National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Ocean Science Data Analytics using Apache Science Data Analytics Platform

Thomas Huang
Technical Group Supervisor

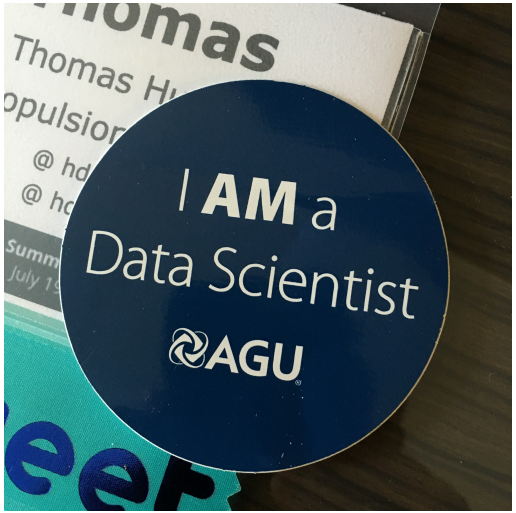
Computer Science for Data-Intensive Applications Group
Instrument Software and Science Data Systems Section

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

[CL #18-1340]

© 2018 California Institute of Technology. Government sponsorship acknowledged.
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.

@NASA/JPL



Technical Group Supervisor
NASA JPL's Computer Science for Data-Intensive Applications Group

Strategic Lead for Interactive Data Analytics
NASA JPL's Information & Data Science Programs

Principal Investigator
NASA AIST OceanWorks – Ocean Science Platform on Cloud

Co-Investigator and Architect
NASA Sea Level Change Portal

Architect
CEOS Ocean Variables Enabling Research and Application for GEOS (COVERAGE)

Cluster Chair
Federation of Earth Science Information Partners (ESIP) Cloud Computing

Lead Editor
2018 Wiley Book: **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**

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

Previously Architect
Tactical Data Science Framework for Naval Research

Previously Principal Investigator / Co-Investigator
Several NASA-funded Big Data Analytic Projects – Big Data Analytics on the Cloud, Anomaly Detection, In Situ and Satellite Matchup, Search Relevancy, and Quality Screening

THUANG/JPL © 2018. All rights reserved.

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

THUANG/JPL © 2018. All rights reserved.

Data Centers

- **Mainly focus on archives and distributions**
- **With additional services**
  - Better searches – faceted, spatial, keyword, ranking, etc.
  - Data subsetting – home grown, OPeNDAP, Webification, 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?"

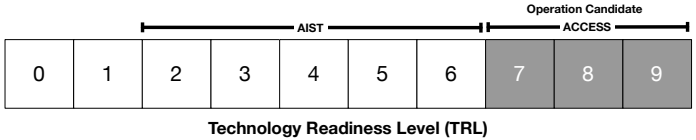
THUANG/JPL © 2018. All rights reserved.

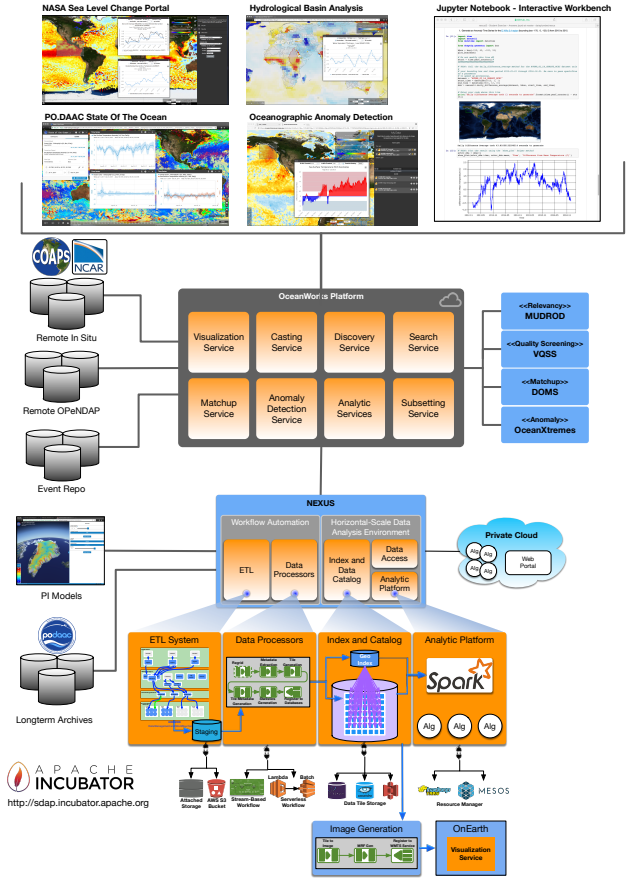
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA OceanWorks
PI: Thomas Huang/JPL

- Sponsored by the NASA's Advanced Information Systems Technology (AIST) Program
- **OceanWorks** is to establish an **Integrated Data Analytics Center** at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) for Big Ocean Science
- Focuses on technology integration, advancement and maturity
- Collaboration between JPL, Center for Atmospheric Prediction Studies (COAPS) at Florida State University (FSU), National Center for Atmospheric Research (NCAR), and George Mason University (GMU)
- Bringing together PO.DAAC-related big data technologies
  - Big data analytic platform
  - Anomaly detection and ocean science
  - Distributed in situ to satellite matchup
  - Dynamic datasets ranking and recommendations
  - Sub-second data search solution and metadata translation and services aggregation
  - Quality-screened data subsetting





THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

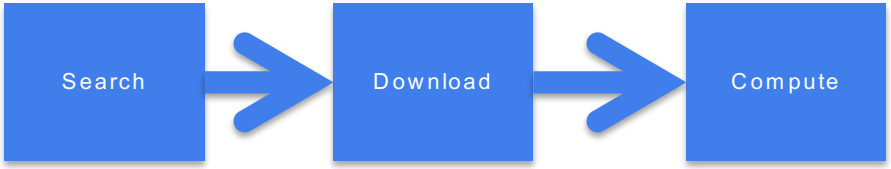
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Big Data Analytics Platform

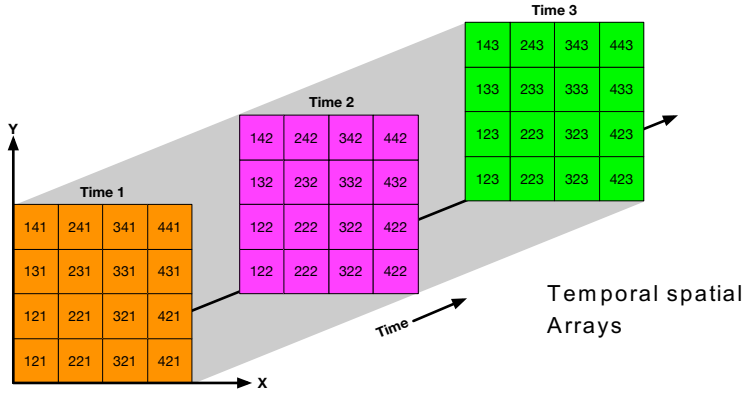
THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Traditional Method for Analyze Satellite Measurements



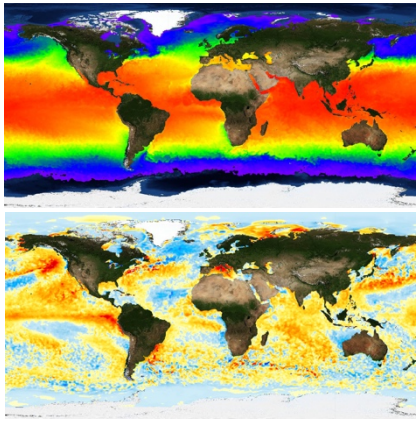
- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced,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

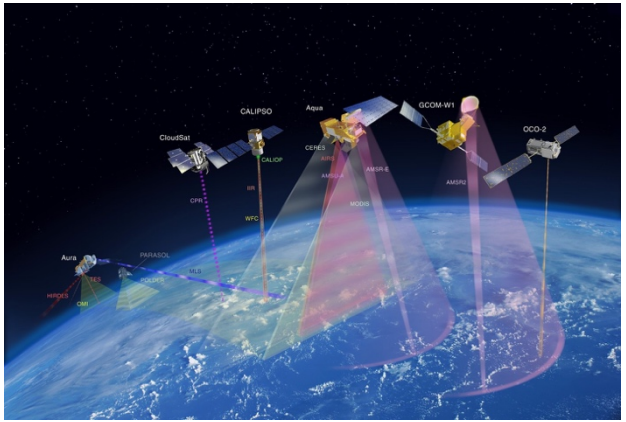


Temporal spatial Arrays

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





THUANG/JPL © 2018. All rights reserved.

NASA's Upcoming Big Data Mission: Surface Water and Ocean Topography (SWOT)

Oceanography: Characterize the ocean mesoscale and sub-mesoscale circulation at spatial resolutions of 10 km and greater.

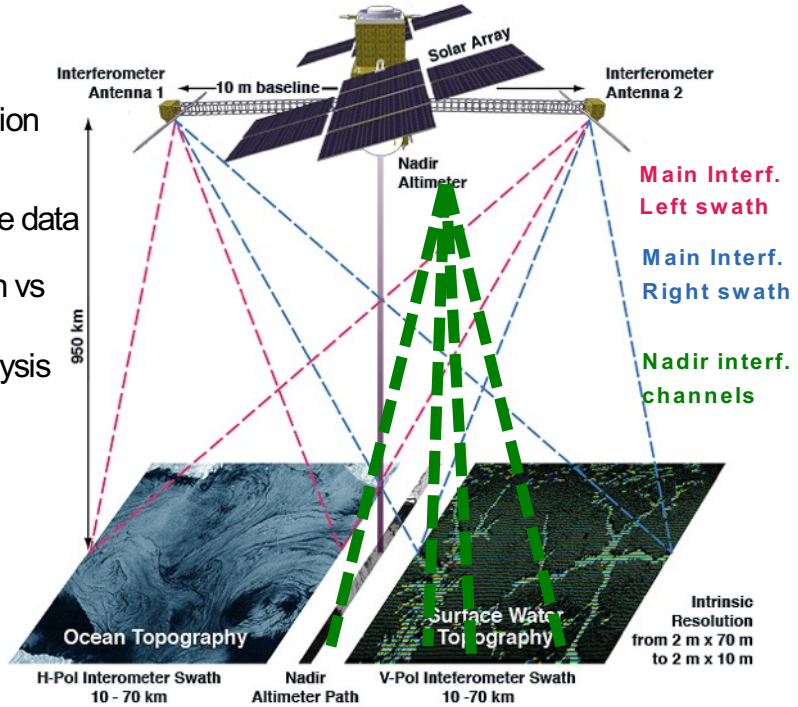
Hydrology: To provide a global inventory of all terrestrial water bodies whose surface area exceeds $(250m)^2$ (lakes, reservoirs, wetlands) and rivers whose width exceeds 100 m (requirement) (50 m goal) (rivers).

- To measure the global storage change in fresh water bodies at sub-monthly, seasonal, and annual time scales.
- To estimate the global change in river discharge at sub-monthly, seasonal, and annual time scales.

- **Data Volume:**
  - 17PB of original data
  - 6 PB of reprocessed data
- **Total of about 23PB for a nominal 3-year mission**
- **Add roughly 450TB/month for any mission extension**

SWOT changes how PO.DAAC operates

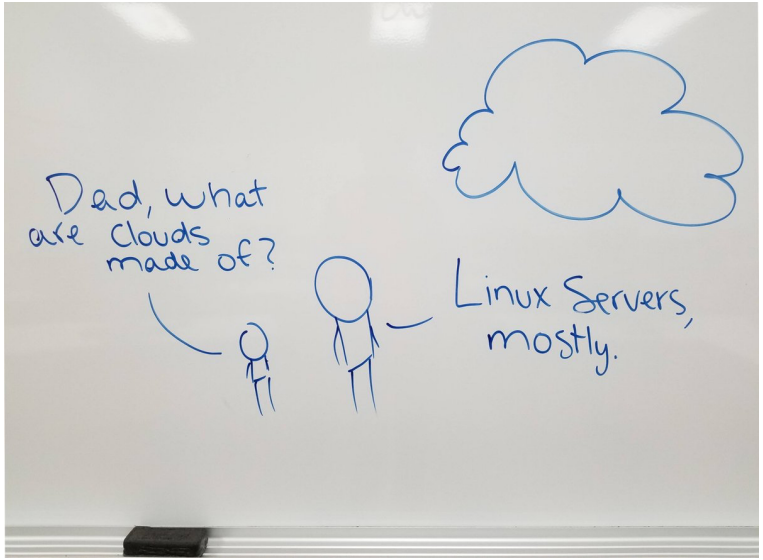
- Infrastructure
- Cloud storage selection (Object store, AWS Glacier)
- Interface with science data system
- On-the-fly generation vs long-term store
- Distribution and analysis services



Launches April of 2021
https://swot.jpl.nasa.gov

THUANG/JPL © 2017. All rights reserved.

Cloud Computing



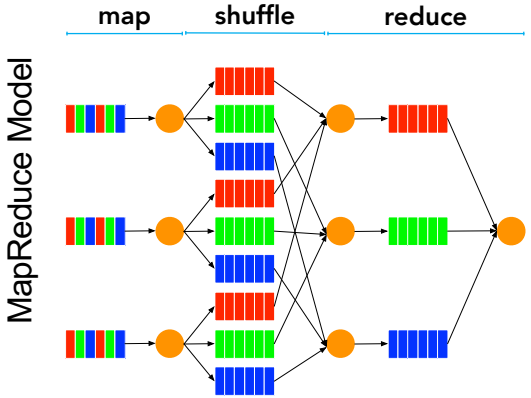
Credit: Twitter

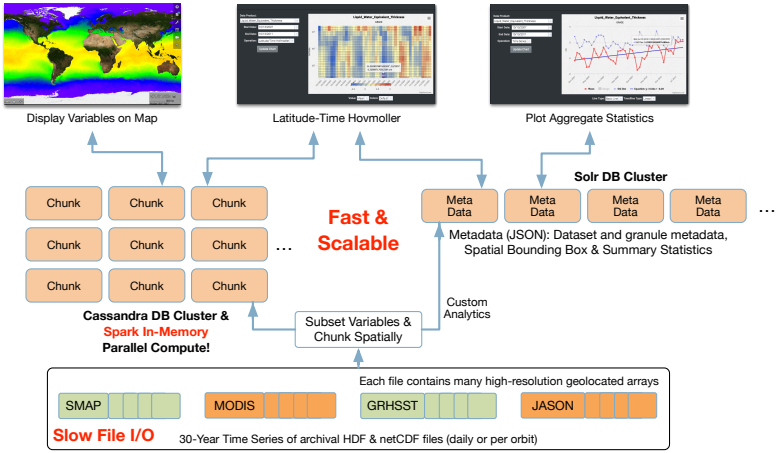
- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured Service

THUANG/JPL © 2018. All rights reserved.

NEXUS: Scalable Data Analytic Solution

- **MapReduce**: A programming model for expressing distributed computations on massive amount of data and an execution framework for large-scale data processing on clusters of commodity servers. - J. Lin and C. Dyer, "*Data-Intensive Text Processing with MapReduce*"
  - **Map**: splits processing across cluster of machines in parallel, each is responsible for a record of data
  - **Reduce**: combines the results from Map processes
- **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



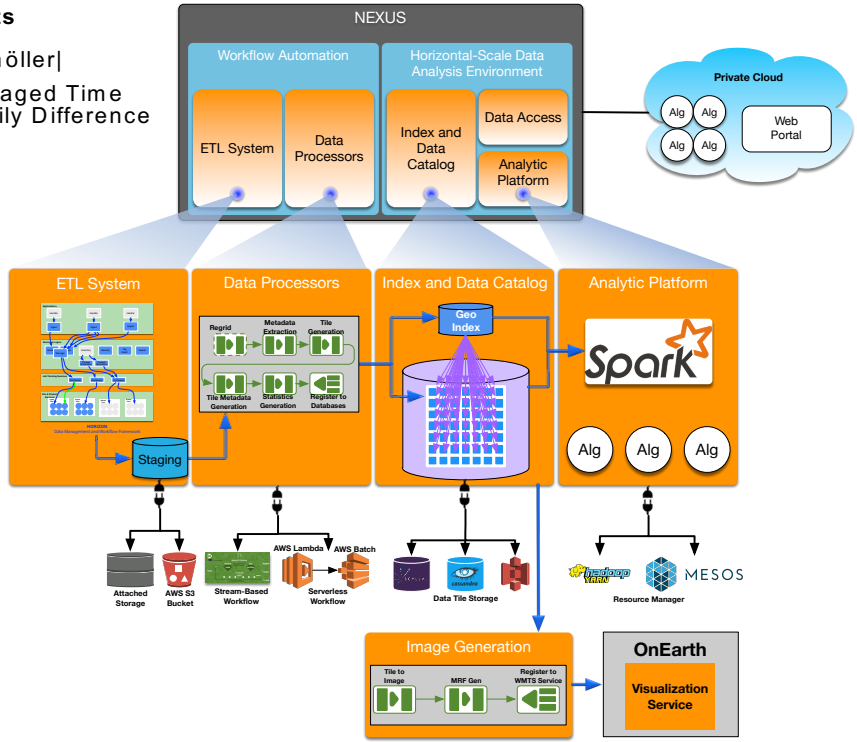


NEXUS' Two-Database Architecture

THUANG/JPL © 2018. All rights reserved.

NEXUS' Pluggable Architecture for different Operation Needs

- **NEXUS supports public/private Cloud and local cluster deployments**
- **It has a growing set of 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
- **It offers several container-based deployment options**
  - Local on-premise cluster
  - Private Cloud
  - Amazon Web Service
- **Automate Data Ingestion with Image Generation**
  - Cluster based
  - Serverless (Amazon Lambda and Batch)
- **Data Store Options**
  - Apache Cassandra
  - ScyllaDB
  - Amazon Simple Storage Service (S3)
- **Resource Management Options**
  - Apache YARN
  - Apache MESOS
- **Analytic Engine Options**
  - Custom Apache Spark Cluster
  - Amazon Elastic MapReduce (EMR)
  - Amazon Athena (work-in-progress)



THUANG/JPL © 2018. All rights reserved.

NEXUS Performance

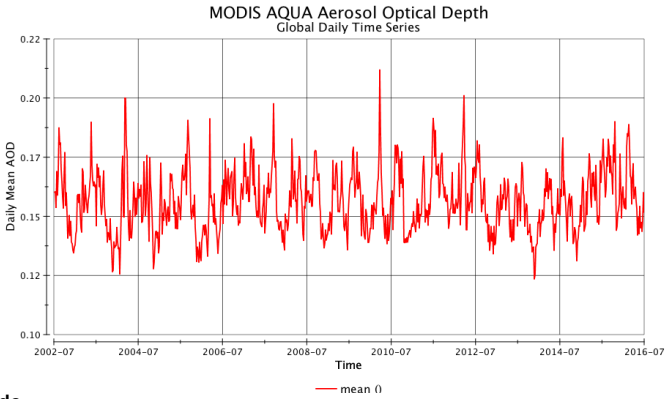
A recent benchmark comparison between **NASA GIOVANNI**, **NEXUS** with **Amazon's Elastic Map Reduce (EMR),** and NEXUS with custom Apache Spark Cluster

- **Giovanni**: 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

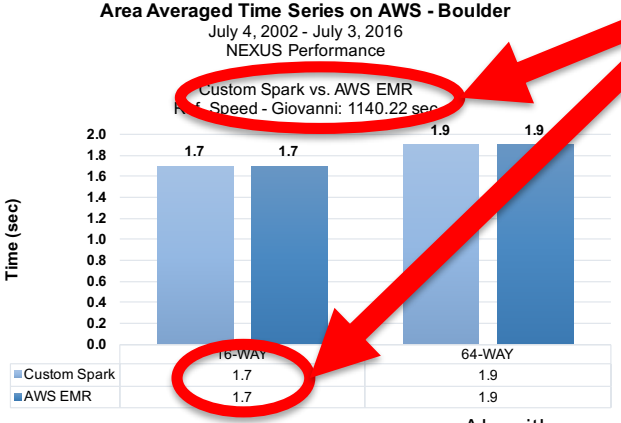
Dataset: 14-years of MODIS AQUA Daily (1 degree daily) Aerosol Optical Depth 550nm (Dark Target) (MYD08_D3v6), Level 3

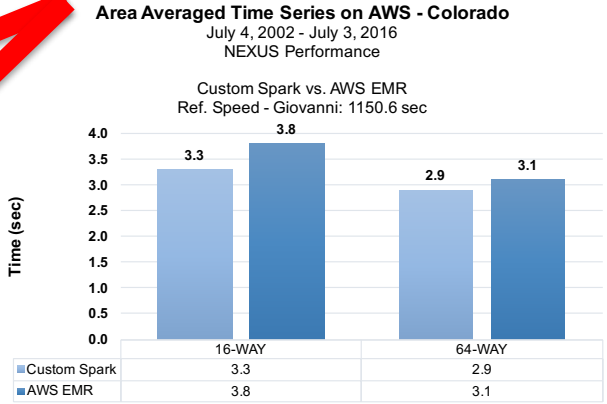
File Count: 5106

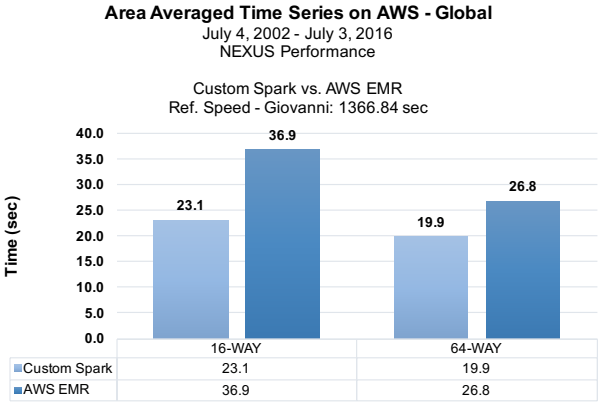
Total 2.6GB



GIOVANNI: 20 min
NEXUS: 1.7 sec







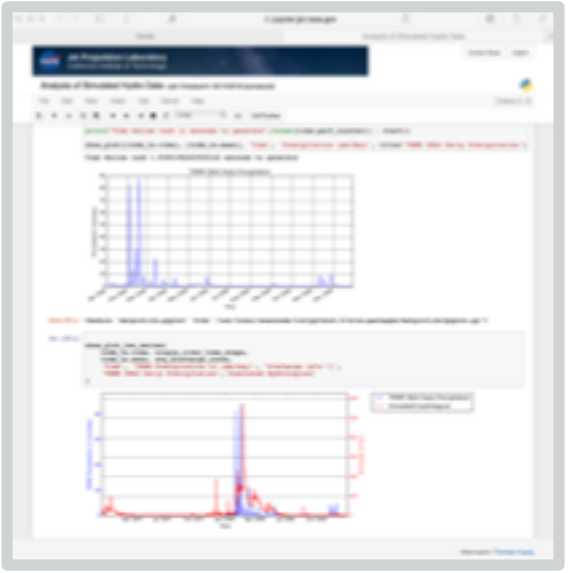
Algorithm execution time. Excludes Giovanni's data scrubbing processing time

THUANG/JPL © 2018. All rights reserved.

Performance example: support for hydrology

Retrieval of a single river time series

Retrieval of time series from 9 rivers



Time series coordination between TRMM and river

- 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

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA Sea Level Change Portal – *https://sealevel.nasa.gov*

Goal for the NASA Sea Level Change Team

- Determine how much will sea level rise by [2100]?
- What are the key sensitivities?
- Where are the key uncertainties? Observables? Model Improvements

Goals for the NASA Sea Level Change Portal

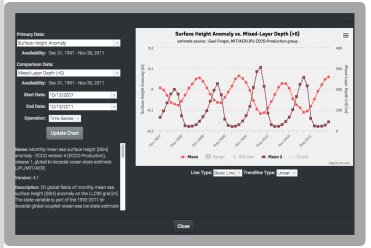
- Provide scientists and the general public with a “one-stop” source for current sea level change information and data
- Provide interactive tools for analyzing and viewing regional data
- Provide virtual dashboard for sea level indicators
- Provide latest news, quarterly report, and publications
- Provide ongoing updates through a suite of editorial products

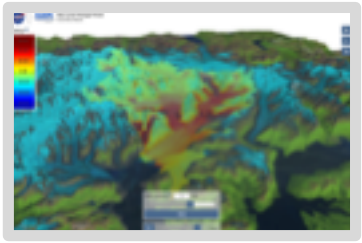
Requires

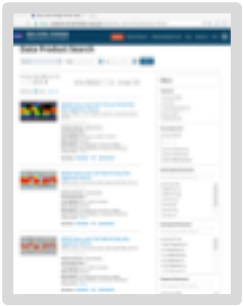
- Interdisciplinary collaboration
- Connect disciplines and evaluate dependencies

Sea Level Change Portal facilitates

- Easy interdisciplinary data comparison
- Access to latest news and information
- Collaboration (data and information exchange)







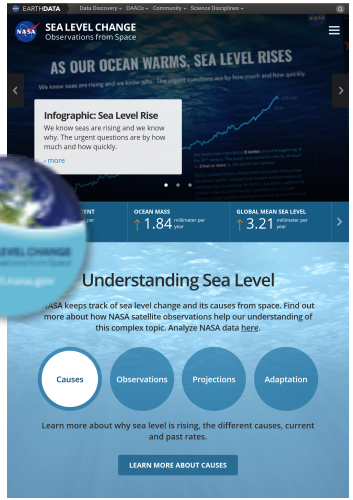
THUANG/JPL © 2018. All rights reserved.

Headliners and Social Media

60,000 monthly page views

50,000 sessions

45,000 users



Over 36,000 Followers

Over 27,000 Followers

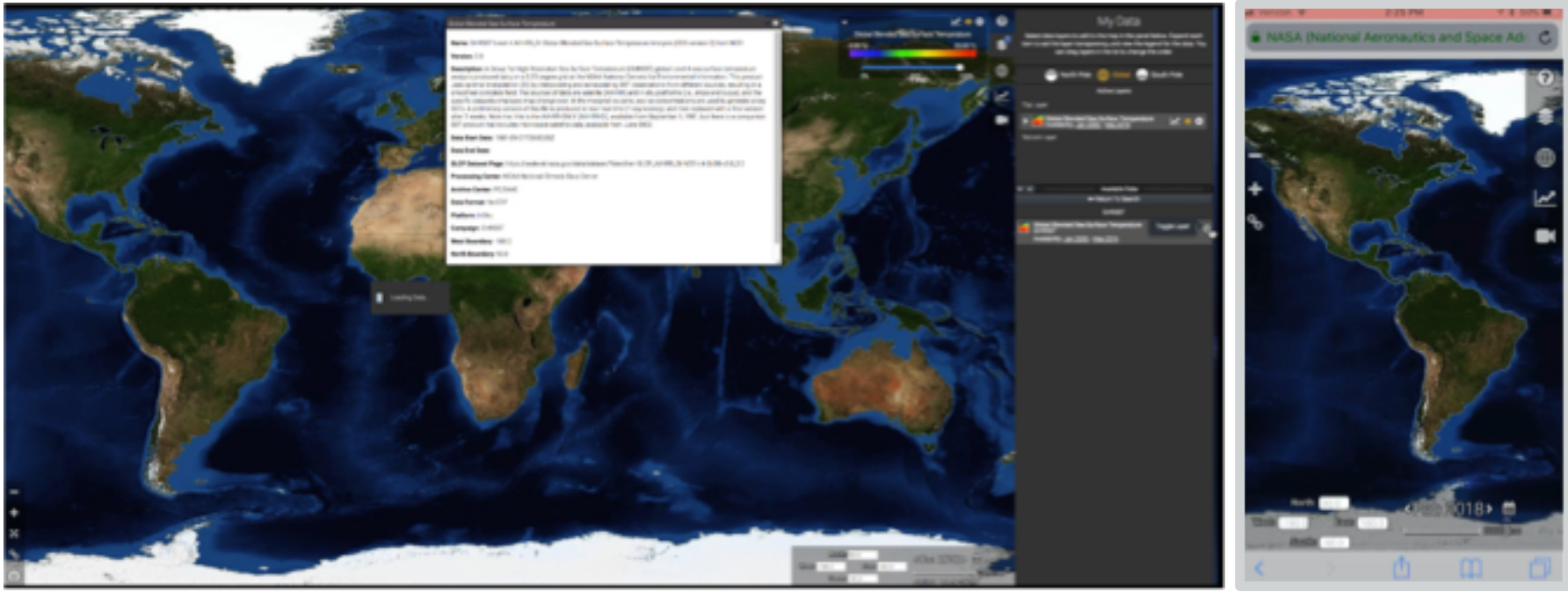
HEADLINES

"NASA Sea Level Change Website Offers Everything You Need to Know About Climate Change"
Tech Times

"NASA's New Sea Level Site Puts Climate Change Papers, Data, and Tools Online"
Tech Crunch

THUANG/JPL © 2018. All rights reserved.

Analyze Sea Level On-The-Fly



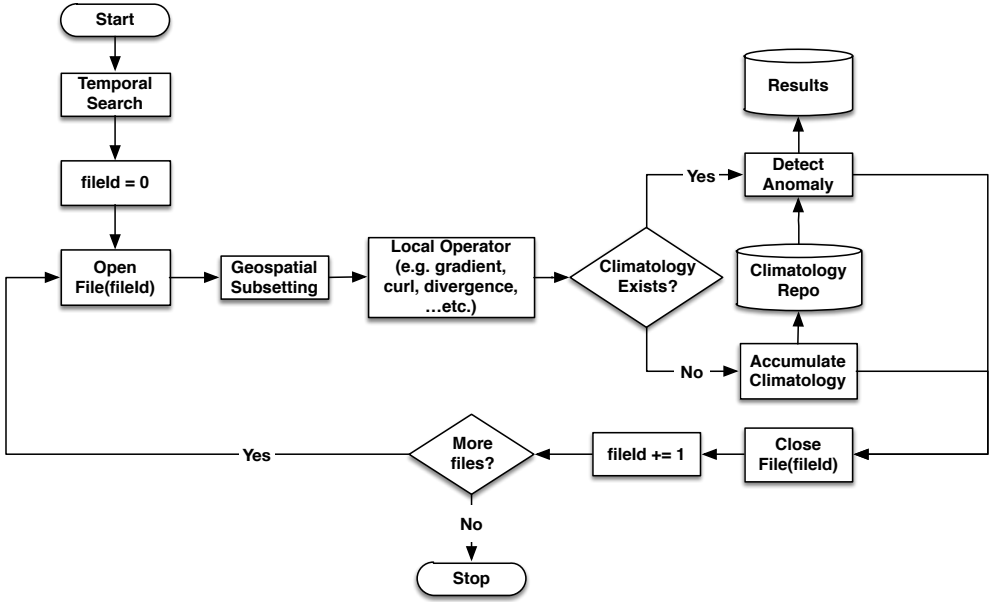
Sea Level Change - Data Analysis Tool

Visualizations | Hydrological Basins | Argo Profile | Time Series | Deseason | Data Comparison | Scatter Plot | Latitude/Time Hovmöller | Etc.

THUANG/JPL © 2018. All rights reserved.

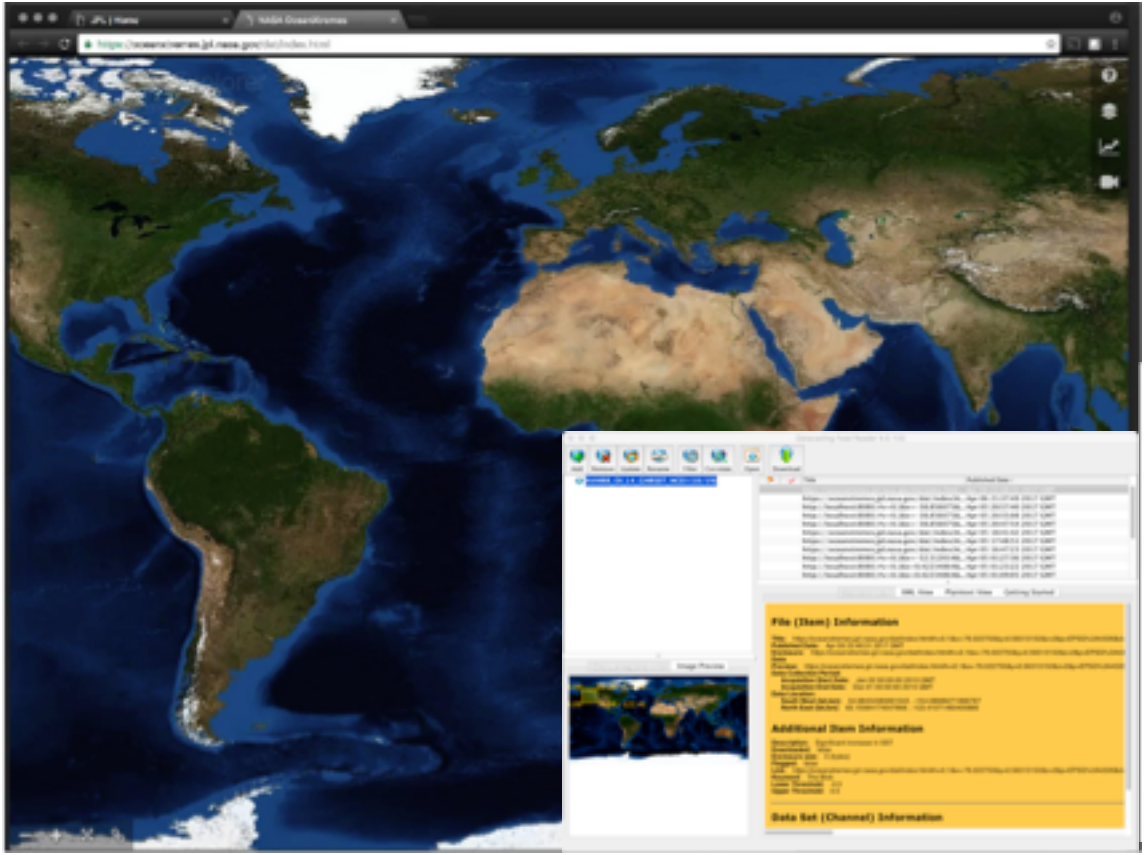
Anomaly Detection

- Anomaly detection is a process of identifying items, events or observations outside the “norm” or expected patterns
- Current and future oceanographic missions and our research communities present us with challenges to rapidly identify features and anomalies in increasingly complex and voluminous observations
- Typically this is a two-stage procedure
  1. Determine a long-term/periodic mean (“climatology”)
  2. Deviations from the mean are searched. Step 1 could be omitted in cases where a climatology data set already exists.



THUANG/JPL © 2018. All rights reserved.

Analyze Ocean Anomaly – “The Blob”

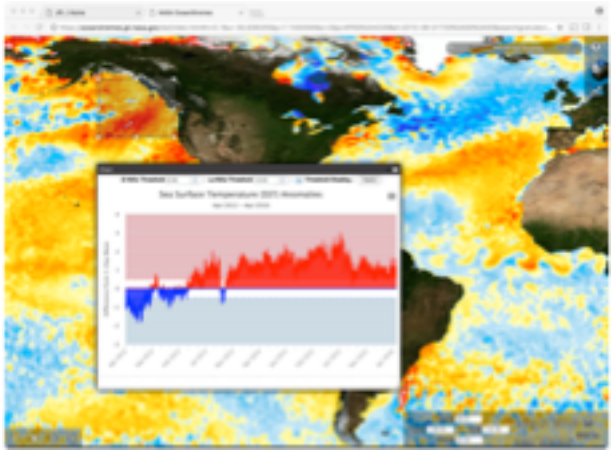


- **Visualize** parameter
- **Compute** daily differences against climatology
- **Analyze** time series area averaged differences
- **Replay** the anomaly and visualize with other measurements
- **Document** the anomaly
- **Publish** the anomaly

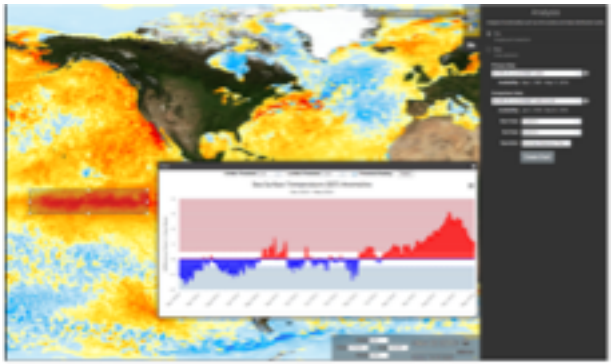
Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." Oceanography 29.

THUANG/JPL © 2018. All rights reserved.

More Anomalies



"The Blob"



El Niño 3.4 regional signal

Recreated identification of "The Blob"

- **The Blob** is the name given to a large mass of relatively warm water in the Pacific ocean off the coast of North America. It was first detected in late 2013 and continued to spread throughout 2014 and 2015.
- SST anomaly = SST – SST Climatology at each location to compare with standard deviation - Chelle Gentemann, Senior Scientist at Earth & Space Research

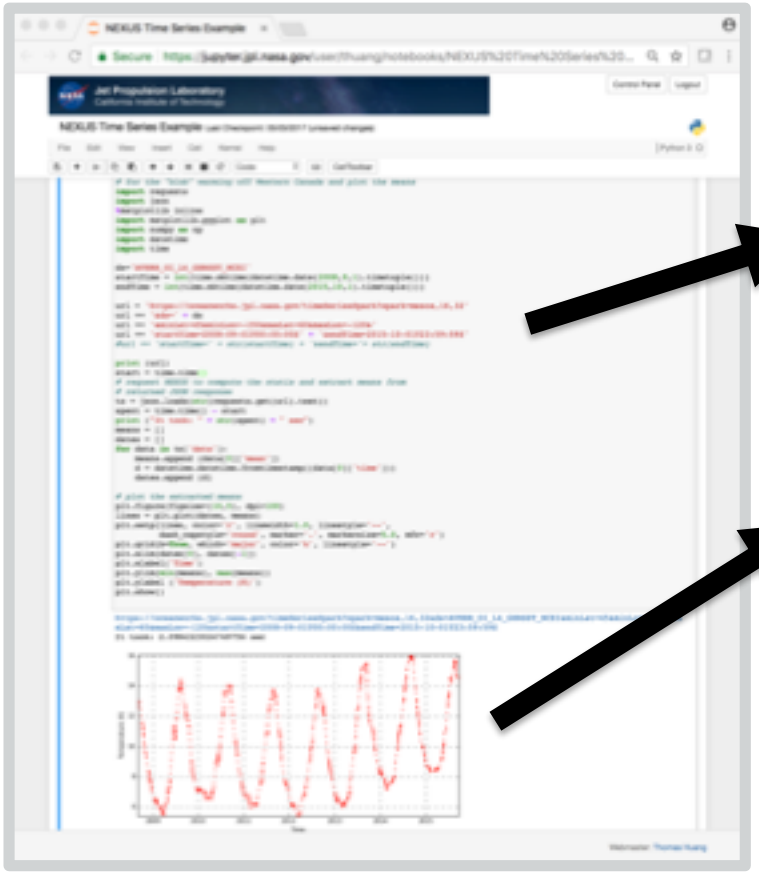
Recreated the El Niño 3.4 regional signal

- **El Niño** is a phenomenon in the equatorial Pacific Ocean characterized by a five consecutive 3-month running mean of sea surface temperature (SST) anomalies in the Niño 3.4 region that is above (below) the threshold of +0.5°C (-0.5°C). This standard of measure is known as the Oceanic Niño Index (ONI).
- https://www.ncdc.noaa.gov/teleconnections/enso/indicators/sst.php

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enable Science without File Download



```
# Request NEXUS to compute SST Time Series 2008/9/1 - 2015/10/1
# for the "blob" warming off Western Canada and plot the means
…
ds='AVHRR_OI_L4_GHRSST_NCEI'

url = … # construct the webservice URL request


# make request to NEXUS using URL request
# save JSON response in local variable
ts = json.loads(str(requests.get(url).text))

# extract dates and means from the response
means = []
dates = []
for data in ts['data']:
    means.append(data[0]['mean'])
    d = datetime.datetime.fromtimestamp((data[0]['time']))
    dates.append(d)

# plot the result
…
```

```
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

It took: 2.0984323024749756 sec
```

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Using IDL with NEXUS

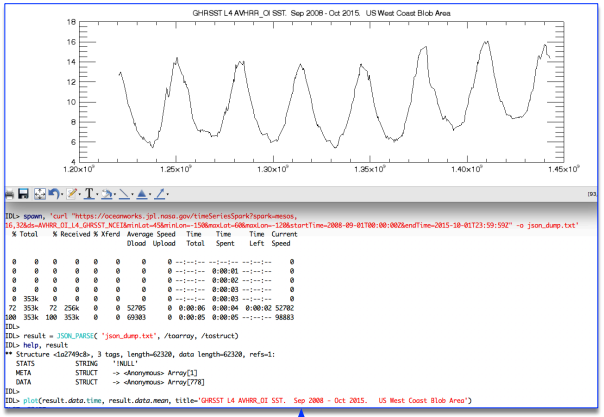
```
IDL> spawn, 'curl
"https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_
GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-
01T00:00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt'
  % Total    % Received % Xferd  Average Speed   Time    Time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



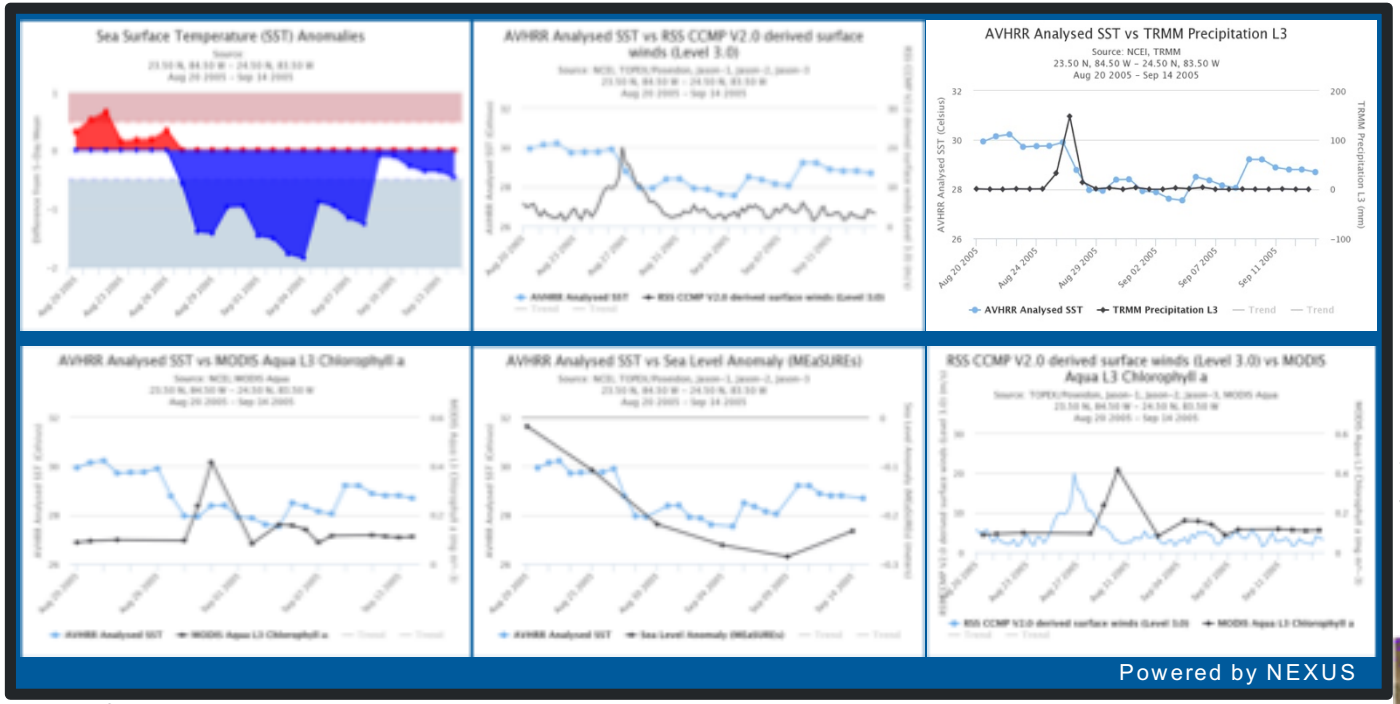
Credit: Ed Armstrong
Jun. 05, 2018

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

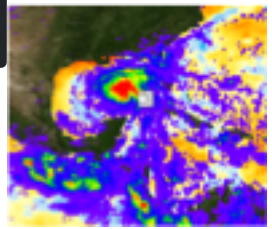
Hurricane Katrina Study



A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi
JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 114, C03023, doi:10.1029/2008JC004934, 2009

Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 ℃ that lingers for several days. Following this was a short intense ocean chlorophyll bloom a few days later. The ocean may have been “preconditioned’ by a cool core eddy and low sea surface height.

The SST drop is correlated to both wind and precipitation data. The Chl-A data is lagged by about 3 days to the other observations like SST, wind and precipitation.

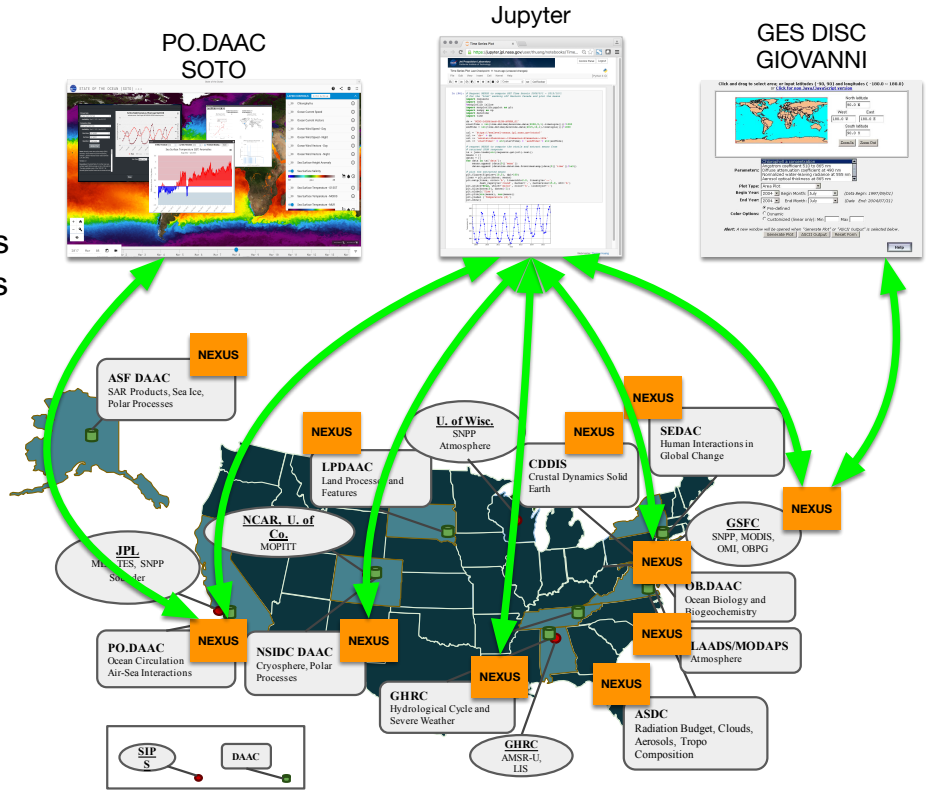


Hurricane Katrina TRMM overlay SST Anomaly

THUANG/JPL © 2018. All rights reserved.

Moving Toward Multi-Variable Analysis

- Public accessible RESTful analytic APIs where computation is next to the data
- NEXUS as the analytic engine infused and managed by the data centers on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files
- Reduce unnecessary data movement and egress charges
- An architecture to enable next generation of scientific applications



THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

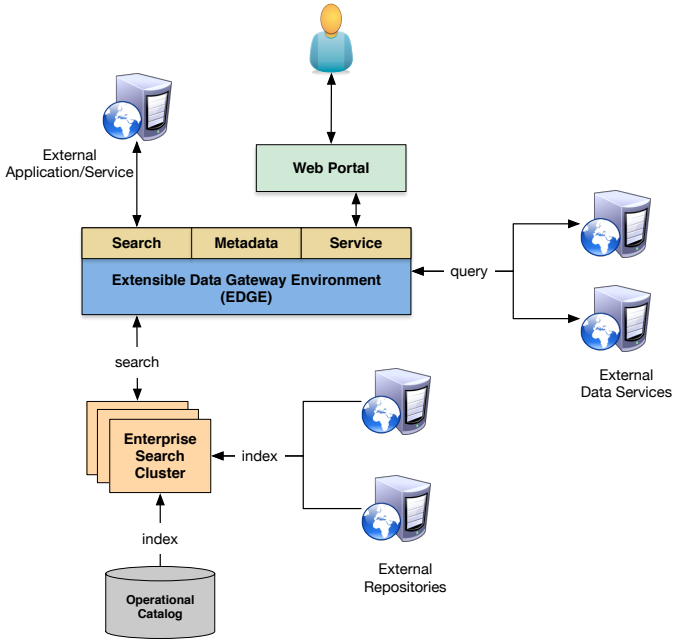
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Search, Relevancy, and Discovery

THUANG/JPL © 2018. All rights reserved.

Extensible Data Gateway Environment (EDGE)

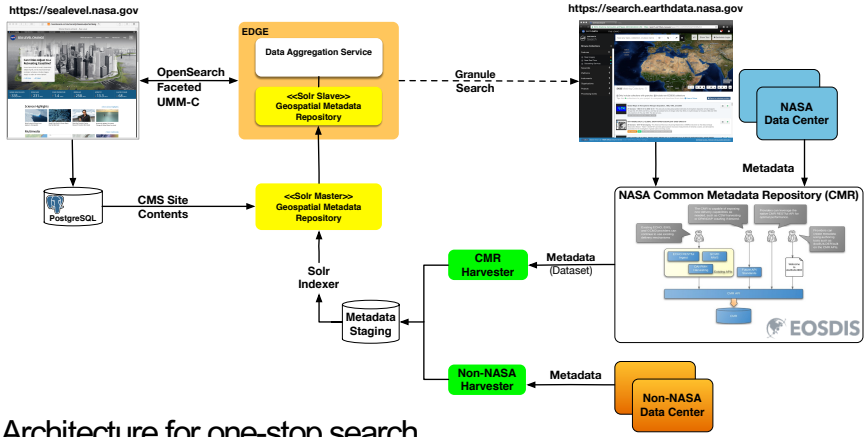
- Open Source high-performance geospatial data search and access
- Delivers sub-second search solution
- Implements the ESIP Federation's Discovery Specification (http://wiki.esipfed.org/index.php/Discovery_Cluster), which is a specialization of the OpenSearch (http://www.opensearch.org) standard (both XML and JSON)
- Platform to support multi-metadata standard specifications including ISO-19115, NASA UMM-C, NASA ECHO-10, NASA Global Change Master Directory (GCMD), Federation Geographic Data Committee (FGDC), and various domain-specific metadata standards
- Two main building blocks: data aggregation service and enterprise geospatial indexed search cluster
- Aggregation – provides a plugin approach to integrate with other external data repositories by proxying to other local/remote data services to reduce the number of interfaces a requestor has to access
- Enterprise geospatial indexed search cluster for fast lookup. Supports Apache Solr (and SolrCloud) and ElasticSearch
- Various production deployments including NASA Sea Level Change Portal, GRACE Web Portal, PO.DAAC, NASA ACCESS and AIST projects, and various Navel Research projects



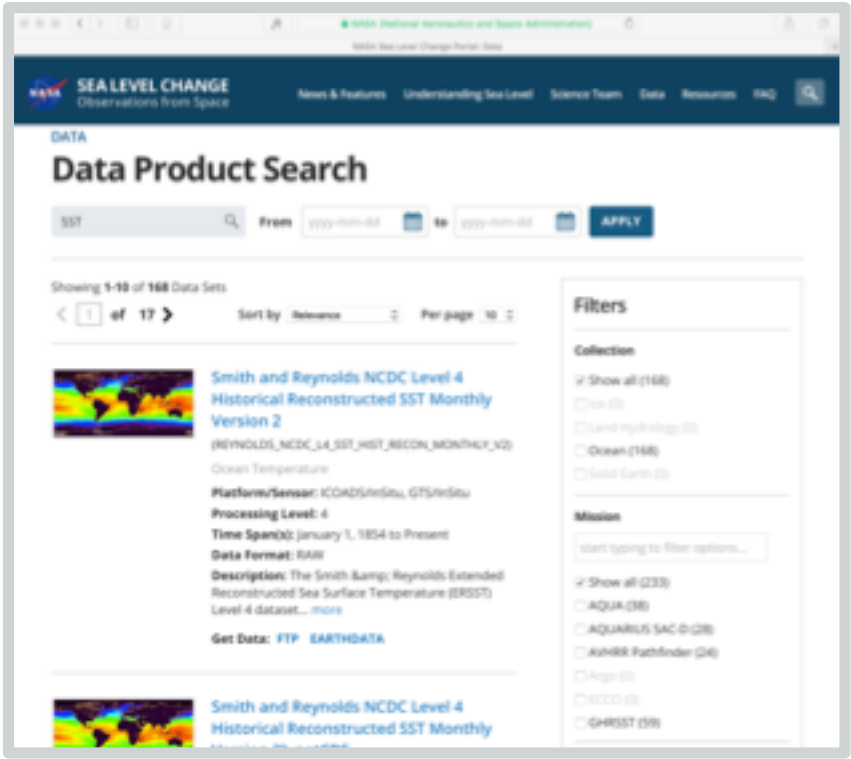
THUANG/JPL © 2018. All rights reserved.

NASA Sea Level Change Portal's One-Stop Search

- Homogenize metadata acquired from different providers
- On-the-fly translation metadata and search results according to the NASA ECHO-10 and UMM-C specification
- Simplify web portal integration by providing one-stop search solution for all Sea Level artifacts – data, news, publications, and multi-media resources



Architecture for one-stop search

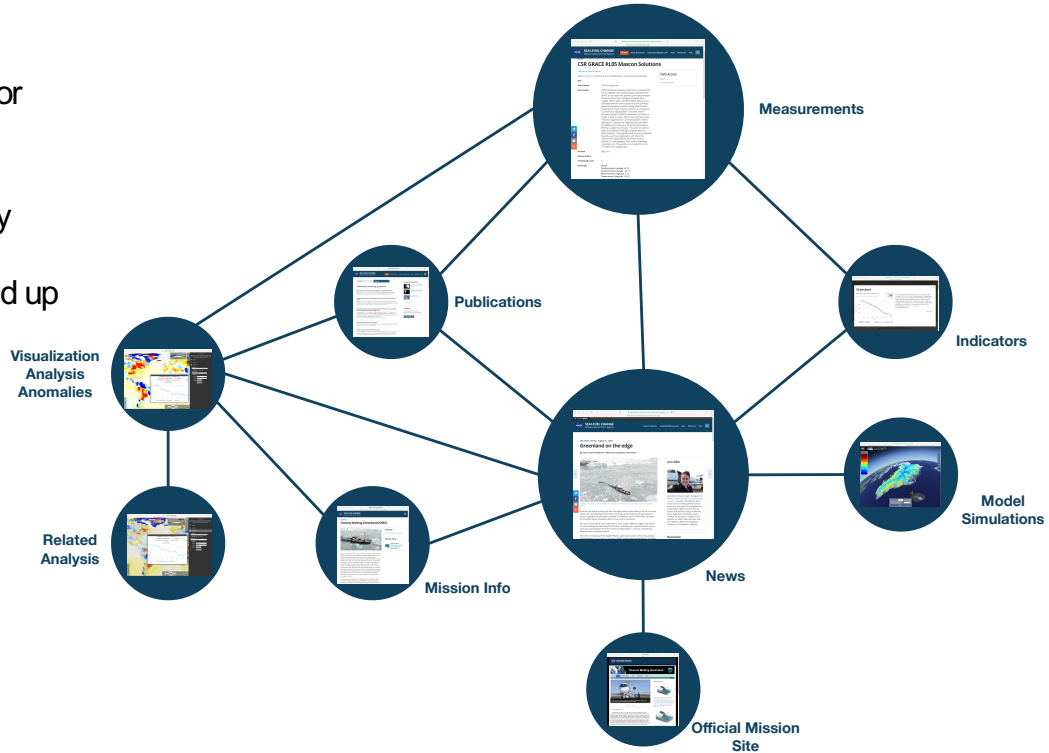


NASA Sea Level Change Portal's One-Stop Search

THUANG/JPL © 2018. All rights reserved.

OceanWorks Tackles Information Discovery

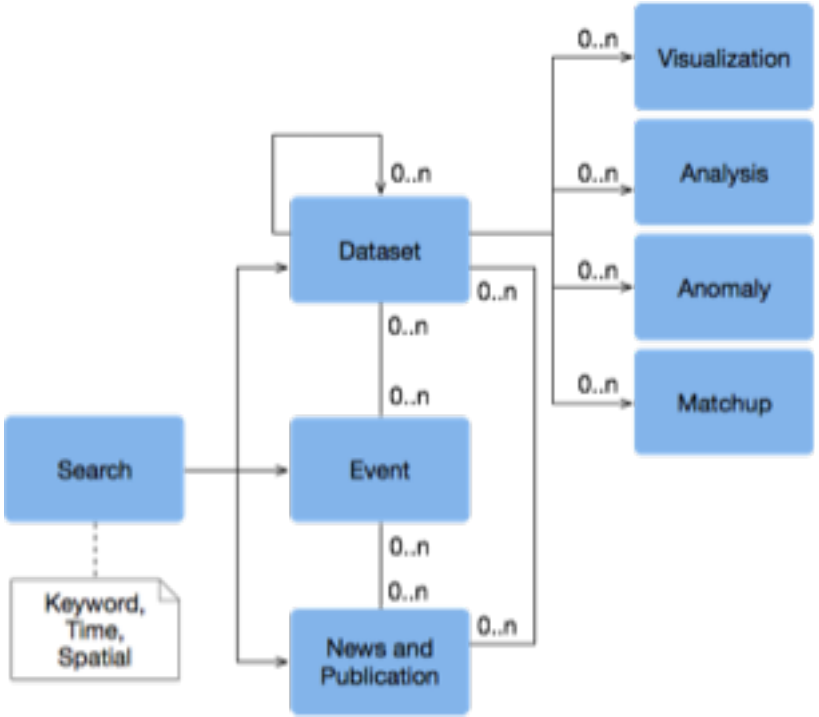
- **Search** is looking for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** is finding something new, or in a new way
  - This is non-trivial
  - Traditional ontological method doesn't quite add up
  - The strength of semantic web is in inference
  - Need method involves
    - Dynamic data ranking
    - Dynamic update to the ontology
    - Mining user interaction and news outlets
- **Relevancy** is
  - Domain-specific
  - Personal
  - Temporal
  - Dynamic



THUANG/JPL © 2018. All rights reserved.

OceanWorks Tackles Information Discovery

- Support for oceanographic events
  - Continuous harvesting active events from Earth Observatory Natural Event Tracker (EONET)
  - Adding ability to register custom events
  - Mapping datasets to events by time and space
- Connecting artifacts
  - Linking datasets with analysis and matchup for recommendation
  - Linking news and publications with events and datasets
- Dynamic ranking of datasets to improve relevancy



Information Model for Data Discovery

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

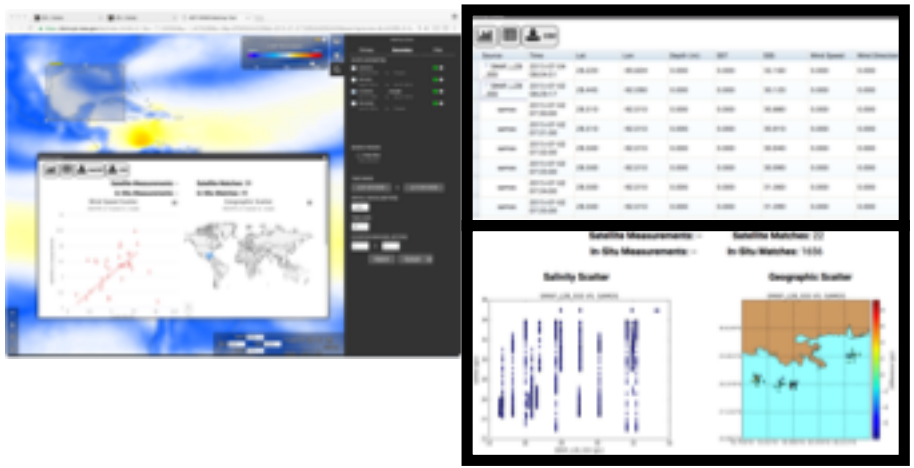
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

In Situ to Satellite Matchup On the Cloud

THUANG/JPL © 2018. All rights reserved.

In-Situ to Satellite Matchup

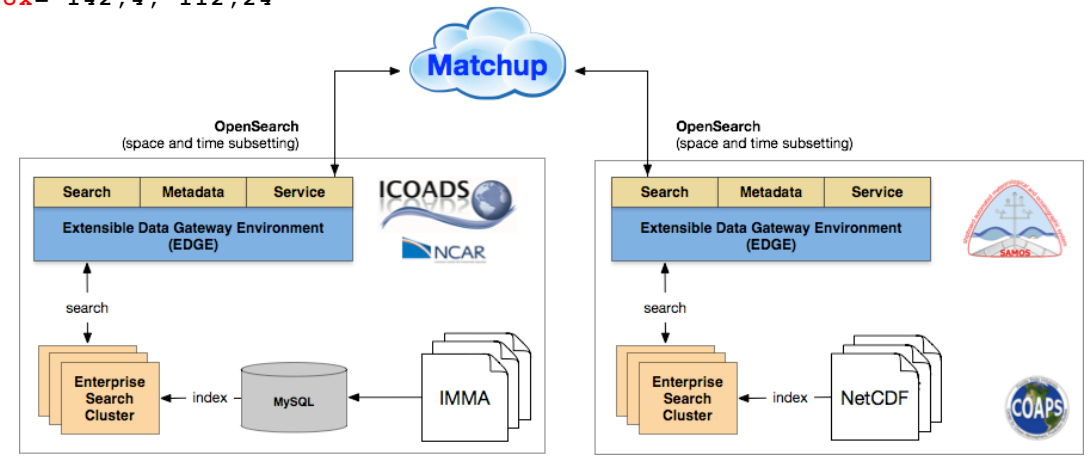
- Typically data matching is done using one-off programs developed at multiple institutions
- Leverage horizontal-scale technology for fast, in-memory execution of matchup algorithm
- Common and open source architecture to reduce in duplicate development and man hours required to match satellite/in situ data
- Satellite measurements. Hosted at the PO.DAAC
  - **GHRSST JPL-L2P-MODIS_A** and **JPL-L2P-MODIS_T**
  - **SMAP L2 Sea Surface Salinity** (JPL Evaluation product) (4/1/2015 – 8/1/2016)
  - **ASCAT ASCATB-L2 Coastal** (10/29/2012 – 06/06/2016)
- In situ data nodes at SPURS/JPL, ICOADS/NCAR, and SAMOS/FSU operational.
  - **Shipboard Automated Meteorological and Oceanographic System (SAMOS)**. Hosted at FSU/COAPS
  - **International Comprehensive Ocean-Atmosphere Data Set (ICOADS).** Hosted at NCAR
  - **Salinity Processes in the Upper Ocean Regional Study: (SPURS-1)** N. Atlantic (2012-13) : salinity max region. **(SPURS-2)** Eastern Equatorial Pacific (15-16): high precipitation/low evaporation region. Hosted at JPL
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Supports on-the-fly in situ to satellite matchup of SST, SSS, Wind parameters



THUANG/JPL © 2018. All rights reserved.

On-The-Fly Subsetting of In-Situ Measurements using OpenSearch

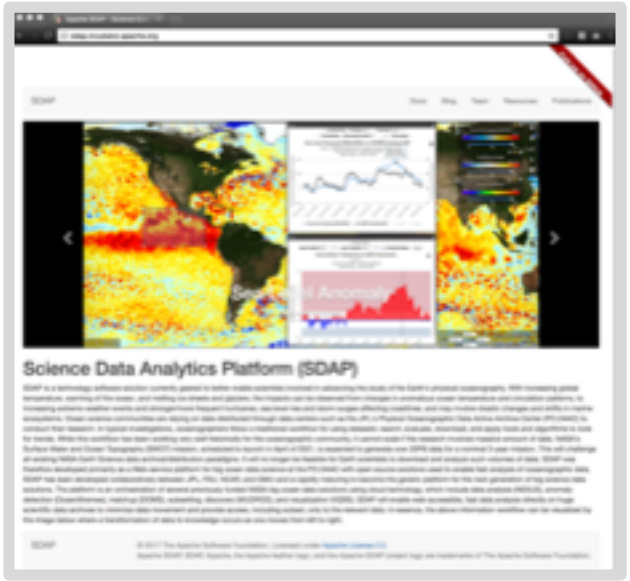
- Using OpenSearch as the standard interface to in-situ data repositories
- Enable distributed, federated search and data subsetting
- Subset in-situ data by time and space using OpenSearch
  - **ICOADS:** `'http://rda-data.ucar.edu:8890/ws/search/icoads?startTime=2012-08-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SAMOS:** `'http://doms.coaps.fsu.edu/edge/samos?startTime=2012-08-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SPURS-1:** `'https://doms.jpl.nasa.gov/spurs?startTime=201208-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SPURS-2:** `'https://doms.jpl.nasa.gov/spurs2?startTime=2016-07-01T00:00:00Z&endTime=2016-07-31T23:59:59Z&bbox=-142,4,-112,24'`



THUANG/JPL © 2018. All rights reserved.

Free and Open Source Software (FOSS)

- October 2017, the OceanWorks project released all of its source code to 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: Jörn Rottmann, Raphael Bircher, and Suneel Marthi
- OceanWorks is now being developed in the open
  - For local cluster and cloud computing platform
  - Fully containerized using Docker (multiple containers)
  - Infrastructure orchestration using Amazon CloudFormation
  - Analyzing satellite and model data
  - In situ data analysis and colocation with satellite measurements
  - Fast data subsetting
  - Data services integration architecture
  - OpenSearch and dynamic metadata translation
  - Mining of user interactions and data to enable discovery and recommendations
  - Streamline deployment through container technology



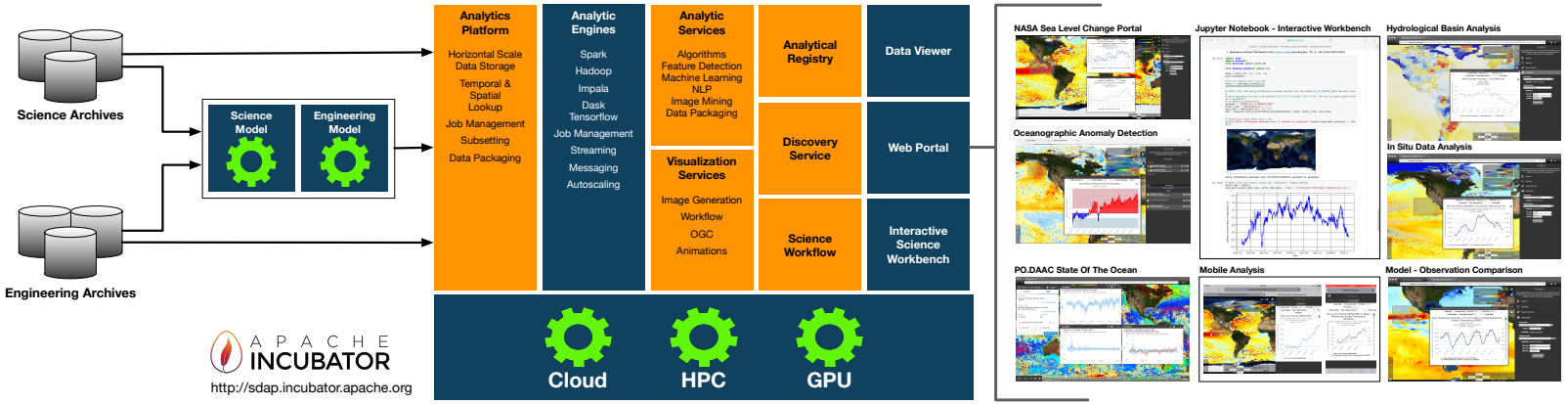
http://sdap.incubator.apache.org

APACHE INCUBATOR

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Integrated Ocean Science Data Analytics Platform



DATA → INFORMATION → Data Analysis → Systematic analysis of massive data

- Building block for an **Integrated Ocean Science Data Analytics Platform**: 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

THUANG/JPL © 2018. All rights reserved.

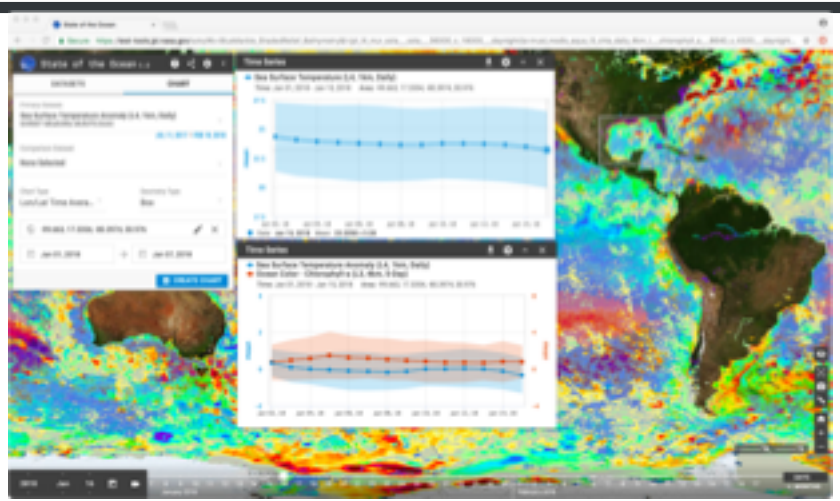
Notable Public Engagements

- Hosted hands-on cloud analytics workshops using Amazon Web Services (AWS) at 2017 Earth Science Information Partners (ESIP) summer meeting
- Invited to speak at the **Space Studies Board of The National Academy of Sciences**
- Invited to present to the **NASA Advisory Council's Ad-Hoc Big Data Task Force (BDTF)**
- Invited to present to the **JPL Deputy Lab Director and Chief Technologist**
- Invited to present to **CNES Chief Technologist and Delegations**

THUANG/JPL © 2018. All rights reserved.

In Summary

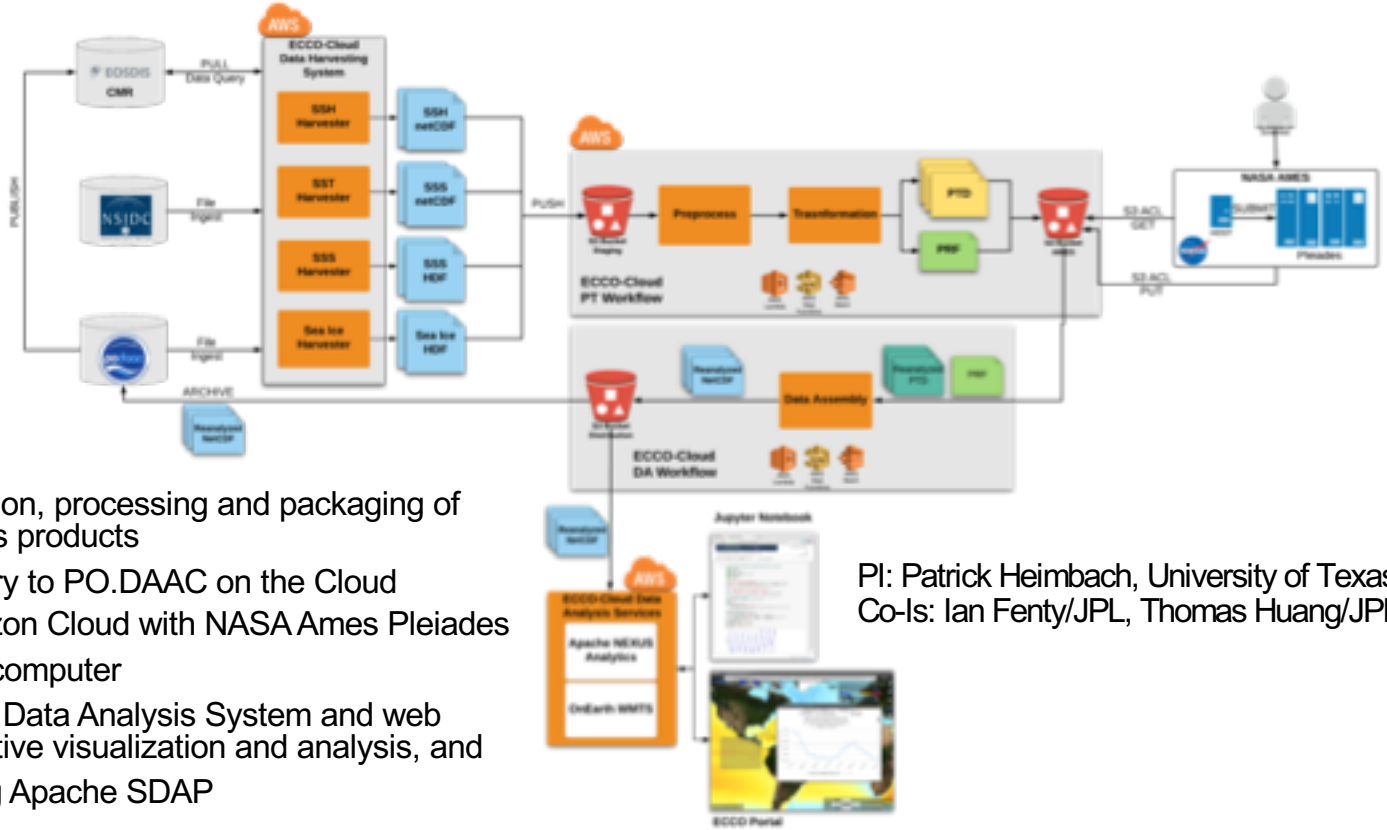
- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Let's think beyond archive and file downloads
- Connected information enables discovery
- Community developed solution through open sourcing
- Investment in data and computational sciences
- Data Centers might want to be in the business of Enabling Science!
- OceanWorks infusion 2018 – 2019
  - Watch for changes to the NASA's Sea Level Change Portal
    - Even faster analysis capabilities
    - More variety of measurements – satellites, in situ, and models
    - Event more relevant recommendations
  - NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)
    - More than just pretty pictures. SOTO will have new analytic capabilities.
- Coming Soon: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**



Coming Soon: PO.DAAC State of the Ocean (SOTO) 2018

THUANG/JPL © 2018. All rights reserved.

NASA ACCESS 2017: ECCO-Cloud (just announced)



- Automate ingestion, processing and packaging of ECCO reanalysis products
- Automate delivery to PO.DAAC on the Cloud
- Integrating Amazon Cloud with NASA Ames Pleiades petascale supercomputer
- Establish ECCO Data Analysis System and web portal for interactive visualization and analysis, and distribution using Apache SDAP

PI: Patrick Heimbach, University of Texas, Austin
Co-Is: Ian Fenty/JPL, Thomas Huang/JPL

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

thomas.huang@jpl.nasa.gov

Thomas Huang, thomas.huang@jpl.nasa.gov
Jet Propulsion Laboratory
California Institute of Technology

JPL Team
Ed Armstrong, Frank Greguska, Joseph Jacob, Lewis McGibbney, Nga Quach, Vardis Tsontos, and Brian Wilson

Florida State University Team
[redacted], Mark A. Bourassa, Jocelyn Elya

National Center for Atmospheric Research Team
[redacted], Tom Cram, Zaihua Ji

George Mason University Team
[redacted], Yongyao Jiang, and Yun Li

THUANG/JPL © 2018. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Backup

THUANG/JPL © 2018. All rights reserved.

Example EDGE API

- OpenSearch (ESIP Federation Discovery Cluster Specification)
  - Getting all the SST datasets from NASA Sea Level Change Portal using OpenSearch
    - `curl -X GET 'https://sealevel.nasa.gov/data/search-passthru?keyword=SST&sort=Relevance_desc&itemsPerPage=200'`

- Dataset Metadata
  - Getting dataset description in ISO-19115 format
    - `curl -X GET 'https://podaac.jpl.nasa.gov/ws/metadata/dataset?shortName=AVHRR_OI-NCEI-L4-GLOB-v2.0&output=iso'`
  - Getting dataset description in NASA UMM-C format
    - `curl -X GET 'https://oceanworks.jpl.nasa.gov/ws/metadata/dataset?shortName=MUR-JPL-L4-GLOB-v4.1&format=umm-json&pretty=true'`

THUANG/JPL © 2018. All rights reserved.

Growing List of Supported Datasets

- **Atmosphere**
  - MODIS Aqua Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
  - MODIS Terra Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)
  - MODIS Aqua Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
  - MODIS Terra Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)
- **Chlorophyll**
  - MODIS Aqua Level 3 Global Daily Mapped 4 km Chlorophyll a
- **Estimating the Circulation and Climate of the Ocean (ECCO)**
  - Monthly Mean Version 4 release 2 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Total Heat Flux, Total Salt Flux
  - Monthly Mean Version 4 release 1 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Ocean Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Actual Sublimation Freshwater Flux, Total Heat Flux, Total Salt Flux
- **Gravity**
  - Center for Space Research (CSR) GRACE RL05 Mascon Solutions
  - JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI filtered Version 2
- **Ocean Temperature**
  - GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (v4.1)
  - GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (25km)
  - GHRSST Level 4 AVHRR_OI Global Blended Sea Surface Temperature Analysis (GDS version 2) from NCEI
  - MODIS Aqua Level 3 SST Thermal IR Daily 4km Nighttime v2014.0
  - MODIS Aqua Level 3 SST Thermal IR Daily 4km Daytime v2014.0

THUANG/JPL © 2018. All rights reserved.

Supported Datasets (+)

- **Salinity**
  - JPL SMAP Level 2B CAP Sea Surface Salinity V2.0 Validated Dataset
  - JPL SMAP Level 3 CAP Sea Surface Salinity Standard Mapped Image Monthly V3.0 Validated Dataset
- **Sea Surface Height Anomalies (SSHA)**
  - JPL MEaSUREs Gridded Sea Surface Height Anomalies Version 1609
- **Wind**
  - Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L3.0 First-Look Analyses
- **Precipitation (non-ocean data)**
  - TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_Daily) at GES DIS
  - TRMM (TMPA-RT) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_RT) at GES DISC
- **In Situ**
  - Shipboard Automated Meteorological and Oceanographic System (SAMOS)
  - International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations
  - Salinity Process in the Upper Ocean Regional Study – 1 (SPURS1)
  - Salinity Process in the Upper Ocean Regional Study – 2 (SPURS2)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (salinity variables)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (temperature variables)

THUANG/JPL © 2018. All rights reserved.