**National Aeronautics and
Space Administration**

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Apache Science Data Analytic Platform (SDAP)

Thomas Huang

Data Scientist | Principal Investigator | Technologist | Architect

thomas.huang@jpl.nasa.gov

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.

Big Data and Data Centers

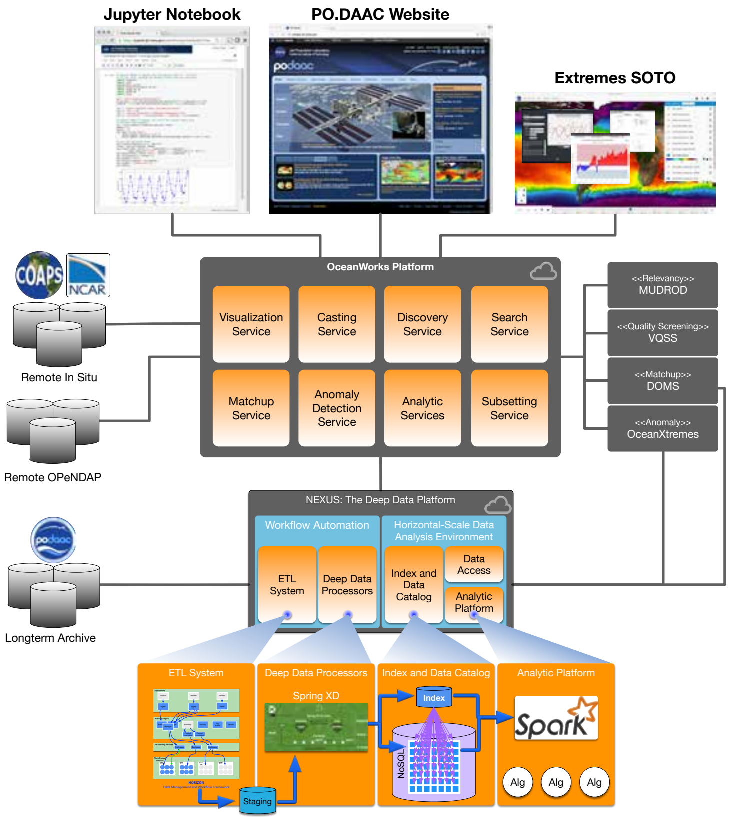
- **Increasing “big data” era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computation and data science across the lifecycle
- **For NASA Data Centers, with large amount of observational and modeling data, downloading to local machine 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
- **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

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

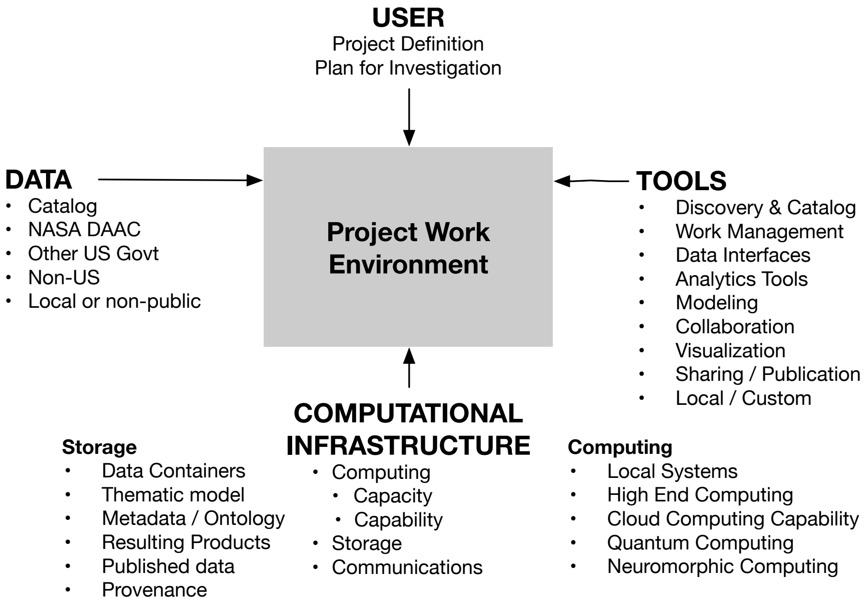
AIST OceanWorks

- **OceanWorks** is to establish an **Integrated Data Analytic 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, FSU, NCAR, and GMU
- Bringing together PO.DAAC-related big data technologies
    - **OceanXtremes –** Anomaly detection and ocean science
    - **NEXUS –** Big data analytic platform
    - **Data Container Studies**
    - **DOMS –** Distributed in-situ to satellite matchup
    - **MUDROD –** Search relevancy and discovery – linking datasets, services, and anomalies through recommendations
    - **VQSS –** Virtualized Quality Screening Service



Integrated Data Analytic Center

- An environment for conducting a Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (ocean, atmospheric, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
  - Reduce the data preparation time to something tolerable
  - Catalog of optional resources
  - Semantic-enabled catalog of resources
  - Relevant publications
  - Provide established training data sets of varying resolution
  - Provide effective project confidentiality, integrity and availability
  - Single sign-on and unified financial tracking



USER
Project Definition
Plan for Investigation

DATA
- Catalog
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

Project Work Environment

TOOLS
- Discovery & Catalog
- Work Management
- Data Interfaces
- Analytics Tools
- Modeling
- Collaboration
- Visualization
- Sharing / Publication
- Local / Custom

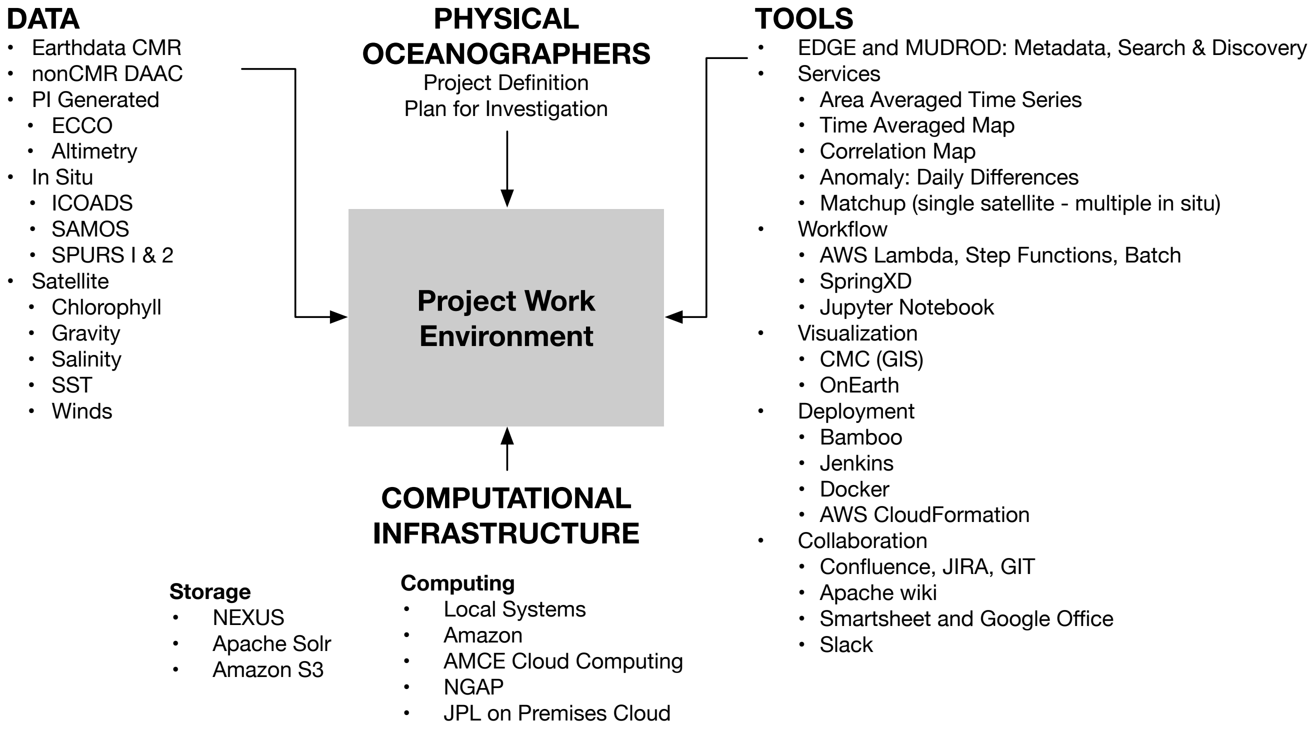
Storage
- Data Containers
- Thematic model
- Metadata / Ontology
- Resulting Products
- Published data
- Provenance

COMPUTATIONAL INFRASTRUCTURE
- Computing
  - Capacity
  - Capability
- Storage
- Communications

Computing
- Local Systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing

Credit: Mike Little, NASA

OceanWorks as an Analytic Center



DATA

- Earthdata CMR
- nonCMR DAAC
- PI Generated
  - ECCO
  - Altimetry
- In Situ
  - ICOADS
  - SAMOS
  - SPURS I & 2
- Satellite
  - Chlorophyll
  - Gravity
  - Salinity
  - SST
  - Winds

PHYSICAL OCEANOGRAPHERS

Project Definition
Plan for Investigation

Project Work Environment

TOOLS

- EDGE and MUDROD: Metadata, Search & Discovery
- Services
  - Area Averaged Time Series
  - Time Averaged Map
  - Correlation Map
  - Anomaly: Daily Differences
  - Matchup (single satellite - multiple in situ)
- Workflow
  - AWS Lambda, Step Functions, Batch
  - SpringXD
  - Jupyter Notebook
- Visualization
  - CMC (GIS)
  - OnEarth
- Deployment
  - Bamboo
  - Jenkins
  - Docker
  - AWS CloudFormation
- Collaboration
  - Confluence, JIRA, GIT
  - Apache wiki
  - Smartsheet and Google Office
  - Slack

COMPUTATIONAL INFRASTRUCTURE

Storage

- NEXUS
- Apache Solr
- Amazon S3

Computing

- Local Systems
- Amazon
- AMCE Cloud Computing
- NGAP
- JPL on Premises Cloud

End User Applications

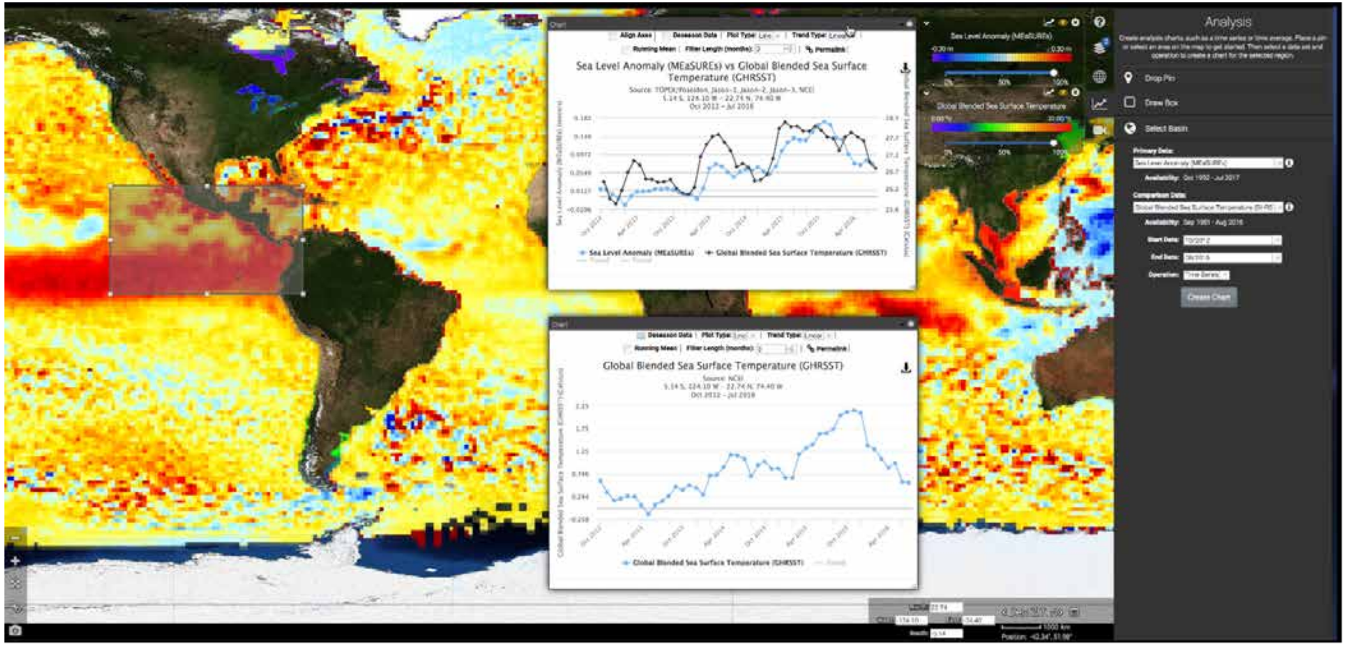
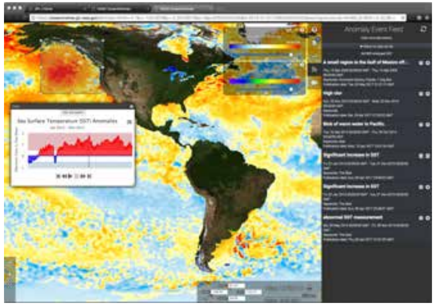
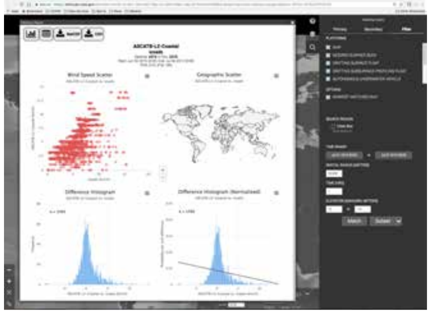
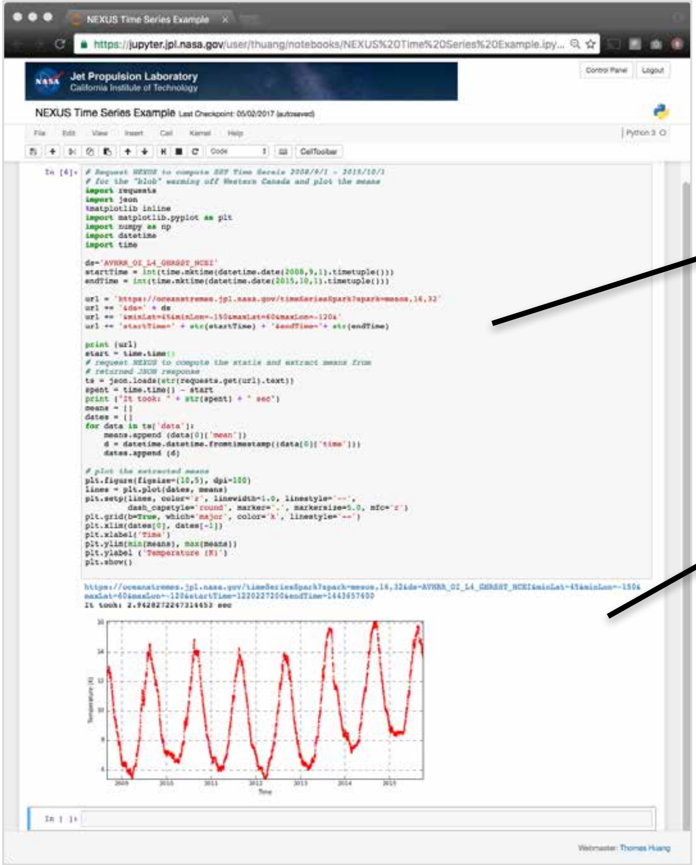


Chart
Align Axes | Deseason Data | Plot Type: Line | Trend Type: Linear
Running Mean | Filter Length (months): 3 | Permalink
Sea Level Anomaly (MEaSUREs) vs Global Blended Sea Surface Temperature (GHRSST)
Source: TOPEX/Poseidon, Jason-1, Jason-2, Jason-3, NCEI
5.14 S, 124.10 W - 22.74 N, 74.40 W
Oct 2012 - Jul 2016
Sea Level Anomaly (MEaSUREs)
Global Blended Sea Surface Temperature (GHRSST)
Global Blended Sea Surface Temperature (GHRSST)
Source: NCEI
Sea Level Anomaly (MEaSUREs)
-0.30 m
0.30 m
0%
50%
100%
Global Blended Sea Surface Temperature
Analysis
Drop Pin
Draw Box
Select Basin
Primary Data:
Sea Level Anomaly (MEaSUREs)
Availability: Oct 1992 - Jul 2017
Comparison Data:
Global Blended Sea Surface Temperature (GHRSST)
Availability: Sep 1981 - Aug 2016
Start Date:
End Date:
Operation: Time Series
Create Chart





Enable Science without File Download

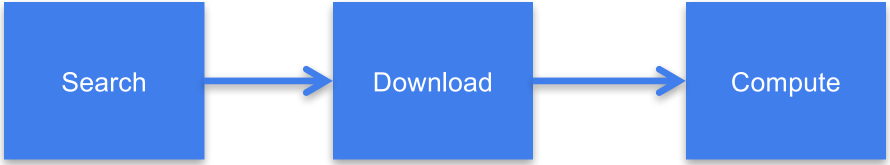

NEXUS Time Series Example
https://jupyter.jpl.nasa.gov/user/thuang/notebooks/NEXUS%20Time%20Series%20Example.ipy...
Jet Propulsion Laboratory
California Institute of Technology
Control Panel
Logout
NEXUS Time Series Example Last Checkpoint: 05/02/2017 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
Python 3
Code
CellToolbar
16
14
12
10
8
6
Temperature (K)
2009
2010
2011
2012
2013
2014
2015
Time
Webmaster: Thomas Huang

```
In [4]: # Request NEXUS to compute SST Time Sereis 2008/9/1 - 2015/10/1
        # for the "blob" warming off Western Canada and plot the means
        import requests
        import json
        %matplotlib inline
        import matplotlib.pyplot as plt
        import numpy as np
        import datetime
        import time

        ds='AVHRR_OI_L4_GHRSST_NCEI'
        startTime = int(time.mktime(datetime.date(2008,9,1).timetuple()))
        endTime = int(time.mktime(datetime.date(2015,10,1).timetuple()))

        url = 'https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32'
        url += '&ds=' + ds
        url += '&minLat=45&minLon=-150&maxLat=60&maxLon=-120&'
        url += 'startTime=' + str(startTime) + '&endTime='+ str(endTime)

        print (url)
        start = time.time()
        # request NEXUS to compute the static and extract means from
        # returned JSON response
        ts = json.loads(str(requests.get(url).text))
        spent = time.time() - start
        print ("It took: " + str(spent) + " sec")
        means = []
        dates = []
        for data in ts['data']:
            means.append (data[0]['mean'])
            d = datetime.datetime.fromtimestamp((data[0]['time']))
            dates.append (d)

        # plot the extracted means
        plt.figure(figsize=(10,5), dpi=100)
        lines = plt.plot(dates, means)
        plt.setp(lines, color='r', linewidth=1.0, linestyle='--',
                 dash_capstyle='round', marker='.', markersize=5.0, mfc='r')
        plt.grid(b=True, which='major', color='k', linestyle='--')
        plt.xlim(dates[0], dates[-1])
        plt.xlabel('Time')
        plt.ylim(min(means), max(means))
        plt.ylabel ('Temperature (K)')
        plt.show()

https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&
maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600
It took: 2.9428272247314453 sec

In [ ]:
```

```
# 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://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=me
sos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-
150&maxLat=60&maxLon=-
120&startTime=1220227200&endTime=1443657600

It took: 2.9428272247314453 sec
```

Traditional Method for Analyze Satellite Measurements

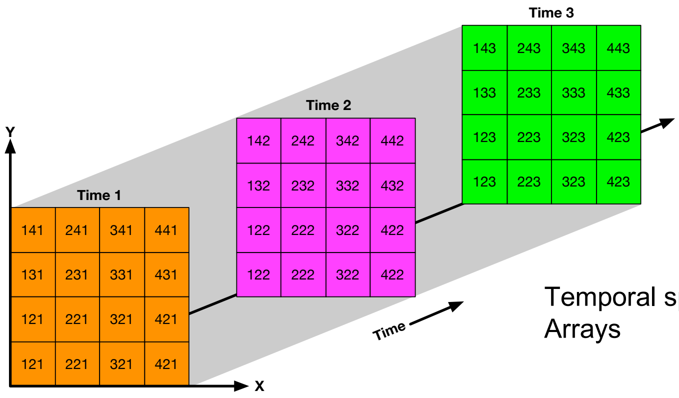


Search → Download → Compute

- 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

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



Time 1

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

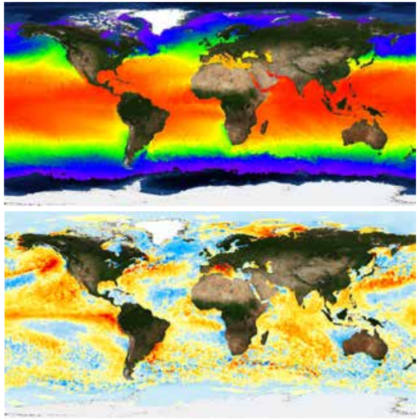
Time 2

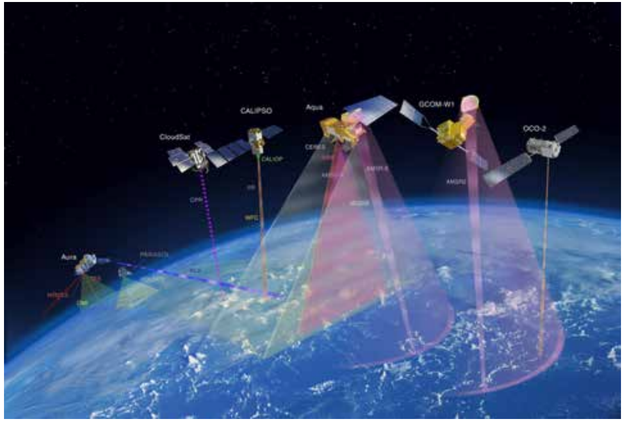
142	242	342	442
132	232	332	432
122	222	322	422
122	222	322	422

Time 3

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

Temporal spatial Arrays





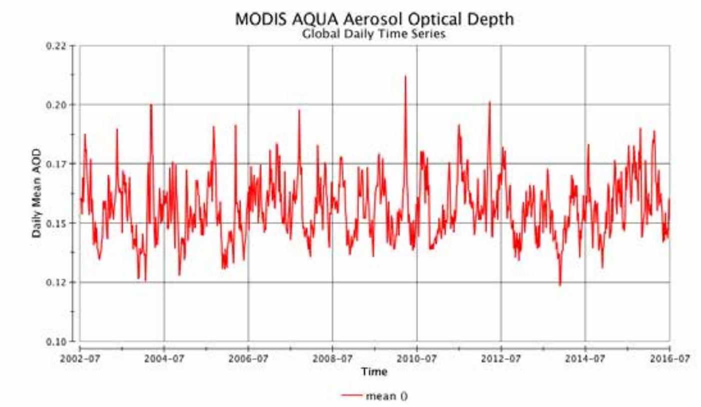
NEXUS Performance: Custom Spark vs. AWS EMR

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

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

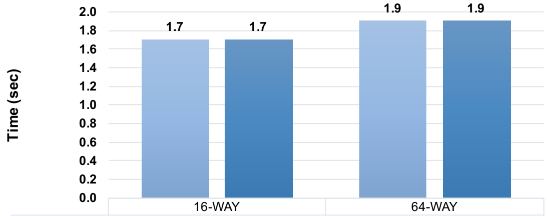
- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

AWS EMR: Amazon's provisioned MapReduce cluster


MODIS AQUA Aerosol Optical Depth
Global Daily Time Series

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

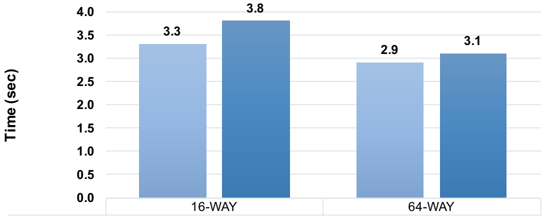
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1140.22 sec



	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

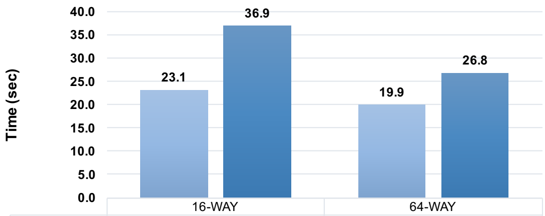
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1150.6 sec



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

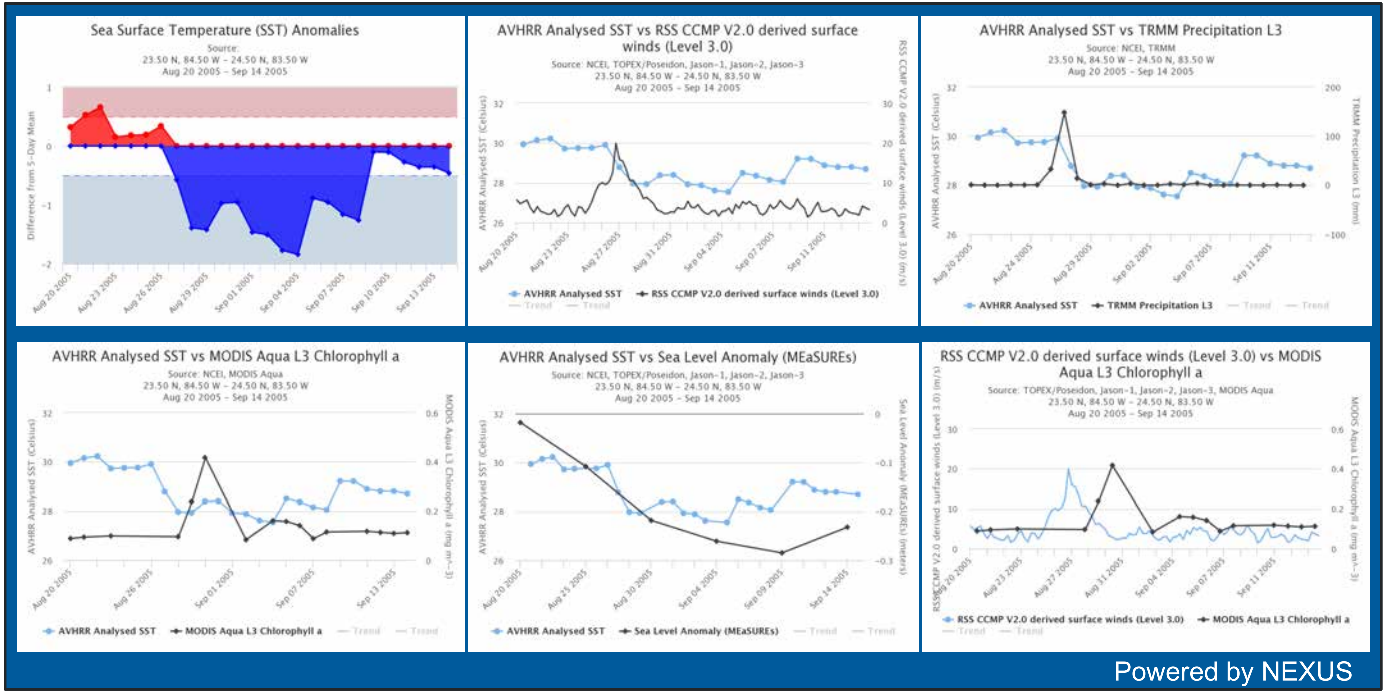
Area Averaged Time Series on AWS - Global
July 4, 2002 - July 3, 2016
NEXUS Performance

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

Hurricane Katrina Study


Sea Surface Temperature (SST) Anomalies
Source:
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

AVHRR Analysed SST vs RSS CCMP V2.0 derived surface winds (Level 3.0)
Source: NCEI, TOPEX/Poseidon, Jason-1, Jason-2, Jason-3
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

AVHRR Analysed SST vs TRMM Precipitation L3
Source: NCEI, TRMM
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

AVHRR Analysed SST vs MODIS Aqua L3 Chlorophyll a
Source: NCEI, MODIS Aqua
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

AVHRR Analysed SST vs Sea Level Anomaly (MEaSUREs)
Source: NCEI, TOPEX/Poseidon, Jason-1, Jason-2, Jason-3
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

RSS CCMP V2.0 derived surface winds (Level 3.0) vs MODIS Aqua L3 Chlorophyll a
Source: TOPEX/Poseidon, Jason-1, Jason-2, Jason-3, MODIS Aqua
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

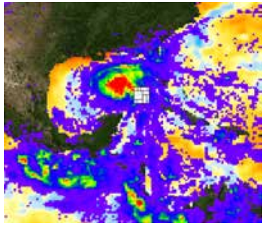
Powered by NEXUS

Difference from 5-Day Mean
AVHRR Analysed SST (Celsius)
RSS CCMP V2.0 derived surface winds (Level 3.0) (m/s)
TRMM Precipitation L3 (mm)
MODIS Aqua L3 Chlorophyll a (mg m^-3)
Sea Level Anomaly (MEaSUREs) (meters)
AVHRR Analysed SST
RSS CCMP V2.0 derived surface winds (Level 3.0)
TRMM Precipitation L3
MODIS Aqua L3 Chlorophyll a
Sea Level Anomaly (MEaSUREs)
Trend
Aug 20 2005
Aug 23 2005
Aug 26 2005
Aug 29 2005
Sep 01 2005
Sep 04 2005
Sep 07 2005
Sep 10 2005
Sep 13 2005

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

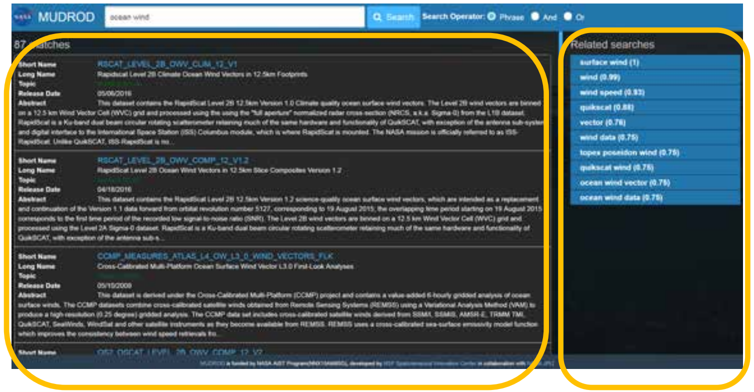
Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 ℃ that lingers for several days. Following this was a short intense ocean 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

Search and Discovery

- **Search** – look for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** – find 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
  - What happen when we have a lot of **`subClassOf`**, **`equivalentClassOf`**, **`sameAs`**?
  - How wide and deep should we go?
- **Relevancy**
  - It is domain-specific
  - It is personal
  - It is temporal
  - It is dynamic



MUDROD | ocean wind | Search | Search Operator: ● Phrase ○ And ○ Or

87 matches

Short Name RSCAT_LEVEL_2B_OWV_CLIM_12_V1
Long Name Rapidscat Level 2B Climate Ocean Wind Vectors in 12.5km Footprints
Topic
Release Date 05/06/2016
Abstract This dataset contains the RapidScat Level 2B 12.5km Version 1.0 Climate quality ocean surface wind vectors. The Level 2B wind vectors are binned on a 12.5 km Wind Vector Cell (WVC) grid and processed using the using the "full aperture" normalized radar cross-section (NRCS, a.k.a. Sigma-0) from the L1B dataset. RapidScat is a Ku-band dual beam circular rotating scatterometer retaining much of the same hardware and functionality of QuikSCAT, with exception of the antenna sub-system and digital interface to the International Space Station (ISS) Columbus module, which is where RapidScat is mounted. The NASA mission is officially referred to as ISS-RapidScat. Unlike QuikSCAT, ISS-RapidScat is no...

Short Name RSCAT_LEVEL_2B_OWV_COMP_12_V1.2
Long Name RapidScat Level 2B Ocean Wind Vectors in 12.5km Slice Composites Version 1.2
Topic
Release Date 04/18/2016
Abstract This dataset contains the RapidScat Level 2B 12.5km Version 1.2 science-quality ocean surface wind vectors, which are intended as a replacement and continuation of the Version 1.1 data forward from orbital revolution number 5127, corresponding to 19 August 2015; the overlapping time period starting on 19 August 2015 corresponds to the first time period of the recorded low signal-to-noise ratio (SNR). The Level 2B wind vectors are binned on a 12.5 km Wind Vector Cell (WVC) grid and processed using the Level 2A Sigma-0 dataset. RapidScat is a Ku-band dual beam circular rotating scatterometer retaining much of the same hardware and functionality of QuikSCAT, with exception of the antenna sub-s...

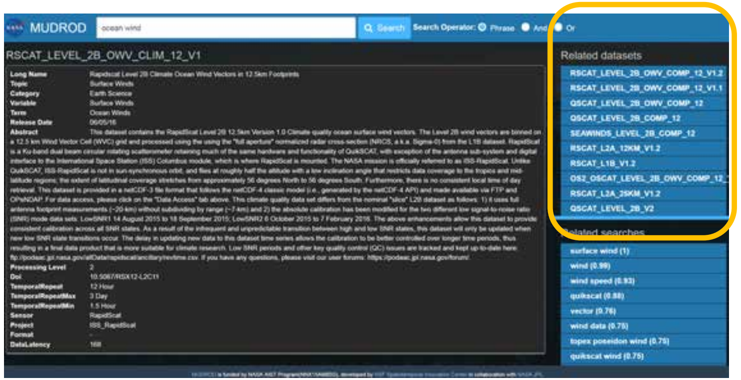
Short Name CCMP_MEASURES_ATLAS_L4_OW_L3_0_WIND_VECTORS_FLK
Long Name Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L3.0 First-Look Analyses
Topic
Release Date 05/15/2009
Abstract This dataset is derived under the Cross-Calibrated Multi-Platform (CCMP) project and contains a value-added 6-hourly gridded analysis of ocean surface winds. The CCMP datasets combine cross-calibrated satellite winds obtained from Remote Sensing Systems (REMSS) using a Variational Analysis Method (VAM) to produce a high-resolution (0.25 degree) gridded analysis. The CCMP data set includes cross-calibrated satellite winds derived from SSMI, SSMIS, AMSR-E, TRMM TMI, QuikSCAT, SeaWinds, WindSat and other satellite instruments as they become available from REMSS. REMSS uses a cross-calibrated sea-surface emissivity model function which improves the consistency between wind speed retrievals fro...

Short Name OS2_OSCAT_LEVEL_2B_OWV_COMP_12_V2

Related searches
- surface wind (1)
- wind (0.99)
- wind speed (0.93)
- quikscat (0.88)
- vector (0.76)
- wind data (0.75)
- topex poseidon wind (0.75)
- quikscat wind (0.75)
- ocean wind vector (0.75)
- ocean wind data (0.75)

MUDROD is funded by NASA AIST Program(NNX15AM85G), developed by NSF Spatiotemporal Innovation Center in collaboration with NASA JPL

Search Ranking
Based on a machine learning model (RankSVM) which takes a number of features, such as vector space model, version, processing level, release date, all-time popularity, monthly-popularity, and user popularity.



MUDROD | ocean wind | Search | Search Operator: ● Phrase ○ And ○ Or

RSCAT_LEVEL_2B_OWV_CLIM_12_V1

Long Name Rapidscat Level 2B Climate Ocean Wind Vectors in 12.5km Footprints
Topic Surface Winds
Category Earth Science
Variable Surface Winds
Term Ocean Winds
Release Date 06/05/16
Abstract This dataset contains the RapidScat Level 2B 12.5km Version 1.0 Climate quality ocean surface wind vectors. The Level 2B wind vectors are binned on a 12.5 km Wind Vector Cell (WVC) grid and processed using the using the "full aperture" normalized radar cross-section (NRCS, a.k.a. Sigma-0) from the L1B dataset. RapidScat is a Ku-band dual beam circular rotating scatterometer retaining much of the same hardware and functionality of QuikSCAT, with exception of the antenna sub-system and digital interface to the International Space Station (ISS) Columbus module, which is where RapidScat is mounted. The NASA mission is officially referred to as ISS-RapidScat. Unlike QuikSCAT, ISS-RapidScat is not in sun-synchronous orbit, and flies at roughly half the altitude with a low inclination angle that restricts data coverage to the tropics and mid-latitude regions; the extent of latitudinal coverage stretches from approximately 56 degrees North to 56 degrees South. Furthermore, there is no consistent local time of day retrieval. This dataset is provided in a netCDF-3 file format that follows the netCDF-4 classic model (i.e., generated by the netCDF-4 API) and made available via FTP and OPeNDAP. For data access, please click on the "Data Access" tab above. This climate quality data set differs from the nominal "slice" L2B dataset as follows: 1) it uses full antenna footprint measurements (~20-km) without subdividing by range (~7-km) and 2) the absolute calibration has been modified for the two different low signal-to-noise ratio (SNR) mode data sets: LowSNR1 14 August 2015 to 18 September 2015; LowSNR2 6 October 2015 to 7 February 2016. The above enhancements allow this dataset to provide consistent calibration across all SNR states. As a result of the infrequent and unpredictable transition between high and low SNR states, this dataset will only be updated when new low SNR state transitions occur. The delay in updating new data to this dataset time series allows the calibration to be better controlled over longer time periods, thus resulting in a final data product that is more suitable for climate research. Low SNR periods and other key quality control (QC) issues are tracked and kept up-to-date here: ftp://podaac.jpl.nasa.gov/allData/rapidscat/ancillary/revtime.csv. If you have any questions, please visit our user forums: https://podaac.jpl.nasa.gov/forum/.
Processing Level 2
Doi 10.5067/RSX12-L2C11
TemporalRepeat 12 Hour
TemporalRepeatMax 3 Day
TemporalRepeatMin 1.5 Hour
Sensor RapidScat
Project ISS_RapidScat
Format -
DataLatency 168

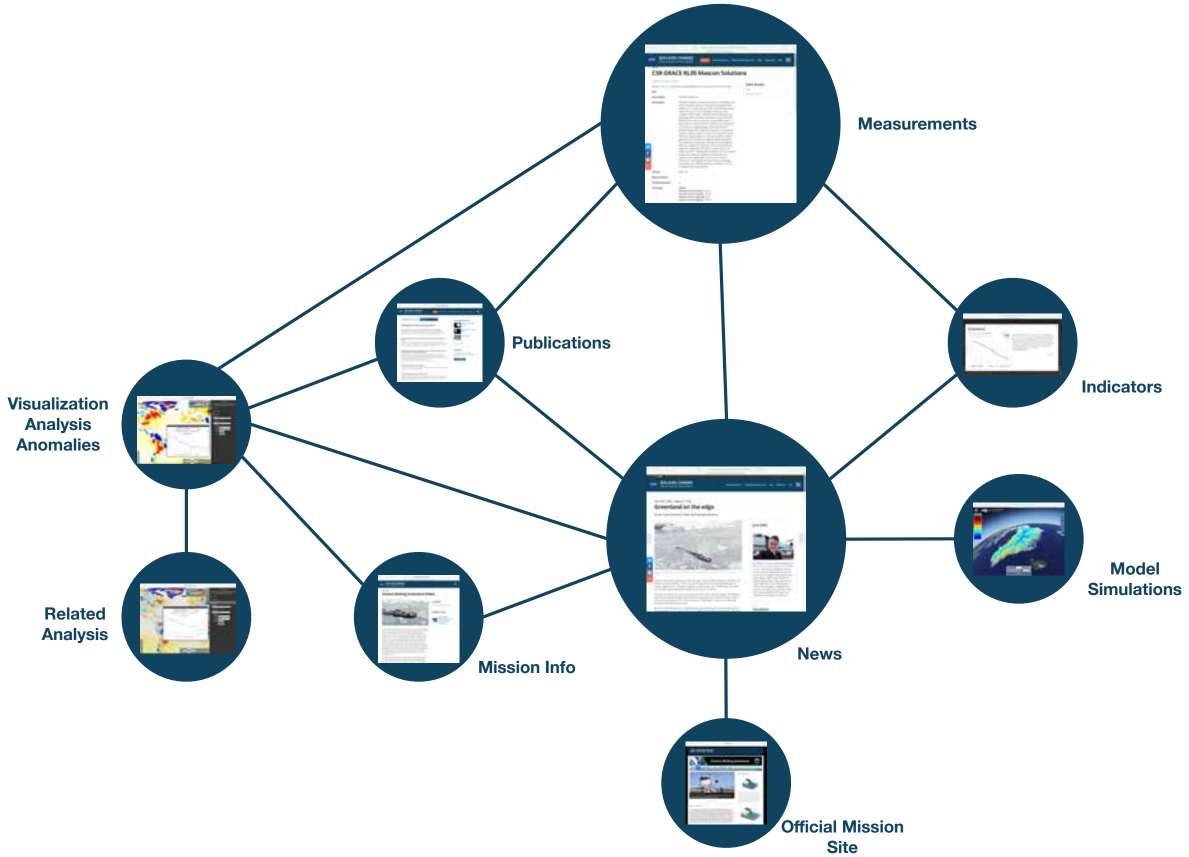
Related datasets
- RSCAT_LEVEL_2B_OWV_COMP_12_V1.2
- RSCAT_LEVEL_2B_OWV_COMP_12_V1.1
- QSCAT_LEVEL_2B_OWV_COMP_12
- QSCAT_LEVEL_2B_COMP_12
- SEAWINDS_LEVEL_2B_COMP_12
- RSCAT_L2A_12KM_V1.2
- RSCAT_L1B_V1.2
- OS2_OSCAT_LEVEL_2B_OWV_COMP_12_
- RSCAT_L2A_25KM_V1.2
- QSCAT_LEVEL_2B_V2

Related searches
- surface wind (1)
- wind (0.99)
- wind speed (0.93)
- quikscat (0.88)
- vector (0.76)
- wind data (0.75)
- topex poseidon wind (0.75)
- quikscat wind (0.75)

MUDROD is funded by NASA AIST Program(NNX15AM85G), developed by NSF Spatiotemporal Innovation Center in collaboration with NASA JPL

Search Recommendation
Based on dataset metadata content and web session co-occurrence

Developing Information Discovery Solutions

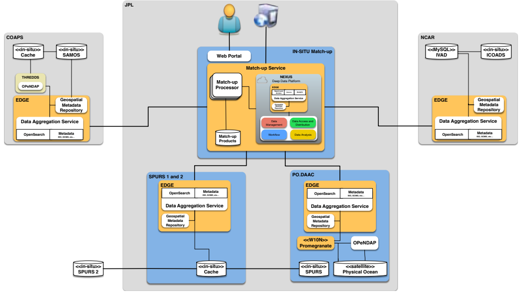


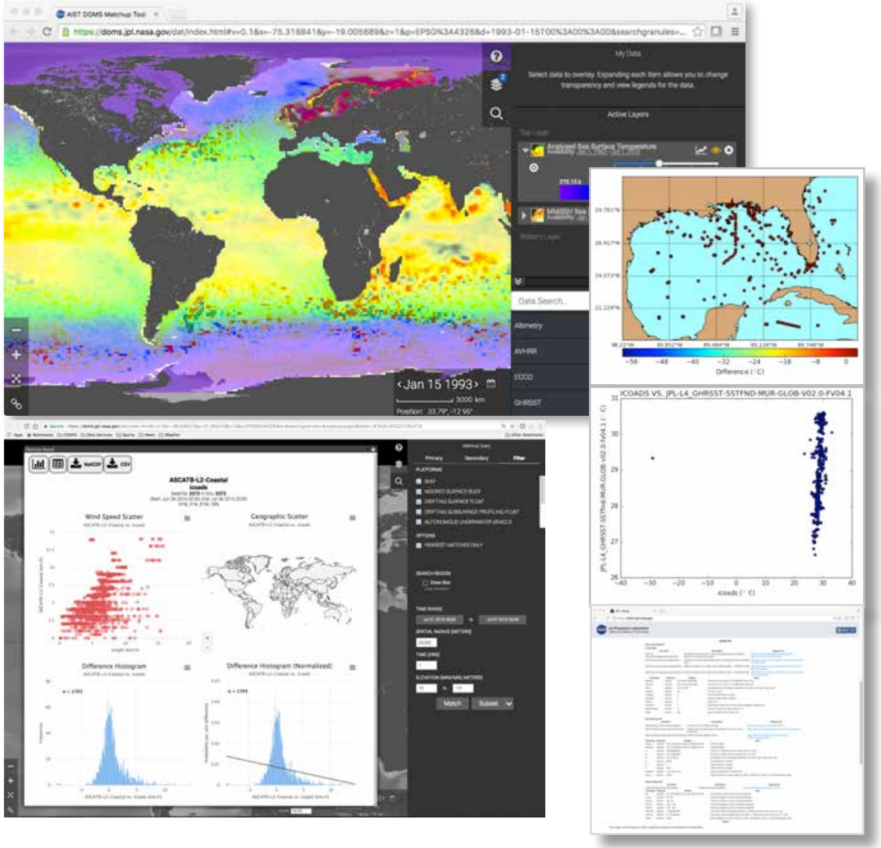
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

In Situ to Satellite Matchup

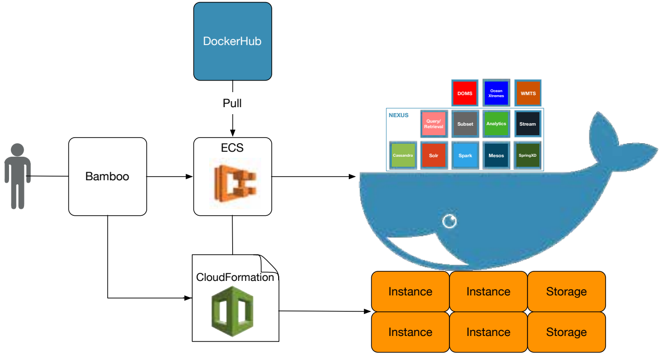
- Distributed Oceanographic Matchup Service
- Typically data matching is done using one-off programs developed at multiple institutions
- A primary advantage of DOMS is the reduction in duplicate development and man hours required to match satellite/in situ data
    - Removes the need for satellite and in situ data to be collocated on a single server
    - Systematically recreate matchups if either in situ or satellite products are re-processed (new versions), i.e., matchup archives are always up-to-date.
- In situ data nodes at JPL, NCAR, and FSU operational.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Prototype graphical user interface (UI) and APIs accessible for external users.
- Plugin architecture for in situ data providers

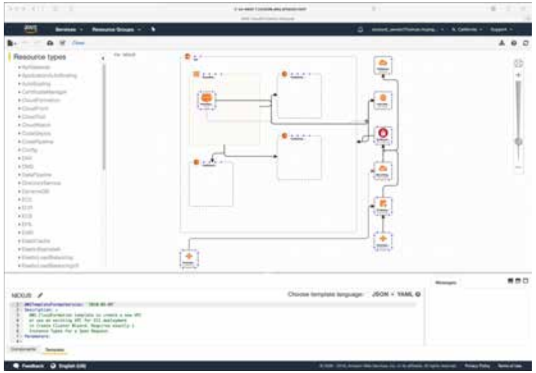


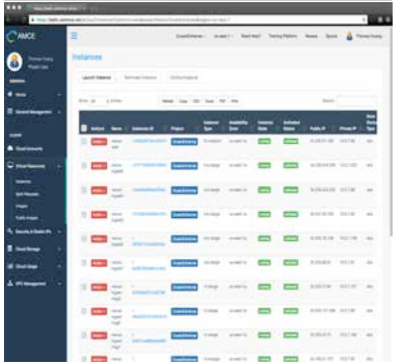


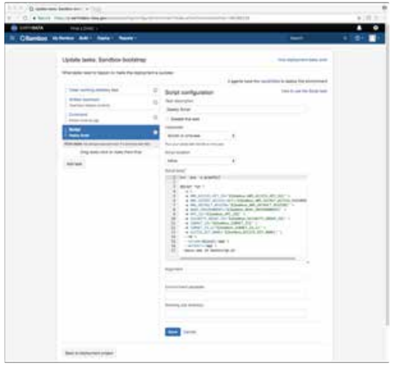
Deployment Automation

- 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)







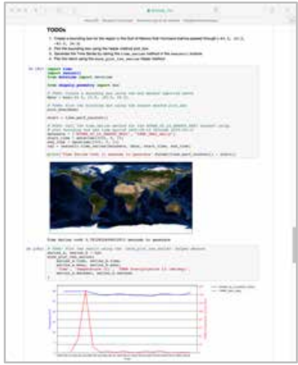


AMCE Deployment

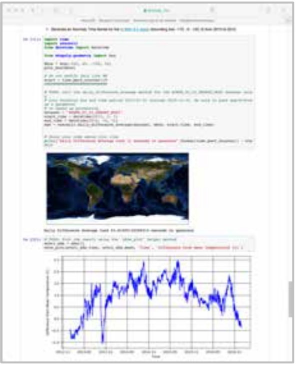
NGAP Deployment

Working with both Science and Informatics Communities

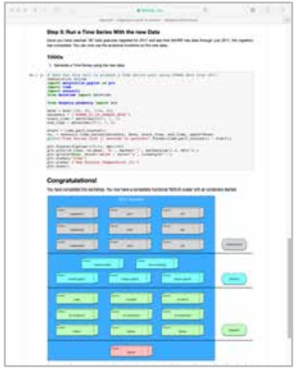
- Established Apache Incubator project
- OceanWorks is developed in the open
- Target Apache top-level project by 2019.
- Public hands-on workshops
- Organize technical sessions at conferences
- Invited speaker and panelist
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**



Analyze Hurricane Katrina by comparing SST and TRMM time series



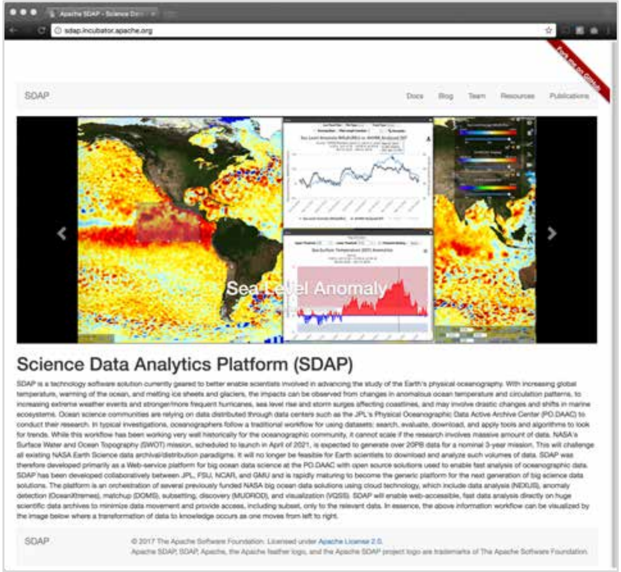
Generate daily difference average
"The Blob" is an oceanographic anomaly



Each participant deployed 3 computing clusters, a total of 24 containers on EC2

Open Source

- Technology sharing through Free and Open Source Software (FOSS)
- Further technology evolution that is restricted by projects / missions
- **Science Data Analytic Platform (SDAP)**, the implementation of **OceanWorks**, in **Apache Incubator**
  - Cloud platform
  - Analyzing satellite and model data
  - In situ data analysis and coordination with satellite measurements
  - Fast data subsetting
  - Mining of user interactions and data to enable discovery and recommendations
  - Streamline deployment through container technology



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

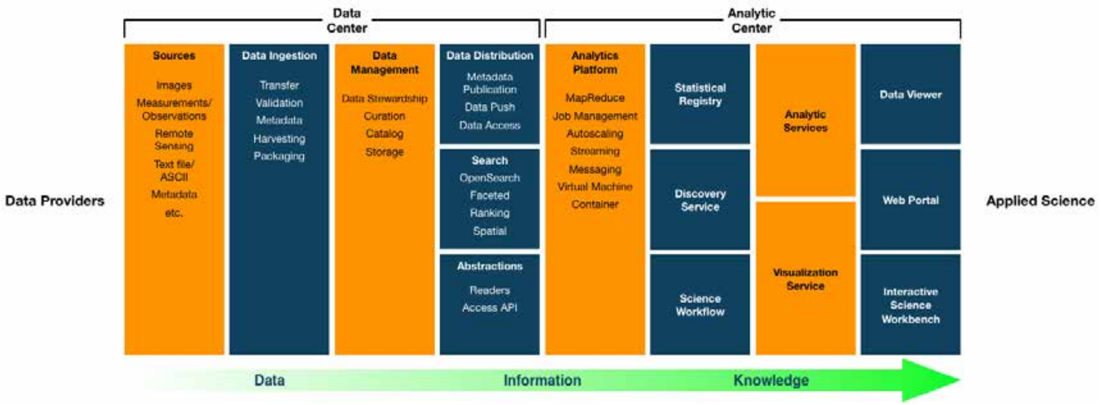
http://sdap.incubator.apache.org

APACHE INCUBATOR

In Summary

- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Think beyond the archive
- Connected information enables discovery
- Community developed solution through open sourcing
- Thanks to the NASA ESTO/AIST and Sea Level Rise programs, and the NASA ESDIS project
- 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 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)

Transforming Data to Knowledge



Data Center

Analytic Center

Data Providers

Sources: Images, Measurements/ Observations, Remote Sensing, Text file/ ASCII, Metadata, etc.

Data Ingestion: Transfer, Validation, Metadata, Harvesting, Packaging

Data Management: Data Stewardship, Curation, Catalog, Storage

Data Distribution: Metadata Publication, Data Push, Data Access

Search: OpenSearch, Faceted, Ranking, Spatial

Abstractions: Readers, Access API

Analytics Platform: MapReduce, Job Management, Autoscaling, Streaming, Messaging, Virtual Machine, Container

Statistical Registry

Discovery Service

Science Workflow

Analytic Services

Visualization Service

Data Viewer

Web Portal

Interactive Science Workbench

Applied Science

Data → Information → Knowledge

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Thomas Huang
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