Advanced Analytics and Big Earth Data

Christopher Lynnes*, NASA/Goddard Space Flight Center

*U.S. Civil Servant
NASA’s Earth Science Data Systems Program

- Actively manages NASA’s Earth science data as a national asset: satellite, airborne, and in situ
- Develops capabilities to support rigorous science research
- Processes instrument data to create high quality long-term Earth science data records.

Land Surface Temperature on a base of Corrected Reflectance from Aqua Moderate Resolution Imaging Spectroradiometer, 16 Jun 2018
Earth Observing System Data and Information System

EOSDIS

Capture and clean

Distribute

Transform*

Archive

Process

Commercial

Research

Applications

Education

*Subset, reformat, reproject
Emerging Trend: Big Data

Data Volume

Small
- Data
- Inspection

Big
- Model-based Analysis
  - Physical
  - Statistical
  - Machine Learning
Emerging Trend: Machine Learning

1. Classification
2. Data Mining for Features
3. Segmentation
4. Data Fusion
5. Super-Resolution / Downscaling
6. Estimation of Geophysical Parameters
7. Data Quality Assessment / Improvement

Impact: Data Enhancement
Emerging Trend: Analysis Software Ecosystems

- Python + Numpy + SciPy + Jupyter + pandas + scikit-learn + dask + …
- Docker Containers
- Microservices
- Anaconda/conda
- github

*Impacts:*

*Lower barrier to entry -> More users -> More collaboration*
Emerging Trend: Archiving Data in the Cloud

- Web Object Store (e.g., Simple Scalable Storage)
- Accessing data via web protocols, not file system
  - Fine-grained access via range-get
  - Abstracted access via Application Program Interfaces

Impact:

*Big Data next to Big Compute*

Re-evaluation of Storage Format + Structure
Emerging Trend: Cloud Computing for Analysis

- Elasticity
  - big analysis jobs
  - short timeframes
- Hardware Choice
  - Virtual Machines: input/output-, storage-, compute-optimized
  - Special Purpose: e.g., Graphical Processing Units, Quantum
- Software-as-a-Service
- Data-proximal analysis

Impact: Analysis at scale
Projected Data Volumes in EOSDIS
Distribution increases similarly to cumulative volume in EOSDIS
Solution: Data-proximal Analysis
Abstract Analytics Workflow

- Extract
- Transform
- Load
- Analyze
- Visualize

12
Extract-and-Transform Preprocessing

Cumulus Archive → Preprocessing as-a-service → Analysis Ready Data → Analytics Optimized Data Store
Analysis-Ready Data for End Users

Preprocessing

Cumulus Archive

as-a-service

Analysis Ready Data

Analytics Optimized Data Store

End-User Algorithm
Cloud-Native Analysis for Data Scientists

Preprocessing as-a-service

Cumulus Archive

Analysis Ready Data

Analytics Optimized Data Store

End-User Cloud-Native Analysis

End-User Algorithm
Analysis-as-a-Service for Interdisciplinary Users

Preprocessing
as-a-service

Cumulus Archive

Analysis Ready Data

Analytics Optimized Data Store

$\int f(x)dx$

Analysis as-a-service

End-User Cloud-Native Analysis

End-User Algorithm

Analysis Ready Data

End-User Interpretation
Service exposure enables access throughout the value chain.
Service consumption opens pipeline to external data
Service consumption opens pipeline to external data
Key Questions

Q: Which data transformations are common enough to include in production of Analysis-Ready Data?

1. Subsetting
2. Regridding and reprojection
3. Quality filtering?
Key Questions

Q: What is the optimal Analytics Optimized Data Store?

1. Highly scalable database?
2. Hadoop File System?
3. Data Cubes in Web Object Storage?
4. Xarray / zarr?
5. It depends...