Reflections on NDSC6

Craig

- It might be better if the first day was a general developer day instead of being so focused on the work coming out of NCSA. It would be a great opportunity for groups to share detailed information about their systems and potentially offer the opportunity for attendees to present what they're working on. The focus on Labs Workbench seems of limited value to the community.
- On a related note, updates on Labs Workbench during the main workshop should be a full presentation, not a demo. The 10-min demo assumes the audience knows what they're looking at, which they don't. If we do this in the future, we should slot more time and actually explain the back-story and build the case for what we're demoing. Joel's interruption was appreciated, but indicates a problem with the demo approach.
- I really like the idea of Labs Workbench as a training platform. This obviously meets a need for the Odum workshops. I spoke with an iSchool data curation instructor, and they would happily incorporate hands-on with popular data repository software (e.g., Dataverse, DSpace, Sufia). This seems like a natural next step for Labs Workbench.
- Docu-thon: I'd really like to see us pursue this before NDSC7 – compiling general architecture descriptions for NDS-related projects.
- Data-discovery: I heard it mentioned that the data discovery pilot from NDSC5 didn't have legs (and the interpretation that this is because it wasn't an NCSA-centric effort). I'd like to pursue this for NDSC7. I'm already involved with the RDA discovery group and have signed up for the bioCaddie Challenge.
- I think there's an opportunity to leverage our relationship with the iSchool more at NCSA – to put together one or more short seminars on research data management from the LIS perspective, intended for NDS developers. We could do this with the iSchool or RDS.
- The AWS presentation highlighted a few things: SCs should be similarly storing public datasets of interest from regional institutions. Also, AWS apparently offers the ability to launch an NDS Share-like CKAN instance:
    - https://aws.amazon.com/government-education/open-data/
- Things we should be aware of
    - OSF SHARE https://share.osf.io/
    - RDA PID group
    - DataQ http://researchdataq.org/
    - ESIP http://www.esipfed.org/
    - Code4lib http://code4lib.org/
    - bioCaddie challenge https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge

---

David's observations and some ideas about paths forward

Workbench:

The most notable interest in the workbench was as a platform for training and education, with strong interest from the I-School constituents as a turn-key system for educating students in hands-on *data-lab* environments. This scenario can easily be expanded to professional training, and in-fact the workbench has been used for small-scale targets workshop education. To implement and support these activities would require the following features, modifications, and additions:

- next-level data support - more realistic in real-world settings but still small-scale
- better tool catalog support – fast self-service creation and update of curated tools to allow very small, activity-specific mini catalogs that could be easily changed by instructors week-by-week
- ability to run multiple independent workbench instances simultaneously
- support for sites to deploy and operate their own workbench infrastructures – we can't be the only provider
- a configurable system for user management, authentication, and access-control by educators that is self-service and adaptable to their on-premise systems (campus course logins, etc.)
- addition of some analytics/usage gathering that feeds back to NDSlabs so we can understand the who/how/what usage of the platform

Implementing this would further the capabilities of the workbench in a real-world, real-data scenario in addition to enabling a useful service in support of data educators. Based on the need and enthusiasm expressed at NDSC, the scope of the work, and the benefits to a targeted community – this could be the centerpiece of a proposal between NDSL and ischools around big data education.

NDS Federation Services:

To support NDS services based on the search/share/publish/re-use model where data, data-tools, and documents are first-class objects across many sites and are orchestrated by a set of NDS federated services, NDSLabs should be working towards providing turn-key *Federation as a Service for Data Applications* - a packaged set of services that can be deployed quickly and easily at NDS sites (on site resources, with site management) that enable NDS capabilities on-site and automatically inter-operate with the larger federation to provide data search and access, compute-near-data, ready-to-use data tools, and ease-of-use. The features of such a system include

- Site self-service - a site can self-deploy and self-manage NDS services
- Loosly coupled – sites and site-services can come/go at will without intervention
- Extend not replace – NDS services work *with* existing infrastructure, leveraging their capabilities and making resources available across the federation – including compute, data, and authentication/access
- Self-monitoring – monitoring is automated for the site, and across the federation
- Auto-remediation – automatic recovery from resource unavailability, network errors, application crashes, etc. without human intervention
- Scalability – sites can easily add more compute or storage as demand requires
- Simple data registration – sites can easily publish data-sets making them usable and searchable
- Policy-based controls – data and compute access can be restricted using simple policy controls
- Data and compute mobility – data can be copied to compute, compute can be deployed near-data, on-demand

To support these features, NDSLabs can develop and provide:

- Turn-key site infrastructure – container-based scalable cluster-OS and installation tools to deploy on existing IaaS (OpenStack), or bare systems.
- Local service adapters - a set of containerized services that adapt and integrate existing site compute and data resources, and provide local control, monitoring, analytics
- Federation services – a set of containerized services that integrate the site with the federation – provides search, data transfer, and local compute

- Interfaces for 3rd party NDS applications – enables development and deployment of new services onto the NDS fabric. Supports the NDS core services and approved 3rd party services

- Some high-level ideas for development Paths to support NDS Federation -
    - Build federated multi-site cluster SDSC/TACC/NCSA and evaluate
        - In cooperation with site-admins using refined deploy tools
        - Not clear if federated clusters in-orchestration or multiple independent clusters with NDSL federation services are better, or some mix.
    - Build/test core per-site with federation integration services:
        - ID/auth/access – dex or keycloak?
    - Data *interface* service: site-run auto-federating data/tool registration
        - Federated extensions for search, use, publish, compute...
    - Determine a re-usable policy architecture for federated services: needs goals in federation, points of policy application, mechanisms for policy enforcement at those points, policy decision engine.
    - Implement site/federated software-defined data-integration services: – interfaces and mechanisms for adapting to local data in low-level: (filesystem, NFS, GPFS,…) and high-level (swift, HFS, …)
        - Perhaps with data movement (globus, zfs-send, …) and dataset caching for multi-site copy.
    - Implement compute near-data service – integrated with local data-set registration and auth/access policy integration
    - Implement turn-key data-publishing service for sites: turnkey full-stack of re-usable site-side publishing services/tools enable deploy on local provisioned infrastructure.