development of the VISSR project, as well as DSAi.

These efforts have resulted in the creation of a fork of DE at: https://github.com/DICE-UNC/DE. This fork expands the CyVerse code to include extended Ansible and Docker resources that represent many of the ancillary technologies in DE, including LDAP, Grouper, CAS, and Condor. This installation is considered a "reference implementation," in the sense that it can be configured with a known set of supporting infrastructure to both demonstrate the middleware and to serve as a basis for functional and integration tests. In addition, the DFC fork of DE upgrades the iRODS version, and efforts are underway to redact CyVerse specific functionality, and to increase the pluggability of the software to include additional configurable components. It is hoped that these efforts will result in a more official and robust "community" edition of the CyVerse stack so that iRODS and the DE architecture can become an installable middleware architecture for collaborative, data-driven research. At this writing, two independent DE instances have been installed at RENCI and at the Odum Institute. Each of these instances is now moving towards a production release, in close collaboration with CyVerse. Future plans for a community release will originate from CyVerse, and it is intended that this collaborative effort will result in a constellation of middleware installations that can share a common architecture, and the ability to federate and manage distributed data at scale.

C. *Virtual Computing Lab*

In addition, the NCSU Apache VCL cloud (e.g., [7]) is available (via Shiboleth authentication) to the whole UNC System, including DSai users. VCL a mature and proven academic cloud environment that offers on-demand resources and a rich collection of analytics and workflow management tools, from software such as SAS, SPSS, and R, to a collection of IBM Watson-type services, to Kepler workflow framework [8], (http://kepler-project.org), to Hadoop and tightly coupled HPC services, etc. (http://vcl.ncsu.edu).

D. *Risk Analytics Discovery Environment (RADE)*

The concept of RADE is a domain-specific example of the type of flexible, extensible, data science as a service described in this paper. The RADE part of the overall project will adapt the generalized version of the DE and "brand" it for the RADE specific set of research communities. The RADE will be connected to DSAi which will serve as the data repository for the scientists using RADE. The three RADE use cases are data and computationally intensive, require execution of complex scientific models, chaining models, as well as developing complex work flows. RADE will allow researchers to upload, archive, and manage complex data, install applications, and develop workflows to carry out analytics required for the use cases as described above. The RADE use cases outlined above will store data in DSAi and then use the RADE DE to develop their analytics. For example, the natural hazards use case will run a storm surge model as an app from within DE and the

model will input and output data into the DSAi iRODS grid. The data being integrated into RADE includes large amounts of text data, large datasets of shipping information, as well as large volumes of complex geographic information system data representing building outlines and plots for an extended geographic region. RADE will also run analytics and computation of such things as storm surge, disease vector models, and NLP. DSAi will also serve as the platform to manage and disseminate results from the RADE use cases.

IV. Comparison to Other Similar tools

Our project represents one approach to the challenges of delivering a data science as a service platform. There are a number of other examples across a variety of science domains that may be thought of as addressing various aspects of the data science as a service concept. In this section we outline a very brief comparison to two related examples, Earth System Grid and science data repositories.

A. Earth System Grid

The Earth System Grid (ESG, http://esgf.llnl.gov/index.html) is a set of federated nodes based on GridFTP used primarily for climate modeling and analysis data. ESG provides the data sharing framework for important international efforts such as the Intergovernmental Panel on Climate Change (IPCC) periodic climate assessments. These climate assessments involve the most advanced operational climate models and disparate types of data; i.e. volume and variety. Data held in ESG nodes may be seemlessly searched and accessed via the ESG search interface. The ESG technology stack is non-proprietary and based on Open Source technologies. ESG is designed to handle petabyte level data holdings. Led by LLNL with Department of Energy-funding, partners included major climate and environment related Federal agencies, including NASA, NOAA. ESG supports services such as THREDDS, OPeNDAP, and GridFTP. ESG is also developing capabilities to run server-side analytical tools [9].

B. [Big Science] Data Repositories

Large scale, or what might be termed ‘big science’, data repositories are information systems designed around expensive science instruments such as earth observing satellites (NASA, https://earthdata.nasa.gov), large physics instruments (http://opendata.cern.ch), and astronomical observing systems (http://www.sdss.org/surveys/). These systems are often purpose built and have traditionally relied on more closed architectures and software. These kinds of repository systems are designed to support search and retrieval as well as rigorous data curation. They may include limited visualization capabilities. With some exceptions, these repositories fall into the paradigm of search, download, and analyze locally.

V. Analysis of RADE v Other Tools

Each science domain tends to have purpose built science data infrastructure. This makes sense as these infrastructures have developed over time to serve their respective communities, incorporating new science and technology in an evolving way, and these systems tend to have funding streams specific to their domains. Further contributing are domain-specific data formats, vocabularies, and analytical requirements. As a result the science data systems that have developed tend to be tightly coupled in terms of architecture, standards, and domain-specific terminologies. To be sure these systems are moving to incorporate new approaches such as virtualization and server-side processing.

In contrast, DSAi and the DE/RADE tools are designed to support any number of protocols, standards, and services. The approach is to create a middleware framework that does not necessarily require the user to adopt a particular approach with which they may not be familiar. The approach is more loosely-coupled and modular allowing users to essentially ‘plug in’ to the system with their particular tools and services. The goal is to abstract the data layer and the application layer. We are also seeking to add networking abstractions, and data curation/publishing to the framework.

DSAi and RADE are designed to strike the balance between narrow purpose-built systems and overly general but complicated systems that are unusable. Once installed, they are also designed so that the scientist need not be an IT systems technical expert in order to use the system.

VI. Lessons Learned

A. Authentication

In the course of developing DSAi, one of the initial difficulties that developers encountered creating DSAi is how to negotiate cross-institutional authentication. Addressing common authentication methods is an essential first step for creating such a cross-institutional system. We did this by integrating CAS (central authentication service) which allows the use of other authentication protocols such as InCommon and Shibboleth. Both InCommon and Shibboleth are identity management services prevalent in academic settings.

B. Integration of Key Technologies Can Be Problematic

Key technologies such as middleware, database, or client server applications are becoming complex and challenging to deploy in an integrated way due to underlying software version dependencies, and variations in operating system type and version. Using virtualized and containerized applications and software stacks can help in this regard.

C. Community Code

Relying on research community code has its benefits and challenges. Benefits include open sharing and community code development. However, research code tends to be hard-wired to support the specific research applications for which it was developed. Generalizing these types of tools can be difficult and often requires coordination with the original developers as well as focused, agile approaches to reengineering. Repositories such as Github are crucial to making this work.

D. Data science as a service

The DSAi and community Discovery Environment approach provides the ability to deploy a generalized infrastructure that can more easily support specific domains and types of research without the need to re-engineer the entire system. These types of approaches that integrate various types of cyberinfrastructure to support data intensive science are going to become even more

important in the future as present scientific, data, and technology trends continue.

E. HPC Centers, Data Repositories, and Domain Science

Even though cyberinfrastructure as represented by academic and research high performance computing centers, scientific data repositories, and domain science cyber-tools have developed as largely separate streams of research and application, data intensive science necessitates that these resources function in a more seemless way. Experts in the relevant domains such as data science, computer science, computational science, and the like will need to work together to develop data science as a service capabilities. It is no longer acceptable to allow these sorts of resources to remain stove-piped.

F. Walled Gardens are Still a Problem

Even with the best of intentions, the lack of resources and the sheer complexity of cyberinfrastructure often result in addressing only a part of the problem domain. The cumulative effect of technical choices can result in systems that need to cooperate to address the full requirements, yet are difficult to integrate. Widely accepted standards and approaches often do not exist in efforts that are on the cutting edge of cyberinfrastructure research, and a drive towards maturation of these concepts is needed.

G. Design for Pluggability and Expansion

A critical enabler of these efforts is the fact that it can leverage a software stack that has been designed with configurability and pluggability in mind. iRODS in particular is designed to be modular, with plug-ins for storage, API, and many other functions. Customizable management policies have allowed custom configuration without altering core source code. DE is built as to allow easy reuse of core API functions. This is illustrated by using their 'apps' approach to plug in new publishing mechanisms as Docker images.

H. Metadata, Data Management and Curation

Metadata, documentation and data management/curation are another afterthought in the context of traditional big data approaches in the context of scientific research. The DSAi and DE approach allows for the potential to more easily integrate automatic or semi-automatic generation of metadata. The potential to add data management rules or modules and the ability to potentially link to repositories and other curation environments. These elements are essential for science data, model, and workflow reuse and reproducibility.

VII. Next Steps and Additional Challenges

The flexible and modular approach used in developing DSAi and DE allow for the option to add additional capabilities to enhance the data science as a service framework. Developers on the project are working to create a new kind of computational resource that allows containerized computational resources to be sent to data at rest for processing.

Developers will also be connecting capabilities for software defined networking into the DSAi/DE framework. This capability, known as ExoGENI, can also be used with iRODS capabilities for rule-based networking needs [9].

Other projects at RENCI leverage iRODS to provide secure workspaces in order to work with confidential data. Adding this sort of capability to the project could widen the potential utility to include domains such as medical and social science research.

While this piece has focused on the technical aspects of building a data science as a service platform, there are two additional aspects to consider as necessary for creating a successful data science as a service platform. First is training end-users in how to use the technology. While the DSAi and RADE environments are meant to ease the technological burden on scientists there is still a learning curve to traverse. Second, it must also be noted that there is a community building aspect to this as well. Communities of potential users must be brought together in order to drive clear use case development but also to inculcate usage. Potential considerations in this context include usability, community buy-in, and stakeholder alignment related to governance of a resource. Just because you build it, the users may not come. The iPlant user community is an example of successful community development [11][12][13].

VIII. Summary

Access to computation and distributed data in the context of scientific research are often dealt with in a piecemeal fashion. Researchers need integrated frameworks that put much of the technical details in the background allowing researchers to be more productive and to facilitate reproducibility, i.e., data science as a service. The set of services described in this article are specifically designed to address these challenges. This paper outlined a three-campus effort to develop a data science as a service platform for researchers. The project integrates various open source components to provide scientists with access to robust storage, data management tools, access to compute resources, and the ability to integrate their particular science models and algorithms as applications that are much more easily manipulated. The infrastructure also provides the means to much more easily share data and resources in a distributed fashion. The integrated technology also provides the option to more easily integrate metadata, data management and curation workflows. We anticipate that these types of approaches will become more prevelant as the need to conduct interdisciplinary research grows to address existing and future grand challenge science questions.

References

[1] R. L. Grossman, A. Heath, M. Murphy, M. Patterson, W. Wells, "A case for data commons: Towards data science as a service, arXiv:1604.02608v1 [cs.CY], unpublished.

[2] R. Dooley, & M. R. Hanlon (2015). Recipes 2.0: building for today and tomorrow," Concurrency and Computation: Practice and Experience, 27(2), 258–270. doi:10.1002/cpe.3285.

[3] S. A. Goff, et al., "The iPlant collaborative: Cyberinfrastructure for plant biology," Frontiers in Plant Science 2 (2011), doi: 10.3389/fpls.2011.00034.

[4] N. Merchant, E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Micklos, et al. (2016), "The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences," PLoS Biol 14(1): e1002342. doi:10.1371/journal.pbio.1002342.

[5] C. Rusbridge, "Create, Curate, Re-Use: The Expanding Life Course of Digital Research Data," 2007. https://www.era.lib.ed.ac.uk/handle/1842/1731, unpublished.

[6] A. S. Huff, and K. M. Möslein, "Framing research on service," Research Methodology in Strategy and Management 5 (2009): 179–212.

[7] Mladen A. Vouk, Samuel F. Averitt, Patrick Dreher, Dennis H. Kekas, Andy Kurth, Marc I. Hoit, Paul Mugge, Aaron Peeler, Henry E. Schaffer, Eric D. Sills, Sarah Stein, John Streck, Josh Thompson and David Wright, "Constructing next generation academic cloud services," Int. J. of Cloud Computing, Vol. 2, Nos. 2/3, July 2013, pp 104-122.

[8] Arie Shoshani, Ilkay Altintas, Alok Choudhary, Terence Critchlow, Chandrika Kamath, Bertram Ludäscher, Jarek Nieplocha, Steve Parker, Rob Ross, Nagiza Samatova, Mladen Vouk , "Scientific Data Management Center Technologies for Accelerating Scientific Discoveries," presented at the SciDac 2007, Dec 2007, published in Journal of Physics, Conference Series 78, paper # 012068, 5 pages

[9] Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., … Schweitzer, R. (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. *Future Generation Computer Systems*, *36*, 400–417. doi:10.1016/j.future.2013.07.002

[10] I. Baldin, et al, "ExoGENI: A Multi-Domain Infrastructure-as-a-Service Testbed?," http://www.exogeni.net/download/exogeni-a-multi-domain-infrastructure-as-a-service-testbed-book-chapter/, unpublished.

[11] J. Brazelton, & G. A. Gorry (2003), "Creating a knowledge-sharing community: If you build it, will they come?," Commun. ACM, 46(2), 23–25. doi:10.1145/606272.606290.

[12] C. Warwick, M. Terras, P. Huntington, and N. Pappa, "If you build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data," Lit Linguist Computing (2008) 23 (1): 85-102, first published online November 20, 2007 doi:10.1093/llc/fqm045.

[13] Cutcher-Gershenfeld, J. et al., (2016). Build it, but will they come? a geoscience cyberinfrastructure baseline analysis. Data Science Journal. 15, p.8. DOI: http://doi.org/10.5334/dsj-2016-008.