Panel Discussion presentation: "Data-Intensive Science with High Performance Computing Leveraging"

John W. Cobb

Oak Ridge National Laboratory
Data-Intensive Science with High Performance Computing leveraging

Presented to
Fifth Annual
University of Massachusetts and New England Area Librarian
e-Science Symposium Afternoon Panel

John W. Cobb, Ph.D.
Physicist
Computer Science & Mathematics Division

Shrewsbury, Massachusetts
April 3, 2013
Acknowledgements

- DataONE project (PI Michener, U. New Mexico)
- Oak Ridge National Laboratory and the Oak Ridge Leadership Computing Facility
- Cornell Lab of Ornithology eBrid project and S. Kelling, D. Fink, K. Webb, T. Damalou, (Cornell)
- Collaborators: M. Jones (UCSB) C. Tenopir (UTK), S. Allard (UTK), B. Wilson (ORNL/UTK), D. Vieglais (Kansas)
DataONE Community
Outline

• Data Begets Science
• The data lifecycle – the workflow of data driven science
• Data at Scale
• HPC at Scale
• Pathfinder exemplar: eBird occurrence maps
• Data management challenges
• DataONE project
• Dryad
• Role of libraries as data repositories
• DMPTool
• Open data movement
Data Gives Birth to Scientific Revolutions

- Kepler’s laws were divined by careful examination of Brahe’s recorded observations.
- Leeuwenhoek’s founding of microbiology was triggered by observations with newly developed microscope.
The data lifecycle: the workflow of science

The conduct of science is collaborative and multidisciplinary

Collect → Analyze → Integrate → Discover → Preserve → Deposit → Assure → Describe → Collect

Refined DataONE internal view
## User Matrix (DataONE)

Different team members care about different things

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Data Service</th>
<th>Investigator Toolkit</th>
<th>Data Management Planning</th>
<th>Best Practices</th>
<th>Tools Database</th>
<th>Training</th>
<th>Curricula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Librarians</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecological Modeler</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resource Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Can we share data along the data lifecycle?

**Demographics**

**Discipline**
- medicine: 2%
- ecology: 18%
- biology: 14%
- atmospheric science: 4%
- environment sciences: 18%
- physical sciences: 12%
- computer science/engineering: 9%
- social sciences: 16%
- other: 7%

**Work Sector**
- academic: 80%
- government: 13%
- non-profit: 3%
- commercial: 2%
- other: 2%


n=1317

n=1315
Many are interested in sharing data

- Willing to share data across a broad group of researchers: 81%
- Willing to place at least some of my data into a central data repository with no restrictions: 78%
- Appropriate to create new datasets from shared data: 76%
- Willing to place all of my data into a central data repository with no restrictions: 41%
What standard do you currently use?

![Metadata language chart]

- DIF: 12
- DwC: 21
- DC: 26
- EML: 95
- FGDC: 95
- Open GIS: 96
- ISO: 97
- My Lab: 266
- none: 676
Answer: Yes!

But: There is a gap between desire and practice.

This indicates an opportunity to improve practice and improve science outcomes

“The spirit is willing but the flesh is weak”
How big is big data?

• Possible answers:
  – the largest of all datasets ever created (>10 PB)
  – The largest of all datasets ever created in each discipline
  – larger than we are comfortable managing
  – larger than what we dealt with last week/year/decade
How big is big data?

• Possible answers:
  – the largest of all datasets ever created (>10 PB)
  – The largest of all datasets ever created in each discipline
  – larger than we are comfortable managing
  – larger than what we dealt with last week/year/decade

• But larger question: what is the measure of data size?
Data Ecosystem

HPC and Data Center

Physical Storage

Replication

Metadata Management

Workflow Provenance

Semantics

Science Leverage

L&IS

I/O Rate
Where are the opportunities?

- Integrating storage management and information management
- Integrating data from different data activities

“Building the Knowledge Pyramid”

90:10 → 10:90

Where are the opportunities?

- Increasing Process Knowledge
- Decreasing Spatial Coverage

Intensive science sites and experiments
Extensive science sites
Volunteer & education networks
Remote sensing

Adapted from CENR-OSTP
HPC at scale – example Titan at OLCF

• Physical plant challenges:
  – Size: 40,000 sq-ft (2 floors)
  – Power: 10’s of MW
  – Cooling: dual loops chilled water
  – Raised floor high-load capacity (36”, 250 lbs/sq-ft)
HPC at scale – example Titan at OLCF

- Named Titan
- 27 Petaflops, 710 TB memory
- Spider storage > 10 PB, 250 GB/s
- 8972 GPU-enabled nodes (Kepler) in 200 cabinets
- Each node contains: One AMD 16-core intelagos CPU, one Nvidia K20x Kepler, 32 GB memory
- Note: NVIDIA offers K20x for desktop
Data and the Long Tail of Science

- As data gets larger, the data tail is now quantifiable: *flocks of black swans*
- Extraordinary events are often the most interesting
  - “500 year storms”
  - Best candidate materials (second place is first loser)
  - Very non-uniform utility functions.
- Conclusion: applying large data analysis can create new breakthroughs
eBird pilot project exploration and visualization

Diverse bird observations and environmental data from 300,000 locations in the US integrated and analyzed using High Performance Computing Resources

Model results
Occurrence of Indigo Bunting (2008)

- Examine patterns of migration
- Infer how climate change may affect bird migration

Spatio-Temporal Exploratory Model identifies factors affecting patterns of migration
Secretary Salazar on Birds (May 3, 2011):

“The State of the Birds report is a measurable indicator of how well we are fulfilling our shared role as stewards of our nation’s public lands and waters.”
HPC centers and data management

- Often HPC focused – cycles (and storage)
- Data and information management may be a foreign culture
- HPC can enable extreme scalability: “What would you do if you had unlimited computing/storage/bandwidth?”
- Bottlenecks:
  - Data management issues
  - Metadata creation and harmonization
  - Data preservation
  - Items not scaling with Moore’s law: metadata, human effort
Data deluge and interoperability
“the flood of increasingly heterogeneous data”

• Data are heterogeneous
  – Syntax
    • (format)
  – Schema
    • (model)
  – Semantics
    • (meaning)

By hand is time-consuming and brittle

Jones et al. 2007
Myriad Metadata Standards

For instance: Metadata Crosswalks
Poor data practice
“data entropy”

In what sense is modern science reproducible?

Time of publication

Specific details

General details

Retirement or career change

Accident

Death

(Michener et al. 1997)
DataONE project (movie with sound)

Depositing Data with DataONE

http://vimeo.com/36383735
DataONE Component Interdependency

**Scientists:**

**Receive:** Access to more data sources and tools

**Provide:** Scientific progress and acknowledgment

**DataONE:**

**Receives:** MN and scientist appreciation, access to MN data

**Provides:** “Glue” services to enable interoperability, communities of best practice, standard interfaces

**Funders:**

**Receive:** More efficient science output, chances for breakthrough advances

**Provide:** Resources to facilitate science

**Member Nodes:**

**Receive:** Additional users, replication, communities of best practice, appreciation

**Provide:** Access to data collections, service interfaces
Current Operational Member Nodes

- Released production CI 10 months ago
- Today: 13 production Member Nodes
- 300,000 Data objects represented

- Near-term 15 more candidates
The Investigator Toolkit

- Developer, end-user tools
- Creation, search, retrieval, management
- Plugins, extensions for analysis tools
Identify objects

**Goal:** Uniquely identify data or metadata objects

- Support the several identifier types widely used
- Identifiers assigned by Member Nodes
- Uniqueness ensured by Coordinating Nodes
- Resolution through Coordinating Nodes

[Images of identifiers: UC3EZID, doi, GUID, LSID, PURL, Handle System]
Provide Credit for Data Publication

- Data citation standards and courtesy customs
- Needs to metrics – how often cited
- Socio-cultural change: include data citations in promotion and tenure
- DataONE needs to nurture Member Node needs not work against them
Identify people: federated identity

- Identity provider selected by the user
- Member nodes define access rules
- Rules propagated by Coordinating Nodes
- Identity and access control consistent across entire infrastructure
- (note similarity with Globus Online approach)
Support for Entire Data Lifecycle

Plan

Collect

Assure

Describe

Preserve

Discover

Integrate

Analyze

MATLAB

DMPTOOL

Excel

Integrate

Kepler

MATLAB

DMPTOOL

Excel
February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather models used by insurance companies and farmers benefit from data generated by federally funded research.

Under the memorandum today (pdf) to Federal agencies that direct those with more than $100 million in research and development expenditures to develop plans to make the results of federally funded research publicly available free of charge within 12 months after original publication.

The memorandum requires that agencies start to address the need to improve upon the management and sharing of scientific data produced with Federal funding.
Building global communities of practice: ... creating long-lived CI enterprises,

- Broad, active community engagement
  - Involvement of library and science educators engaging new generations of students in best practices
  - Existing outreach and education programs
- Transparent, participatory governance
- Adoption/creation of innovative and sustainable business and organizational models
Libraries and museums: value

• **As Member Nodes:**
  – Facilitate the teaching and research mission of institution
  – Build data collections for the 21st century

• **In support of Data Librarians:**
  – Provide access to data management plans
  – Provide best practices for faculty and students
  – Cyberinfrastructure supporting the data lifecycle
Data Management Planning Tool

- Create ready-to-use data management plans for specific funding agencies
- Meet funder requirements for data management plans
- Get step-by-step instructions and guidance for your data management plan as you build it
- Learn about resources and services available at your institution to help fulfill the data management requirements of your grant
- Released: Oct. 2011
- Support for NIH requirements added 2/22/2012
- Other similar efforts now also underway at institutional levels or with other entities.

https://dmp.cdlib.org/
Plug: DMPTool next rev upcoming
Save The Date

DataONE Users Group Meeting
July 7-8th 2013, Chapel Hill, NC

Co-located with ESIP Federation Meeting.
Question & Discussion

John W. Cobb
865.576.5439
cobbjw@ornl.gov
http://www.slideshare.net/johnwcobb/cobb-u-massnealesciencev06