National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OceanWorks:
Ocean Science Data Analytics using Apache Science Data Analytics Platform

Thomas Huang, PI

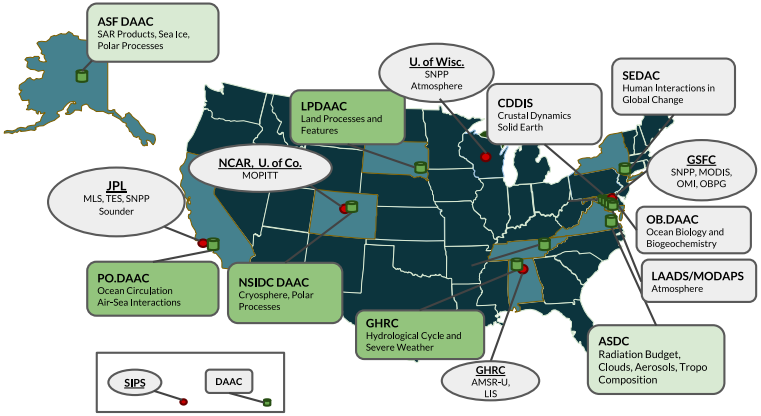
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

ESTO
Earth Science Technology Office

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

NASA's Physical Oceanography Data Center

- The **NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC)** at Jet Propulsion Laboratory is an element of the **Earth Observing System Data and Information System (EOSDIS)**. The EOSDIS provides science data to a wide communities of user for NASA's Science Mission Directorate.
- Archives and distributes data relevant to the physical state of the ocean
- **The mission of the PO.DAAC is to preserve NASA's ocean and climate data and make these universally accessible and meaningful.**



ESTO
Earth Science Technology Office

Data Centers

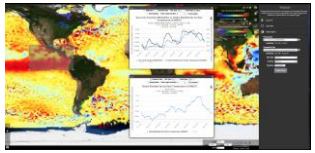
- **Mainly focus on** archives and distributions
- **With additional services**
  - Better searches – faceted, spatial, keyword, ranking, etc.
  - Data subsetting – home grown, OPeNDAP, Webification, etc.
  - Visualization – visual discovery, PO.DAAC's SOTO, NASA Worldview, etc.
- **Limitations**
  - 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?"

National Aeronautics and Space Administration

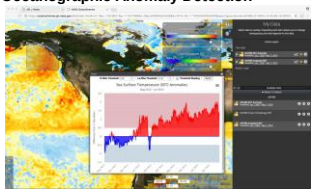
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enabling Next Generation of Ocean Science Tools and Services

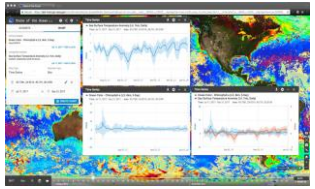
NASA Sea Level Change Portal



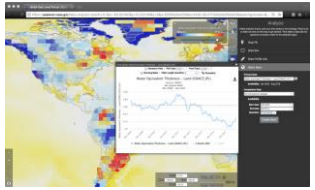
Oceanographic Anomaly Detection



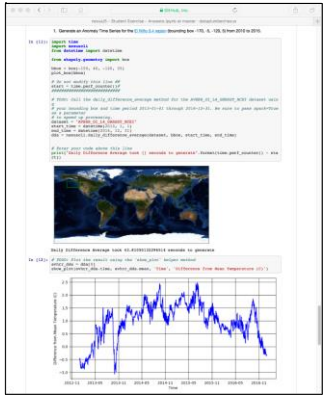
PO.DAAC State Of The Ocean



Hydrological Basin Analysis

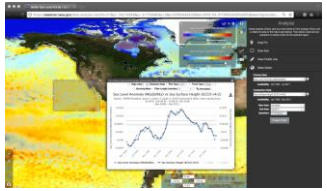


Jupyter Notebook - Interactive Workbench

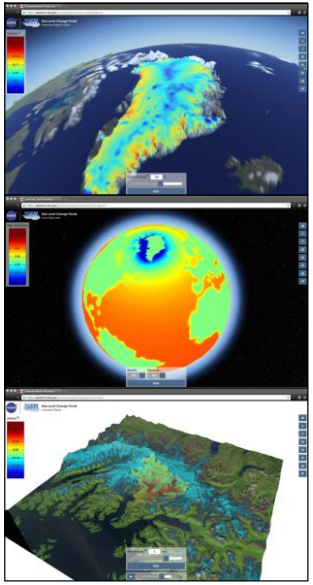


Mobile Analysis

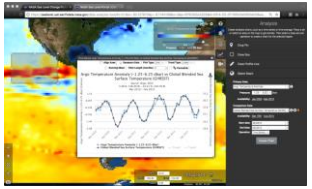
In Situ Data Analysis



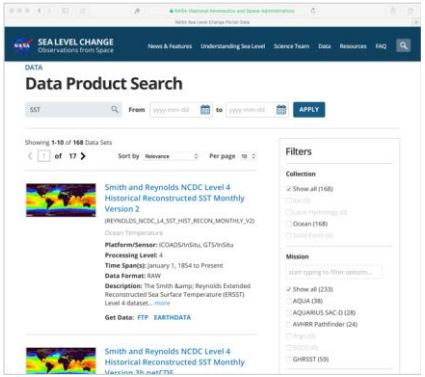
Model Simulations



Model - Observation Comparison



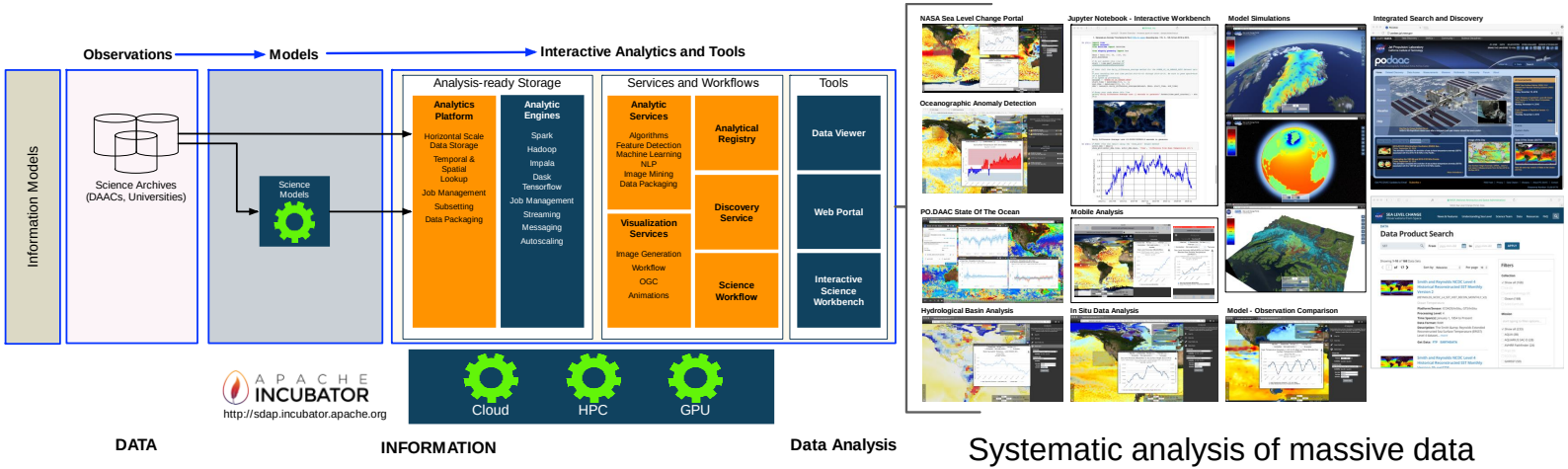
Integrated Search and Discovery



ESTO
Earth Science Technology Office

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

Integrated Ocean Science Data Analytics Platform



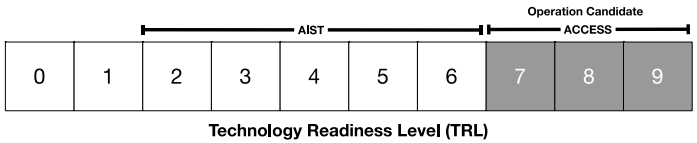
Systematic analysis of massive data

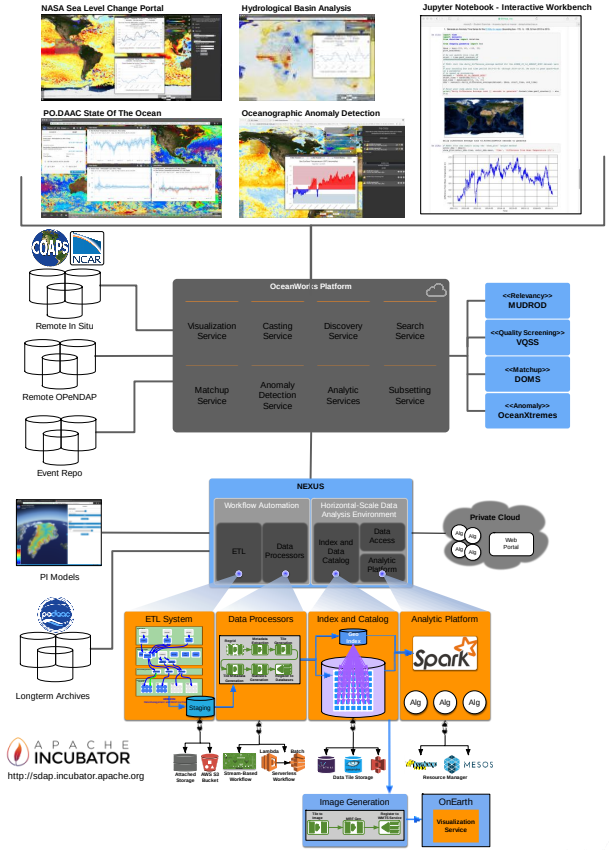
- An **Integrated Ocean Science Data Analytics Platform**: an environment for conducting a Ocean Science investigation
  - Confluence of resources for that investigation
  - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to 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

NASA AIST OceanWorks

- **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



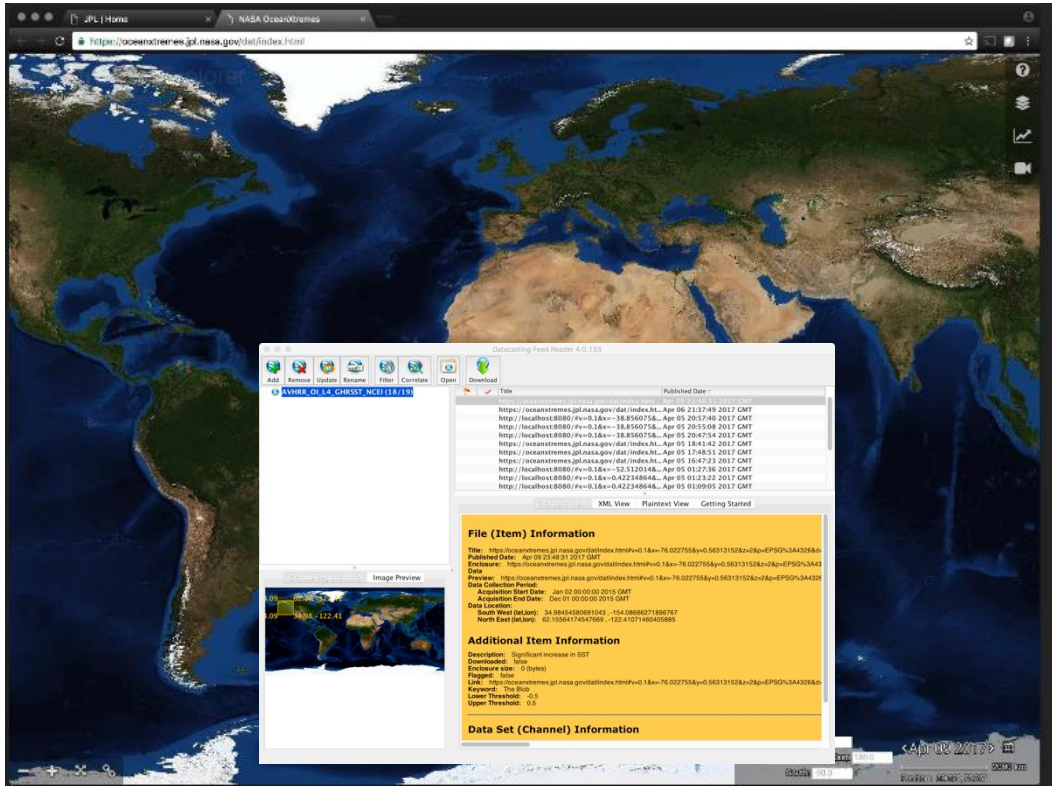


ESTO Earth Science Technology Office

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

Interactive Ocean Science Analysis

Interactive Anomaly Detection



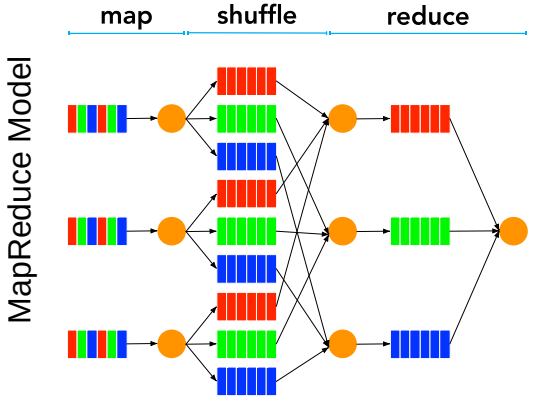
Sea Level Analysis

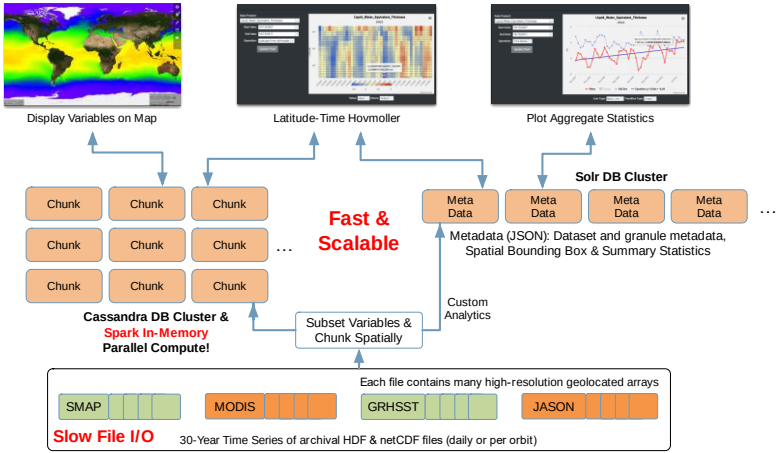
ESTO Earth Science Technology Office

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

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



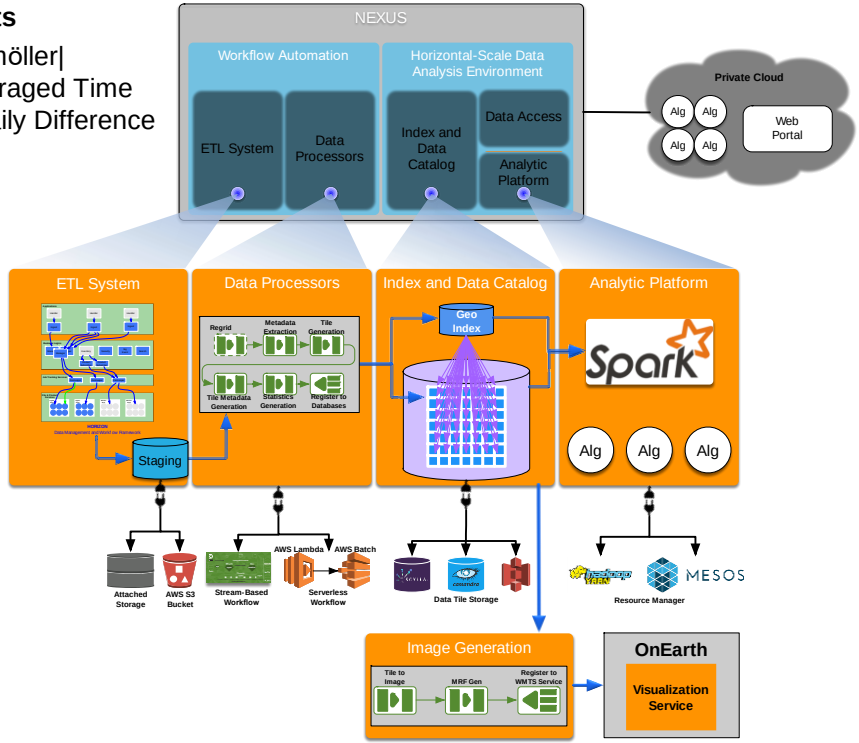


NEXUS' Two-Database Architecture

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

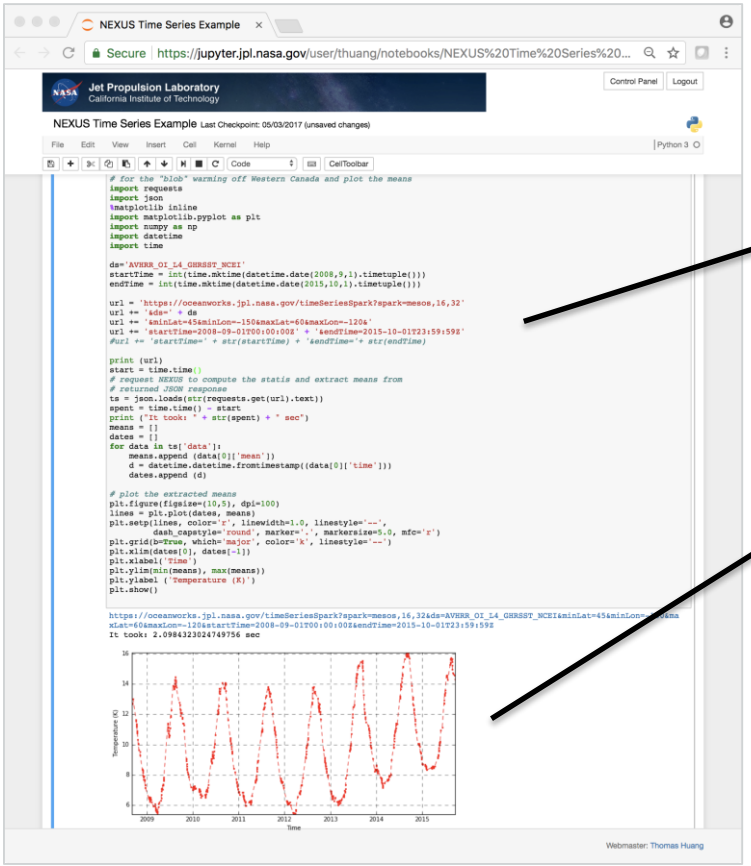
NEXUS' Pluggable Architecture for different Operation Needs

- **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)



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

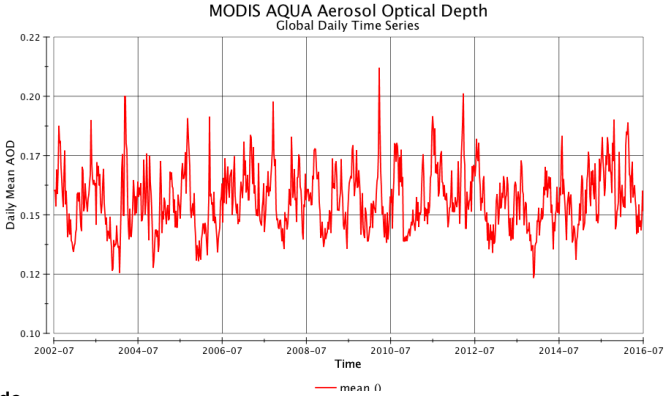
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NEXUS Performance

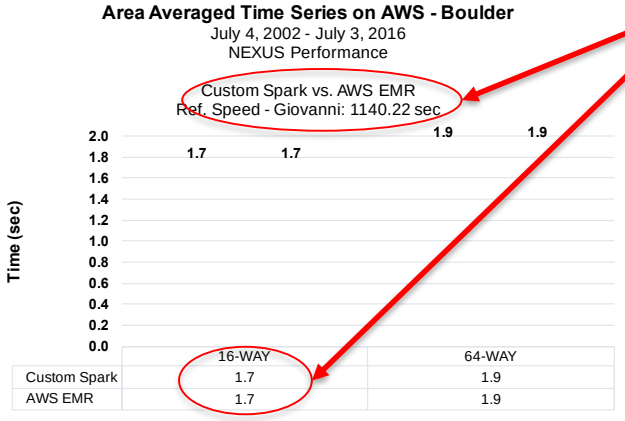
A recent benchmark comparison between **NASA GIOVANNI**, **NEXUS** with **Amazon's Elastic Map Reduce (EMR),** and NEXUS with custom Apache Spark Cluster

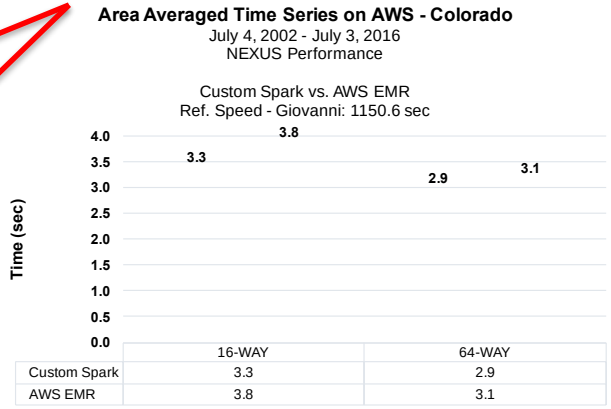
- **Giovanni**: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.
  - Represents current state of data analysis technology, by processing one file at a time
  - Backed by the popular NCO library. Highly optimized C/C++ library
- **AWS EMR**: Amazon's provisioned MapReduce cluster

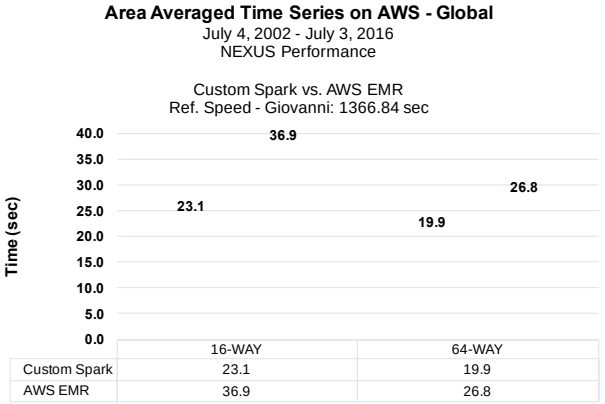
Dataset: 14-years of MODIS AQUA Daily (1 degree daily) Aerosol Optical Depth 550nm (Dark Target) (MYD08_D3v6), Level 3
File Count: 5106
Total 2.6GB



GIOVANNI: 20 min
NEXUS: 1.7 sec







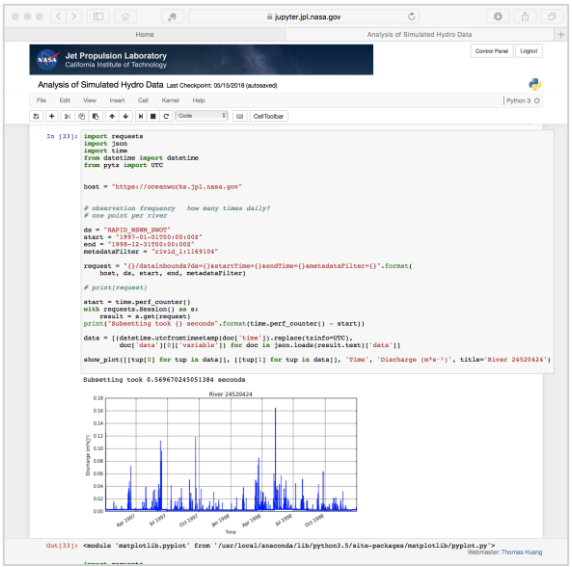
Algorithm execution time. Excludes Giovanni's data scrubbing processing time

ESTO Earth Science Technology Office

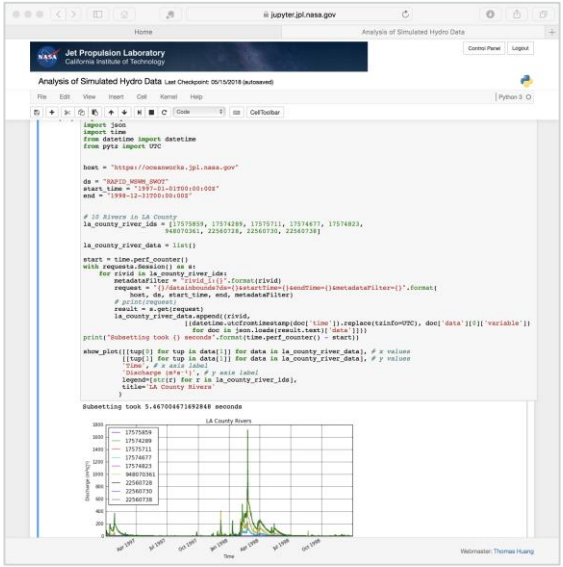
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

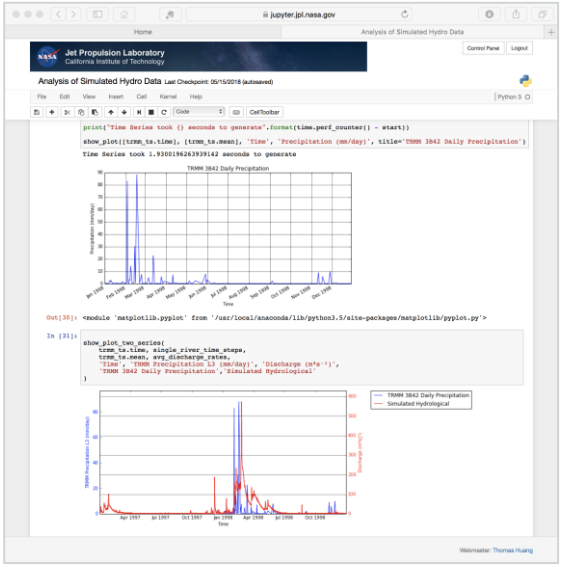
Performance example: support for hydrology



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

ESTO
Earth Science Technology Office

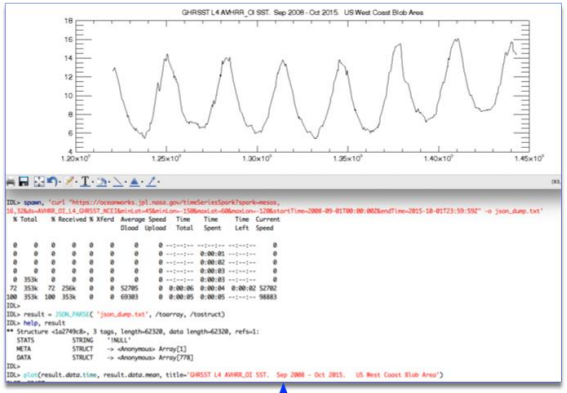
National Aeronautics and Space Administration

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>
```



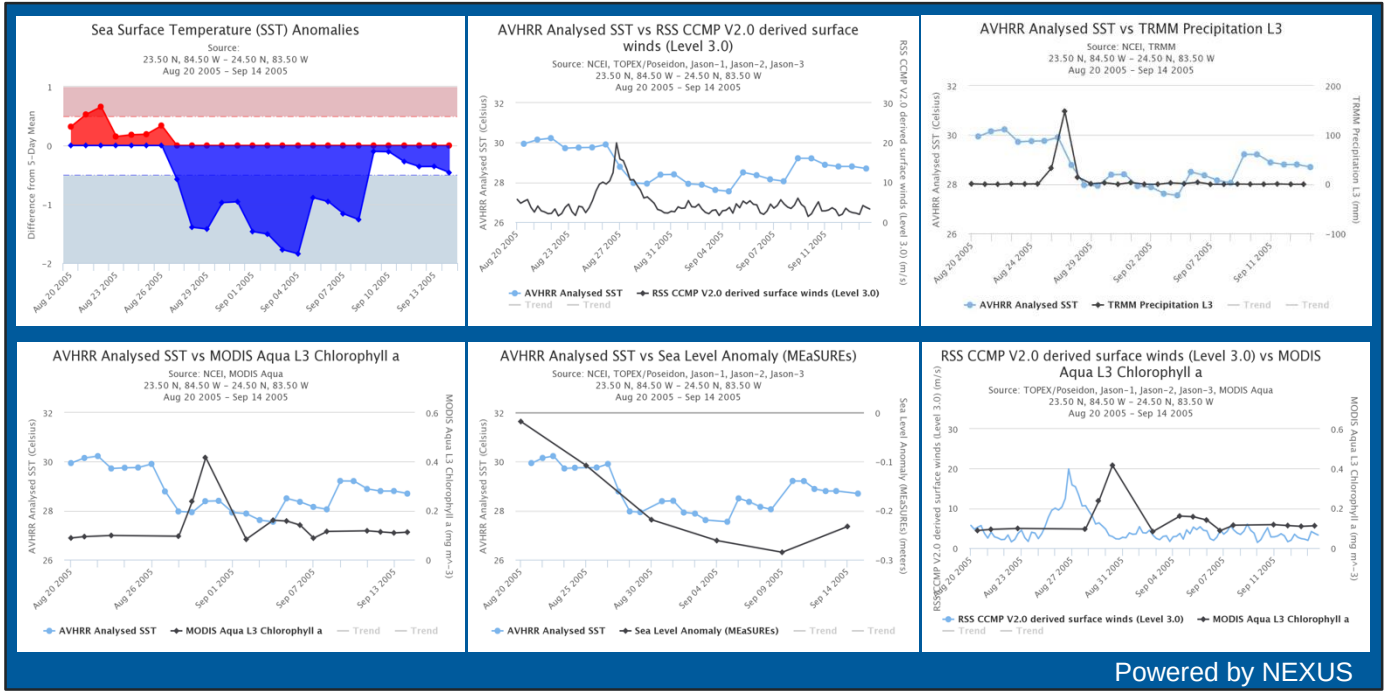
Credit: Ed Armstrong
Jun. 05, 2018

ESTO Earth Science Technology Office

National Aeronautics and Space Administration

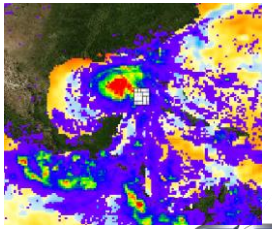
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Hurricane Katrina Study



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

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

ESTO
Earth Science Technology Office

Growing List of 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

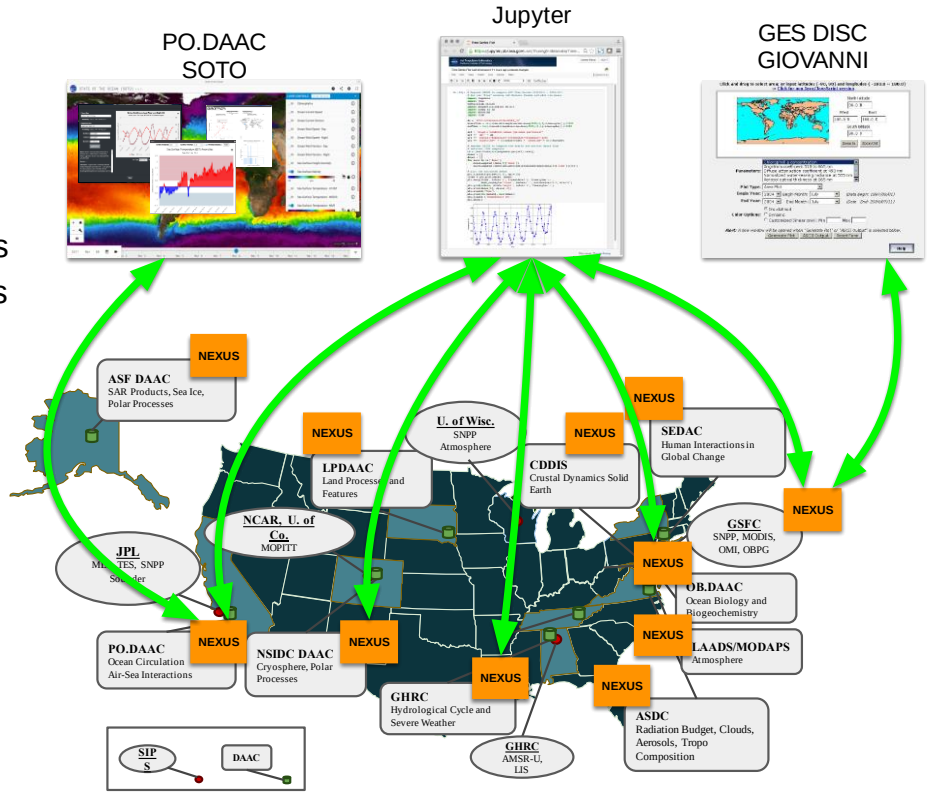
Growing List of Datasets (+)

- **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)

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

Moving Toward Multi-Variable Analysis

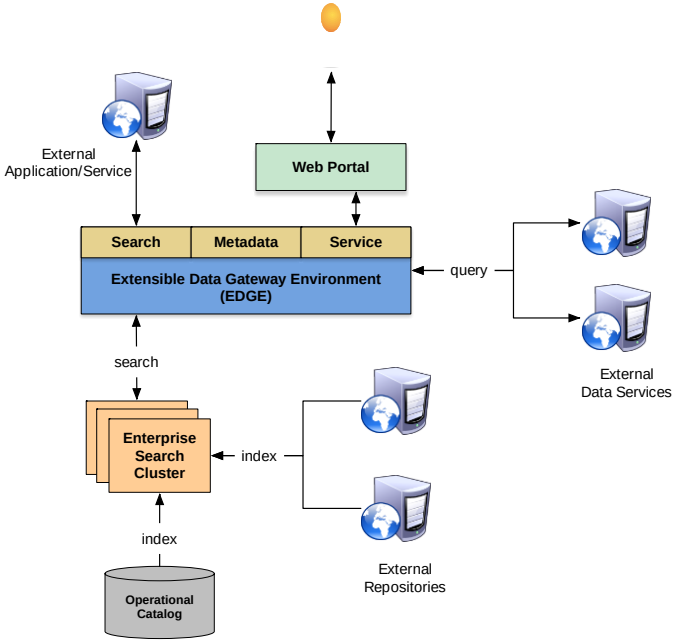
- Public accessible RESTful analytic APIs where computation is next to the data
- NEXUS as the analytic engine infused and managed by the data centers on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files
- Reduce unnecessary data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



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

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

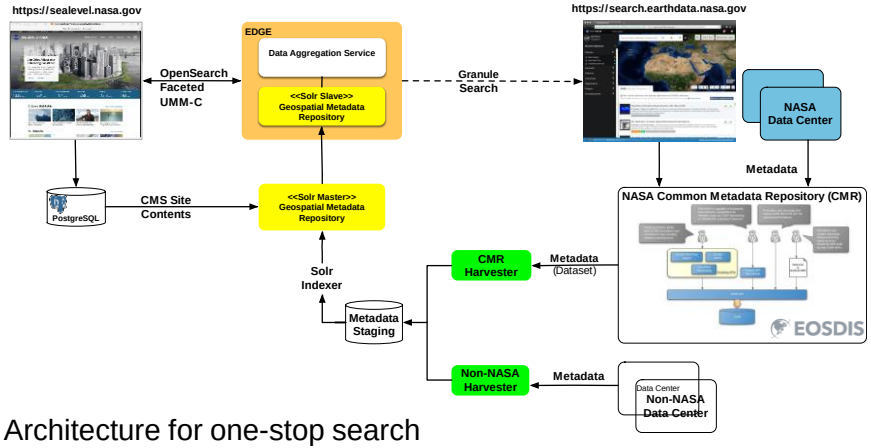


ESTO Earth Science Technology Office

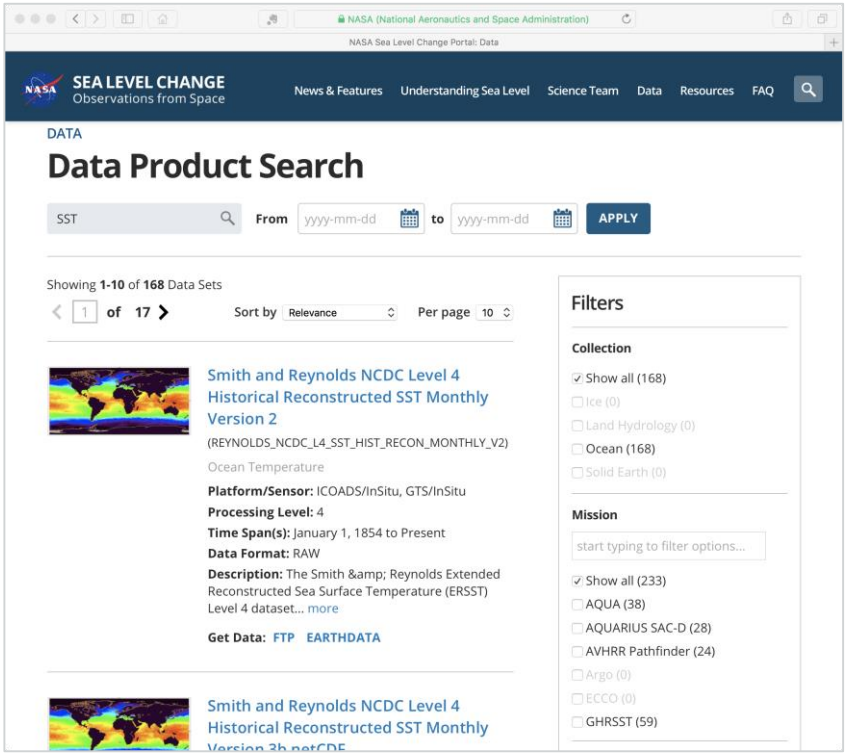
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

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



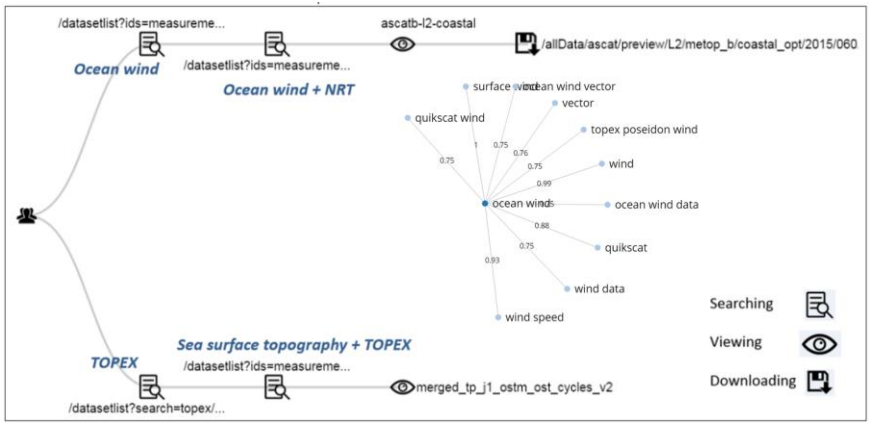
NASA Sea Level Change Portal's One-Stop Search

ESTO Earth Science Technology Office

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

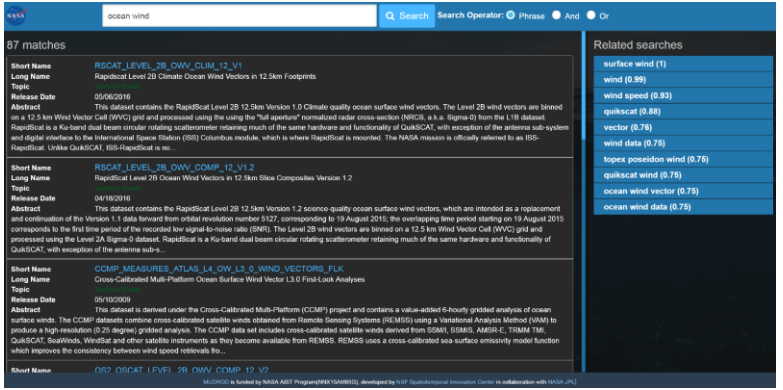
Dynamic Search Ranking

- Adjust search result according how user search and retrieval
- Use machine learning approach to adjust search ranking by taking a number of features into consideration – version, processing level, release date, all-time popularity, monthly-popularity, and user popularity
- Semantically mind dataset metadata to identify relationship

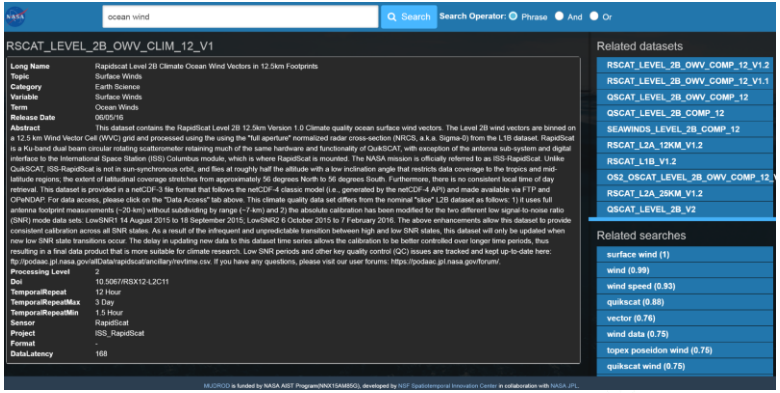


Session Reconstruction

Search Ranking



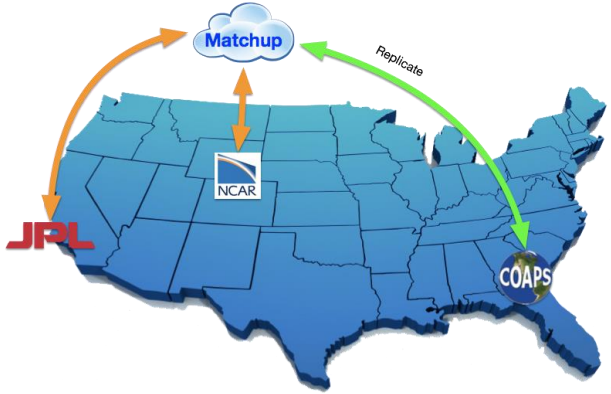
Search Recommendation



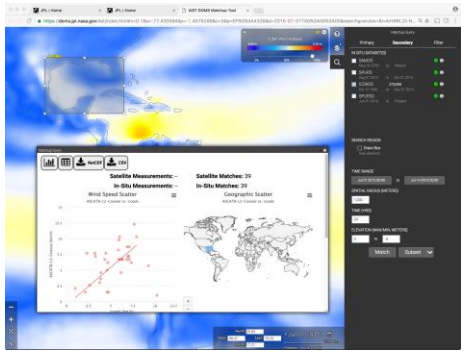
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

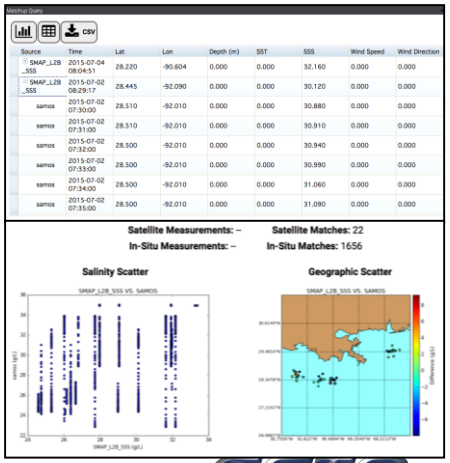
In-Situ to Satellite Matchup

- 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



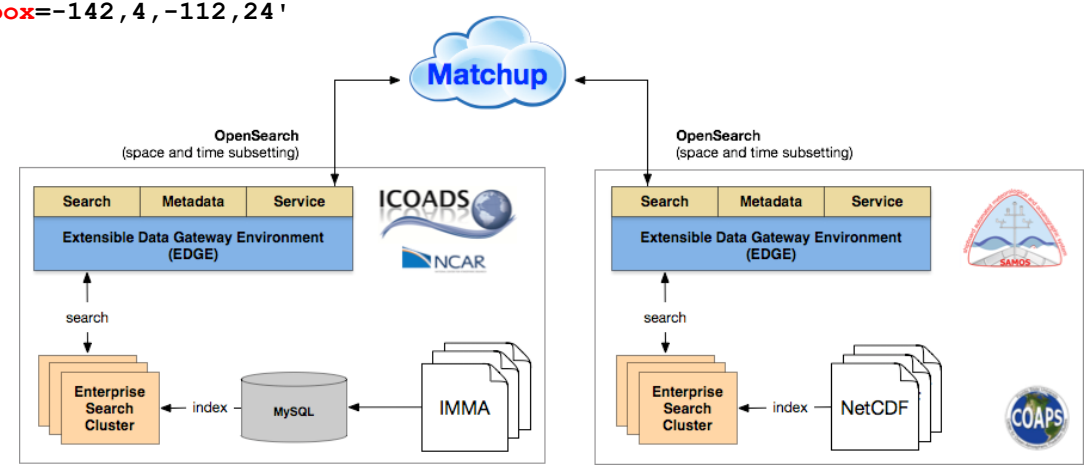


ESTO
Earth Science Technology Office

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

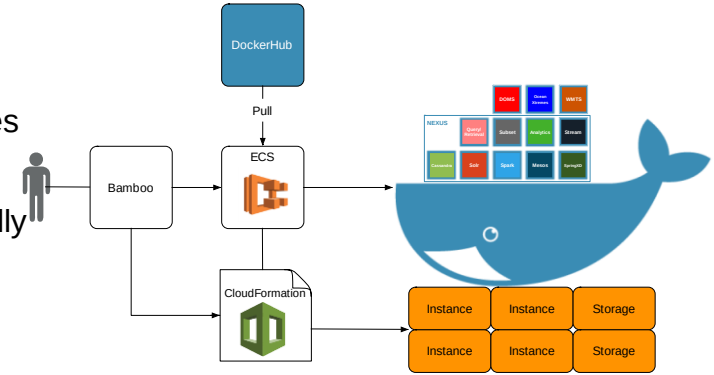
On-The-Fly Subsetting of In-Situ Measurements using OpenSearch

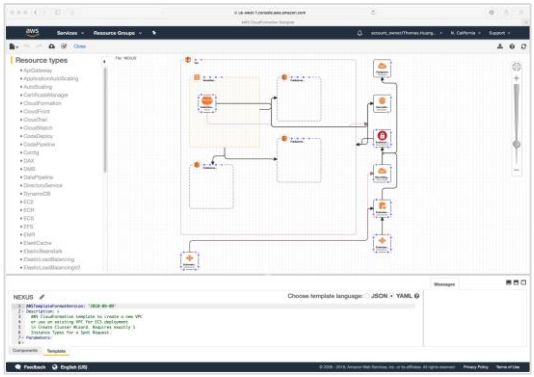
- Using OpenSearch as the standard interface to in-situ data repositories
- Enable distributed, federated search and data subsetting
- Subset in-situ data by time and space using OpenSearch
  - **ICOADS:** `'http://rda-data.ucar.edu:8890/ws/search/icoads?startTime=2012-08-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **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'`



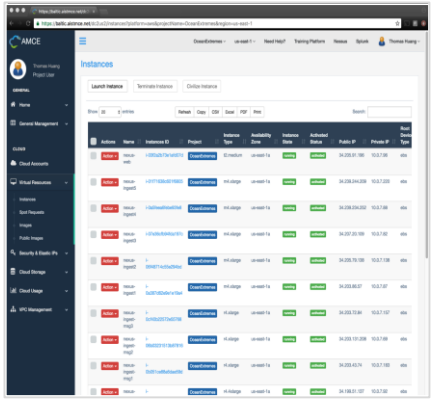
Deployment Automation

- Cloud Deployment is nontrivial
- Infrastructure Definition
  - Various machine instances
  - Storage and buckets
- Software Deployment.. manually
  - Build
  - Package
  - Install
  - Configure
  - Shell login (security issues)
- Best Practice: Deployment Automation
  - Script Infrastructure Definition (e.g. Amazon CloudFormation)
  - Container-based Deployment (e.g. Amazon ECS and DockerHub)

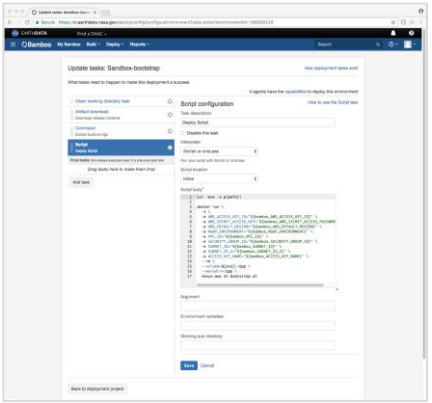




AWS CloudFormation Template Editor



AIST Managed Cloud Environment



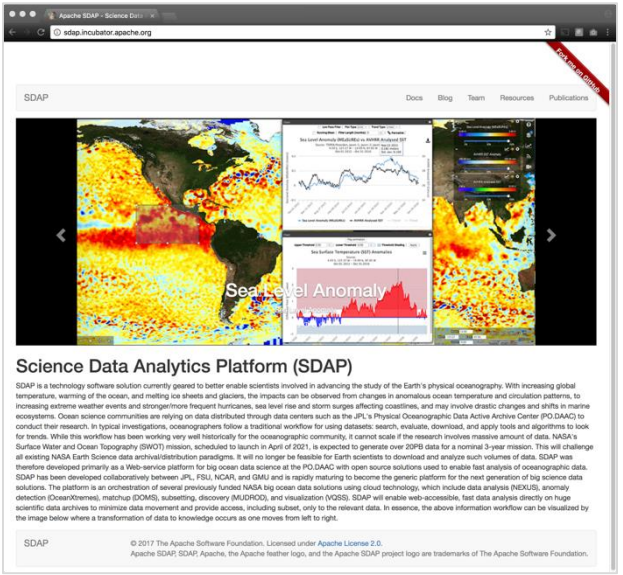
ESDIS NGAP

ESTO Earth Science Technology Office

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

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

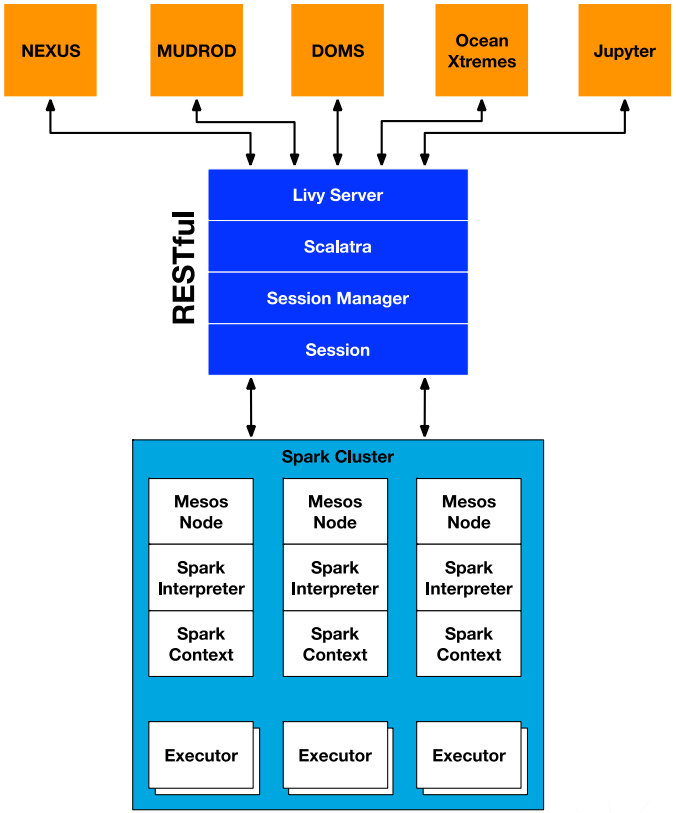
ESTO Earth Science Technology Office

Current Development

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

Architecture for Apache Spark Integration and Sharing

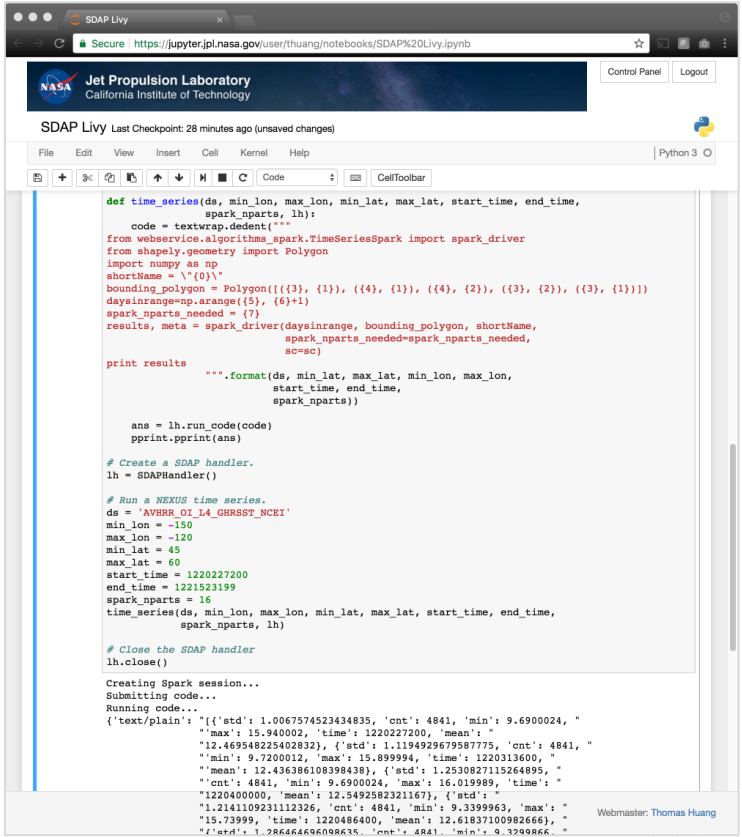
- Developed independently, all the major services in OceanWorks require Apache Spark cluster
- If OceanWorks simply deploy these services to Amazon, it will require dedicated Apache Spark cluster for each
- Too many cluster and very costly, since Apache Spark recommends high memory machine instances
- Looking at the Amazon's EMR model. It is designed to be a job execution solution, and the jobs could from different applications
- Apache Livy provides a RESTful interface to Apache Spark cluster. It is a drop-in service to enable applications to interact with Spark cluster using RESTful api.
- The Apache Livy API also allows users to submit ad hoc map and reduce logics to be handled by the Spark cluster
- Through Apache Livy, scientists could use Jupyter environment to design their analytic algorithms that will be executed in the OceanWorks' Spark Cluster



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

Push Python/Scala Code to Spark Cluster

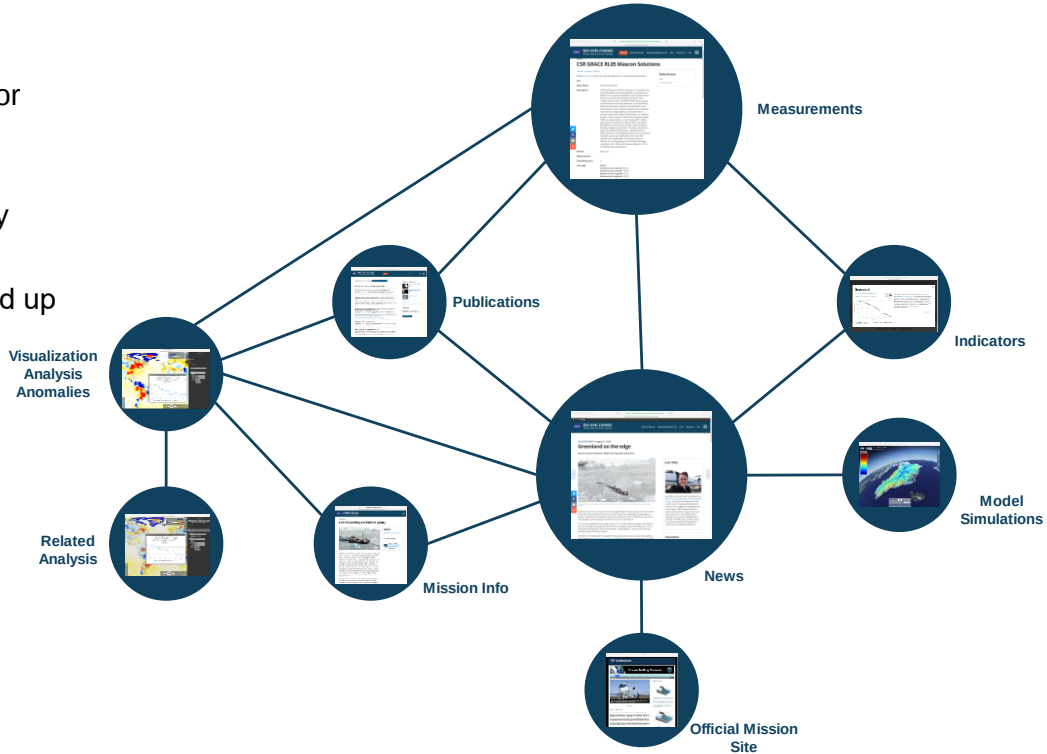
- Provide scientist a platform to develop algorithms to execute in OceanWorks' Spark cluster
- A new OceanWorks' RESTful service to offer flexible environment for researchers to experiment with their algorithms and our data, without having to deal with the complexity of Cloud and job management



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

OceanWorks Tackles Information Discovery

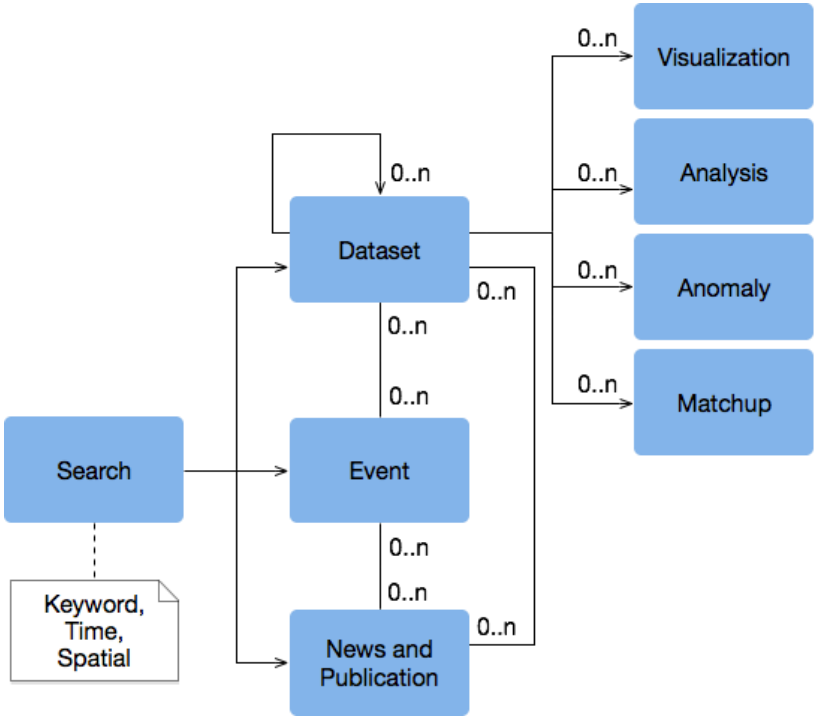
- **Search** is looking for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** is finding something new, or in a new way
  - This is non-trivial
  - Traditional ontological 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



ESTO Earth Science Technology Office

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

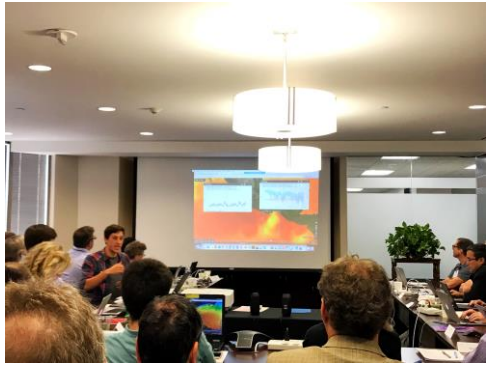
Notable Public Engagements

- 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**

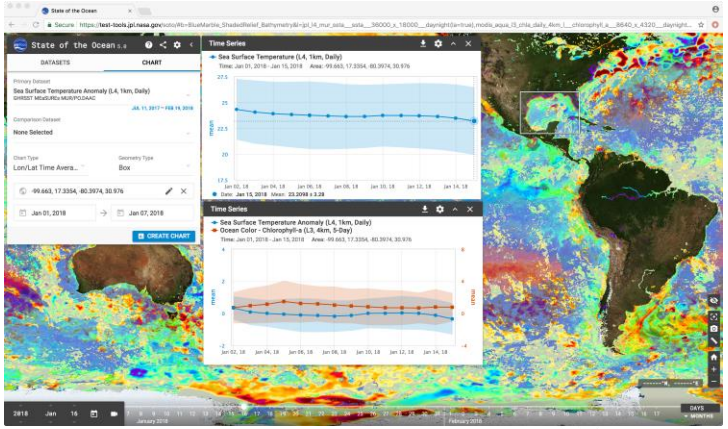
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PO.DAAC UWG 2018

- Demonstrated initial PO.DAAC SOTO integration with OceanWorks' analysis platform on Amazon Web Service
- UWG request immediate access to the new SOTO analytic features
- User Acceptance Testing (UAT) planed in August 2018
- Presented the current state of the OceanWorks' project and map to UWG 2017 recommendations
  - 2017-11.3 Several different projects were coordinated into OceanWorks (i.e. OceanXtremes, NEXUS, DOMS, MUDROD, VQSS)
  - 2016-15 Advanced search capabilities
  - 2016-27 Cloud Computing
  - 2016-36 In situ faceted search



PO.DAAC State of the Ocean (SOTO)
Coming to User Acceptance Testing 2018



ESTO Earth Science Technology Office

In Summary

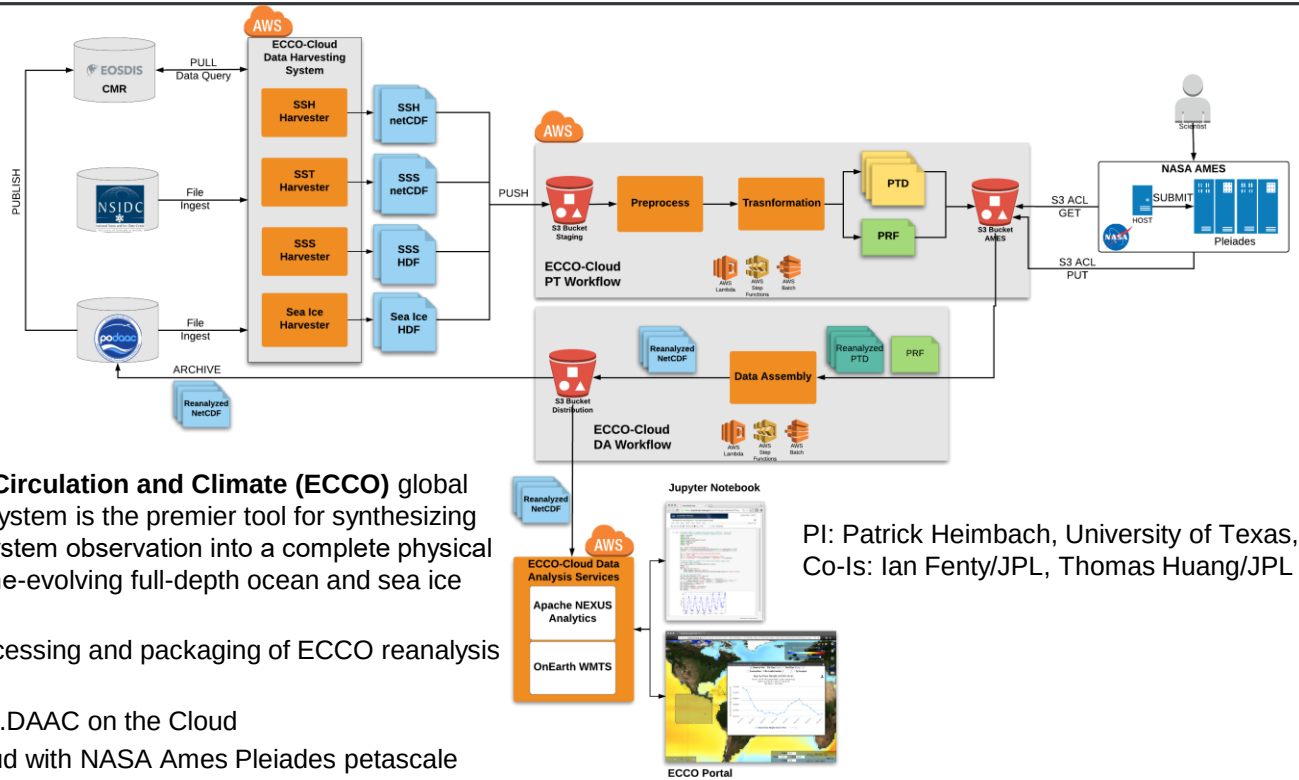
- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Let's think beyond archive and file downloads
- Connected information enables discovery
- Community developed solution through open sourcing
- Investment in data and computational sciences
- Data Centers might want to be in the business of Enabling Science!
- OceanWorks infusion 2018 – 2019
  - Turnkey AMCE deployment
  - Solution for large job management
  - Integration with Amazon Athena
  - Integration with Pangeo
  - 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**

Current OceanWorks Portfolio

- NASA Sea Level Change Portal - https://sealevel.nasa.gov (Production)
- JPL GRACE Website – https://grace.jpl.nasa.gov (Production - June 2018)
- NASA PO.DAAC SOTO (UAT – August 2018)
- CEOS Ocean Variables Enabling Research and Application for GEOS (COVERAGE) (Workshop – Sept. 2018)
- Being review by
  - NASA LARC
  - NOAA
  - Australian Government Bureau of Meteorology (BOM)
  - IFREMER (Institut français de recherche pour l'exploitation de la mer)
  - EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites)

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

NASA ACCESS 2017: ECCO-Cloud (just announced)



- **Estimating the Ocean Circulation and Climate (ECCO)** global ocean state estimation system is the premier tool for synthesizing NASA's diverse Earth system observation into a complete physical description of Earth's time-evolving full-depth ocean and sea ice system.
- 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 Services 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

ESTO Earth Science Technology Office

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

THANK YOU

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
Shawn Smith, Mark A. Bourassa, Jocelyn Elya

National Center for Atmospheric Research Team
Steve J. Worley, Tom Cram, Zaihua Ji

George Mason University Team
Chaowei (Phil) Yang, Yongyao Jiang, and Yun Li

ESTO Earth Science Technology Office