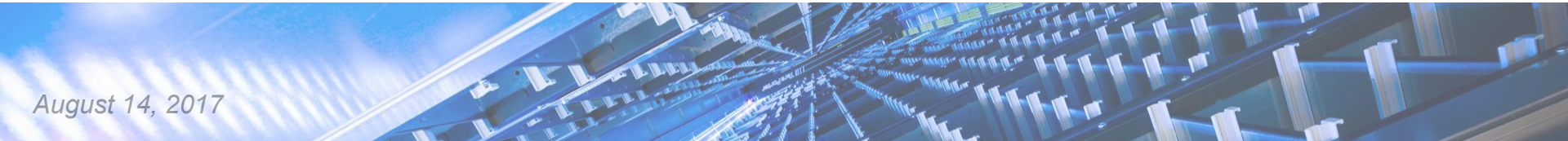
NATIONAL DATA SERVICE

NDS Labs Workbench + DataDNS

Craig Willis
willis8@illinois.edu
NCSA/NDS

NDS Team: Mike Lambert, Kenton McHenry, Kandace Turner, Kevin Coakley, Charles McKay, David Raila, Christine Kirkpatrick

Container Analysis Environments Workshop
August 14, 2017



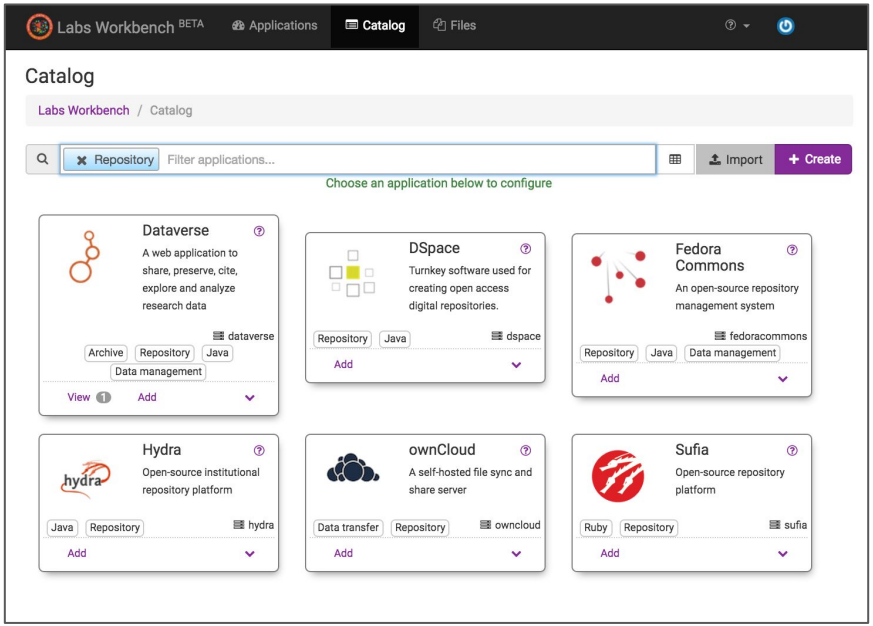
NDS Labs Workbench (beta)

https://www.workbench.nationaldataservice.org

- NDSC initiative started in January 2016
- Community-driven platform to share, discover, evaluate, develop, and test research data management and analysis tools
- Open platform -- community members recommend and/or contribute tools

NDS Labs Workbench

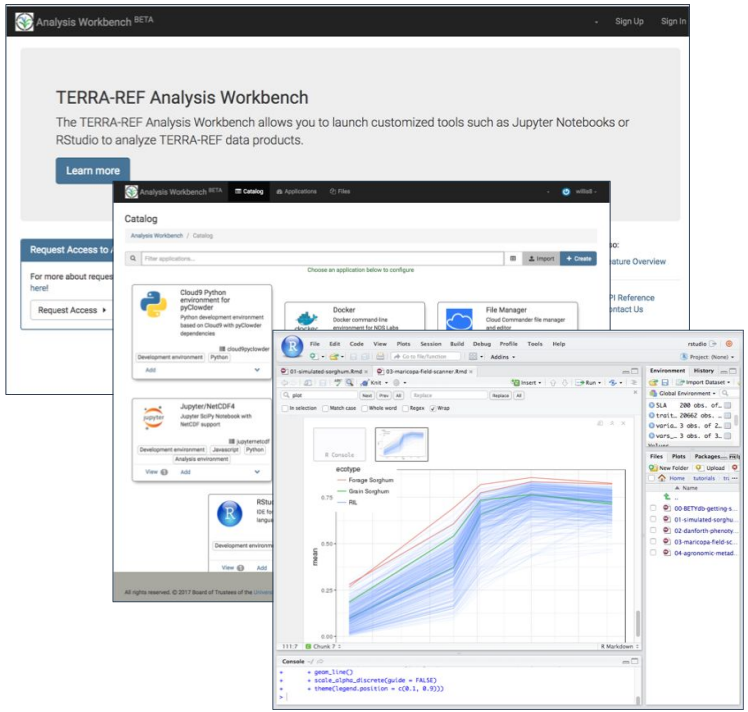
- Scalable platform (Kubernetes/OpenStack)
- Web-based, turn-key system for the deployment of data tools
- Includes catalog of community-contributed tools
- A personal test environment or sandbox
- Personal catalogs
- Console access to all applications



Driving user stories

1. A **developer** is assigned to a new data-related project and tasked with 1) learning, 2) evaluating or 3) extending data-related services.
2. An **instructor** wants to incorporate hands-on experience data management or analysis tools into a workshop or course.
3. **Students** or **workshop** participants need access to specialized software environments and real datasets (over hotel wifi, on tablets).
4. **Researchers** (and developers) associated with the same project need ongoing **remote access to data and software** required to access it.

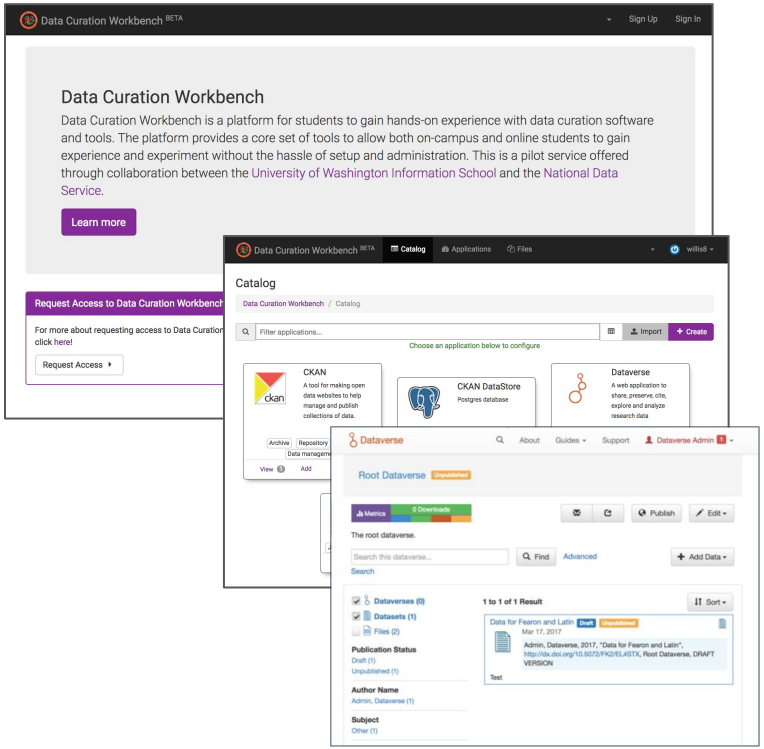
Use case: TERRA-REF Analysis Workbench



- Custom analysis and development environments with direct access to TERRA-REF data.
  - Jupyter, RStudio, Cloud9, Postgres, Xpra, NCO/NetCDF
- Used for workshop tutorials to provide consistent environments scaling to support 30+ participants.

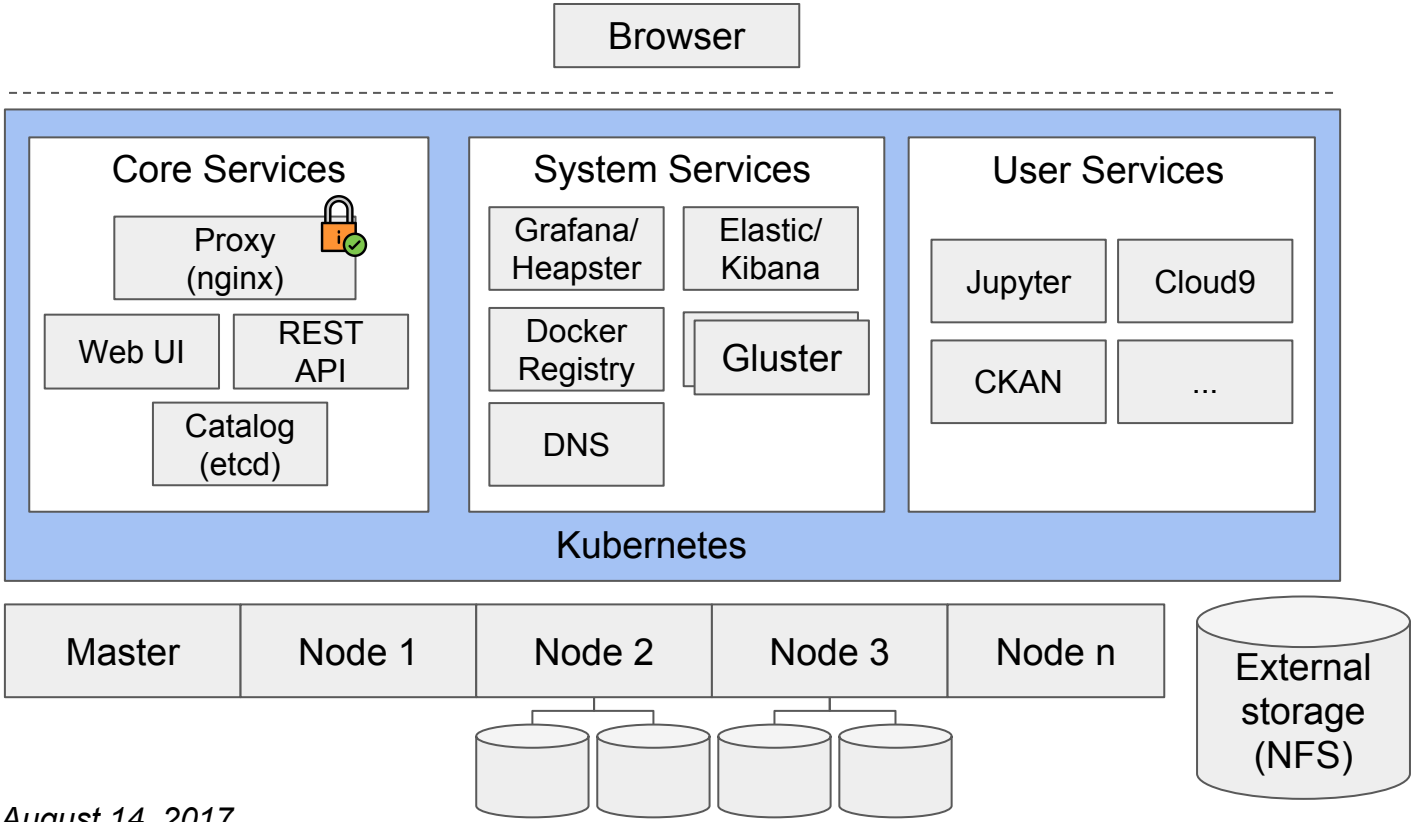
TERRAPHENOTYPING REFERENCE PLATFORM

Use case: Data Curation Education Workbench



- Ongoing pilot project led by University of Washington iSchool
- Customized instance of Labs Workbench hosted on SDSC Cloud
- Sandbox for students to get hands-on experience with data curation systems and tools including CKAN and Dataverse.

Labs Workbench Architecture Overview



Technologies:

- Gluster
- Kubernetes
- Docker
- CoreOS
- OpenStack

Why Docker and Kubernetes?

- Why Docker?
  - Installing/configuring many of these applications is not easy
  - Simplifies dependency management
  - Fast
  - Opportunity for preservation
- Why Kubernetes?
  - Scalable, reliable
  - Runs everywhere (OpenStack, AWS, GCE, Azure)
  - Resource constraints, scheduling, batch/job support
  - Many "free" services (NGINX proxy, LMA, cluster local DNS)

Other features

- Other features
  - Multiple-service applications (web + database + etc)
  - X-Windows application support via NoVNC/Xpra
  - Configurable service timeouts
  - Customizable UI
  - Account approval

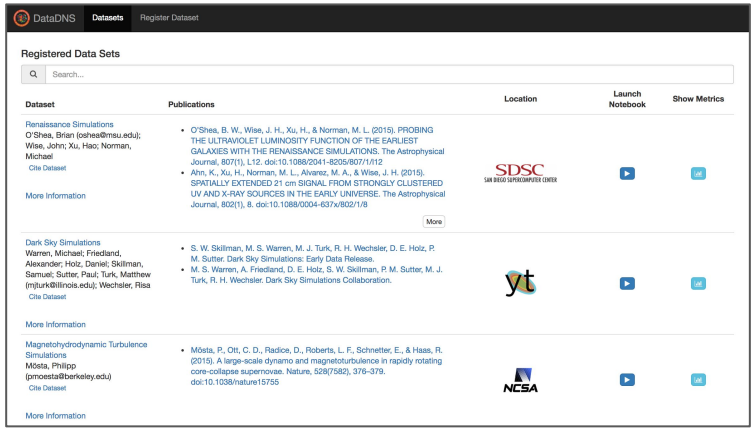
Challenges

- Storage
  - Shared storage -- particularly performance
  - Permissions, ownership
- Getting contributors to create their own images
- Single sign-on
- Versioning/management of images/dependencies
- Requires wildcard DNS/TLS
- Kubernetes -- complex and fast moving
  - Not easy to install on OpenStack

On the horizon:

- Single sign-on (CILogon)
- Integration with cloud/HPC systems
  - Head node replacement
  - Shared authentication, filesystems
- Simplified installation (simplicity + scalability)
- Support commercial cloud environments

DataDNS: Discovery, access, and in-place analysis for large scale data

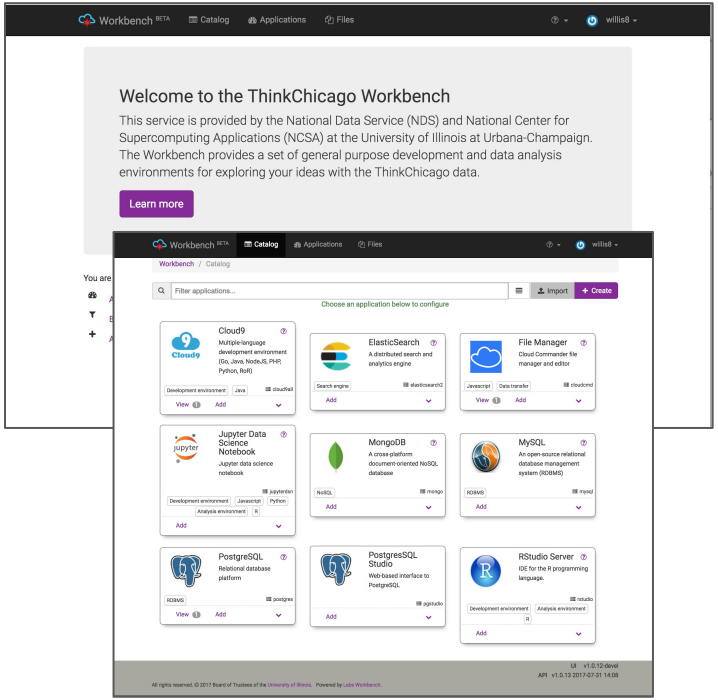


- Emerging vision from NDSC
- Data analysis and compute discovery engine.
- Connecting services that provide in-place access to data analysis environments.
- Defines common interface for heterogeneous environments.

Labs Workbench Platform

- Software packaged as containers
- Enables rapid deployment on scalable infrastructure
- To facilitate:
  - In-place data access
  - Research/development
  - Education/training

Use case: ThinkChicago Tech Challenge



- Hackathon supported by University of Illinois System and City of Chicago.
- 200 students (in teams of 10) present ideas centered around government data.
- Each team had access to common development tools and data via customized Labs Workbench.