National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

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.

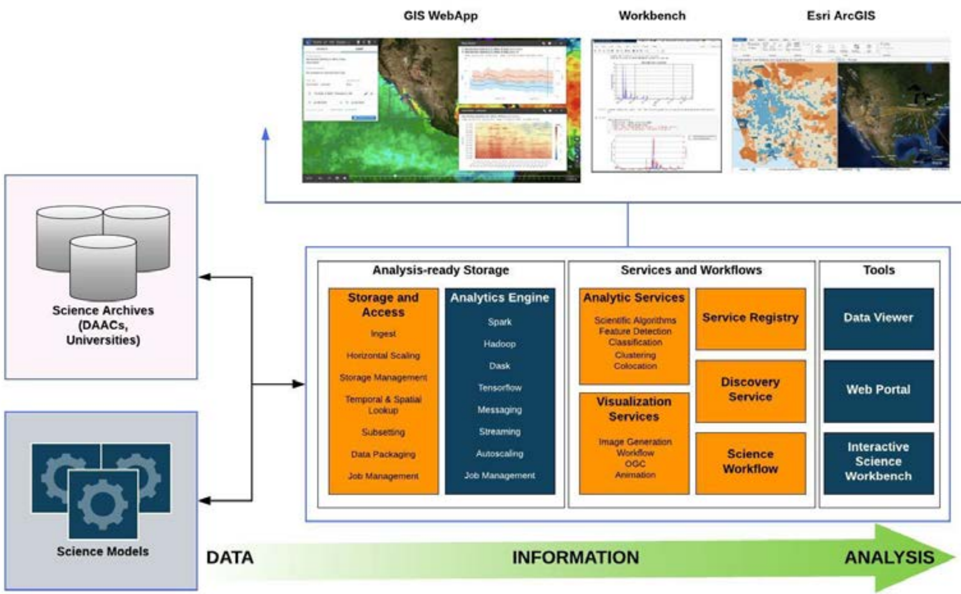
Big Data calls for…

- 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

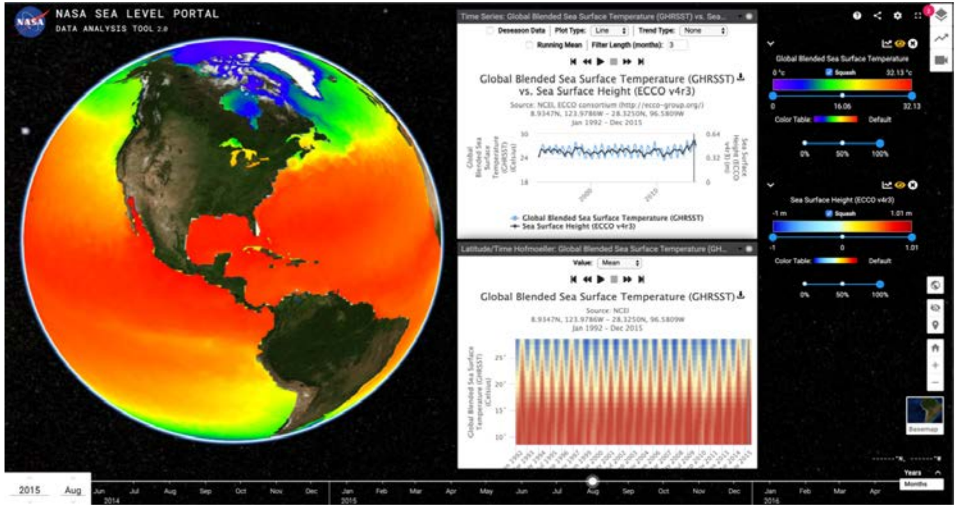
- **Analytics Center Framework**: an environment for conducting a Ocean Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the ocean 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

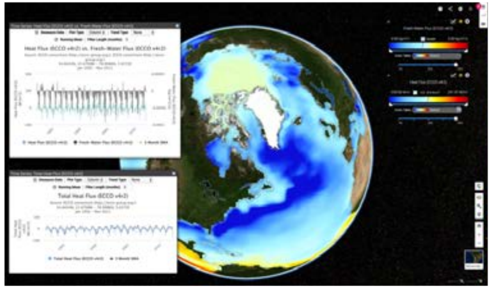


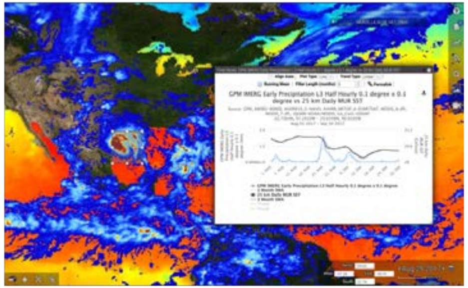
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Not Just Open Source … Professional Open Source!

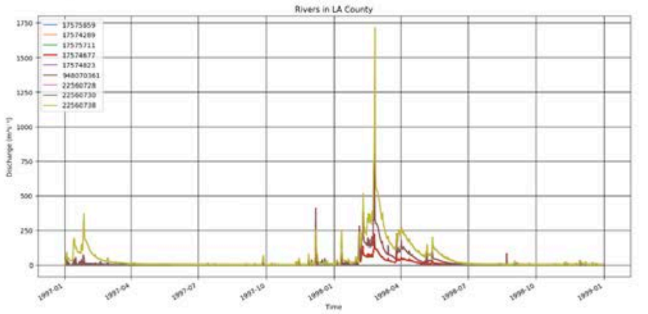






```python
import requests
import json
import time
import nexuscli
from datetime import datetime

nexuscli.set_target("https://doms.jpl.nasa.gov", use_session=False)

# River IDs for the 10 largest (by max discharge rate) Rivers in LA County
#la_county_river_ids = [
#    20351643, 20357290, 20357300, 20357284, 20357292,
#    948070444, 20351637, 20357240, 20357296, 20351677]
la_county_river_ids = [17575859, 17574289, 17575711, 17574677, 17574823,
                       948070361, 22560728, 22560730, 22560738]

ds = "RAPID_WSWM"
start_time = datetime(1997, 1, 1)
end_time = datetime(1998, 12, 31, 23, 59, 59)
la_county_river_data = list()

start = time.perf_counter()
for river_id in la_county_river_ids:
    metadataFilter = "rivid_i:{}".format(river_id)
    result = nexuscli.subset(ds, None, start_time, end_time, None, metadataFilter)
    la_county_river_data.append(result)
print("Subsetting took {} seconds".format(time.perf_counter() - start))

show_plot([[point.time for point in river] for river in la_county_river_data], # x values
          [[point.variable['variable'] for point in river] for river in la_county_river_da
          'Time', # x axis label
          'Discharge (m³s⁻¹)', # y axis label
          legend=[str(r) for r in la_county_river_ids],
          title='Rivers in LA County'
         )
```

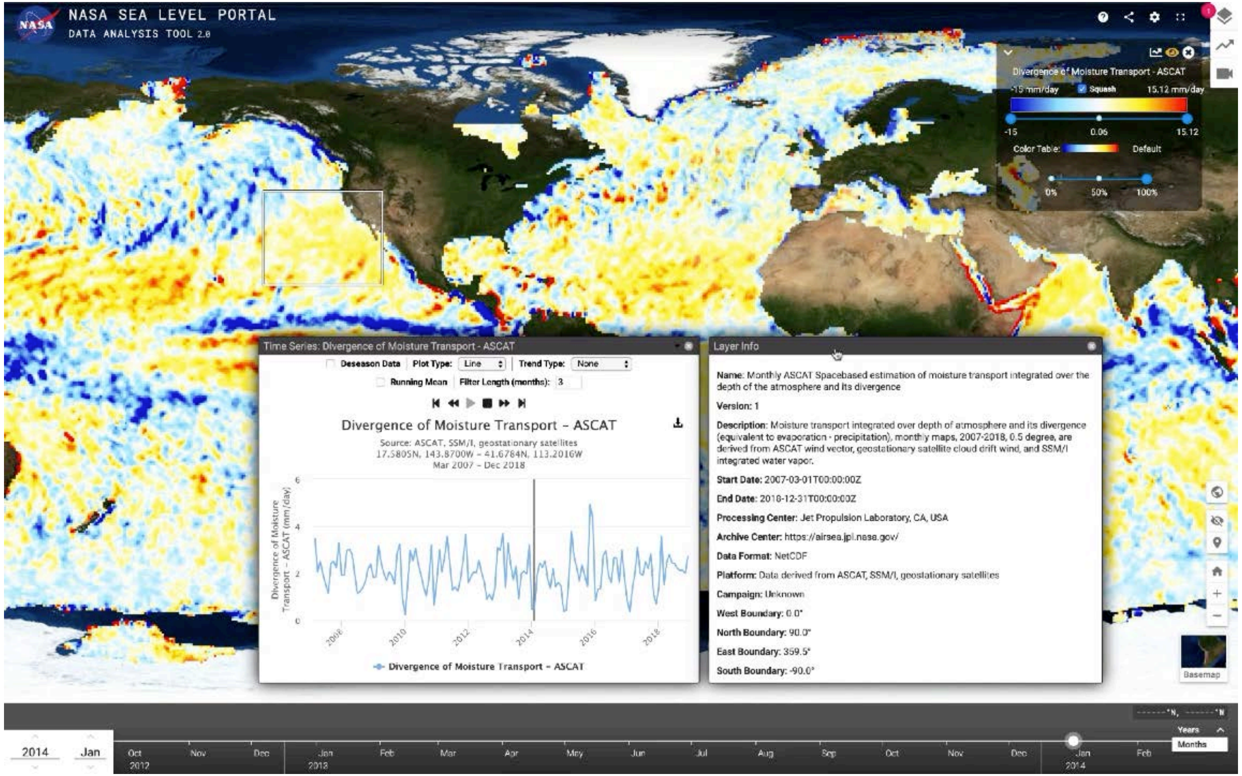
```
Target set to https://doms.jpl.nasa.gov
Subsetting took 4.413320103660226 seconds
```



National Aeronautics and Space Administration

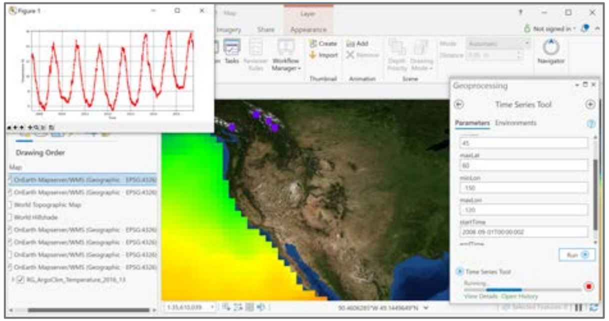
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

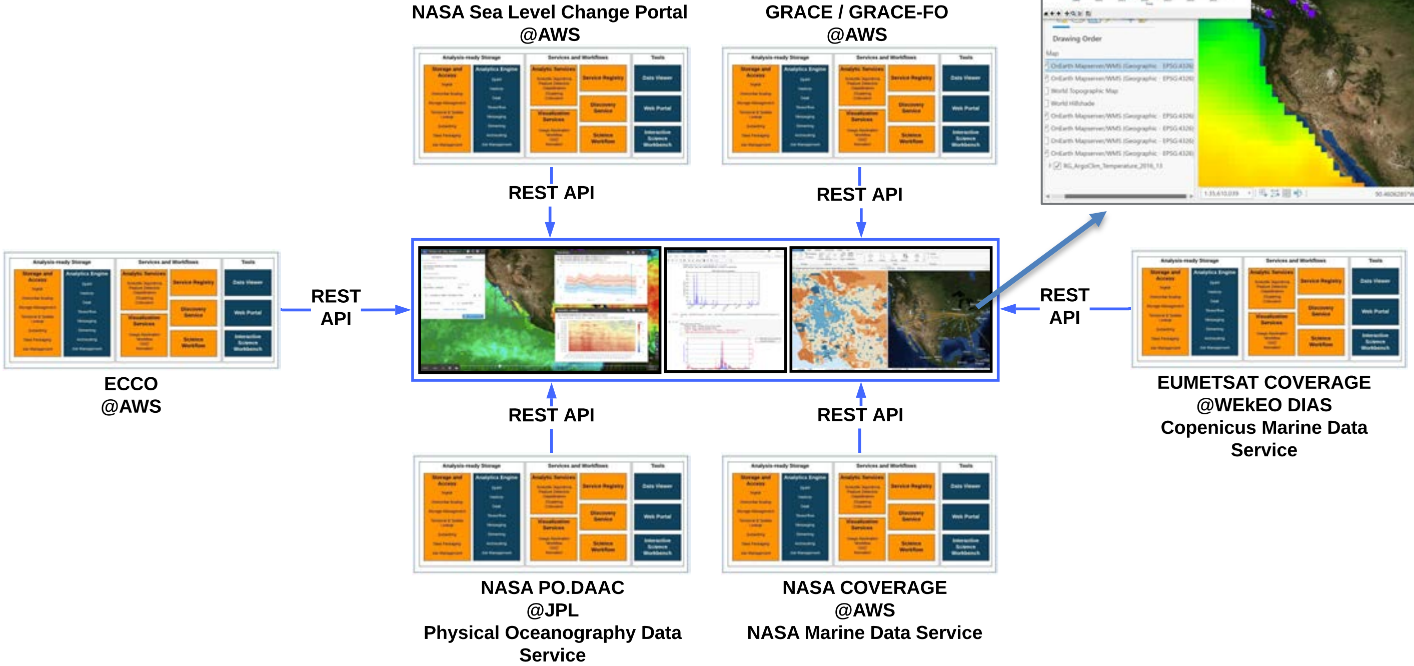
Platform for Production-Quality Interactive Analytics



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



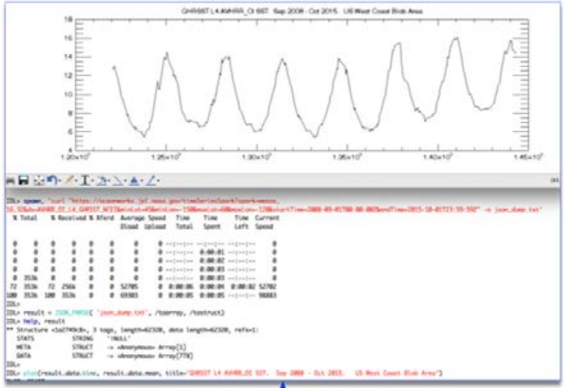


National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Interact with Analytics Platform using any Programming Language

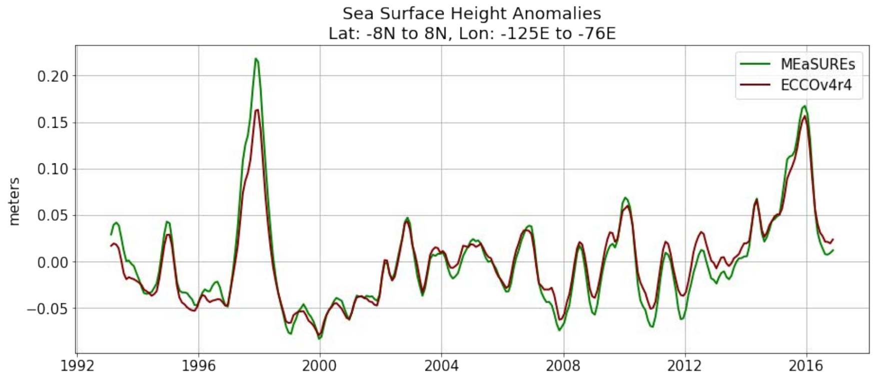
```
IDL> spawn, 'curl
"https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_
GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-
01T00:00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt'
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed

  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:01 --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:02 --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
  0  353k    0     0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
 72  353k   72  256k    0     0  52705      0  0:00:06  0:00:04  0:00:02 52702
100  353k  100  353k    0     0  69303      0  0:00:05  0:00:05 --:--:-- 98883
IDL>
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>
IDL> plot(result.data.time, result.data.mean, title='GHRSST L4 AVHRR_OI SST.  Sep
2008 - Oct 2015.   US West Coast Blob Area')
PLOT <29457>
```



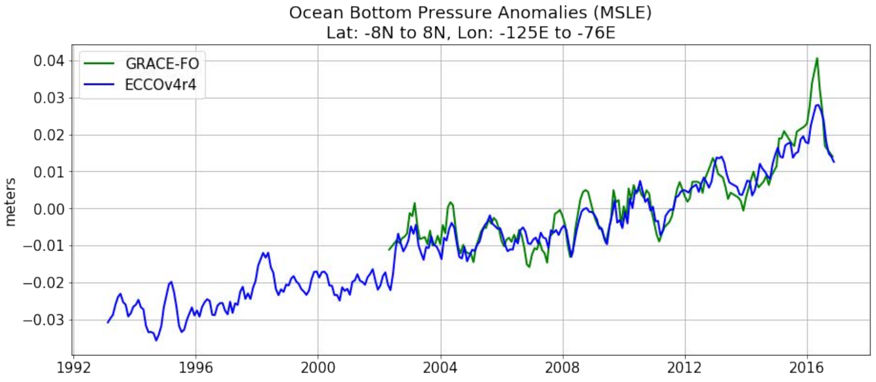
Credit: Ed Armstrong
Jun. 05, 2020

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

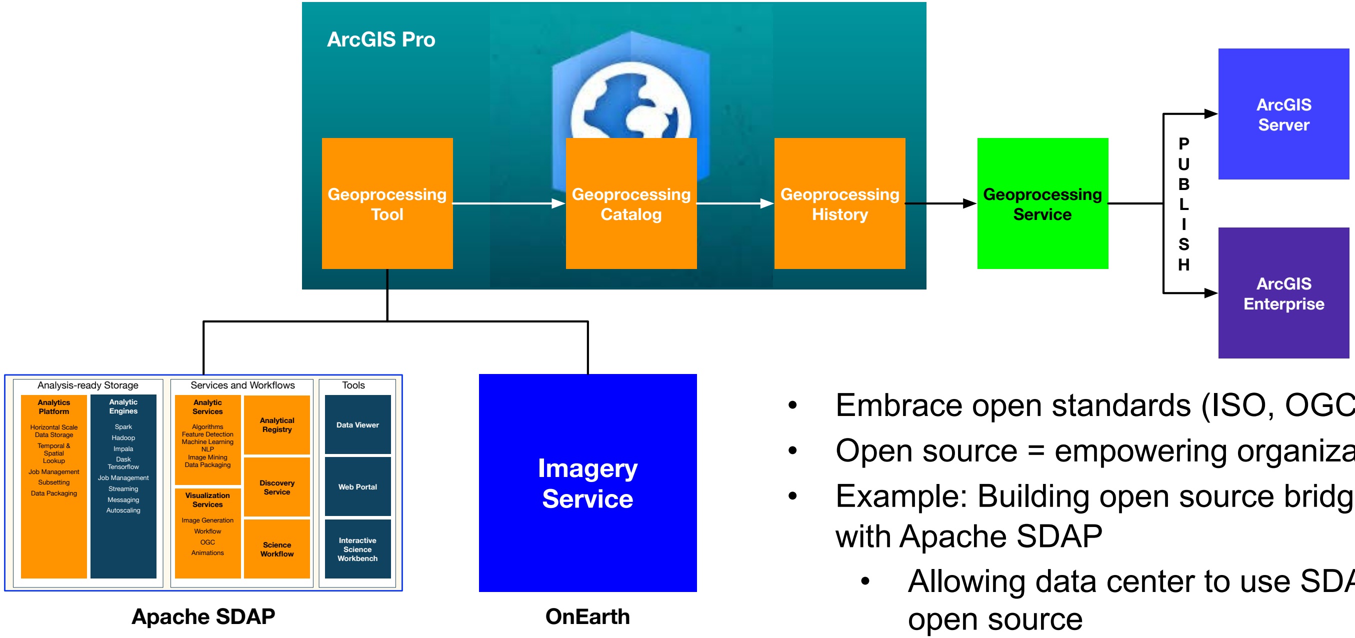


Zero Data Movement, Zero egress, All computed on distributed instances of SDAP on the Cloud

National Aeronautics and
Space Administration

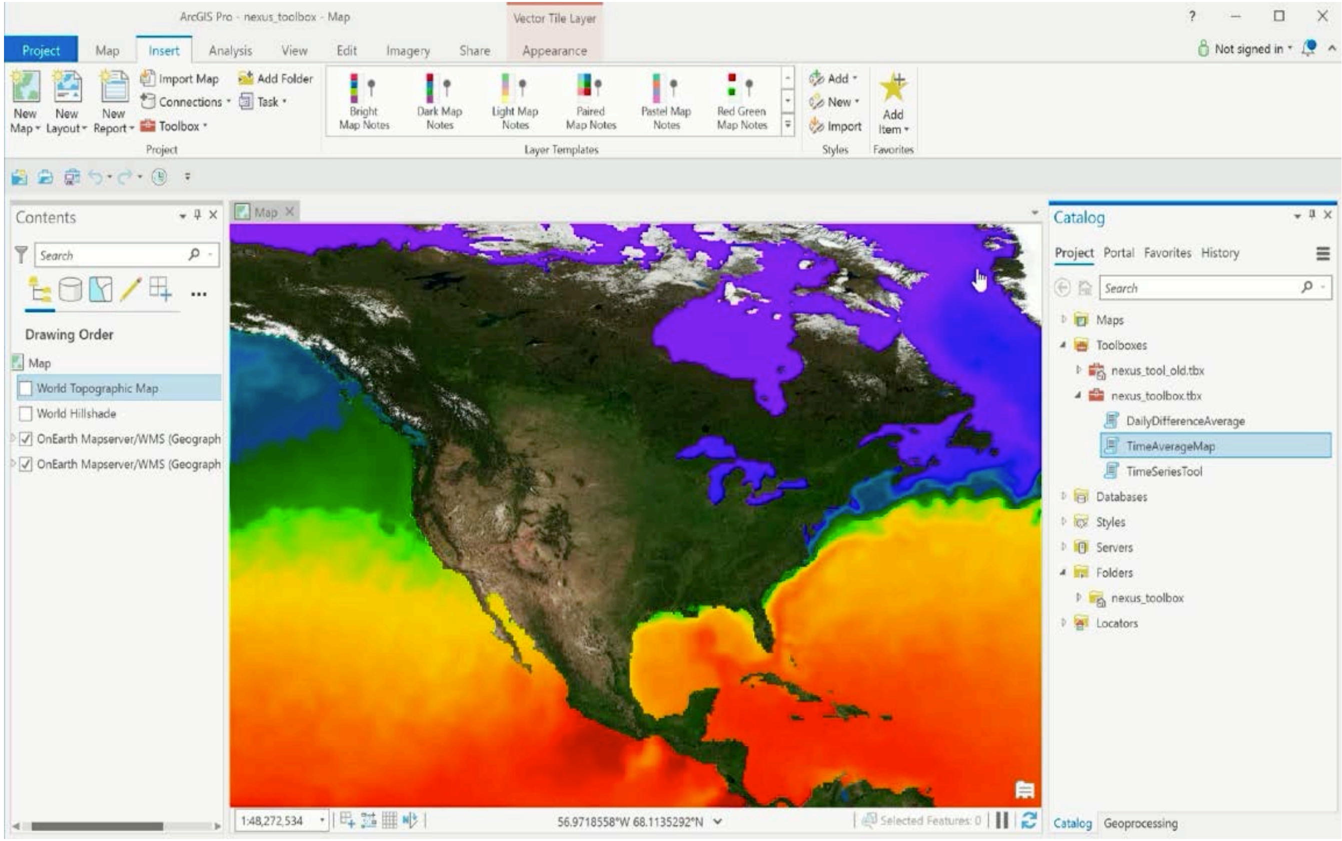
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enabling the Private Sector and its Community



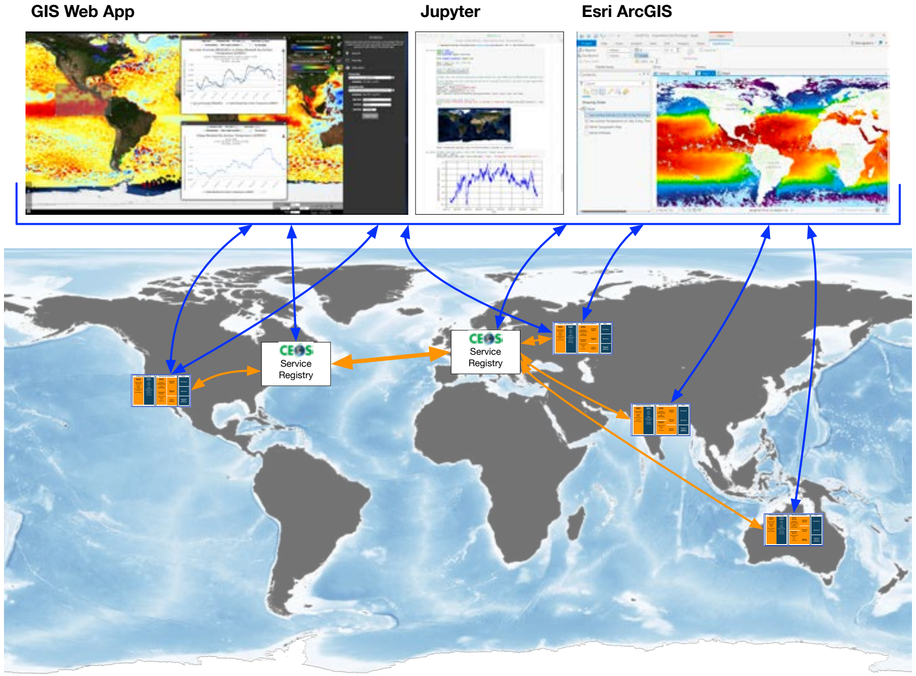
- Embrace open standards (ISO, OGC, etc.)
- Open source = empowering organizations and community
- Example: Building open source bridge between Esri's ArcGIS with Apache SDAP
  - Allowing data center to use SDAP, which is free and open source
  - Allowing Esri user community to directly access and analyze 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

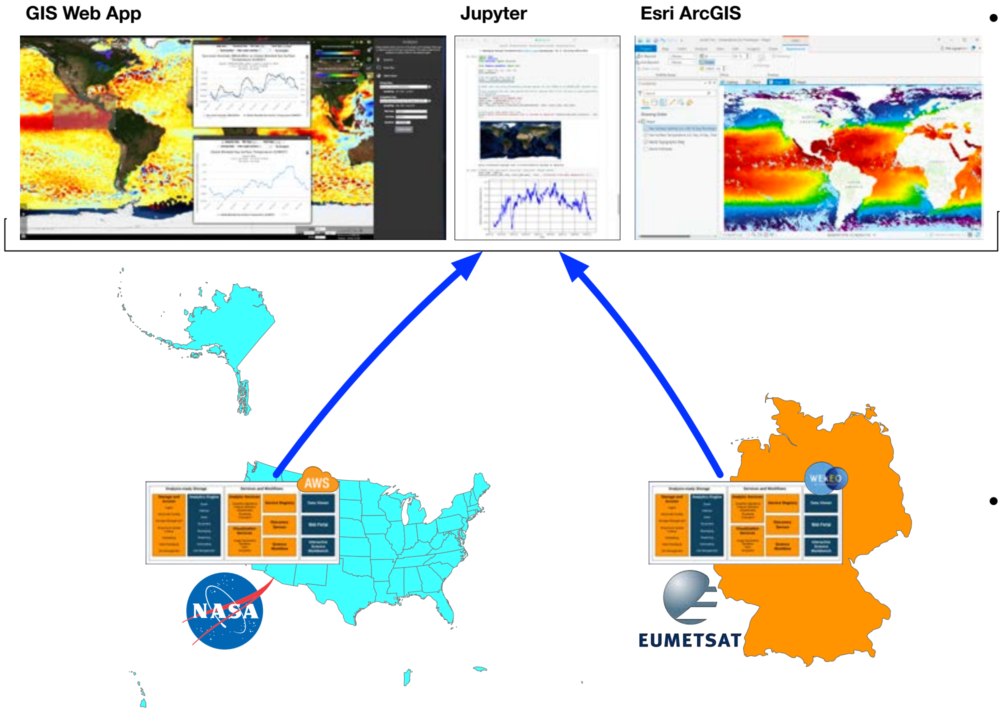


Distributed Analytics Center Architecture

- **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

National Aeronautics and Space Administration

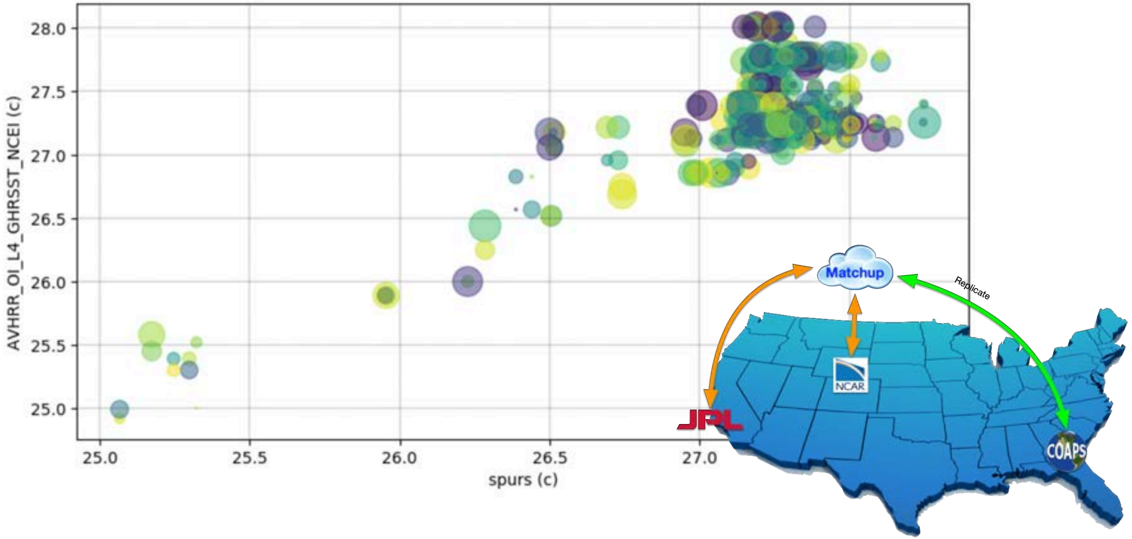
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributed Data Matchup Technology

- 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 re-processed (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

```
satellite = []
in_situ = []
for data in ts ['data']:
    for matches in data ['matches']:
        satellite.append(data['sea_water_temperature'])
        in_situ.append(matches['sea_water_temperature'])

show_plot(in_situ, satellite, secondary+' (c)', primary+' (c)')
```

```
https://oceanworks.jpl.nasa.gov/match_spark?primary=AVHRR_OI_L4_GHRSST_NCEI&matchup=spurs&startTime=2013-10-01T00:00:00Z&endTime=2013-10-30T23:59:59Z&tt=86400&rt=10000.0&b=-30,15,-45,30&platforms=1,2,3,4,5,6,7,8,9&parameter=sst&matchOne=true&depthMin=0&depthMax=5
It took: 2.746319055557251 sec
```



National Aeronautics and
Space Administration

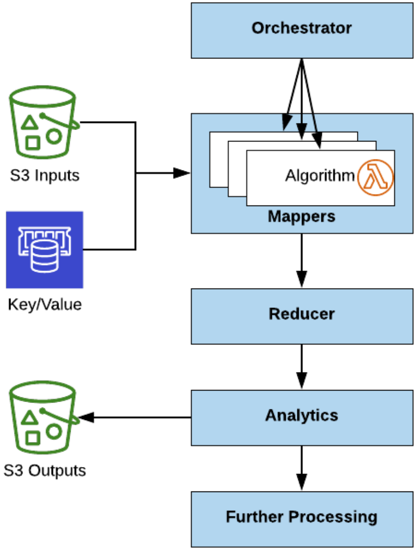
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

From Generalization to Specialization

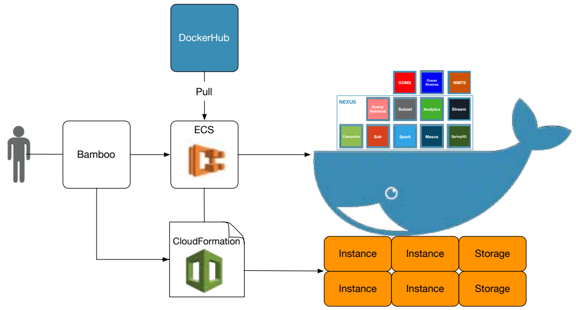
Software architecture that is sustainable needs to have generalized interface and information model and extensible to address domain-specific specialization

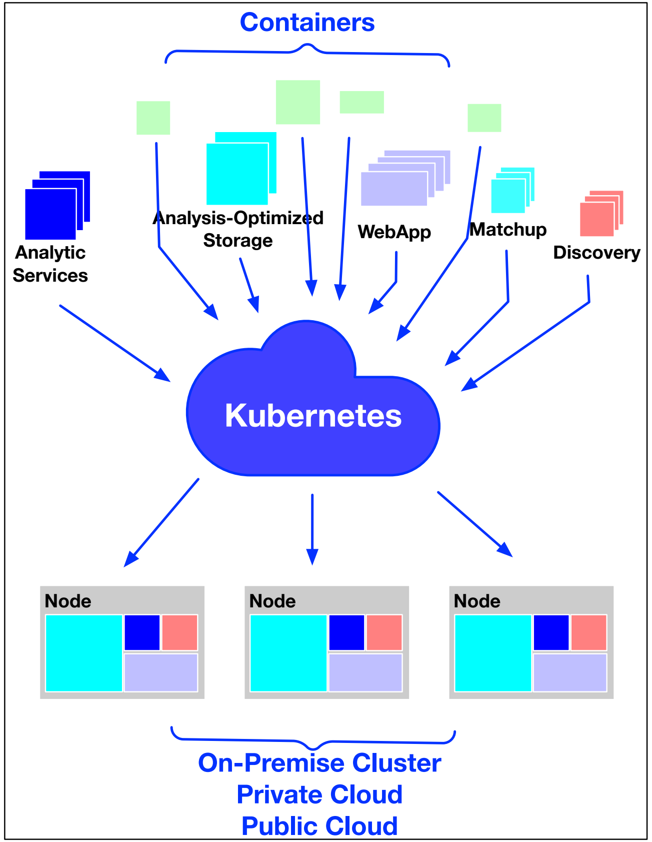
Example: Multi-mission Data Management and Distribution Architecture

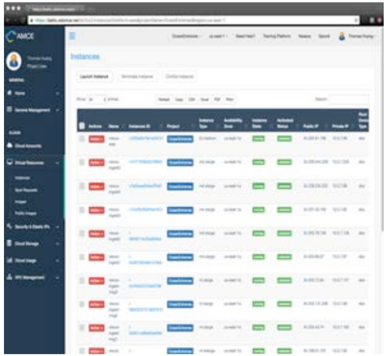
- Parallel Map (a.k.a. Parmap) is a new addition to the Apache SDAP platform as an analytic engine that is scalable for parallel analytics and climate model evaluation on cloud and HPC (Wilson and Jacob, 2019)
    - Enable simple coding for computing analytics in parallel
    - Extends Apache SDAP with 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)

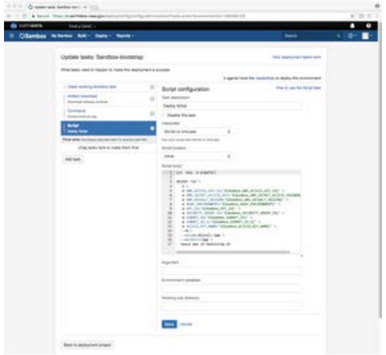


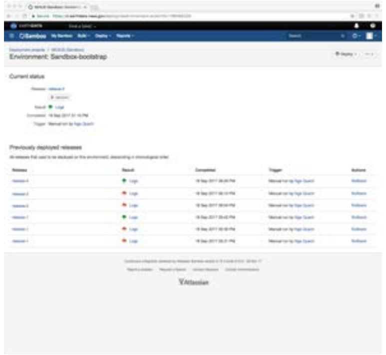
Automated, Container-based Deployment









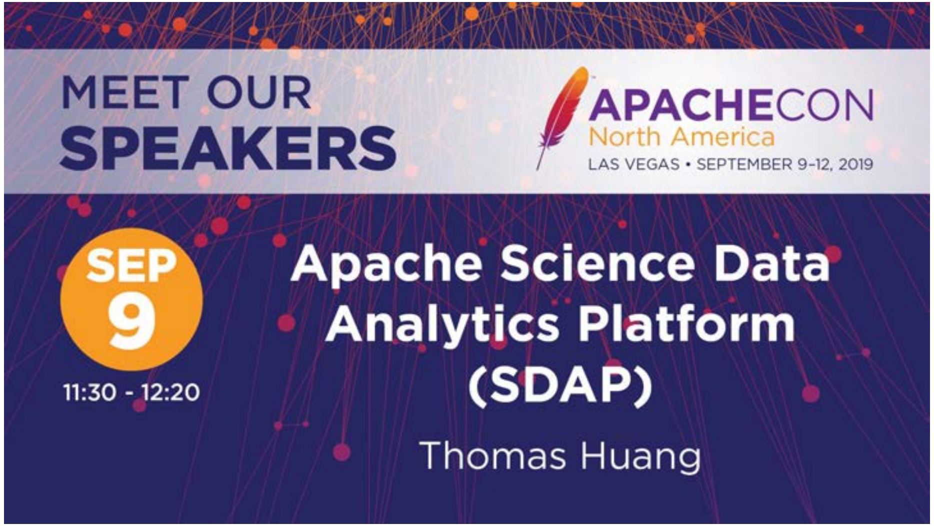


NASA AIST Managed Cloud Environment

NASA Next Generation Application Platform (NGAP)

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



Apache SDAP Acknowledgement

Ed Armstrong/JPL
Jason Barnett/LARC
Andrew Bingham/JPL
Carmen Boening/JPL
Mark Bourassa/FSU
Mike Chin/JPL
Marge Cole/NASA
Tom Cram/NCAR
Dan Crichton/JPL
Maya DeBellis/JPL
Rich Doyle/JPL
Jocelyn Elya/FSU
Ian Fenty/JPL
Eamon Ford/JPL
Kevin Gill/JPL
Frank Greguska/JPL
Patrick Heimbach/UT Austin
Ben Holt/JPL
Thomas Huang/JPL
Joe Jacob/JPL
Zaihua Ji/NCAR
Yongyao Jiang/Esri
Felix Landerer/JPL
Yun Li/GMU
Eric Lindstrom/NASA
Mike Little/NASA
Thomas Loubrieu/JPL
Chris Lynnes/NASA
Lewis McGibbney/JPL
David Moroni/JPL
Kevin Murphy/NASA
Charles Norton/JPL
Jean-Francois Piolle/IFREMER
Nga Quach/JPL
Brandi Quam/NASA
Shawn Smith/FSU
Ben Smith/JPL
Adam Stallard/FSU
Rob Toaz/JPL
Vardis Tsontos/JPL
Suresh Vannan/JPL
Jorge Vazquez/JPL
Ou Wang/JPL
Brian Wilson/JPL
Steve Worley/NCAR
Elizabeth Yam/JPL
Phil Yang/GMU
Alice Yepremyan/JPL

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

Big Data Analytics and Cloud Computing Workshop, 2017 ESIP Summer Meeting, Bloomington, IN

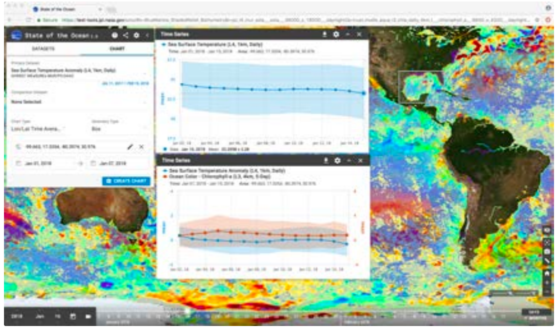
2019 JPL Data Science Showcase

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 Sea Level Change Team

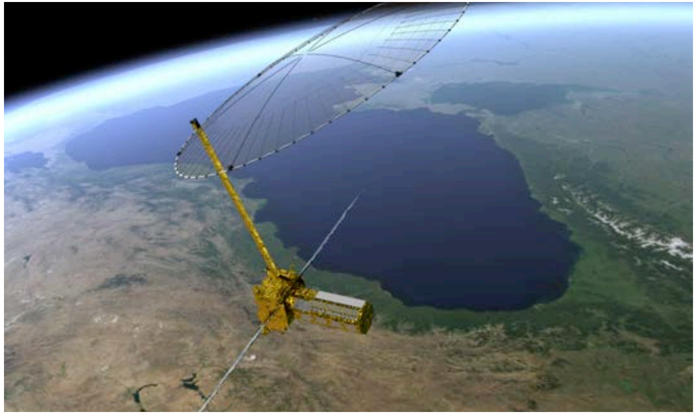
CEOS SIT Technical Workshop

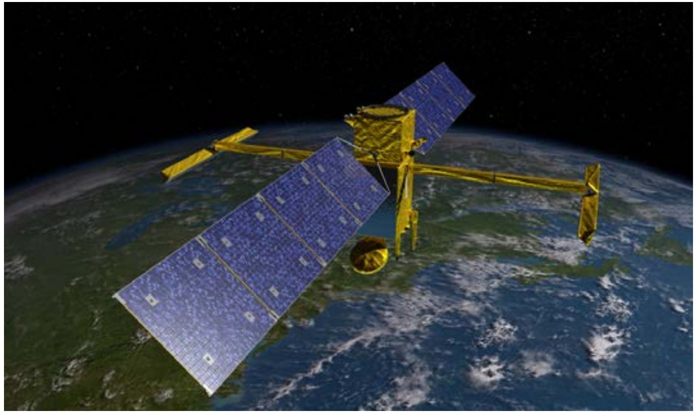


NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)

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