National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Analysis Optimized Data Storage in Apache Science Data Analytics Platform

Thomas Huang
thomas.huang@jpl.nasa.gov

Group Supervisor - Computer Science for Data-Intensive Applications
Strategic Lead - Interactive Data Analytics

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

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsements by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

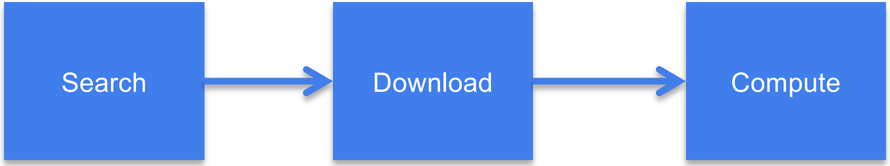
Background

- **NASA has historically focused on systematic capture and stewardship of data for observational Systems**
- **With large amount of observational and modeling data, finding and downloading is becoming inefficient**
- **Reality with large amount of observational and modeling data**
  - Downloading to local machine is becoming inefficient
  - Search has gotten a lot faster. Too many matches.
  - Finding the relevant measurement has becoming a very time consuming process ”*Which SST dataset I should use?*”
  - Analyze decades of regional measurement is labor-intensive and costly
- **Increasing “big data” era is driving needs to**
  - 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

Data Centers

- **Mainly focus on archives and distributions**
- **With additional services**
  - Better searches – faceted, spatial, keyword, ranking, etc.
  - Data subsetting – home grown, OPeNDAP, etc.
  - Visualization – visual discovery, PO.DAAC's SOTO, NASA Worldview, etc.
- **Limitations**
  - Little to no interoperability between tools and services: metadata standard, keyword, spatial coverage (0-360 or -180..180), temporal representation, etc.
  - Making sure the most relevant measurements return first
  - Visualization is nice, but it doesn't provide enough information about the event/phenomenon captured in the image.
  - With large amount of observational data, data centers need to do more than just storing bits
    - "Is the red blob in the middle of Pacific normal this time of the year?"
    - "Any relevant news and publications relate to what I am looking at?"
    - ”What other measurements, phenomena, news, publications relate to the period and location I am looking at?"
    - "I can see the observation from satellite, are there any relevant in situ data I can look at?"

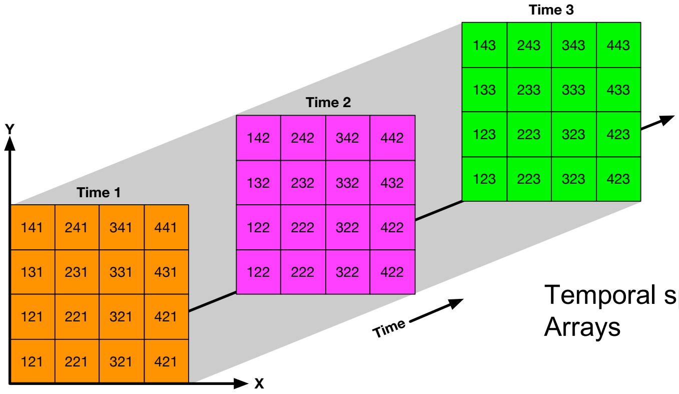
Traditional Method for Analyze Satellite Measurements



Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck

- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files



Time 1

141	241	341	441
131	231	331	431
121	221	321	421
121	221	321	421

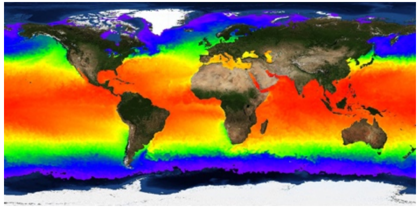
Time 2

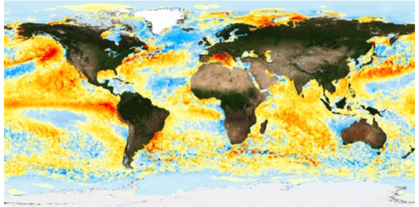
142	242	342	442
132	232	332	432
122	222	322	422
122	222	322	422

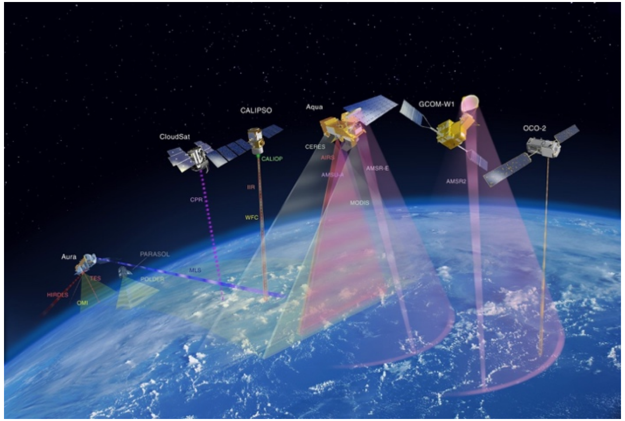
Time 3

143	243	343	443
133	233	333	433
123	223	323	423
123	223	323	423

Temporal spatial Arrays

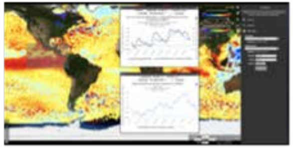




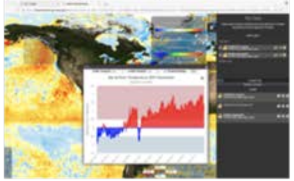


Enabling Next Generation of Earth Science Tools and Services

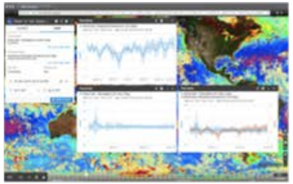
NASA Sea Level Change Portal



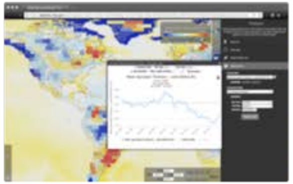
Oceanographic Anomaly Detection



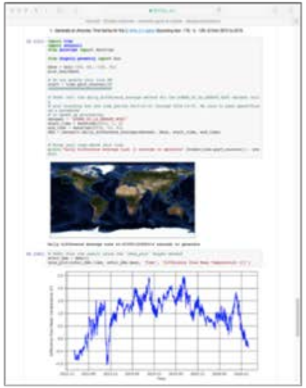
PO.DAAC State Of The Ocean



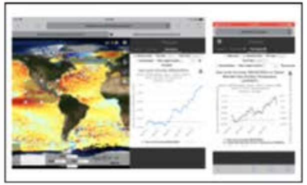
Hydrological Basin Analysis



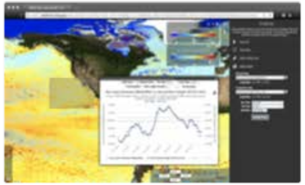
Jupyter Notebook - Interactive Workbench



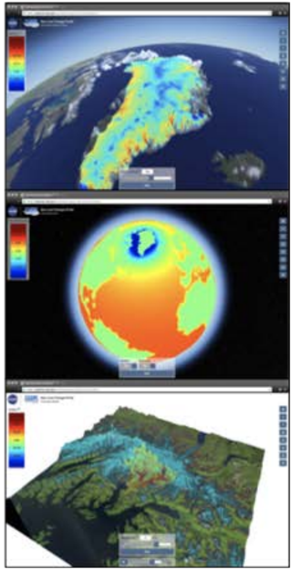
Mobile Analysis



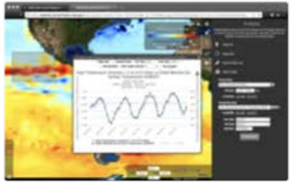
In Situ Data Analysis



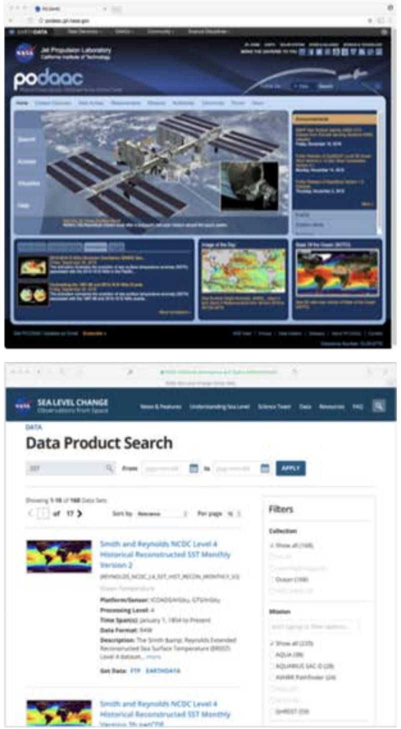
Model Simulations



Model - Observation Comparison



Integrated Search and Discovery



DATA

Data Product Search

SST From to

Showing 1-10 of 168 Data Sets

of 17 Sort by Relevance Per page 10

Smith and Reynolds NCDC Level 4 Historical Reconstructed SST Monthly Version 2

Platform/Sensor: ICOADS/InSitu, GTS/InSitu

Processing Level: 4

Time Span(s): January 1, 1854 to Present

Data Format: RAW

Description: The Smith & amp; Reynolds Extended Reconstructed Sea Surface Temperature (ERSST) Level 4 dataset... more

Get Data: FTP EARTHDATA

Smith and Reynolds NCDC Level 4 Historical Reconstructed SST Monthly

Collection

✓ Show all (168)

Ocean (168)

Mission

✓ Show all (220)

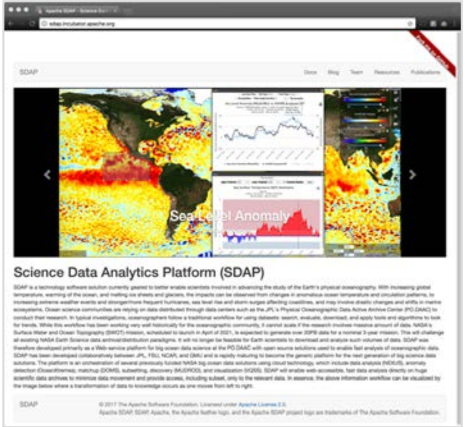
AQUA (38)

AQUARIUS SAC-D (28)

AVHRR Pathfinder (24)

Managing Open Source

- After more than two years of active development, on October 2017 the **NASA ESOT/AIST OceanWorks** team established Apache Software Foundation and established the **Science Data Analytics Platform (SDAP)** in the **Apache Incubator**
- 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

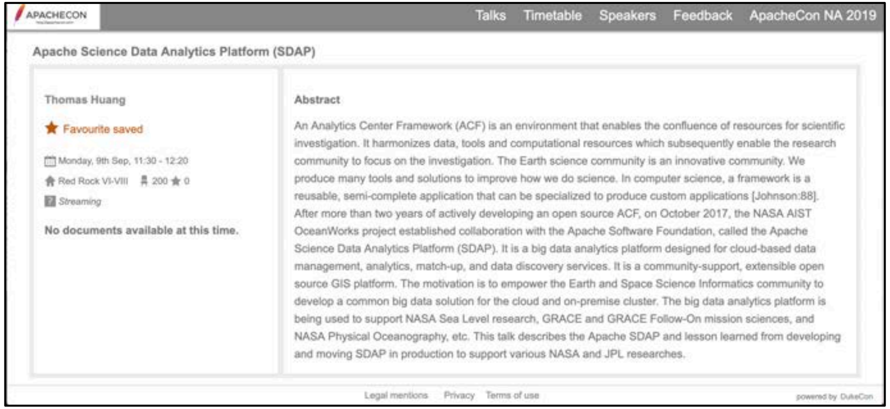


Science Data Analytics Platform (SDAP)

SDAP is a technology software solution currently geared to better enable scientists involved in advancing the study of the Earth's physical oceanography. With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, the impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and stronger/more frequent hurricanes, sea level rise and storm surges affecting coastlines, and may involve drastic changes and shifts in marine ecosystems. Ocean science communities are relying on data distributed through data centers such as the JPL's Physical Oceanographic Data Active Archive Center (PO.DAAC) to conduct their research. In typical investigations, oceanographers follow a traditional workflow for using datasets: search, evaluate, download, and apply tools and algorithms to look for trends. While this workflow has been working very well historically for the oceanographic community, it cannot scale if the research involves massive amount of data. NASA's Surface Water and Ocean Topography (SWOT) mission, scheduled to launch in April of 2021, is expected to generate over 20PB data for a nominal 3-year mission. This will challenge all existing NASA Earth Science data archival/distribution paradigms. It will no longer be feasible for Earth scientists to download and analyze such volumes of data. SDAP was therefore developed primarily as a Web-service platform for big ocean data science at the PO.DAAC with open source solutions used to enable fast analysis of oceanographic data. SDAP has been developed collaboratively between JPL, FSU, NCAR, and GMU and is rapidly maturing to become the generic platform for the next generation of big science data solutions. The platform is an orchestration of several previously funded NASA big ocean data solutions using cloud technology, which include data analysis (NEXUS), anomaly detection (OceanXtremes), matchup (DOMS), subsetting, discovery (MUDROD), and visualization (VQSS). SDAP will enable web-accessible, fast data analysis directly on huge scientific data archives to minimize data movement and provide access, including subset, only to the relevant data. In essence, the above information workflow can be visualized by the image below where a transformation of data to knowledge occurs as one moves from left to right.

SDAP © 2017 The Apache Software Foundation. Licensed under Apache License 2.0.
Apache SDAP, SDAP, Apache, the Apache feather logo, and the Apache SDAP project logo are trademarks of The Apache Software Foundation.

http://sdap.apache.org



Apache Science Data Analytics Platform (SDAP)

Thomas Huang

★ Favourite saved

Monday, 9th Sep, 11:30 - 12:20

Red Rock VI-VIII 200 ★ 0

Streaming

No documents available at this time.

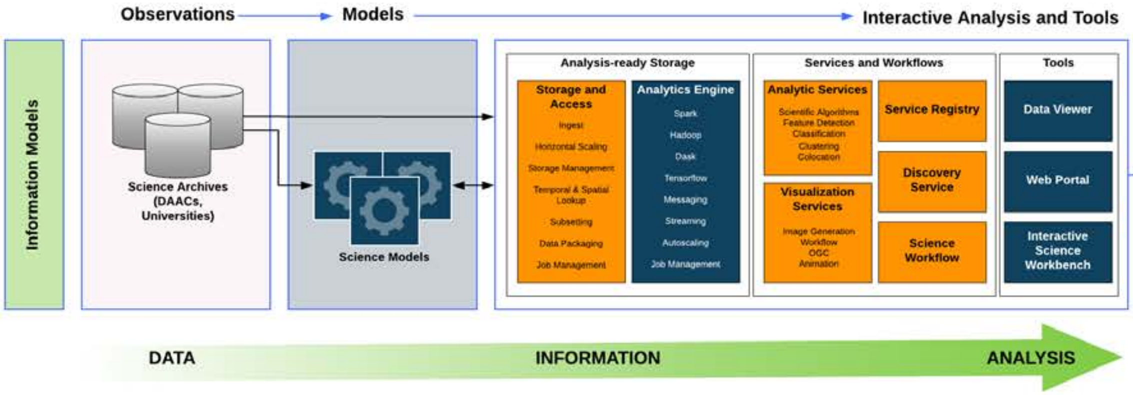
Abstract

An Analytics Center Framework (ACF) is an environment that enables the confluence of resources for scientific investigation. It harmonizes data, tools and computational resources which subsequently enable the research community to focus on the investigation. The Earth science community is an innovative community. We produce many tools and solutions to improve how we do science. In computer science, a framework is a reusable, semi-complete application that can be specialized to produce custom applications [Johnson:88]. After more than two years of actively developing an open source ACF, on October 2017, the NASA AIST OceanWorks project established collaboration with the Apache Software Foundation, called the Apache Science Data Analytics Platform (SDAP). It is a big data analytics platform designed for cloud-based data management, analytics, match-up, and data discovery services. It is a community-support, extensible open source GIS platform. The motivation is to empower the Earth and Space Science Informatics community to develop a common big data solution for the cloud and on-premise cluster. The big data analytics platform is being used to support NASA Sea Level research, GRACE and GRACE Follow-On mission sciences, and NASA Physical Oceanography, etc. This talk describes the Apache SDAP and lesson learned from developing and moving SDAP in production to support various NASA and JPL researches.

Legal mentions Privacy Terms of use powered by DukeCon

Integrated Science Data Analytics Platform

SaaS and PaaS for Science Tools and Services

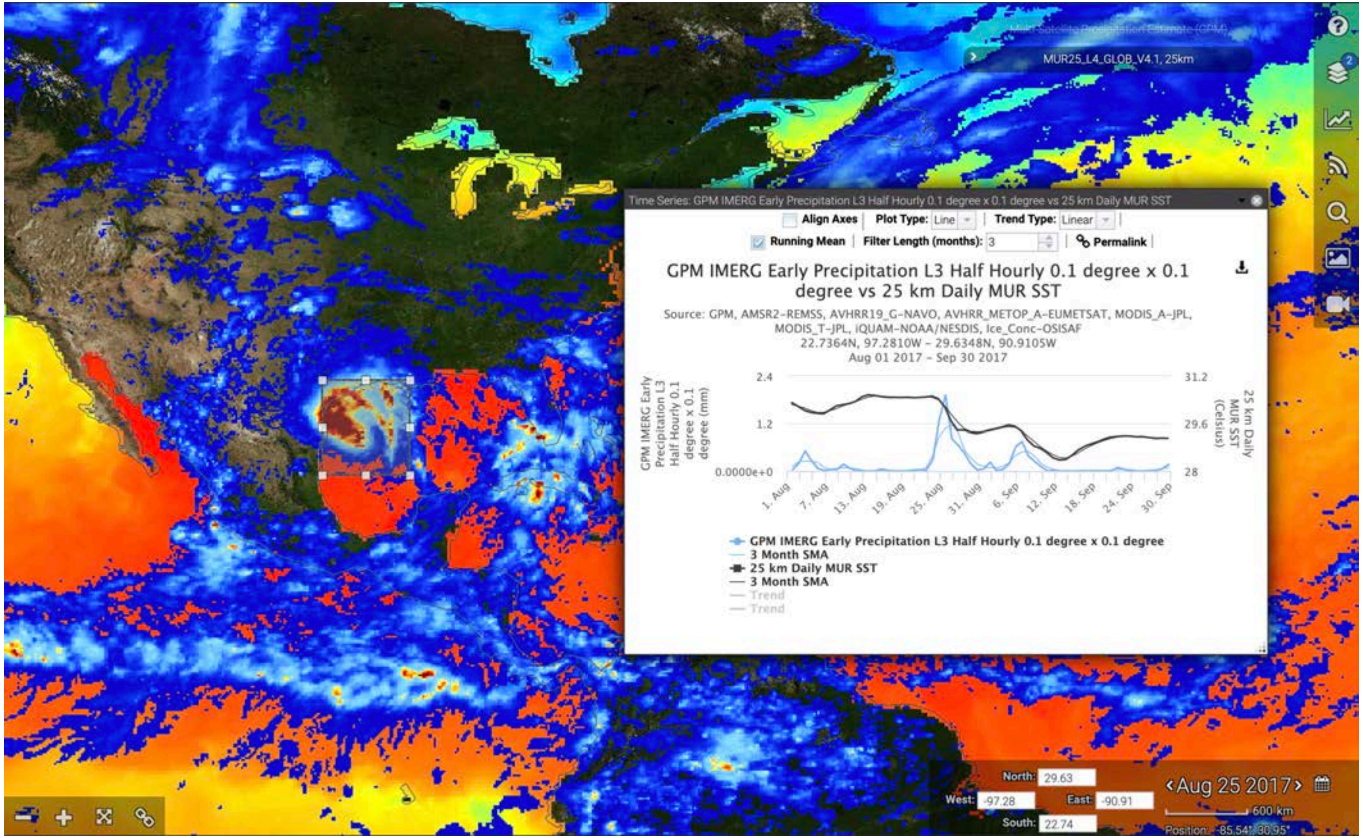


Systematic analysis of massive data

- **An effort funded by the NASA's Advanced Information Systems Technology (AIST) program**
- **Integrated Science Data Analytics Platform**: an analytic center framework to provide an environment for conducting a 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 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

Analyze Hurricane Harvey using GPM and SST

Aug 17, 2017 – Sept. 2, 2017



Multi-Satellite Precipitation Estimate (GPM)

MUR25_L4_GLOB_V4.1, 25km

Time Series: GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree vs 25 km Daily MUR SST

□ Align Axes | Plot Type: Line | Trend Type: Linear |

☑ Running Mean | Filter Length (months): 3 | Permalink |

GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree vs 25 km Daily MUR SST

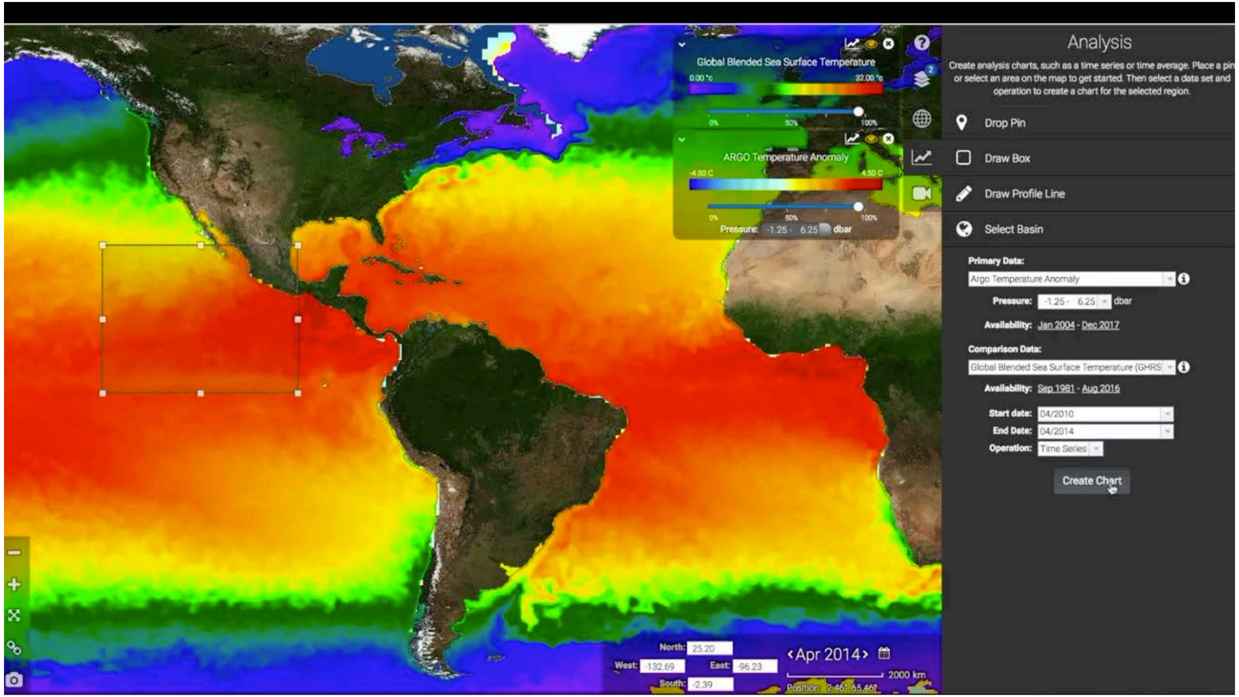
Source: GPM, AMSR2-REMSS, AVHRR19_G-NAVO, AVHRR_METOP_A-EUMETSAT, MODIS_A-JPL, MODIS_T-JPL, iQUAM-NOAA/NESDIS, Ice_Conc-OSISAF
22.7364N, 97.2810W - 29.6348N, 90.9105W
Aug 01 2017 - Sep 30 2017

GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree (mm)
2.4
1.2
0.0000e+0
31.2
29.6
28
25 km Daily MUR SST (Celsius)
1. Aug
7. Aug
13. Aug
19. Aug
25. Aug
31. Aug
6. Sep
12. Sep
18. Sep
24. Sep
30. Sep
GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree
3 Month SMA
25 km Daily MUR SST
3 Month SMA
Trend
Trend

North: 29.63
West: -97.28
East: -90.91
South: 22.74

‹Aug 25 2017›

Example Interactive Analysis: Visualize and Analyze Sea Level



North: 25.20
West: -132.69
East: -96.23
South: -2.39

Analysis

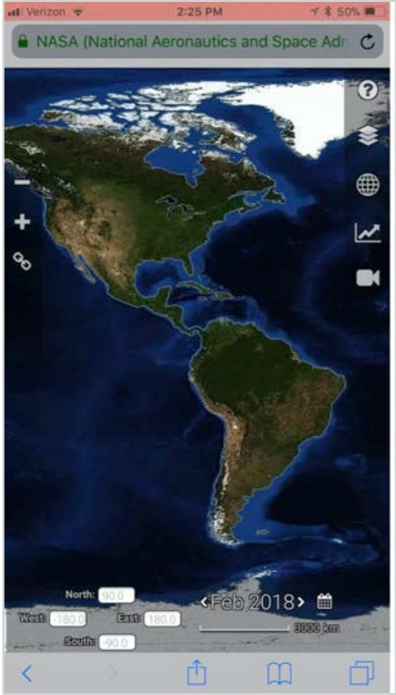
Create analysis charts, such as a time series or time average. Place a pin or select an area on the map to get started. Then select a data set and operation to create a chart for the selected region.

Primary Data: Argo Temperature Anomaly
Pressure: -1.25 - 6.25 dbar
Availability: Jan 2004 - Dec 2017

Comparison Data: Global Blended Sea Surface Temperature (GHRS
Availability: Sep 1981 - Aug 2016

Start date: 04/2010
End Date: 04/2014
Operation: Time Series

Analyze *in situ* and satellite observations

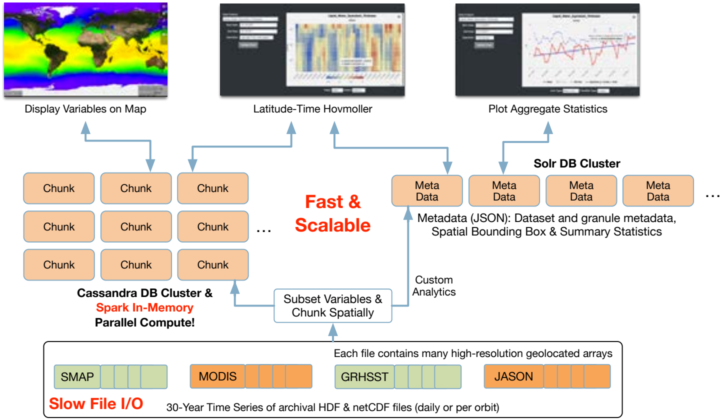


North: 90.0
West: -180.0
East: 180.0
South: -90.0

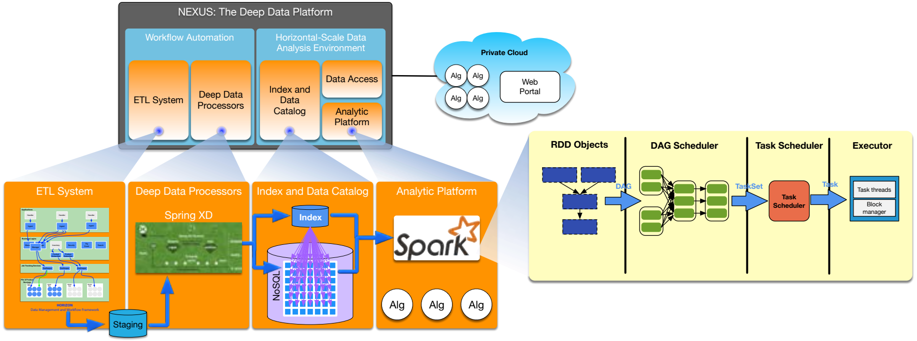
Analyze Sea Level
on mobiles

Scalable Data Analytic Solution

- SDAP’s analytics engine (a.k.a NEXUS) is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
- Streaming architecture for horizontal scale data ingestion
- Scales horizontally to handle massive amount of data in parallel
- Provides high-performance geospatial and indexed search solution
- Provides tiled data storage architecture to eliminate file I/O overhead
- A growing collection of science analysis webservices using Apache Spark: parallel compute, in-memory map-reduce framework
- Pre-Chunk and Summarize Key Variables
  - Easy statistics instantly (milliseconds)
  - Harder statistics on-demand using Spark (in seconds)
  - Visualize original data (layers) on a map quickly (Cassandra store)
- **Algorithms** – Time Series | Latitude/Time Hovmöller| Longitude/Time Hovmöller| Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average



Two-Database Architecture



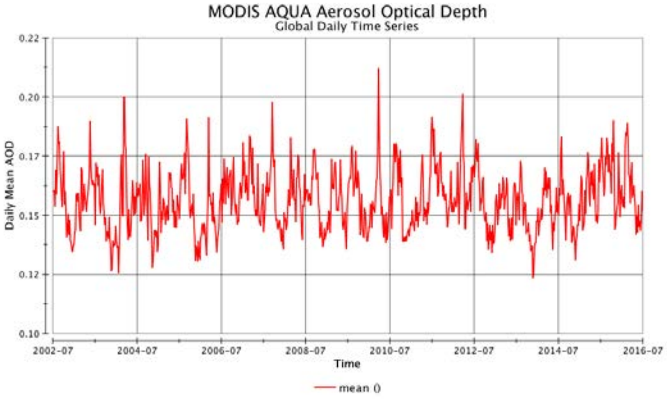
NEXUS Performance: File-based vs. Custom Spark vs. AWS EMR

Dataset: MODIS AQUA Daily
Name: Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
File Count: 5106
Volume: 2.6GB
Time Coverage: July 4, 2002 – July 3, 2016

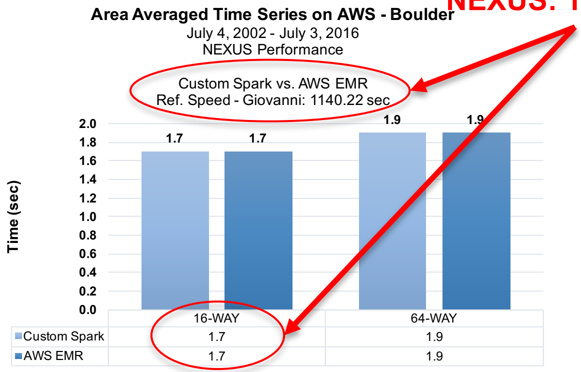
File-based: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

AWS EMR: Amazon’s provisioned MapReduce cluster

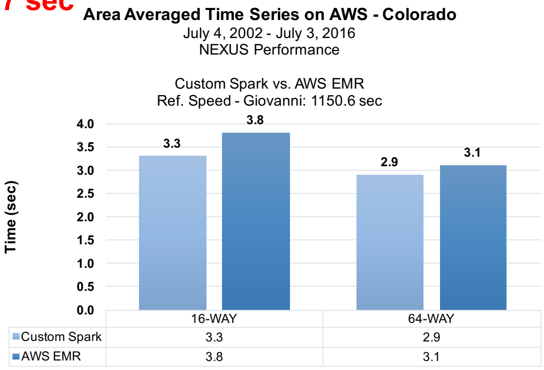


File-based: 20 min
NEXUS: 1.7 sec


Area Averaged Time Series on AWS - Boulder
July 4, 2002 - July 3, 2016
NEXUS Performance

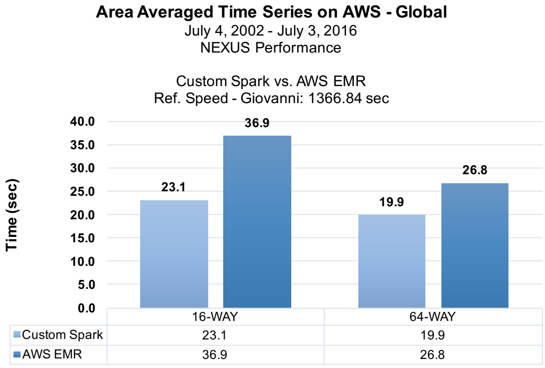
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1140.22 sec

	16-WAY	64-WAY
Custom Spark	1.7	1.9
AWS EMR	1.7	1.9


Area Averaged Time Series on AWS - Colorado
July 4, 2002 - July 3, 2016
NEXUS Performance

Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1150.6 sec

	16-WAY	64-WAY
Custom Spark	3.3	2.9
AWS EMR	3.8	3.1

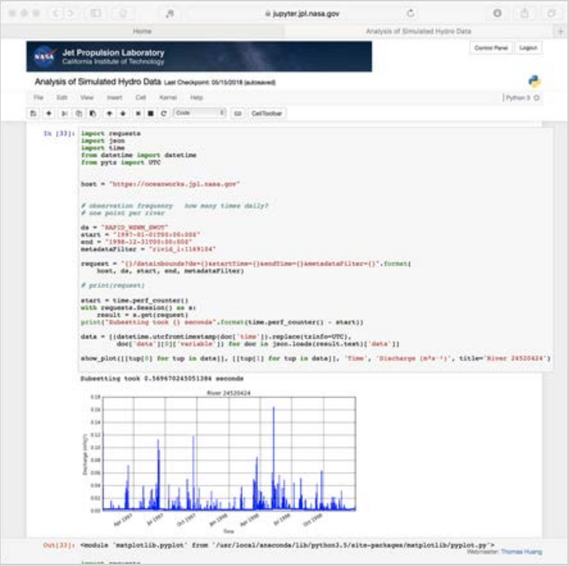

Area Averaged Time Series on AWS - Global
July 4, 2002 - July 3, 2016
NEXUS Performance

Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1366.84 sec

	16-WAY	64-WAY
Custom Spark	23.1	19.9
AWS EMR	36.9	26.8

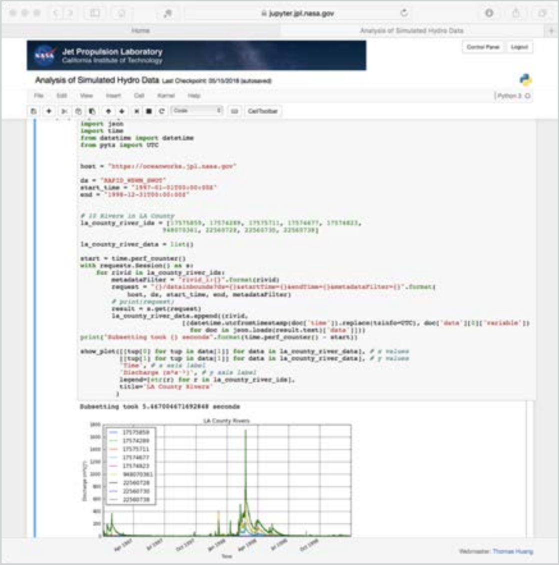
Algorithm execution time

Analyze Large Collection of Observational Data Directly … across the ocean


jupyter.jpl.nasa.gov
Home
Analysis of Simulated Hydro Data
Jet Propulsion Laboratory
California Institute of Technology
Control Panel
Logout
Analysis of Simulated Hydro Data Last Checkpoint: 05/15/2018 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
Python 3
Code
CellToolbar
River 24520424
Time
Discharge (m³s⁻¹)
Webmaster: Thomas Huang

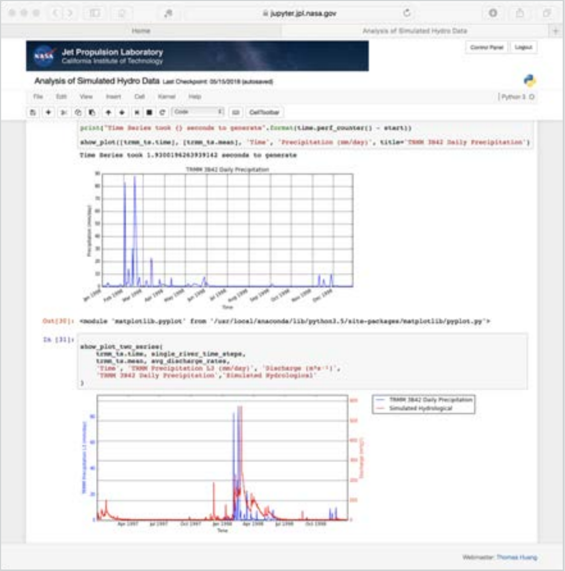
```
In [33]: import requests
         import json
         import time
         from datetime import datetime
         from pytz import UTC


         host = "https://oceanworks.jpl.nasa.gov"


         # observation frequency   how many times daily?
         # one point per river

         ds = "RAPID_WSWM_SWOT"
         start = "1997-01-01T00:00:00Z"
         end = "1998-12-31T00:00:00Z"
         metadataFilter = "rivid_i:1169104"

         request = "{}/datainbounds?ds={}&startTime={}&endTime={}&metadataFilter={}".format(
             host, ds, start, end, metadataFilter)

         # print(request)

         start = time.perf_counter()
         with requests.Session() as s:
             result = s.get(request)
         print("Subsetting took {} seconds".format(time.perf_counter() - start))

         data = [(datetime.utcfromtimestamp(doc['time']).replace(tzinfo=UTC),
                  doc['data'][0]['variable']) for doc in json.loads(result.text)['data']]

         show_plot([[tup[0] for tup in data]], [[tup[1] for tup in data]], 'Time', 'Discharge (m³s⁻¹)', title='River 24520424')

Subsetting took 0.569670245051384 seconds

Out[33]: <module 'matplotlib.pyplot' from '/usr/local/anaconda/lib/python3.5/site-packages/matplotlib/pyplot.py'>
```

Retrieval of a single river time series



```
import json
import time
from datetime import datetime
from pytz import UTC


host = "https://oceanworks.jpl.nasa.gov"

ds = "RAPID_WSWM_SWOT"
start_time = "1997-01-01T00:00:00Z"
end = "1998-12-31T00:00:00Z"


# 12 Rivers in LA County
la_county_river_ids = [17575859, 17574289, 17575711, 17574677, 17574823,
                       948070361, 22560728, 22560730, 22560738]

la_county_river_data = list()

start = time.perf_counter()
with requests.Session() as s:
    for rivid in la_county_river_ids:
        metadataFilter = "rivid_i:{}".format(rivid)
        request = "{}/datainbounds?ds={}&startTime={}&endTime={}&metadataFilter={}".format(
            host, ds, start_time, end, metadataFilter)
        # print(request)
        result = s.get(request)
        la_county_river_data.append((rivid,
                                     [(datetime.utcfromtimestamp(doc['time']).replace(tzinfo=UTC), doc['data'][0]['variable'])
                                      for doc in json.loads(result.text)['data']]))
print("Subsetting took {} seconds".format(time.perf_counter() - start))

show_plot([[tup[0] for tup in data[1]] for data in la_county_river_data], # x values
          [[tup[1] for tup in data[1]] for data in la_county_river_data], # y values
          'Time', # x axis label
          'Discharge (m³s⁻¹)', # y axis label
          legend=[str(r) for r in la_county_river_ids],
          title='LA County Rivers'
         )

Subsetting took 5.467004671692848 seconds
```

Retrieval of time series from 9 rivers



```
print("Time Series took {} seconds to generate".format(time.perf_counter() - start))

show_plot([trmm_ts.time], [trmm_ts.mean], 'Time', 'Precipitation (mm/day)', title='TRMM 3B42 Daily Precipitation')

Time Series took 1.9300196263939142 seconds to generate

Out[30]: <module 'matplotlib.pyplot' from '/usr/local/anaconda/lib/python3.5/site-packages/matplotlib/pyplot.py'>

In [31]: 
         show_plot_two_series(
             trmm_ts.time, single_river_time_steps,
             trmm_ts.mean, avg_discharge_rates,
             'Time', 'TRMM Precipitation L3 (mm/day)', 'Discharge (m³s⁻¹)',
             'TRMM 3B42 Daily Precipitation','Simulated Hydrological'
         )
```

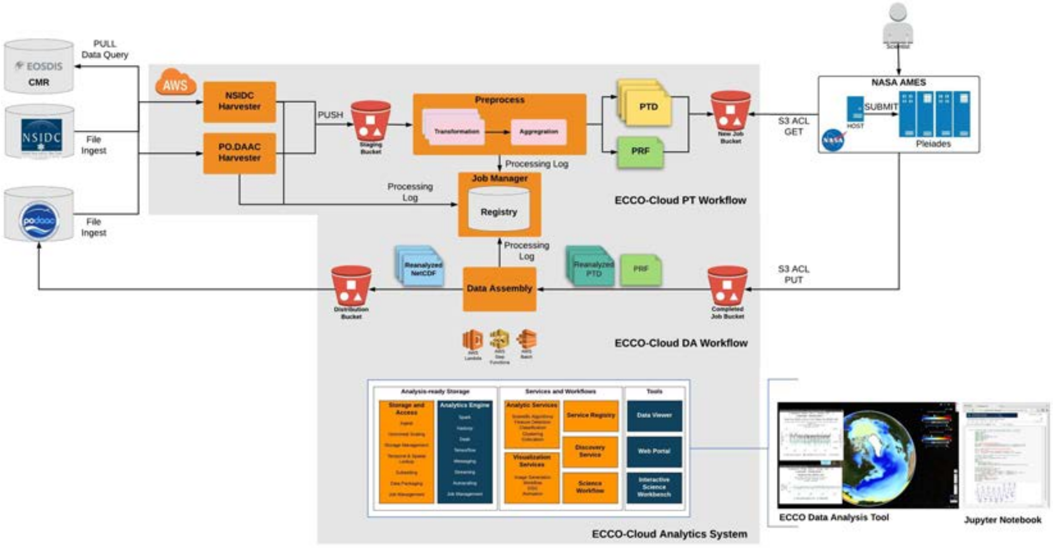
Time series coordination between TRMM and river

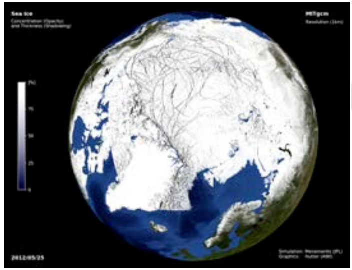
- Running Jupyter from Germany and interacts with analytics services hosted on Amazon and at JPL
- Simulated hydrology data in preparation for SWOT hydrology
- **River data**: ~3.6 billion data points. 3-hour sample rate. Consists of measurements from ~600,000 rivers
- **TRMM data**: 17 years, .25deg, 1.5 billion data points
- Sub-second retrieval of river measurements
- On-the-fly computation of time series and generate coordination plot

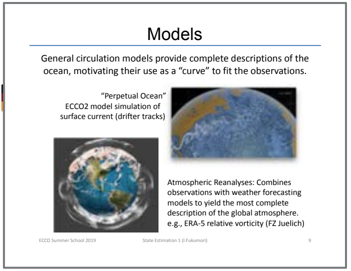
Data Access and the ECCO Ocean and Ice State Estimate

NASA ACCESS Program | PI: Patrick Heimbach; Co-Is: Ian Fenty and Thomas Huang

- **Estimating the Circulation and Climate of the Ocean** (ECCO) is a consortium endeavors to produce the best possible estimates of ocean circulation and its role in climate
- Combining state-of-the-art ocean circulation models with global ocean and sea-ice data in a physically and statistically consistent manner
- ECCO products are being used in studies on ocean variability, biological cycles, coastal physics, water cycle, ocean-cryosphere interactions, and geodesy
- **Goals**
  - Expand and accelerate in a sustainable and scalable manner the integration of NASA Earth system data into ECCO through automated preprocessing and transformation
  - Automate generation of ECCO reanalysis products into CF-compliant NetCDF products
  - Radically streamline the integration of updated ECCO products into NASA's Earth Observing System Data and Information System (EOSDIS)







Models

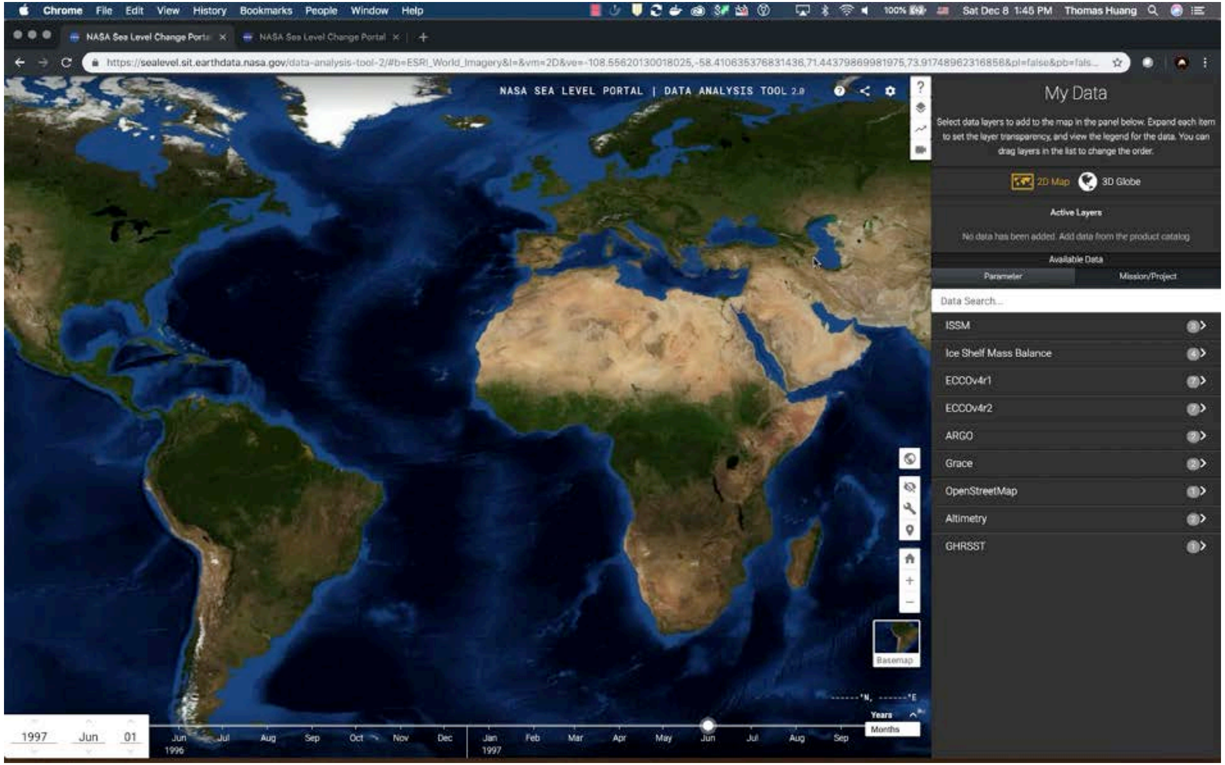
General circulation models provide complete descriptions of the ocean, motivating their use as a "curve" to fit the observations.

"Perpetual Ocean"
ECCO2 model simulation of
surface current (drifter tracks)

Atmospheric Reanalyses: Combines observations with weather forecasting models to yield the most complete description of the global atmosphere.
e.g., ERA-5 relative vorticity (FZ Juelich)

ECCO Summer School 2019 State Estimation 1 (I.Fukumori) 9

Interactive Analysis of ECCO Products


Chrome File Edit View History Bookmarks People Window Help
100% Sat Dec 8 1:45 PM Thomas Huang
NASA Sea Level Change Portal
NASA Sea Level Change Portal
https://sealevel.sit.earthdata.nasa.gov/data-analysis-tool-2/#b=ESRI_World_Imagery&l=&vm=2D&ve=-108.55620130018025,-58.410635376831436,71.44379869981975,73.91748962316858&pl=false&pb=fals...
NASA SEA LEVEL PORTAL | DATA ANALYSIS TOOL 2.0
Basemap
Years
Months
1997 Jun 01
Jun 1996 Jul Aug Sep Oct Nov Dec Jan 1997 Feb Mar Apr May Jun Jul Aug Sep

My Data

Select data layers to add to the map in the panel below. Expand each item to set the layer transparency, and view the legend for the data. You can drag layers in the list to change the order.

2D Map | 3D Globe

Active Layers

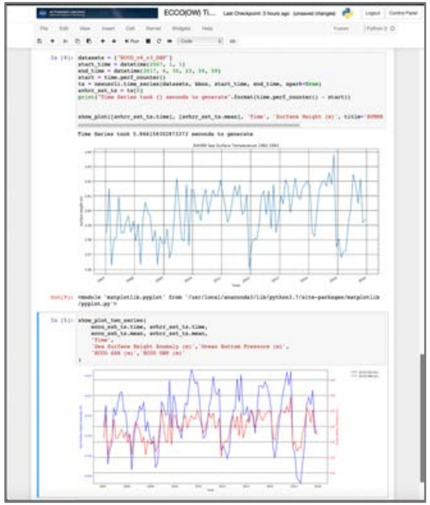
No data has been added. Add data from the product catalog

Available Data

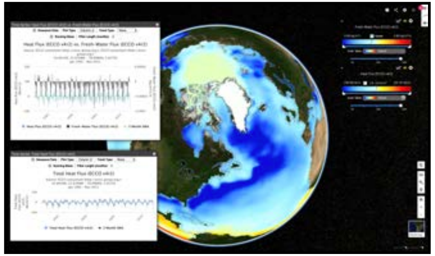
Parameter | Mission/Project

Data Search...

- ISSM
- Ice Shelf Mass Balance
- ECCOv4r1
- ECCOv4r2
- ARGO
- Grace
- OpenStreetMap
- Altimetry
- GHRSST

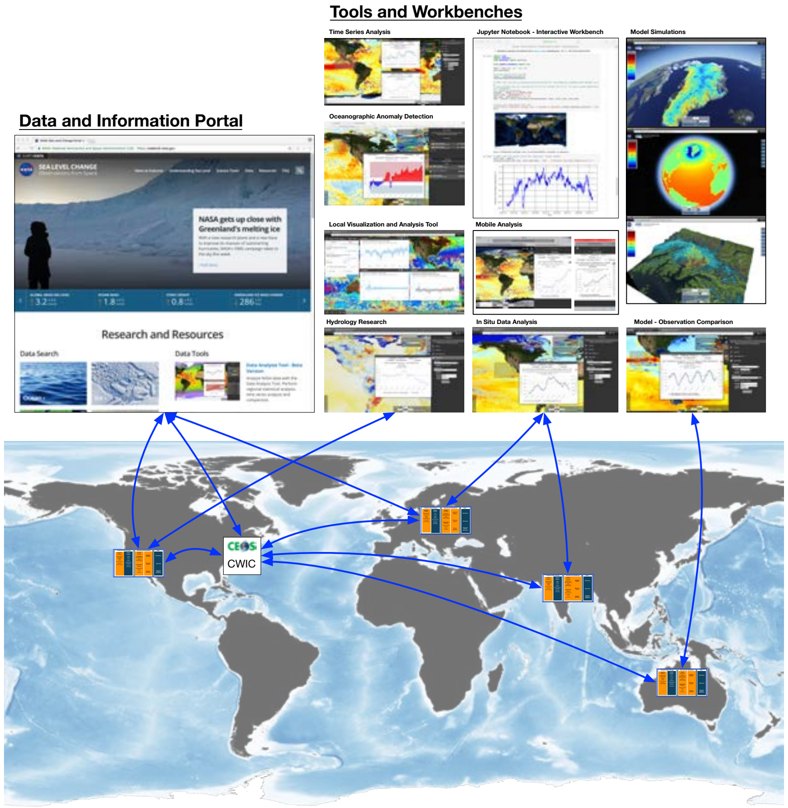

ECCO(OW) Ti... Last Checkpoint: 3 hours ago (unsaved changes)
Logout Control Panel
File Edit View Insert Cell Kernel Widgets Help
Trusted Python 3
Run Code

```
In [9]: datasets = ["ECCO_v4_r3_OBP"]
        start_time = datetime(2007, 1, 1)
        end_time = datetime(2017, 4, 30, 23, 59, 59)
        start = time.perf_counter()
        ts = nexuscli.time_series(datasets, bbox, start_time, end_time, spark=True)
        avhrr_sst_ts = ts[0]
        print("Time Series took {} seconds to generate".format(time.perf_counter() - start))

        show_plot([avhrr_sst_ts.time], [avhrr_sst_ts.mean], 'Time', 'Surface Height (m)', title='AVHRR

        Time Series took 5.866158302873373 seconds to generate

Out[9]: <module 'matplotlib.pyplot' from '/usr/local/anaconda3/lib/python3.7/site-packages/matplotlib
        /pyplot.py'>

In [5]: show_plot_two_series(
            ecco_ssh_ts.time, avhrr_sst_ts.time,
            ecco_ssh_ts.mean, avhrr_sst_ts.mean,
            'Time',
            'Sea Surface Height Anomaly (m)','Ocean Bottom Pressure (m)',
            'ECCO SSH (m)','ECCO OBP (m)'
        )
```


Heat Flux (ECCO v4r2) vs. Fresh Water Flux (ECCO v4r2)
Total Heat Flux (ECCO v4r2)

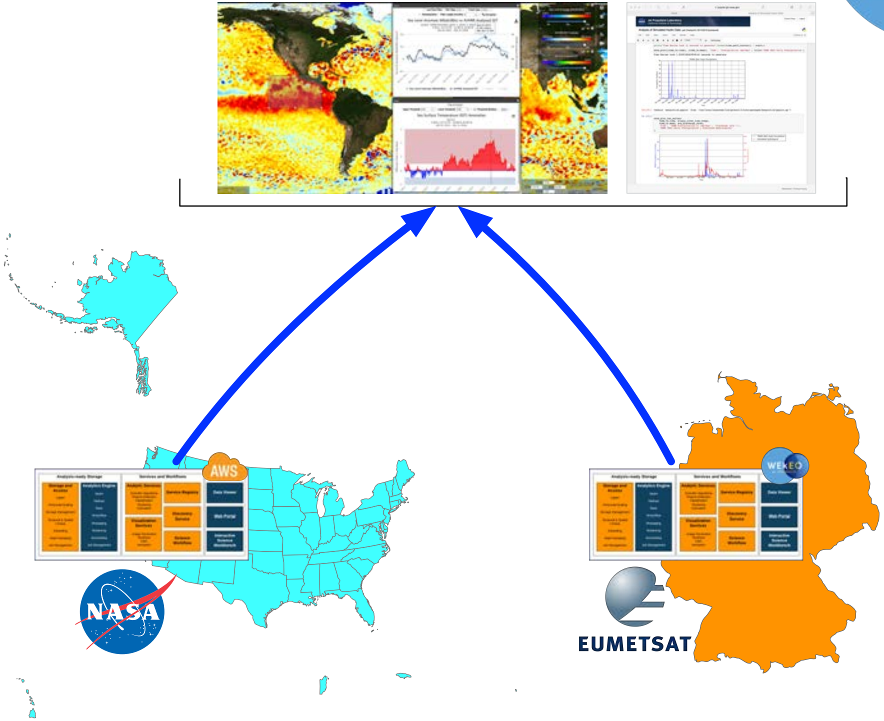
Distributed Analytics Center Architecture

- **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
- Public accessible RESTful analytic APIs where computation is next to the data
- 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



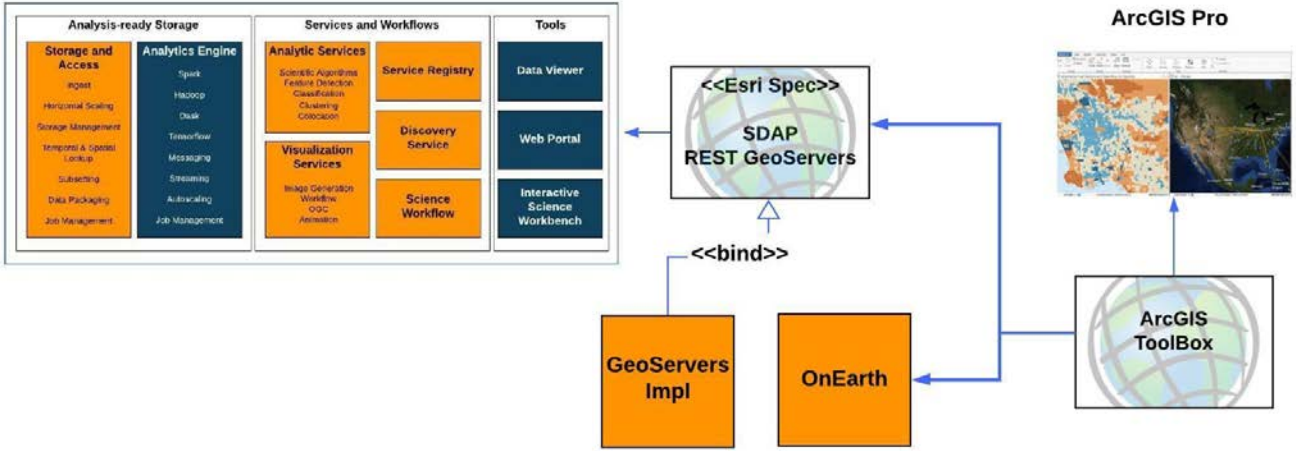
COVERAGE Phase B

- 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

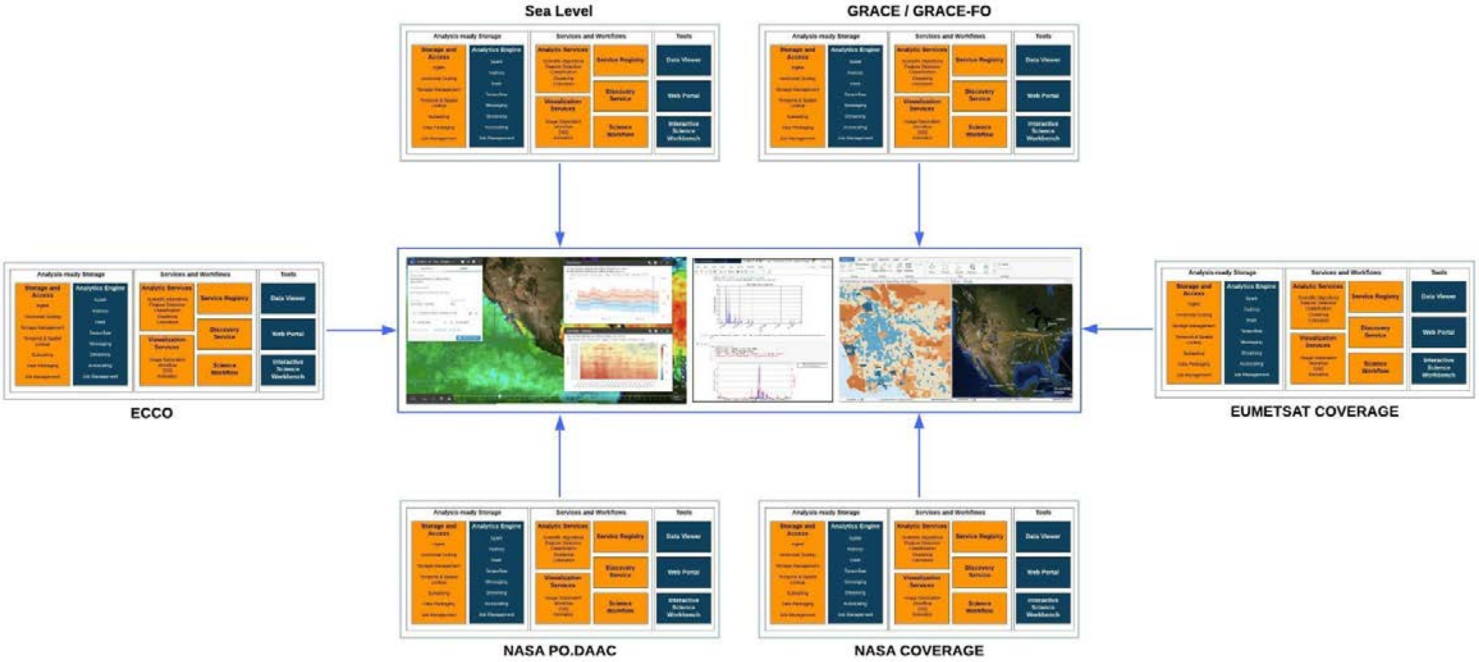


Integration with Esri ArcGIS

- NASA’s Advanced Information System Technology (AIST) effort for their big ocean science analytics solution using Apache SDAP (a.k.a OceanWorks)
- Enable scientists to use OceanWorks data and analytics within ArcGIS.
- Many scientists already use ArcGIS for their day-to-day work. Enabling them to use OceanWorks data and cloud analytics from within their familiar ArcGIS tools will enable them to perform analyses that cut across several NASA Ocean science datasets, and will expand the reach and impact of the OceanWorks data analytic system within the Ocean Science community



Connecting to a Federation of SDAPs



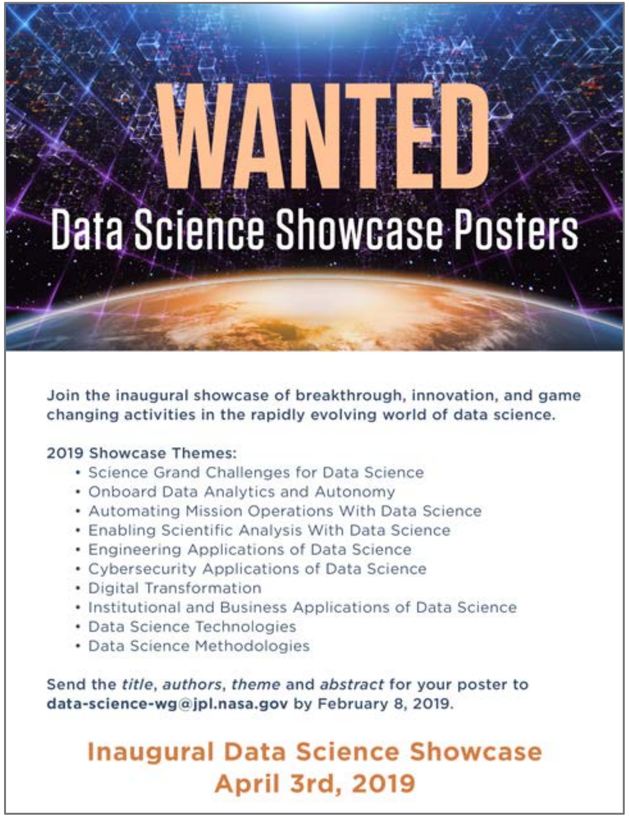
Instances of Apache SDAP already in production or will be in production within the next 12 months

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

2019 EGU – NASA Hyperwall

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



WANTED

Data Science Showcase Posters

Join the inaugural showcase of breakthrough, innovation, and game changing activities in the rapidly evolving world of data science.

2019 Showcase Themes:

- Science Grand Challenges for Data Science
- Onboard Data Analytics and Autonomy
- Automating Mission Operations With Data Science
- Enabling Scientific Analysis With Data Science
- Engineering Applications of Data Science
- Cybersecurity Applications of Data Science
- Digital Transformation
- Institutional and Business Applications of Data Science
- Data Science Technologies
- Data Science Methodologies

Send the *title*, *authors*, *theme* and *abstract* for your poster to **data-science-wg@jpl.nasa.gov** by February 8, 2019.

**Inaugural Data Science Showcase
April 3rd, 2019**

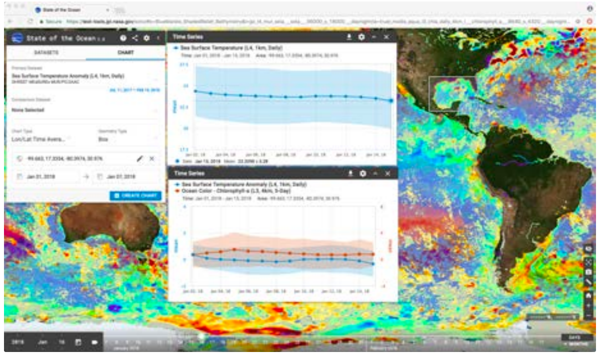
2019 JPL Data Science Showcase

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

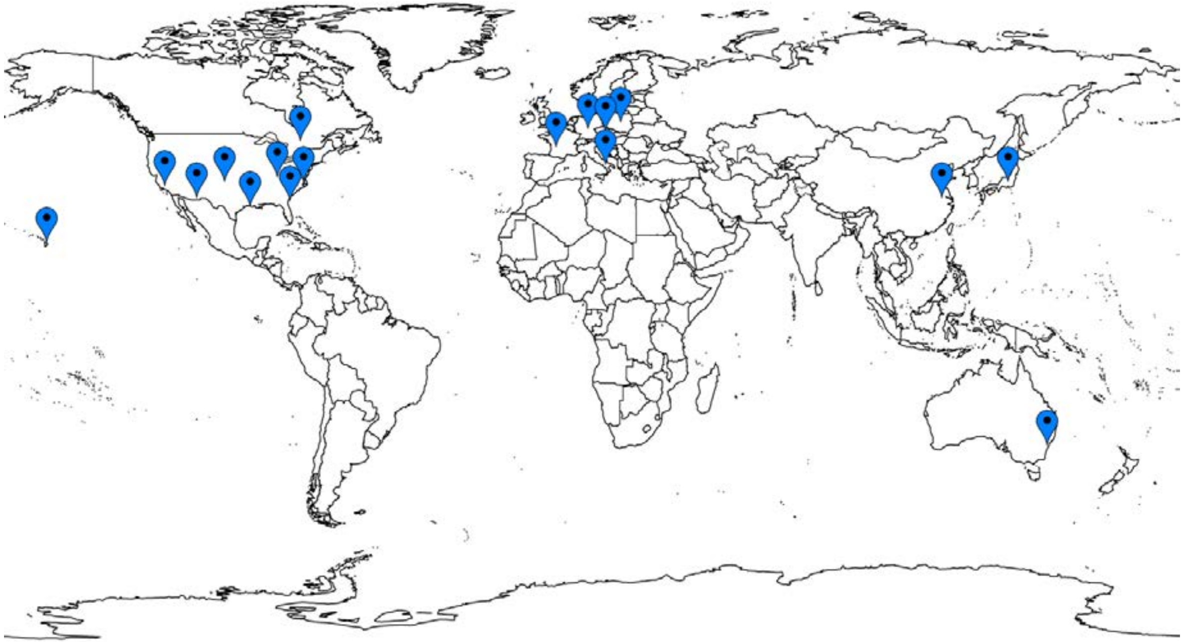
NASA's Sea Level Change Team

CEOS SIT Technical Workshop


State of the Ocean
DATASETS
CHART
Sea Surface Temperature Anomaly (L4, 1km, Daily)
None Selected
Jan 01, 2018
CREATE CHART
Time Series
Sea Surface Temperature (L4, 1km, Daily)
Time Series
Sea Surface Temperature Anomaly (L4, 1km, Daily)
Ocean Color - Chlorophyll-a (L3, 4km, 5-Day)

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

Building Bridges through Open Sourcing



Presentation | Evaluation | Collaboration

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
- Open source and a web of analytics centers should be the architecture for climate science
- Focus on delivering professional quality open source solutions that enables end-to-end data and computation architecture, and the total cost of ownership
- 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

If you want to go fast, go Alone. If you want to go far, go Together.

African Proverb

Thomas Huang

thomas.huang@jpl.nasa.gov

Jet Propulsion Laboratory

California Institute of Technology