Big Data in the Earth Observing System Data and Information System

Chris Lynnes (System Architect, Code 586)
Katie Baynes (System Architect, Code 586)
Mark McInerney (Deputy Project Manager, Code 423)
NASA/GSFC, ESDIS Project
Earth Observing System Data and Information System (EOSDIS)

data downlink

capture and clean

distribute

subset

archive

process

Research

Applications

Education

Users
Earth Observing System Data and Information System (EOSDIS)

data downlink

capture and clean

distribute

subset

archive

process

Research

Applications

Education

Users
1. Cloud prototypes are underway to tackle the *Volume* challenge of Big Data...

2. ...But advances in computer hardware or cloud won’t help (much) with *Variety*

3. Interoperability standards, conventions, and community engagement are the key to addressing *Variety*
V is for...

...Volume

![Graph showing Archive and Distribution Volume over time in petabytes from 2000 to 2016.](image)
## Big Data Indicators

### EOSDIS FY2015 Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Data Products</td>
<td>9,462</td>
</tr>
<tr>
<td>Distinct Users of EOSDIS Data and Services</td>
<td>2.6 M</td>
</tr>
<tr>
<td>Average Daily Archive Growth</td>
<td>16 TB/day</td>
</tr>
<tr>
<td>Total Archive Volume (as of Sept. 30, 2015)</td>
<td>14.6 PB</td>
</tr>
<tr>
<td>End User Distribution Products</td>
<td>1.42 B</td>
</tr>
<tr>
<td>End User Average Daily Distribution Volume</td>
<td>32.1 TB/day</td>
</tr>
</tbody>
</table>
Archive Cloud Prototypes

Benefits from Archive in the Cloud

- Cost savings for storage of Big Data?
- Avoid data downloading and local data mgmt

- Alaska Satellite Facility Web Object Storage prototype
  - Distribute Sentinel radar data from Amazon storage
- Global Imagery Browse Service in the Cloud
- Ingest and Archive management prototype
Cloud Analytics Prototypes

Benefits from Cloud Analytics

- Analyze data at scale
- Analyze datasets together easily
- Avoid data downloading and local mgmt

Analysis support toolbox to attract users to cloud analytics

- Community open source tools
- DAAC-developed tools
- Cloud analytics examples and recipes
- Initial cross-DAAC proof of concept in progress based on Python + Jupyter Hub
Terra Incognita

1. Vendor Lock-in
2. Future storage costs
3. Uncapped egress costs
4. Security Restrictions
5. Network trust
V is for...

...Variety

Distinct Science Products Distributed

- 1998
- 2000
- 2002
- 2004
- 2006
- 2008
- 2010
- 2012
- 2014
- 2016

Graph showing the increase in distinct science products distributed from 1998 to 2016.
## Instrument Variety

<table>
<thead>
<tr>
<th>Spectrometer</th>
<th>Microwave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiometer</td>
<td>Infrared</td>
</tr>
<tr>
<td>Sounder</td>
<td>Visible</td>
</tr>
<tr>
<td>Interferometer</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>Polarimeter</td>
<td></td>
</tr>
<tr>
<td>GPS</td>
<td></td>
</tr>
<tr>
<td>Radar</td>
<td></td>
</tr>
<tr>
<td>Lidar</td>
<td></td>
</tr>
<tr>
<td>Accelerometer</td>
<td></td>
</tr>
<tr>
<td>Limb Scanner</td>
<td>Microwave</td>
</tr>
<tr>
<td>Imager</td>
<td>Infrared</td>
</tr>
<tr>
<td>Microwave</td>
<td>Visible</td>
</tr>
<tr>
<td>Synthetic Aperture</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>Precipitation</td>
<td></td>
</tr>
<tr>
<td>Scatterometer</td>
<td></td>
</tr>
<tr>
<td>Profiler</td>
<td></td>
</tr>
<tr>
<td>Doppler</td>
<td></td>
</tr>
<tr>
<td>Altimeter</td>
<td></td>
</tr>
</tbody>
</table>
Satellite Instrument “Footprints”

Example Imaging Footprint

![Example Imaging Footprint](image1)

Example LIDAR footprint

![Example LIDAR footprint](image2)

Example Limb Scanning Footprint

![Example Limb Scanning Footprint](image3)

Microwave Limb Scanner (from Algorithm Theoretical Basis Document, Livesey and Wu, 1999)
Aircraft and In Situ
Same Instrument, Different Satellite

Aqua

Terra

Aqua - Terra
Same Instrument+Satellite, Different Algorithm

MODIS on Aqua
Aerosol Optical Depth

Dark Target Algorithm

Deep Blue Algorithm
Processing Levels

AIRS data for 2011-08-11

Level 1B
Calibrated radiance at a pixel

Level 2
Carbon monoxide for one scene

Level 3
Global carbon monoxide for one night
Projections

SurfAirTemp_D

SurfAirTemp_D ()

Data Min = 212.125, Max = 306.688, Mean = 285.636
Time Aggregation

Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer

Daily

Monthly
Spatial Aggregation

SeaWiFS Deep-Blue Aerosol Optical Depth 2006-10-06

1.0 Degree Resolution

0.5 Degree Resolution
Data Formats

● Self-Describing API-Based
  ■ Hierarchical Data Format (HDF)
  ■ network Common Data Form (netCDF)

● Additional conventions
  ■ HDF-EOS
  ■ Climate-Forecast coordinates

● Other Standards
  ■ Gridded Binary (GRIB)
  ■ ICARTT (Airborne)

● Binary
● ASCII
Solutions to the Variety Problem

1. Interoperable discipline-focused DAACs
2. Common Metadata Repository
3. OPeNDAP* data services
4. Community engagement

*Open-source Project for a Network Data Access Protocol
Discipline-Focused Distributed Active Archive Centers (DAACs)

- Alaska Satellite Facility (Synthetic Aperture Radar)
- Land Processes DAAC
- Socio-economic Data Archive Center
- Ocean Biology DAAC
- National Snow and Ice Data Center DAAC
- Crustal Dynamics Data and Information System
- Level 1 and Atmosphere Archive and Distribution System (MODIS)
- Oak Ridge National Laboratories DAAC (Biogeochemistry)
- Goddard Earth Sciences Data and Information Service Center
- Physical Oceanography DAAC
- Global Hydrology Resource Center
- Atmospheric Sciences Data Center
Different DAACs have different “-Spheres of Influence”

<table>
<thead>
<tr>
<th>DAAC</th>
<th>Atmo</th>
<th>Hydro</th>
<th>Cryo</th>
<th>Litho</th>
<th>Bio</th>
<th>Anthropo</th>
<th>Sub-Specialty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska Satellite Facility</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>SAR</td>
</tr>
<tr>
<td>Atm. Sciences Data Center</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crustal Dynamics Data Info Sys</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>Space geodesy</td>
</tr>
<tr>
<td>Global Hydrology Resource Ctr</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Weather events</td>
</tr>
<tr>
<td>Goddard Earth Sciences DISC</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land Processes DAAC</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 and Atm Archive &amp; Dist Sys</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MODIS, VIIRS</td>
</tr>
<tr>
<td>Nat. Snow Ice Data Ctr DAAC</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oak Ridge Nat Lab DAAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>Field experiments</td>
</tr>
<tr>
<td>Ocean Biology DAAC</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Oceanography DAAC</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socioeconomic Data Arch Ctr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
The Common Metadata Repository presents a consistent catalog for discovery of data from multiple DAACs.

One Metadata System to rule them all,
One Metadata System to find them,
One Metadata System to bring them all
And in cyberspace bind them.
OPeNDAP

- Open-source Project for a Network Data Access Protocol
- High-performance network access protocol for complex science data
- Well-supported in Earth science community tools
  - Free: Panoply, IDV, McIDAS-V, nco,…
  - Commercial: ArcGIS, Matlab, IDL,…
OPeNDAP* access to data smoothes out format heterogeneity and supports subsetting

*Although the Hyrax implementation is shown, other OPeNDAP servers such as GrADS Data Server and THREDDS Data Server have similar capabilities but different architectures.
Data transformation options of several kinds can help with Variety and Volume

Data transformation applies fundamental changes and conversions to attributes of the original data to suit the application requirements of end-users.

Courtesy of B. Ramachandran, MODAPS/LAADS
Big Earth Data Initiative (BEDI)

- OSTP-driven multi-agency effort
- Focus on datasets in Societal Benefit Areas
- Several interoperability aspects...
BEDI in EOSDIS

- Improve dataset consistency across EOSDIS
  - Metadata in Common Metadata Repository
  - Data in OPeNDAP
- Improve machine access to EOSDIS
  - Developers’ portal
    - How To Access Common Metadata Repository
    - How to Access OPeNDAP-served Data
  - OPeNDAP performance
  - OPeNDAP use with Cloud storage
Community Engagement on Big Data

● Earth Science Information Partners (ESIP)
  ○ Variety: Clusters on Discovery, Information Quality
  ○ Volume: Clusters on Earth Science Data Analytics and Cloud Computing

● Earth Science Data Systems Working Groups
  ○ Formed of DAACs, ACCESS and MEaSUREs award winners
  ○ Variety: Working Groups on Dataset Interoperability, Search Relevancy
  ○ Volume: OPeNDAP Best Practices, Cloud Computing

● User Needs efforts
  ○ DAAC User Working Groups
  ○ American Customer Satisfaction Index survey
  ○ EOSDIS User Needs Analysis group
Big-Data-Community Engagement

- Big Data Theme for both ESIP 2016 Meetings
- Co-Convening [AGU 2016 session on Big Data Analytics](https://agu.org/)
- Program committee for [IEEE Workshop on Big Data in Earth and Planetary Sciences](https://ieeexplore.ieee.org/)
- ESA’s Big Data from Space (BiDS) workshops
  - “Improving Earth Science Data Discoverability And Use Through Metadata Relationship Graphs, Virtual Collections, And Search Relevancy”
User Needs from Community Sources

Sample Size

Shallow

Large

Small

ACSI Survey Scores
ACSI Survey Comments
Applications Workshops
Advisory Groups (ASAC, UWG)
Webinar Feedback
Help Tickets

Depth & Detail of Insight

Shallow

Deep

33
Take Home Message

1. Cloud prototypes are underway to tackle the Volume challenge of Big Data...

2. But advances in computer hardware or cloud won’t help (much) with Variety

3. Standards, conventions, and community engagement are the key to addressing Variety
Backup Slides
OPeNDAP Enhancements from the Big Earth Data Initiative

- More OPeNDAP for EOSDIS data
- More aggregation along time for data in OPeNDAP
  - Improved performance for aggregation in Hyrax