Open Source Platform for Federated Spatiotemporal Analysis

Thomas Huang
thomas.huang@jpl.nasa.gov

Group Supervisor – Data Product Generation Software
Strategic Lead - Interactive Data Analytics

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.
Big Data calls for…

- Automation and Sustainable technologies
- Scale computational and data infrastructures
- Support new methods for deriving scientific inferences
- Shift towards integrated data analytics
- Apply computational and data science across the lifecycle

**Scalable Data Management**
- Capture well-architected and curated data repositories based on well-defined data/information architectures
- Architecting automated pipelines for data capture

**Scalable Data Analytics**
- Access and integration of highly distributed, heterogeneous data
- Novel statistical approaches for data integration and fusion
- Computation applied at the data sources
- Algorithms for identifying and extracting interesting features and patterns

- How to quickly deploy to the cloud or local cluster?
- How to manage software versioning?
- How to upgrade without complete shutdown?
- How to manage operating cost?
- How to deploy a truly scalable solution within budget?
- How to manage data priority?
- How to on-board and off-board data?
- How to manage job priority?
- How to deliver a cloud-based solution without overwhelming our users about the nuts and bolts of cloud?
ACF – Solution to our Big Earth Science Analytics Challenges

- **Analytics Center Framework**: an environment for conducting an Ocean Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the ocean research community to focus on the investigation
- Scale computational and data infrastructures
- Shift towards integrated data analytics
- Algorithms for identifying and extracting interesting features and patterns
Not Just Open Source … Professional Open Source!
Platform for Production-Quality Interactive Analytics
Federated Data Analytics

- **Federated ACF instances** to enable distributed analytics without requiring massive data download and transfer. Clients can be Jupyter Notebook, GIS Web Applications, and Esri ArcGIS.
```
IDL> spawn, 'curl "https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-01T00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt"

<table>
<thead>
<tr>
<th>% Total</th>
<th>% Received</th>
<th>% Xferd</th>
<th>Average</th>
<th>Time</th>
<th>Time</th>
<th>Time</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Speed</td>
<td>Total</td>
<td>Spent</td>
<td>Left</td>
<td>Speed</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

IDL> result = JSON_PARSE('json_dump.txt', /toarray, /tostruct)
IDL> help, result
** Structure <1a2749c8>, 3 tags, length=62320, data length=62320, refs=1:
| STATS STRING '!NULL'
| META STRUCT -> <Anonymous> Array[1]
| DATA STRUCT -> <Anonymous> Array[778]

IDL> plot(result.data.time, result.data.mean, title='GHRST L4 AVHRR_OI SST. Sep 2008 - Oct 2015. US West Coast Blob Area')
```

Credit: Ed Armstrong
Jun. 05, 2020
Examples of Distributed Analytic Centers

- All done interactively using jupyter
- Using the ECCO’s SDAP and the Sea Level Change Portal’s SDAP
- Use ECCO’s SDAP to compute time series for ECCO’s Surface Height Anomaly product
- Use Sea Level Change Portal’s SDAP to compute time series for MEaSUREs Sea Level Anomaly product
- Plot the two time series using matplotlib

- All done interactively using jupyter
- Using the ECCO’s SDAP and the GRACE Follow-On’s SDAP
- Use ECCO’s SDAP to compute time series for ECCO’s Ocean Bottom Pressure Anomaly product
- Use GRACE Follow-On’s SDAP to compute time series for GRACE Follow-On’s Ocean Bottom Pressure Anomaly product
- Plot the two time series using matplotlib

Zero Data Movement, Zero egress, All computed on distributed instances of SDAP on the Cloud
Enabling the Private Sector and its Community

- Embrace open standards (ISO, OGC, etc.)
- Open source = empowering organizations and community
- Example: Building open source bridge between Esri’s ArcGIS with Apache SDAP
  - Allowing data center to use SDAP, which is free and open source
  - Allowing Esri user community to directly access and analyze satellite observational data directly using Esri applications without having to download massive collection of data to their local computers
Connecting to ACF from Esri ArcGIS
• Committee of Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEO (COVERAGE) Initiative
  • Seeks to provide **improved access** to multi-agency ocean remote sensing data that are **better integrated** with in-situ and biological observations, in support of oceanographic and decision support applications for societal benefit.
  • A community-support open specification with common taxonomies, information model, and API (maybe security)
  • Putting value-added services next to the data to eliminate unnecessary data movement
  • Avoid data replication. Reduce unnecessary data movement and egress charges
  • Analytic engine infused and managed by the data centers perhaps on the Cloud
  • Researchers can perform multi-variable analysis using any web-enabled devices without having to download files
COVERAGE – Phase B

- **WEkEO**
  - Copernicus Data and Information Access Services (DIAS)
    1. Copernicus Data
    2. Virtual Environment and Tools
    3. User Support
  - Harmonized Data Access for Satellite data and Services
  - Virtualized infrastructure for personal sandboxes
  - Pre-configured tools

- **COVERAGE Phase B**
  - Establish US Node on Amazon Cloud
  - Establish EU Node on WEkEO at EUMETSAT
  - Establish COVERAGE data portal and analysis tool powered by the COVERAGE Nodes at US and EU
Typically data matching is done using one-off programs developed at multiple institutions.

A primary advantage of SDAP’s matchup service is the reduction in duplicate development and man hours required to match satellite/in situ data.

- Removes the need for satellite and in situ data to be collocated on a single server.
- Systematically recreate matchups if either in situ or satellite products are reprocessed (new versions), i.e., matchup archives are always up-to-date.

Through the AIST Distributed Oceanographic Matchup Service (DOMS), we established in situ data nodes at JPL, NCAR, and FSU operational.

Cloud-based data querying, subset, and match-up services.
Software architecture that is sustainable needs to have generalized interface and information model and extensible to address domain-specific specialization

Example: Multi-mission Data Management and Distribution Architecture

• Parallel Map (a.k.a. Parmap) is a new addition to the Apache SDAP platform as an analytic engine that is scalable for parallel analytics and climate model evaluation on cloud and HPC (Wilson and Jacob, 2019)
  • Enable simple coding for computing analytics in parallel
  • Extends Apache SDAP with streamlined workflow optimized for data products on a regular coordinated grid
  • Operates on the original data granule files, so no data duplication is needed
  • No persistent services required means it is easy to deploy
  • Simple to deploy since the MapReduce operations can execute over files or object store
  • Plug-in architecture to support various parallel analytics infrastructure: Multicore (single node) | PySparkling (pure Python, single node) | Full Apache PySpark Cluster (multi-node) | Dask Cluster (multi-node) | GPU | AWS Lambda (serverless)
Managing Professional Open Source

• Technology sharing through Free and Open Source Software (FOSS)
• Why? Further technology evolution that is restricted by projects / missions
• It is more than GitHub
  • Quarterly reporting
  • Reports are open for community review by over 6000 committers
  • SDAP has a group of appointed international mentors
• SDAP and many of its affiliated projects are now being developed in the open
  • Support local cluster and cloud computing platform support
  • Fully containerized using Docker and Kubernetes
  • Infrastructure orchestration using Amazon CloudFormation
  • Satellite and model data analysis: time series, correlation map,
  • In situ data analysis and collocation with satellite measurements
  • Fast data subsetting
  • Upload and execute custom parallel analytic algorithms
  • Data services integration architecture
  • OpenSearch and dynamic metadata translation
  • Mining of user interaction and data to enable discovery and recommendations
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ed Armstrong/JPL</td>
<td>Rich Doyle/JPL</td>
</tr>
<tr>
<td>Jason Barnett/LARC</td>
<td>Jocelyn Elya/FSU</td>
</tr>
<tr>
<td>Andrew Bingham/JPL</td>
<td>Ian Fenty/JPL</td>
</tr>
<tr>
<td>Carmen Boening/JPL</td>
<td>Eamon Ford/JPL</td>
</tr>
<tr>
<td>Mark Bourassa/FSU</td>
<td>Kevin Gill/JPL</td>
</tr>
<tr>
<td>Mike Chin/JPL</td>
<td>Frank Greguska/JPL</td>
</tr>
<tr>
<td>Marge Cole/NASA</td>
<td>Patrick Heimbach/UT Austin</td>
</tr>
<tr>
<td>Tom Cram/NCAR</td>
<td>Ben Holt/JPL</td>
</tr>
<tr>
<td>Dan Crichton/JPL</td>
<td>Thomas Huang/JPL</td>
</tr>
<tr>
<td>Maya DeBellis/JPL</td>
<td>Joe Jacob/JPL</td>
</tr>
<tr>
<td>Zaihua Ji/NCAR</td>
<td>Kevin Murphy/NASA</td>
</tr>
<tr>
<td>Yongyao Jiang/Esri</td>
<td>Charles Norton/JPL</td>
</tr>
<tr>
<td>Felix Landerer/JPL</td>
<td>Jean-Francois Piolle/IFREMER</td>
</tr>
<tr>
<td>Yun Li/GMU</td>
<td>Nga Quach/JPL</td>
</tr>
<tr>
<td>Eric Lindstrom/NASA</td>
<td>Brandi Quam/NASA</td>
</tr>
<tr>
<td>Mike Little/NASA</td>
<td>Shawn Smith/FSU</td>
</tr>
<tr>
<td>Thomas Loubrieu/JPL</td>
<td>Ben Smith/JPL</td>
</tr>
<tr>
<td>Chris Lynnes/NASA</td>
<td>Adam Stallard/FSU</td>
</tr>
<tr>
<td>Lewis McGibbney/JPL</td>
<td>Rob Toaz/JPL</td>
</tr>
<tr>
<td>David Moroni/JPL</td>
<td>Vardis Tsontos/JPL</td>
</tr>
<tr>
<td>Suresh Vannan/JPL</td>
<td>Jorge Vazquez/JPL</td>
</tr>
<tr>
<td>Ou Wang/JPL</td>
<td>Brian Wilson/JPL</td>
</tr>
<tr>
<td>Steve Worley/NCAR</td>
<td>Elizabeth Yam/JPL</td>
</tr>
<tr>
<td>Phil Yang/GMU</td>
<td>Alice Yepremyan/JPL</td>
</tr>
</tbody>
</table>
Building Community-Driven Open Data and Open Source Solutions

- Deliver solutions to establish coherent platform solutions
- Embrace open source software
- Community validation
- Evolve the technology through community contributions
- Share recipes and lessons learned
- Technology demonstrations
- Host webinars, hands-on cloud analytics workshops and hackathons
Partner with NASA and non-NASA Projects – Deliver to Production

- The gap between visionary to pragmatists is significant. – Geoffrey Moore
- Become an expert in the production environment and devote resources in automations
- Give project engineering team early access to the PaaS
- Deliver all technical documents and work with project system engineering
- Provide project-focused trainings

NASA’s Physical Oceanography Distributed Active Archive Center (PO.DAAC)

NASA’s Sea Level Change Team

CEOS SIT Technical Workshop
In Summary

• You’ve got to think about big things while you’re doing small things, so that all the small things go in the right direction – Alvin Toffler

• Climate research requires Autonomously Sustainable Solutions

• Focus on delivering professional quality open source solutions

• Enables end-to-end data and computation architecture, and the total cost of ownership

• Start with system architecture aiming for simple interfaces and information model

• From generalization to specialization

• Apache SDAP is a multi-cloud, multi-cluster, multi-data-center, and multi-agency platform

• Open source should not be a destination, it should be in place from the beginning

• How a technology is being managed will determine how far it can go
Thomas Huang
thomas.huang@jpl.nasa.gov
Jet Propulsion Laboratory
California Institute of Technology