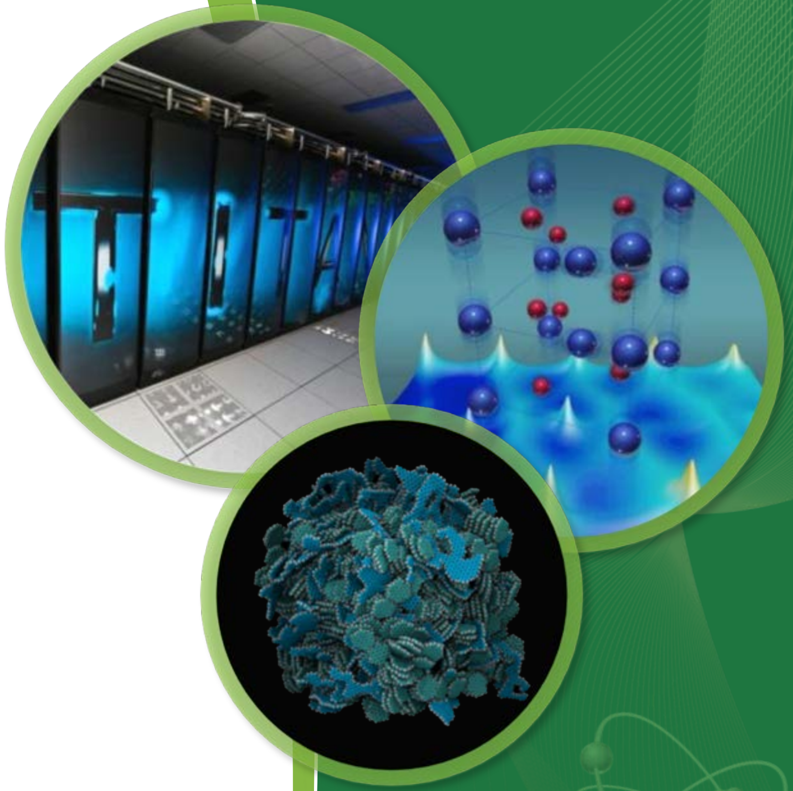


OLCF Research Data Initiatives

Exploring requirements and needs

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Motivation

Motivation - Exascale Crosscut Report

- ASCR and OLCF are guided by science needs that increasingly a community effort.
- We are generating data at unprecedented scales, both from observations and simulations.
    - From 10^1 PB to 10^2 PB in the next few years
- Users expect multi-year commitments from facilities, like OLCF.
- Scientific needs include real-time modeling and simulations during experiments, requiring exascale computational resources.
- Exascale ecosystems will include high-end data capabilities.
- Increasing complexity in everything - computing, data, workflows and management.

A few common themes

- Large-scale data analysis, long-term data storage, and community reuse.
- Integrated experimental and simulation workflows.
- Develop effective data management solutions and best practices.
- Support for data life cycle management activities, including archiving and curations.
- Sharing of data and provisioning remote access to data.
- Facilitate efficient and fast data transfer mechanisms.
- Improve IO support for simulation and data analysis at scale.
- Facilitate reuse of techniques.
- Support community standards.

Current status

- Increasing user needs and requests to retain data at OLCF for continued analysis and future projects.
- Requests for data services and tools to exploit data for science deliverables.
- The few data-only projects are *ad hoc*
  - Evaluated on their own merit and requirements.
  - Resource utilization varies vastly.
  - We do not (yet) have consistent policies for long-term storage, publication and data management.
  - We lack the tools and services to support longer term projects.

2017 OLCF User Survey

- Increased need for tools for data analysis [4% => 18%].
- Long-term data retention is extremely important [69%].
- Need data curation [47%].
- Access to data via portals [43% INCITE PIs].
- Support for jupyter notebooks [26%].
- Satisfaction with our offerings [74%].
- Other
  - Improved awareness of visualization and analytics tools.
  - Re-evaluation of purge policy: advance notifications.
  - Remote visualization capabilities.
  - Issues with HPSS and lustre.

Tentative Approach

Path to Formalizing Typical User Data Needs

- We are formulating programmatic focus areas, categorized as **Types**.
- **Type 1** data repository program for ***"data-only"*** projects.
  - Large volume of data challenging to move back to host institutions.
  - Need more time to complete analysis and publish.
  - Opportunity for a follow-on project.
- **Type 2** data services program for user communities.
  - Data collections that benefit the broader domain science community.
  - Forcing and parameter data; validation data; reference data.
  - Data publication and utilize DOI as a service.
- **Type 3** computational and data science end station program.
  - Goal is to enable discovery science.
  - Enable analytics at largest scales.

Type 1 Project Ideas

- Currently supported an *ad hoc* basis.
- Computational requirements are none to low.
- Storage requirements moderate to high.
- Data service requirements are minimal.
- Require efficient data transfer mechanism.
- Project duration variable.
- Some data may need to persist beyond project duration.
  - May be useful for INCITE and/or Type 3 projects in the future.
- Need to prioritize existing resources.
- New allocation unit for storage (say Ebyte-years).

Type 2 Project Ideas

- Purpose: serve distributed project team and/or domain user communities.
- Data collections likely to include input data for simulations, forcing / parameter data, validation data and other reference data.
- Computational requirements minimal.
- Storage needs low to high.
- Relatively longer retention period.
- Data services include portals, databases, containers, data transformation, data fusion, data catalogs & publication (DOI services), data transfer and other TBD.
- Projects need a well-defined data curation & lifecycle management process.
- Workflows need to be initiated via NCCS-Open.

Type 3 Project Ideas

- Projects may leverage existing (shared) collections from Type 1 or Type 2 projects.
- Data collections can be analogous to beam lines at experimental facilities offering opportunities for discovery science via data intensive computing.
- Enabling domain-dependent analytics (e.g., machine/deep learning/AI)
- Computational needs low - high, possibly computing at scale.
- Project duration relatively shorter (say < 1 year).
- Some projects may be preparation for future INCITE competition.
- Possibly transition to Type 1 or Type 2 upon completion.

Constraints and Considerations

Facility resource management / operational / policy considerations.

- Need to leverage already available resources.
- Disk, HPSS and other services are finite.
- Data duplication and movement is expensive.
- Need to understand access patterns to plan for growth.
- Need the ability to estimate and forecast capacity and bandwidth - near-term as well as the future.
- Existing resources need to be rationalized.

Proposal elements

- Scientific impacts
  - DOE SC mission: *“deliver scientific discoveries … to transform our understanding of nature …”*
- Ownership of data and access considerations.
- Target community and consumers and mode of usage.
- File size distribution, type, volume, etc.
- Metadata and provenance.
- Software and tools.
- Availability (disk, tape) and access requirements.
- Data lifecycle management plan.
- Disposition of data upon completion.
- OLCF acknowledgement.

Example guidance for Type 1 projects (preliminary)

- Expected scientific outcomes and impact.
- Analysis plans and requirements (software, tools & libraries, etc.).
- Duration of award.
- Source of data (in not at OLCF) and ingress plans.
- Resource utilization:
  - HPSS & disk: volume, file size distribution, growth rate, retention needs (scratch/project/tape).
  - Data transfer
  - Analysis: allocation, typical job size, wallclock, etc.
- Allocation & utilization currency: EB-years (HPSS), PB-years (online)
- Engagement with OLCF liaisons.
- Reporting requirements.
- Proposal review via RUC.

Summary of Data Project Types for Discussion

Requirements & Characteristics	Type 1	Type 2	Type 3
Project duration	< 1 Y	> 3 Y	< 3 Y
Renewable?	Y (rarely) | N (mostly)	Y (mostly)	Y (sometimes) | N (mostly)
Storage volume	Moderate - High (mostly)	Low - High	Med - High
Production velocity	Static	Static - Low	Low - High
Online storage duration & persistence	< 1 Y	> 3 Y	< 3 Y
Persistence (archive)	N	Y	N
Compute	None - Low	None - Low	Med - High

Requirements & Characteristics	Type 1	Type 2	Type 3
Compute	None - Low	None - Low	Med - High
Workflow complexity	Low	Low - High	Med - High
NCCS Open	N	Y	Y | N
CADES	N	Y	Y | N
Period of Performance	< 1 Y	> 3 Y	< 1 Y
Primary beneficiary	PI	Community	PI
Risks	Low	Moderate	Unknown
Implementation	< 6 months (Q2)	FY19	Unknown or FY19
Training	None	Low - Med	Med - High

Acknowledgement

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

Contact: Valentine Anantharaj <vga@ornl.gov>

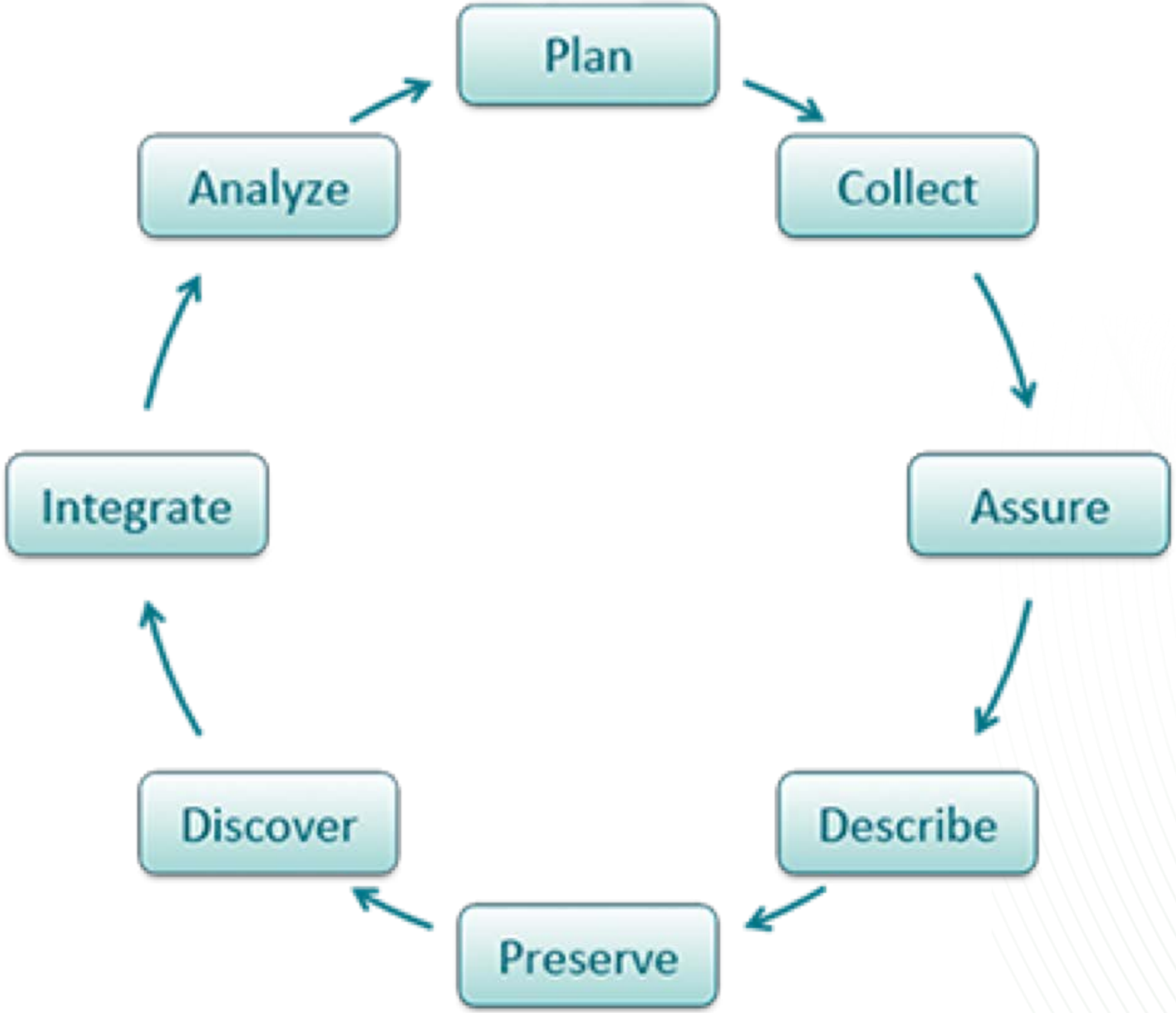
Additional Topics

Research Data Record Essential Elements

- Software readiness
- Metadata
- Documentation
- Validation
- Access
- Applications and utility

Bates et al., 2016

Data Lifecycle



Source: DataOne

Data Management Best Practices