Earthdata Cloud Analytics Project

Chris Lynnes and Rahul Ramachandran
Goals

1. Big compute next to big data: the big win in cloud data archives*

2. User adoption of cloud for analytics

3. Maximum analytics capability at minimum cost
   a. Use internal capabilities more effectively and efficiently
   b. Leverage external analytics capabilities

*Vis2020: “Data Analysis at Scale”
Main Constraints

1. ERT Recommendation #2*:
   a. Right Tools + Less Duplication
   b. Reuse + Open Source
   c. Common web services

2. Do NOT affect the main Earthdata Cloud goals
   a. NGAP rearchitecture
   b. Cumulus development
   c. Archive migrations

*Vis2020: Combining Tools: Tools and services within the community are easy to combine.
Key Features

1. Satisfy a diverse user community
2. Support analysis in the cloud without egressing data*
   a. Legacy algorithms
   b. Cloud-native algorithms
3. Facilitate multi-dataset comparison and fusion**
4. Support different modes of data interaction (Batch, Interactive, Streaming)
5. Support different modes of data storage (file systems, indexed storage..)
6. Support cost constraints and cost-sharing

*Vis2020: Mobile Data and Processing: Data and processing move transparently as necessary to achieve optimal performance. (kinda)
**Vis2020: Combining Data: NASA data can be combined with data from other agencies, nations and other entities
Earthdata Cloud Analytics Guiding Principles

1. **Provide building blocks** for innovation and infuse rapidly evolving technologies
2. Eschew monolithic systems
3. Enable openness: code and service exposure
4. Encourage/Enforce interoperability and reuse
5. Eschew *unnecessary* duplication in the form of undifferentiated heavy lifting
Architectural Concept

Earth Science Data Analytics the Cloud-Native Way: Everything is a Service

This approach produces key important benefits for the user community and EOSDIS
Abstract Analytics Workflow

data → Extract → Transform → Load → Analyze → Visualize
Earthdata Cloud Analytics Reference Architecture

1. **Preprocessing as-a-service**
2. **ARD\(^1\) as-a-service**
3. **Analysis as-a-service**
4. **Visualization as-a-service**

\(^1\) Analysis Ready Data
Supports Interactive, Batch and Streaming Modes
Interactive Mode: Analytics-Optimized Storage

- Cumulus as-a-service
- Preprocessing as-a-service
- ARD¹ as-a-service
- Analysis as-a-service
- Visualization as-a-service

¹ Analysis Ready Data
Batch Mode

Cumulus

Preprocessing
as-a-service

Analysis
as-a-service

Visualization
as-a-service

ARD\(^1\)
as-a-service

\(^1\) Analysis Ready Data
Streaming Mode

Cumulus

Preprocessing as-a-service

Event Analytics as-a-service

Visualization as-a-service
Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base
Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base

Preprocessing as-a-service

ARD\(^1\) as-a-service

Analysis as-a-service

Visualization as-a-service

End-User-Specific Cloud-Native Analysis

End-User Cloud-Native Analysis

Cumulus

Analysis Ready Data
Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base
Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base

Preprocessing as-a-service

ARD\textsuperscript{1} as-a-service

Analysis as-a-service

Visualization as-a-service

End-User-Specific Analysis

End-User Cloud-Native Analysis

End-User Interpretation

Data Exploration
Open Pipeline Enables Integration with Other Data, Scripts, and Workflows

Preprocessing as-a-service
End-User-Specific Analysis

ARD\textsuperscript{1} as-a-service
End-User Cloud-Native Analysis

Analysis as-a-service
End-User Interpretation

Visualization as-a-service
Data Exploration
Open Pipeline Enables Integration with Exploitation Platforms
### Where EOSDIS Stops and the User Begins

**Strawman: Open for Discussion!**

<table>
<thead>
<tr>
<th>EOSDIS Responsibility</th>
<th>User Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>General-purpose preprocessing</td>
<td>User-specific preprocessing</td>
</tr>
<tr>
<td>Common summary statistics algorithms</td>
<td>User-defined analysis algorithms</td>
</tr>
<tr>
<td>Budget-limited processing for analysis</td>
<td>User-pays processing beyond budget limits</td>
</tr>
<tr>
<td>Provide analysis capability to user (e.g., AMI)</td>
<td>Run analysis</td>
</tr>
<tr>
<td>Interactive visualization for exploration</td>
<td>Publication-quality data graphics</td>
</tr>
<tr>
<td>EOSDIS standard projections</td>
<td>Other projections</td>
</tr>
</tbody>
</table>

---

**Earthdata Cloud Analytics**
HOW: Implementation
Earthdata Cloud Analytics Project

- Architecture and Design
- Current Capability Implementation in ECA Architecture
- Technology Infusion
Earthdata Cloud Analytics Project

- Technology Infusion
- Current Capability Implementation in ECA Architecture
- Architecture and Design
- Design Patterns
- Architecture
- Standards

Business Model

Earthdata Cloud Analytics
Seeking to Form an Architecture Team

- **Tasks**
  - Refine Reference Architecture
  - Develop Design Patterns
  - Work with applicable standards groups (ESO, ESDSWG, WGISS)
  - Work on Business Model
  - Design User Experience

- **Membership**
  - ESDIS architects
  - DAAC architects
  - EED2 architects
  - EOSDIS Standards Office
  - Affiliated project architects: e.g., H. Hua (Mission Analysis Platform*)
  - Experts in the field as available

- **Vis2020: Sharing:** Scientists are able to share all scientific resources (data, tools, results, workflows, contextual knowledge)
Earthdata Cloud Analytics Project

- Technology Infusion
  - Current Capability Implementation in ECA Architecture
    - Giovanni
  - Design Patterns
- Architecture and Design
Earthdata Cloud Analytics Project

- New Technology Infusion
- Cloud Analytics Workshop
- Current Capability Implementation in ECA Architecture
- Architecture and Design
- Community Adoption
- Interoperability Demonstrations
Technology Infusion Suggestion:

● Informal ESDIS RFP to integrate analytics capabilities into Earthdata Cloud Analytics
● Timeline
  ○ RFP: July 2018
  ○ Selection: Sept. 2018
To Do

➢ Get Go-Ahead from Kevin (Done)
➢ Brief DAAC Managers (Feb. 8)
➢ Develop Project Plan
➢ Recruit Architecture Team
1. Architecture and Design

1.1 Earthdata Cloud Analytics Architecture
   a. Service-based architecture
   b. Design patterns
   c. User Experience

1.2 Earthdata Cloud Analytics Standards
   a. Service APIs
   b. Data Formats and Conventions

1.3 Business Model
   a. Nonelastic $\Rightarrow$ Partially Elastic $\Rightarrow$ Fully Elastic
   b. Cost prediction
   c. Cost allocation (ESDIS vs. User)
   d. Cost control
   e. Governance model
Prototype Implementation

- **Earthdata Cloud Analytics ("powered by Giovanni")**
  - High profile win with extensive user base
  - Clear use cases for cloud (e.g., ARSET)
  - Already demonstrated in the cloud
  - NGAP porting underway

- **Success Target:**
  - Cloud-based instance provisioned for ARSET training
  - Three (3) disciplines
  - Most commonly used workflows
    - Time Averaged Map
    - Area-Averaged Time Series

- **Jan - Dec. 2018**
Dedicated Funding vs. Steering Existing Activities

- **Architecture & Design**
  - Funding: EED2 architects
  - Steering: ESO, ESDSWG

- **Current Implementation**:
  - Funding: Giovanni
  - Steering: 

- **Technology Infusion**
  - Steering: Active Scouting, Leveraging Partnerships
  - Funding: Infusion and interoperability demonstrations, Selected infusions
# Communications Plan

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Communications</th>
</tr>
</thead>
<tbody>
<tr>
<td>HQ</td>
<td>Report up through Earthdata Cloud Project Briefings* at HQ</td>
</tr>
<tr>
<td>DAACs</td>
<td>Briefing at DAAC Managers Meeting and Telecons Inreach Webinars System Engineers Meeting Analytics Working Group? Mail Lists?</td>
</tr>
<tr>
<td>User and Partner Communities</td>
<td>Webinars and Seminars ARSET Trainings Recipes DAACs Articles (Eos, Big Earth Data Analytics book) Conferences (BiDS, IGARSS, ESIP, AGU, WGISS…)</td>
</tr>
</tbody>
</table>
## Stakeholders and Benefits

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Key Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert science users</td>
<td>Customized (transformed) data for easy analysis with adjacent computing power</td>
</tr>
<tr>
<td>Cloud-native users</td>
<td>Data reorganized and stored in cloud-analytics-friendly fashion</td>
</tr>
<tr>
<td>Interdisciplinary users</td>
<td>Data statistics computations with full provenance</td>
</tr>
<tr>
<td>Applications</td>
<td>Open API for developing data-to-decision-support chains Ready access to machine learning applied to NASA data</td>
</tr>
<tr>
<td>EOSDIS archive managers</td>
<td>End-user migration to the cloud (=egress-free access) Building blocks and process for constructing rich analysis support for community</td>
</tr>
<tr>
<td>NASA HQ</td>
<td>Full exploitation of complete datasets AIST / ACCESS infusion</td>
</tr>
</tbody>
</table>
2. Port Current Implementation

2.1 Giovanni: Architecture Pathfinder and High Profile Win
   a. Port to NGAP
   b. Refactor for full standards-based service exposure
   c. Business model implementation
3. Technology Infusion

3.1 Cloud Analytics Workshop
   a. Algorithms
   b. Systems
   c. Architecture refinement

3.2 New Technology Infusion
   a. Active Scouting: ACCESS, AIST, ECF, SBIR, NSF, AGU...
   b. Partnerships: MEP, GeoGLAM, WGISS, COVERAGEs...

3.3 Infusion and interoperability demonstrations
   a. Modeled on OGC Interoperability Experiments
   b. Call for Proposals to EOSDIS elements (and ACCESS, AIST?)

3.4 Community Adoption
   a. Communications: Papers, Webinars, How-Tos, Jupyter notebooks...
   b. Partnerships: AGU, OGC, WGISS ...

Earthdata Cloud Analytics