National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Taming the Big Ocean Data

Thomas Huang
Project Technologist

NASA Physical Oceanography Distributed Active Archive Center

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099, United States of America

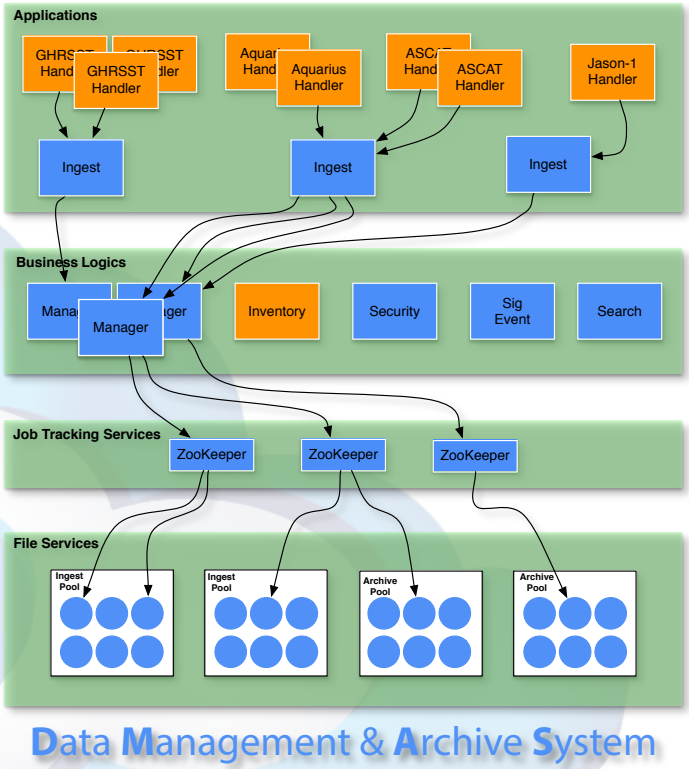
NASA's PO.DAAC

- 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**.

http://podaac.jpl.nasa.gov

PO.DAAC's Mission and Cloud Computing

- **Data Preservation**
  - Archive on the Cloud
  - Storage redundancy
  - Hardware reliability
  - Elastic storage
- **Data Accessibility**
  - Data services availability
  - Platform for
    - Spatial Searches
    - Spatial subsetting
    - Quality screening, etc.
  - Zone replication (with additional costs)
- **Data Analysis**
  - Climatology
  - Data re-gridding
  - Relevancy
  - Anomaly detections



Cloud Options

- **Commercial Cloud: public Cloud** (e.g. Amazon)
  - ~$30/TB/month, additional cost for I/O (PUT, COPY, POST…)
  - Transfer Out: ~$90/TB/Month
  - What do we get?
    - Potential cost reduction
    - No hardware maintenance
    - No System Administration required
    - Reliable storage
    - Pay-by-the-drink

- **On-premise Cloud: private Cloud**
  - Closer to the physical archive
  - Fixed storage cost
  - No transfer / I/O cost
  - Need trained System Administrator
  - Provide elastic computing infrastructure and programming model

- **Bursting Cloud: hybrid Cloud**
  - Bursting computing jobs to external (public/private Clouds)
  - Ability to leverage additional computing resources
  - Gotchas
    - Depending on the computing problem, it might require data replication, hence storage cost if the external Cloud is a commercial Cloud
    - Cost for On-premise Cloud and possible external costs

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Some PO.DAAC-related Cloud Computing Activities

Technology Infusion – Cloud Computing Study

- Amazon
- NASA Nebula
- Apache Hadoop and HBase
- Built Climatology Service

NASA ITLabs Cloudbursting

- Job busting between three NASA centers

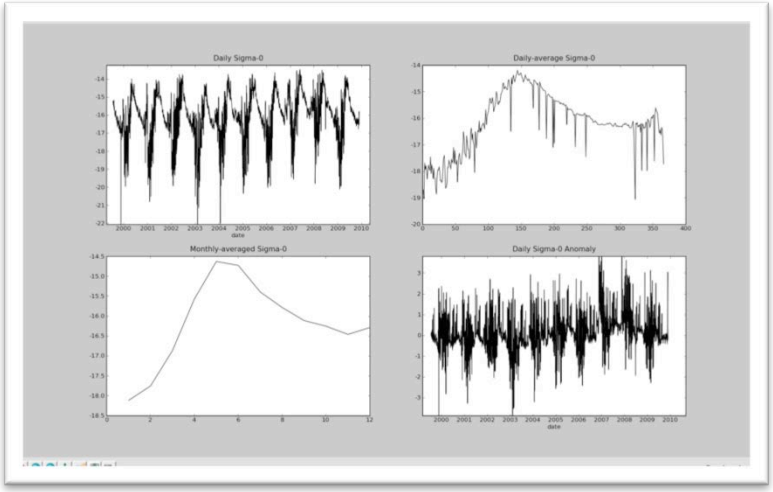
Pilot JPL CIO OpenStack (Nebula Inc.)

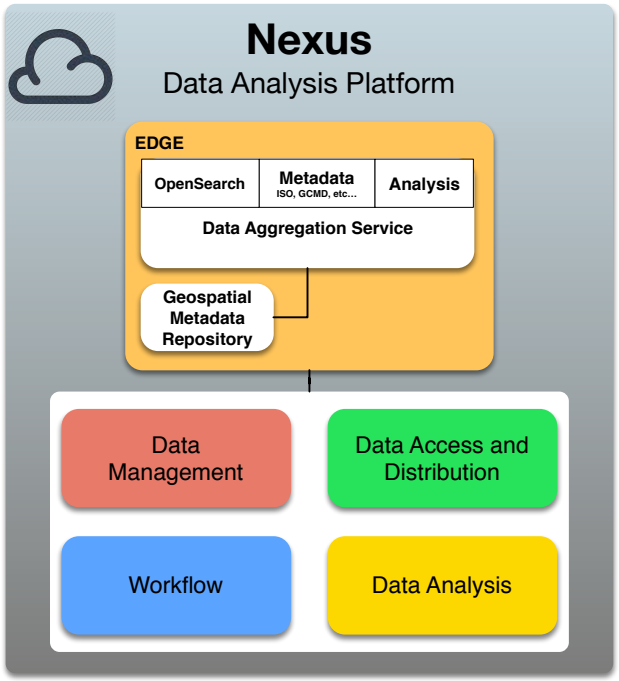
Nexus – Science Data Analysis Platform

- Spark, Hadoop, NoSQL, Solr, etc.
- Turnkey deployment
- Target projects: Sea Level Change Portal, ACCESS, AIST

Working with Big Data and Cloud Computing Communities

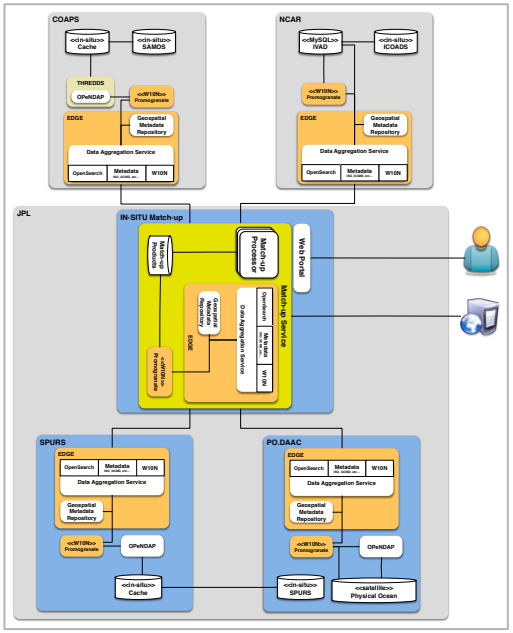
- Chair, ESIP Federation Cloud Computing Cluster
- Chair, NASA ESDSWG Data-Intensive Architecture
- Active Contributors, ESDSWG Cloud Computing
- Active Contributors, NIST Big Data Working Group
- JPL Selection Committee for Cloud RFP

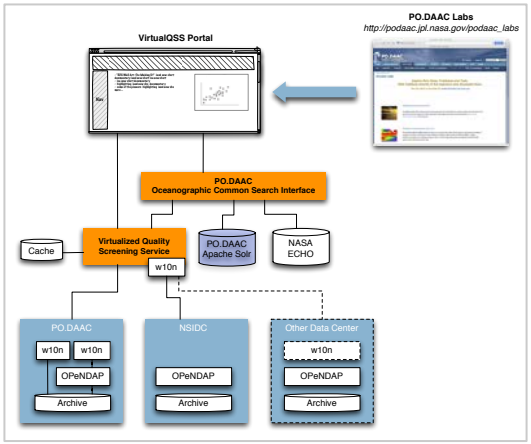


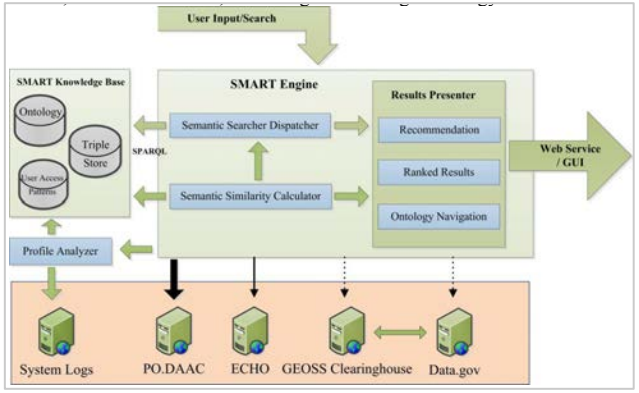


Funded Cloud Computing Efforts

- **2013 Sea Level Rise**
  - PI: C. Boening/JPL, *A NASA Web portal for Sea Level Change*
- **2013 ROSES ACCESS**
  - PI: E. Armstrong/JPL, *Enhanced Quality Screening for Earth Science Data*
  - PI: C. Lynnes/GSFC, *Federated Giovanni*
- **2014 ROSES AIST**
  - PI: T. Huang/JPL, *OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analytics Portal*
  - PI: S. Smith/FSU, *A Service to Match Satellite and In-situ Marine Observations to Support Platform Intercomparisons, Cross-calibration, Validation, and Quality Control*
  - PI: C. Yang/GMU, *Mining and Utilizing Earth Science Dataset Metadata, Usage Metrics, and User Feedback to Improve Dataset Relevancy*

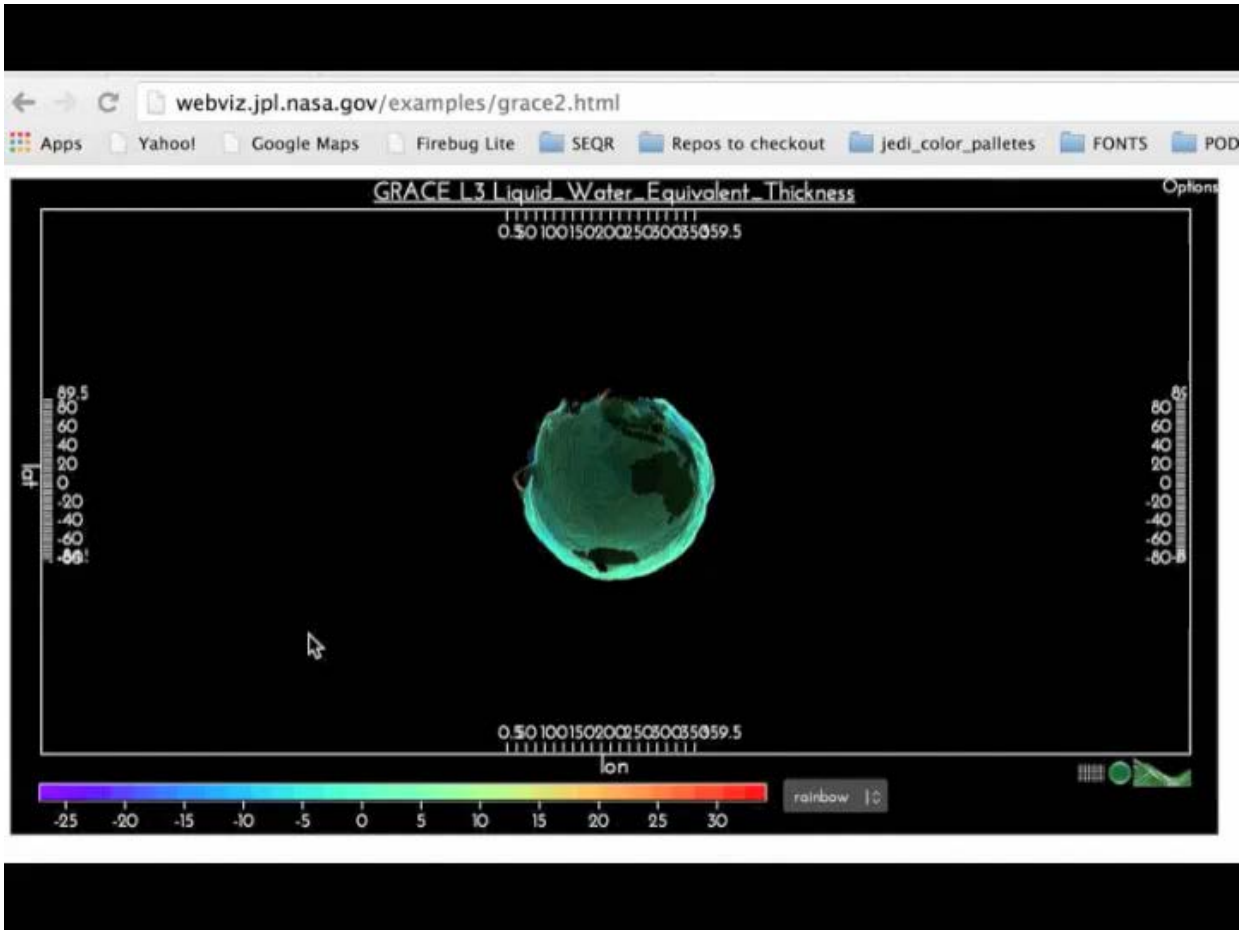






Visualization as a Service (VaaS)

- We are also working on visualization widgets that can be embedded into any HTML pages. **Visualization as a Service (VaaS)**
- Visualizing L3 GRACE data from a NetCDF file located somewhere within PO.DAAC.

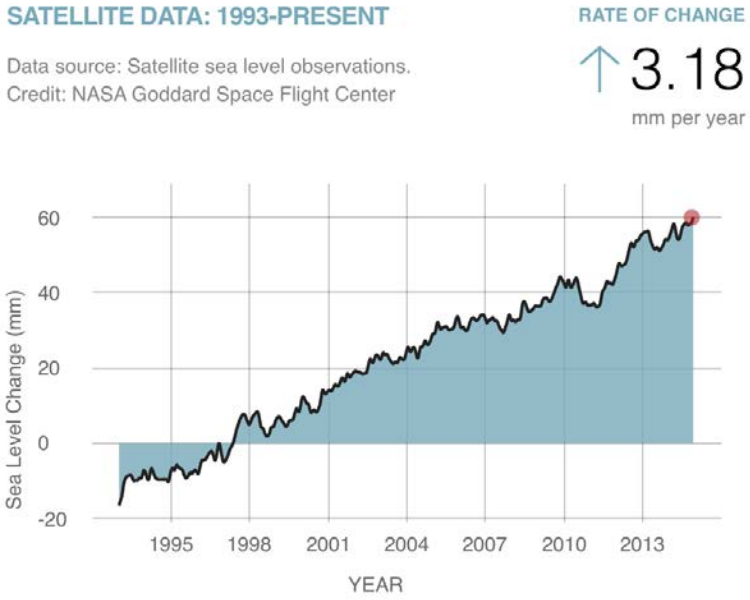


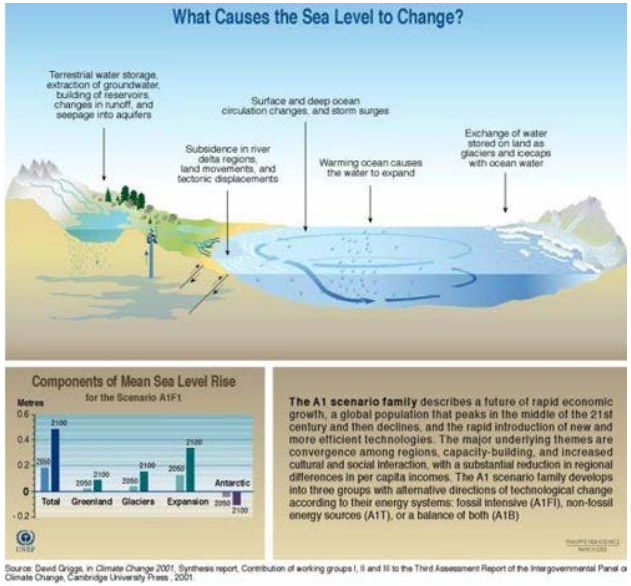
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA Sea Level Change Portal

- The Sea Level Change Portal will serve as central hub for enabling collaboration between the NASA Sea Level Science Team
- The ultimate goal is to provide sciences and general public with “one-stop” source for current sea level change information and data, including interactive tools for accessing and viewing regional data, a virtual dashboard of sea level indicators, and ongoing updates through a suite of editorial products that include content articles, graphics, video, and animations.





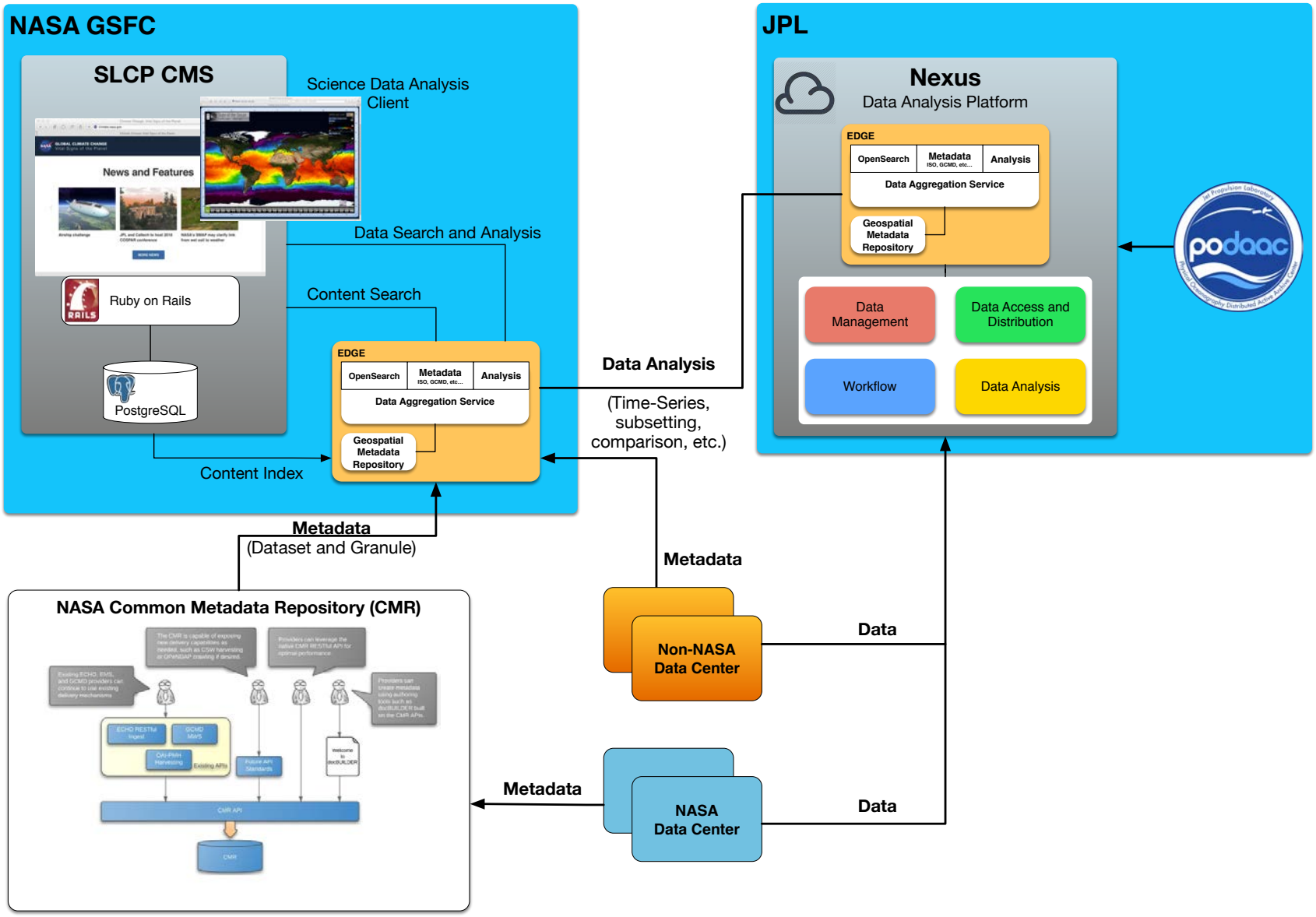
Source: David Griggs, in *Climate Change 2001*, Synthesis report, Contribution of working groups I, II and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press , 2001.

DATA ACCESS AND ANALYSIS

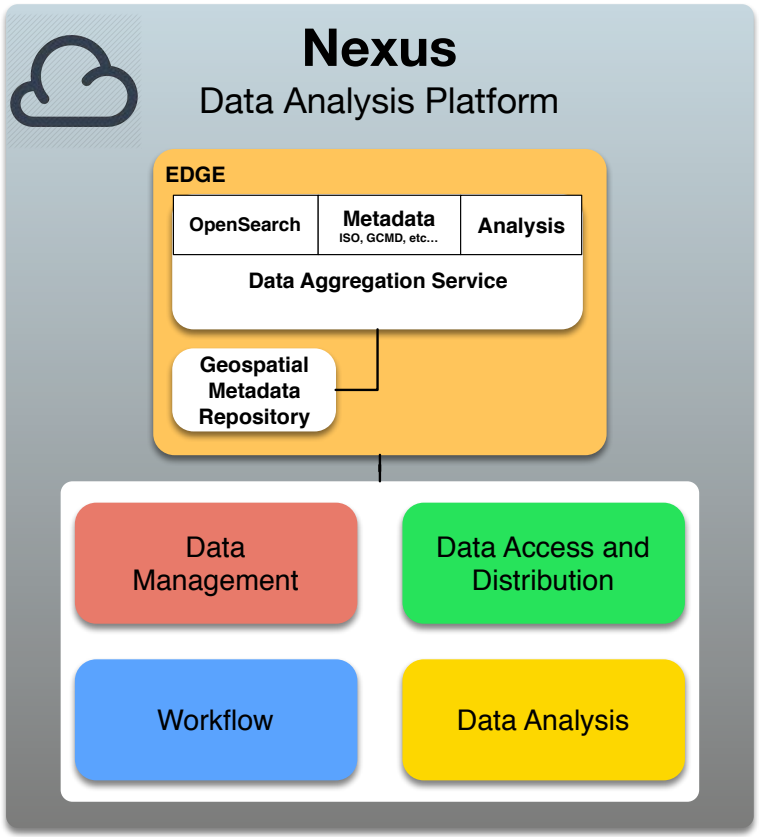
- **Goal**: Enable easy access to multi-disciplinary data sets and facilitate quick online analyses. Features include
  - Global data selection for geographical maps
    - Spatial
    - Temporal
  - Data analysis
    - Regional averages (time series)
    - Basic statistical analysis (RMS, correlation, PDF, spectral analysis, …)
    - Model/data comparison
  - Data subscription
    - Define search and receive “data alert” once new data matching this search arrives.

NASA Sea Level Change Portal

ARCHITECTURE



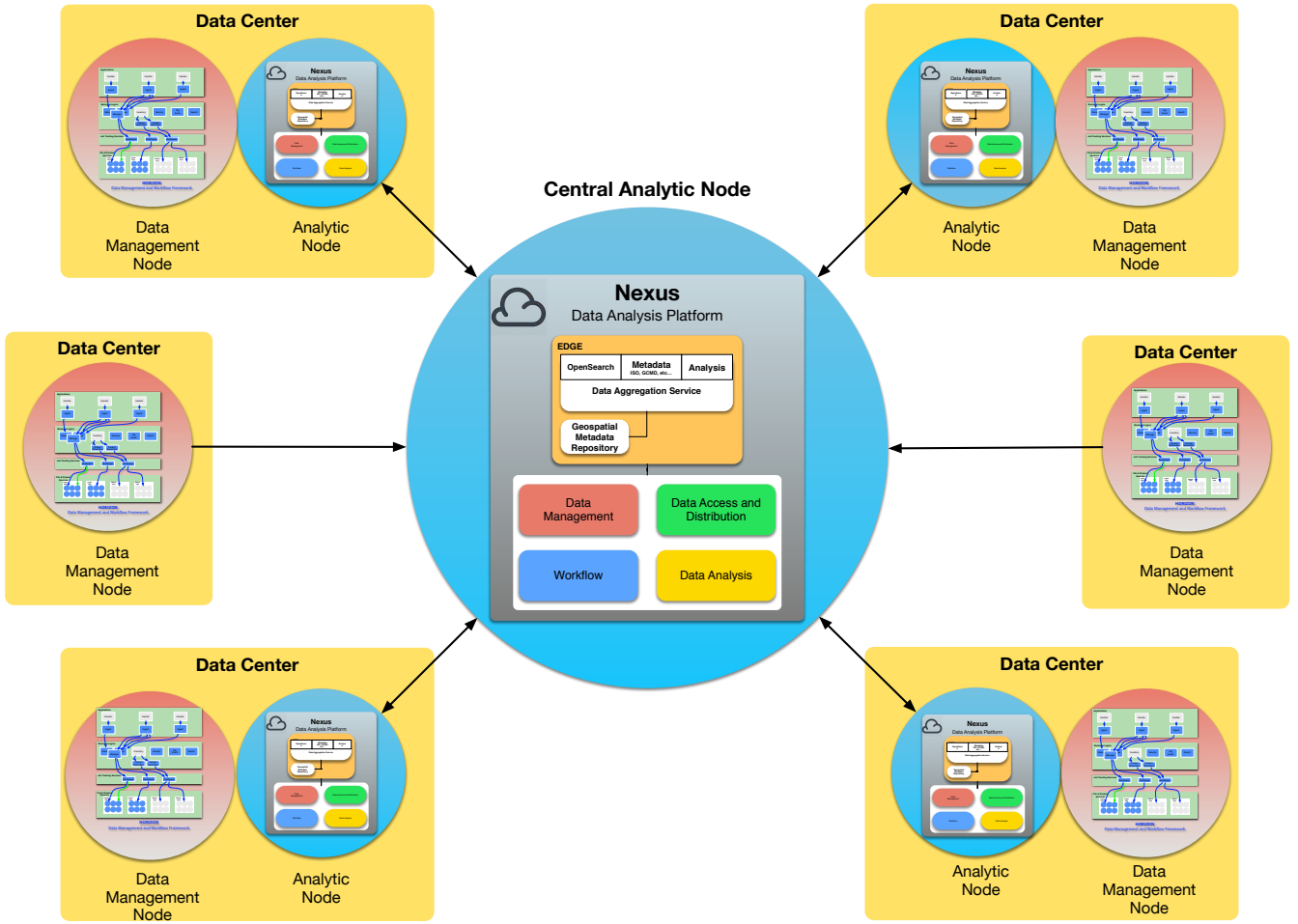
NEXUS: DATA ANALYSIS PLATFORM



- Data analysis platform on the Cloud
- Data management and transformation
- Multi-disciplinary data coordination
- On-the-fly analysis services
  - Time series
  - Correlation
  - Re-gridding
  - Data subsetting
  - Data visualization service
- RESTful access to geospatial array data

Federated, Multi-Cloud Architecture for Big Data

- Recognizing not all data centers are equal
- Providing a common, portal software solution stack improves interoperability by providing distributed data analysis solution for Big Earth Science Data.



Summary

- Start with the architecture. Can't build with just Cloud.
- Don't jump into Cloud because it is popular
- Use Cloud because it makes sense
  - Cost?
  - Reliability?
  - Platform to improve data access and analysis?
- Might need to rethink existing software solutions when moving to Cloud
  - Truly leverage the elasticity of the Cloud?

Summary

- User automation deployment – Puppet, Chef, Salt, etc.
- Bring the computing close to the data make sense – On-Premise Cloud (currently)
  - Need local experts
  - Governance
- For Commercial Cloud
  - Simplified (fixed) costing model for Amazon resources
  - Reduced storage and transfer out pricing, and zone replication
- For On-premise Cloud
  - Export controlled: public data, ITAR software
  - Suggest standardize Cloud stack
  - Federated, multi-Cloud environment

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

THANKS

Questions, and more information

thomas.huang@jpl.nasa.gov