



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Horizon: The Portable, Scalable, and Reusable Framework for Developing Automated Data Management and Product Generation Systems

Thomas Huang, Christian Alarcon, Nga T. Quach

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



Data Management in Big Data Era

“Scientific data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or recommendation for action.”

Deming, William Edwards
Journal of the American Statistical Association
On a Classification of the Problems of Statistical Inference
Volume 37, Number 218, June 1942 (p. 173)



NASA PO.DAAC

- PO.DAAC 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.
- 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>

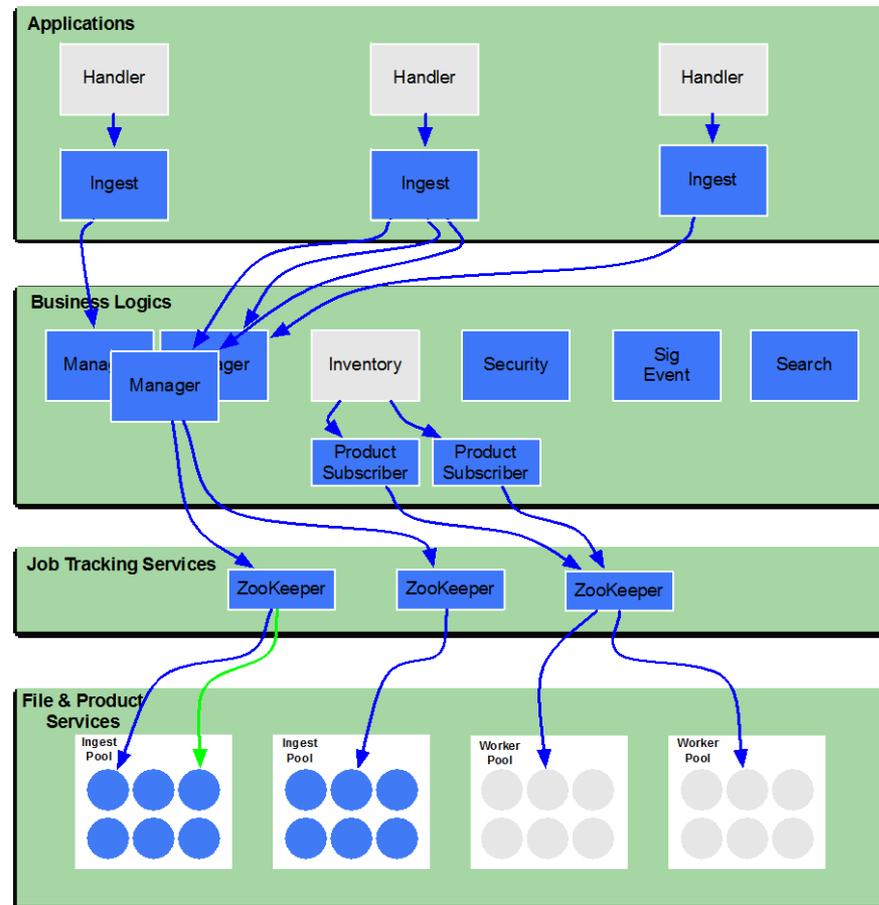




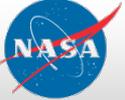
HORIZON

Data Management and Workflow Framework

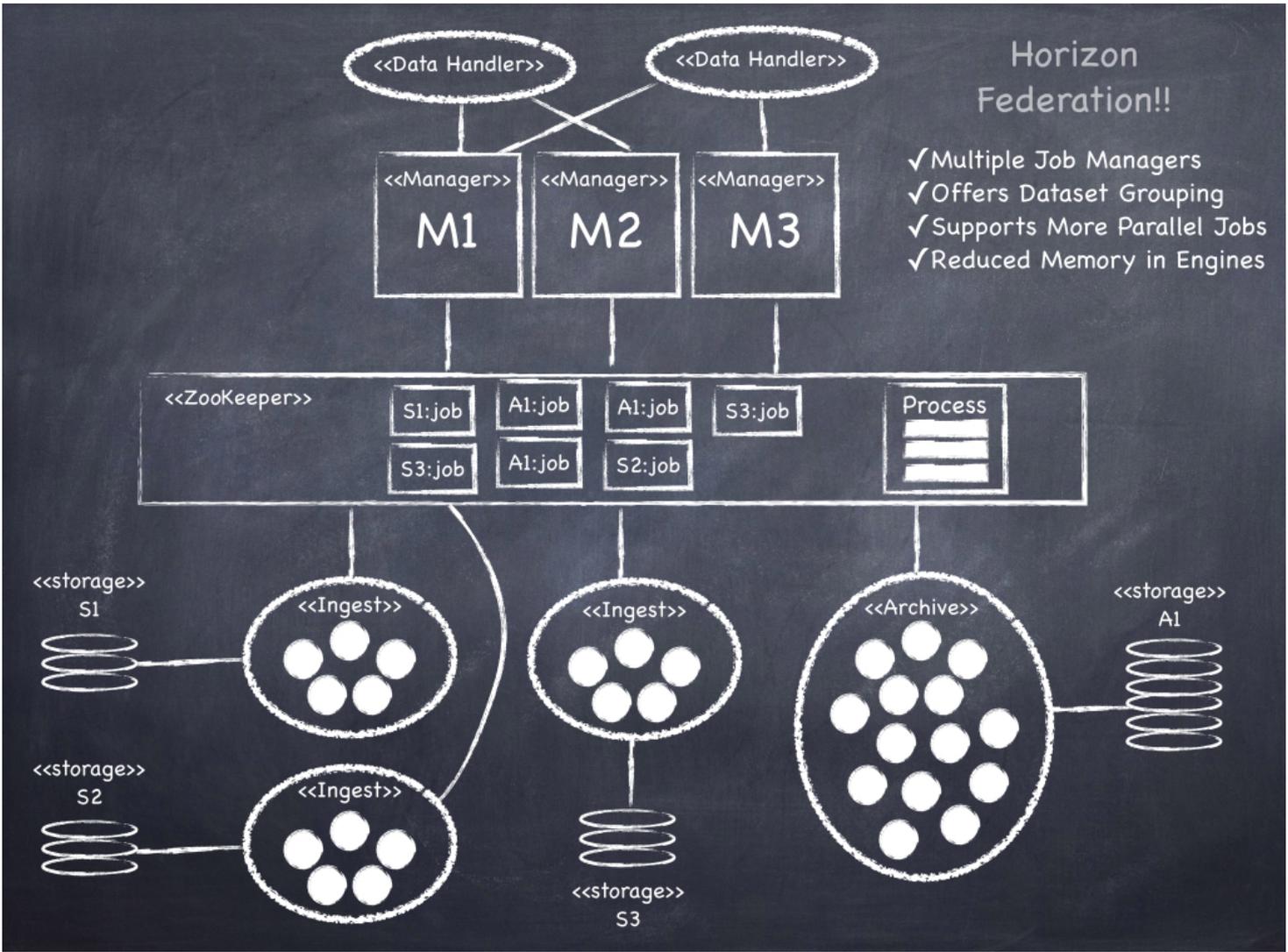
- Developed at Jet Propulsion Laboratory
- Core data infrastructure to manage and automate the data ingestion, archival, and distribution process for several highly visible projects
- A reusable data management and workflow framework
- Automation
- Horizontally scalable – scale down to a single box (a.k.a “DAAC in a Box”) or scale up with pools of ingest, archive, and product generator nodes.
- Self monitoring and tuning
- Service event tracking and notification
- Security service with pluggable authentication and authorization support
- Metrics
- Portable – pure Java core with some Python components



HORIZON
Data Management and Workflow Framework



Federated Architecture





Significant Event Service

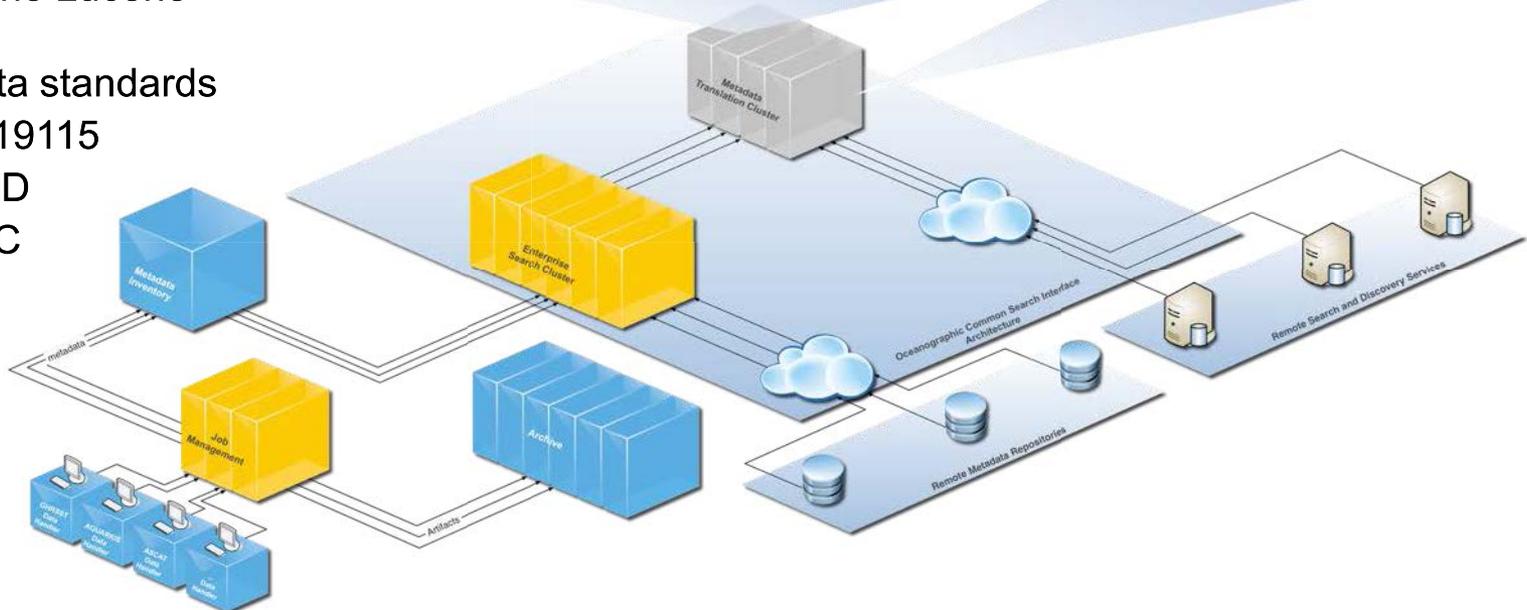
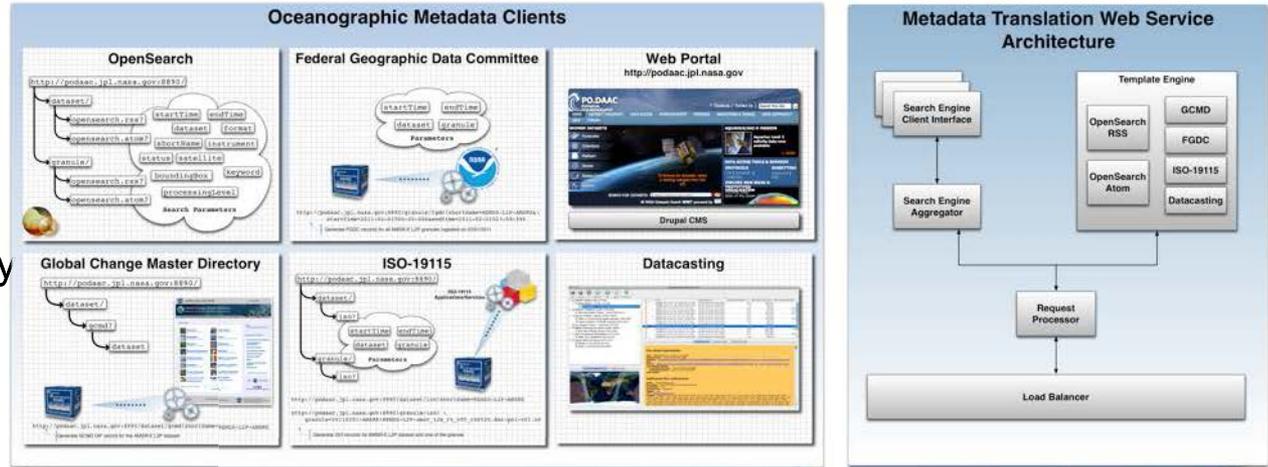
RESTful centralized Publish/Subscribe service for a heterogeneous distributed system.





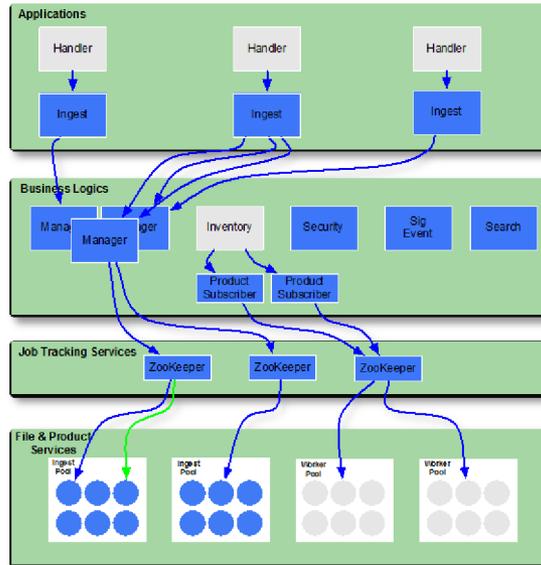
Extensible Data Gateway Environment

- Metadata and Discovery Service
- Pluggable architecture to interface with other repositories and discovery services
- Search standards
 - OpenSearch
 - Apache Lucene
- Metadata standards
 - ISO-19115
 - GCMD
 - FGDC

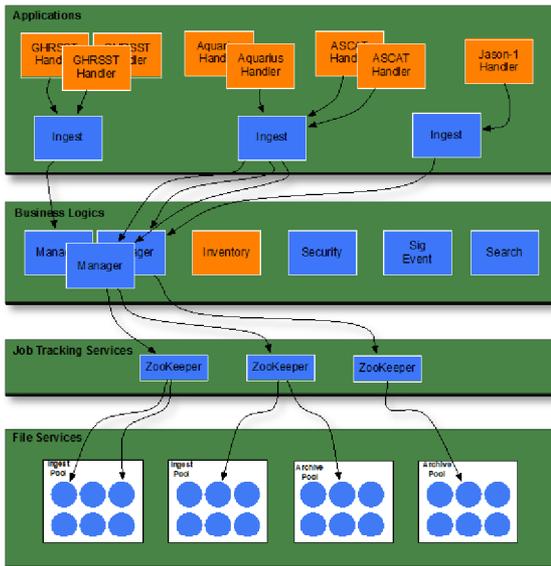
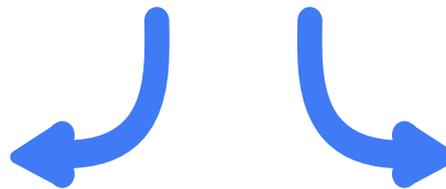




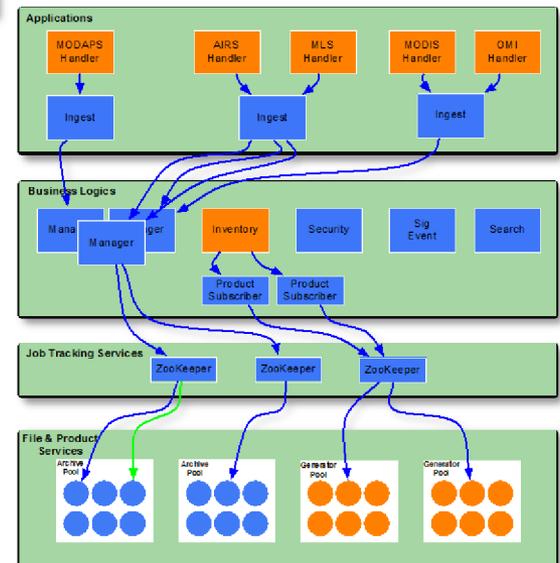
Example Applications of Horizon



HORIZON Data Management and Workflow Framework



Data Management & Archive System



The Imagery Exchange



National Aeronautics and
Space Administration

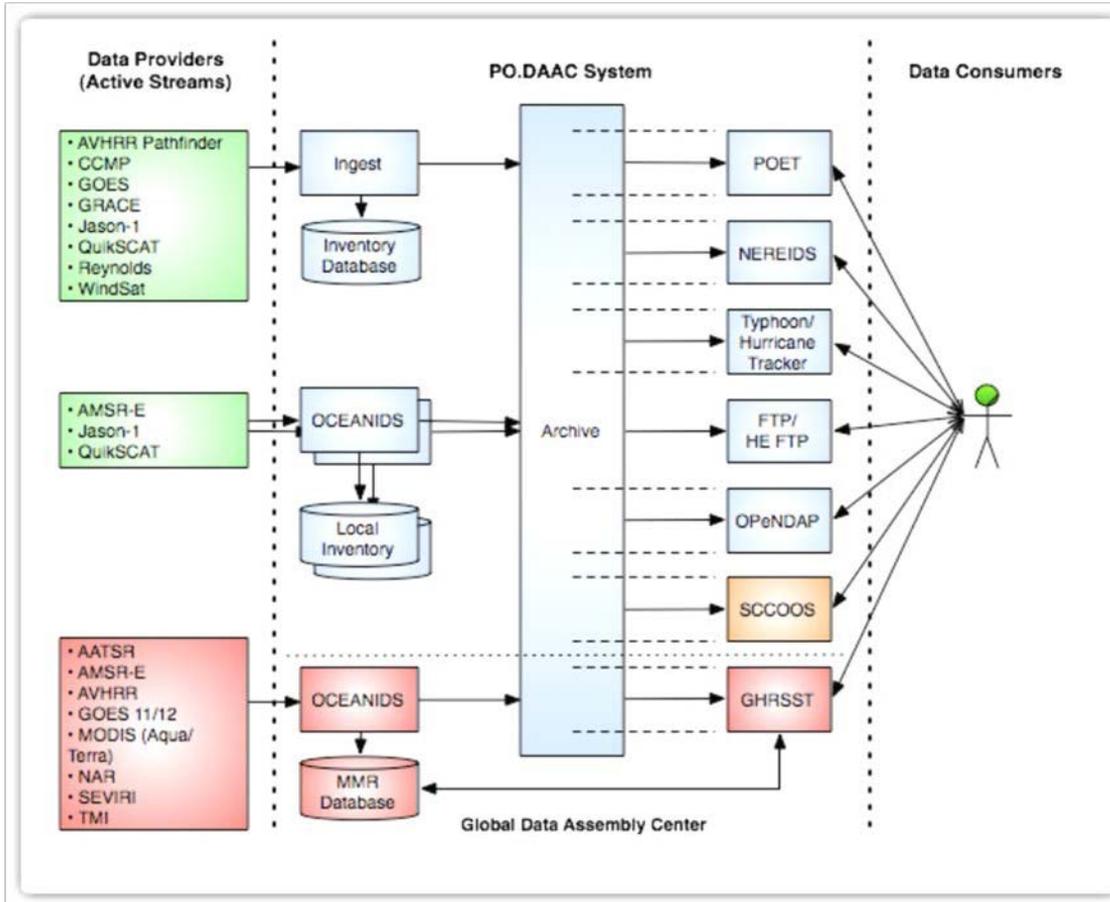
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA PO.DAAC

DATA MANAGEMENT AND ARCHIVE SYSTEM (DMAS)

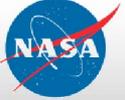


Legacy Data Systems



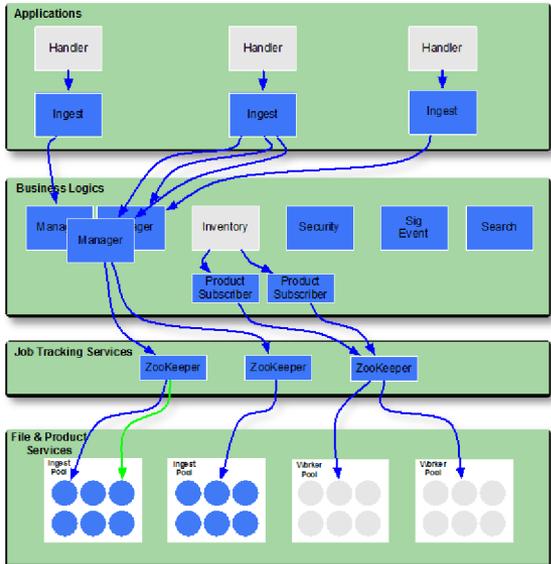
- It works, but...
- 3 different data systems
- Deployed in multiple instances
- Mostly consists of one-off scripts
 - Limited traceability
 - Limited reusability
 - Limited portability
 - Scalability?
 - Reliability?



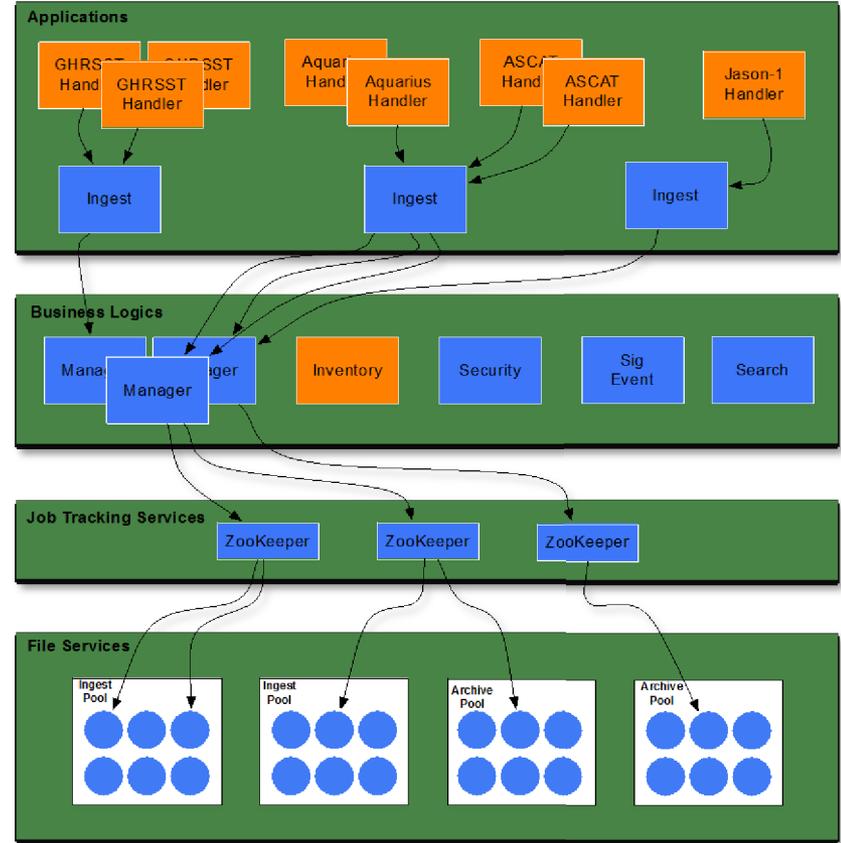
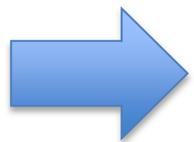


PO.DAAC's Data Management and Archive System

- The Data Management and Archive System (DMAS)
- Extension to Horizon with
 - Specialized data handlers
 - Specialized Inventory data model and service
 - Specialized data curation tools
 - Specialized workflow policy



HORIZON
Data Management and Workflow Framework

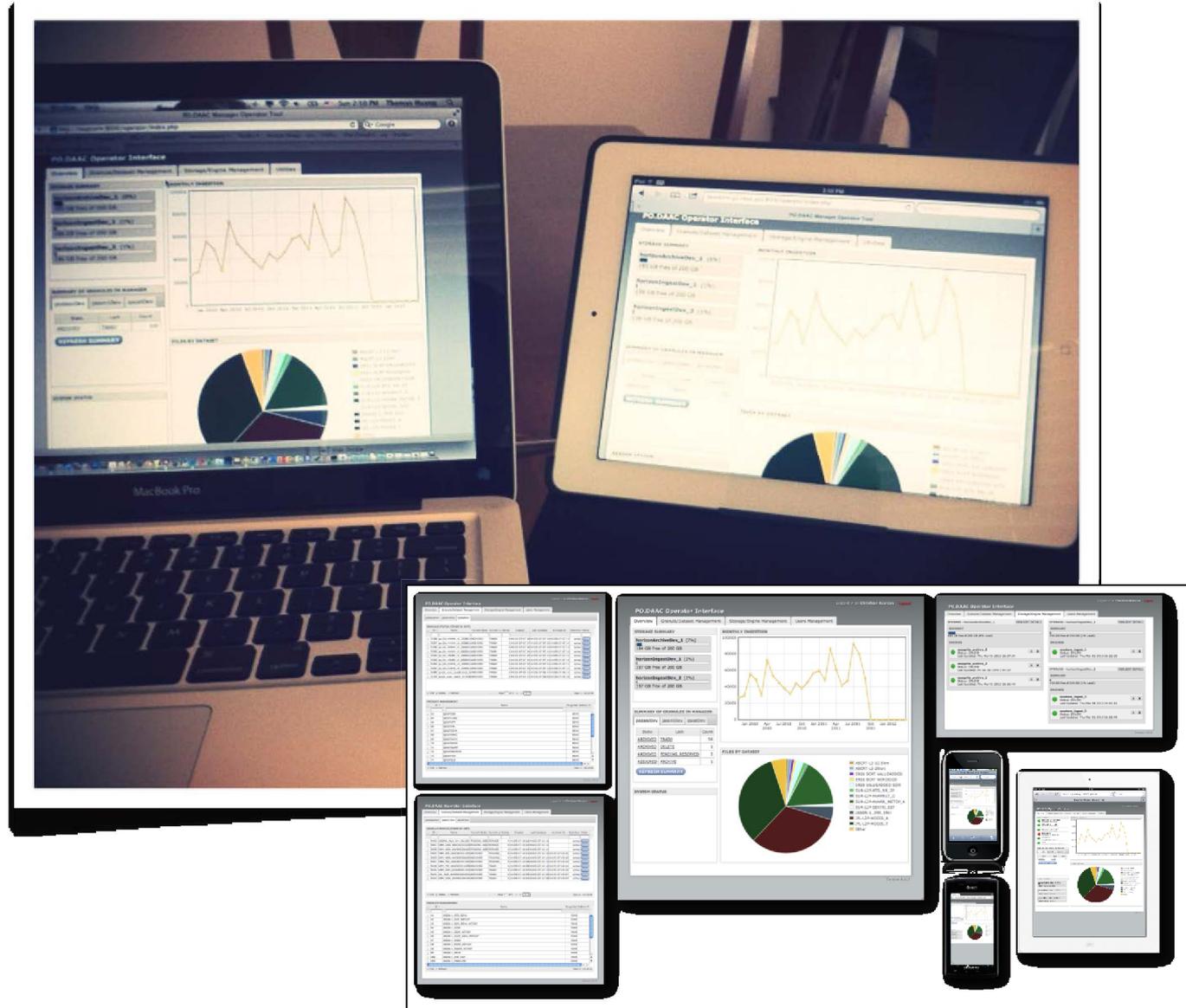


Data Management & Archive System



Service Monitoring and Administration Tool

- Services availability
- Ingest and Archive workflow
- Storage Management
- Metrics

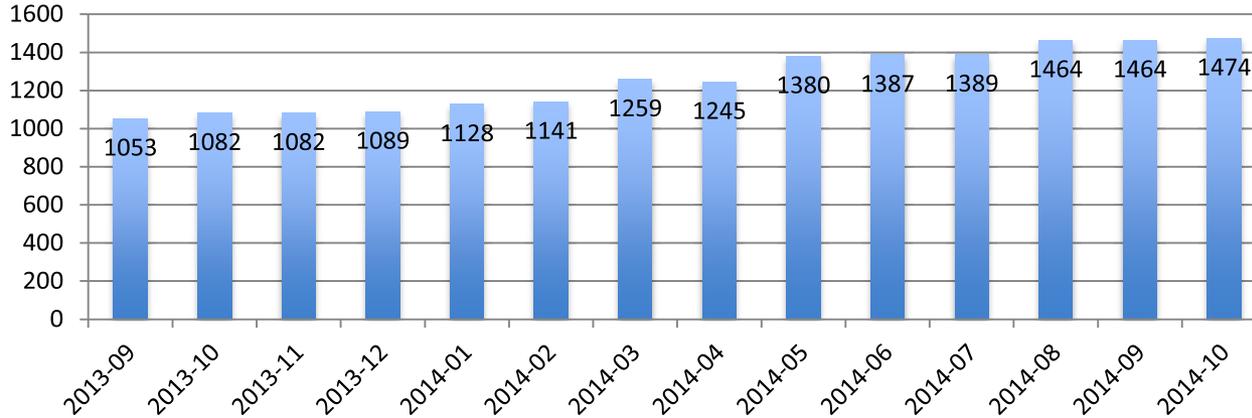




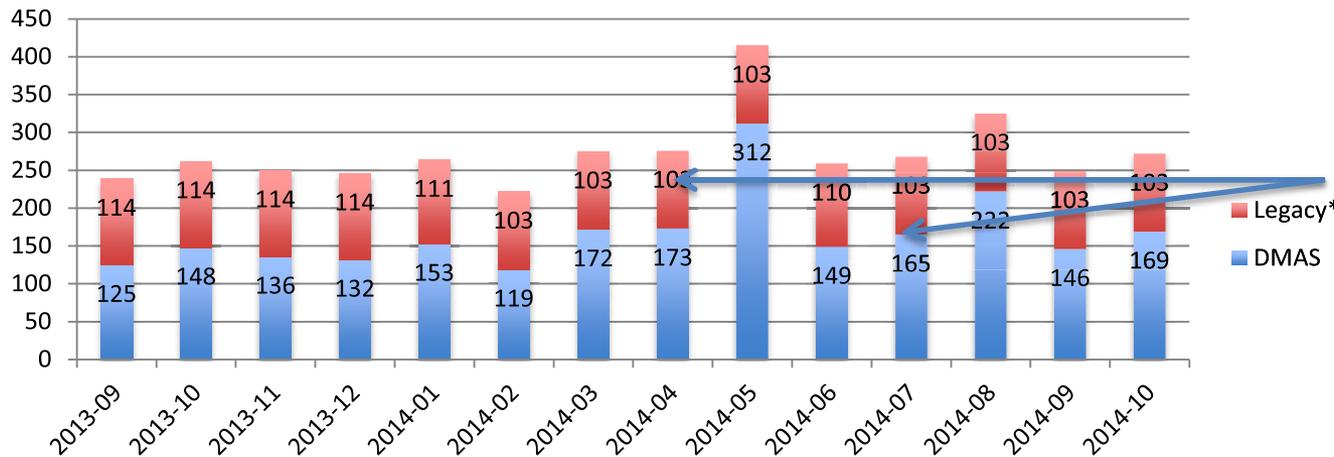
PO.DAAC Dataset Metrics

DMAS Total Datasets

(exclude Dormant)



Actively Ingesting Datasets



*Legacy is best estimate;

Spike in May & Aug (DMAS) is due to Aquarius V2.10.1, V3 and V3.1 addition.



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Global Imagery Browse Services

THE IMAGERY EXCHANGE (TIE)



Vision of the EOSDIS Global Imagery Browse Services (GIBS)

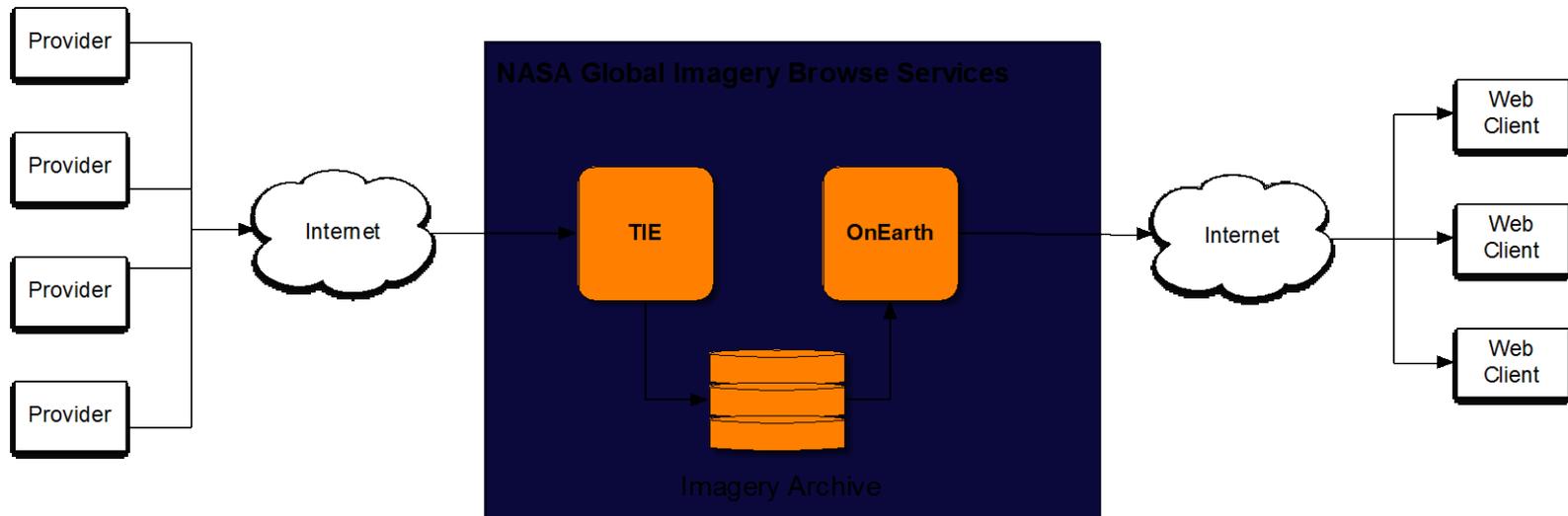
“to transform how end users interact with EOSDIS’ inter-disciplinary data through enhanced visual discovery.”

- Inter-NASA collaboration between JPL and GSFC
- Fast and responsive visualization for modern map clients
- Set of standard services to deliver global, full-resolution satellite imagery (Tiled WMS, WMTS)
- Enable interactive exploration of NASA’s Earth imagery for a broad range of users



GIBS Architecture @10,000-Foot View

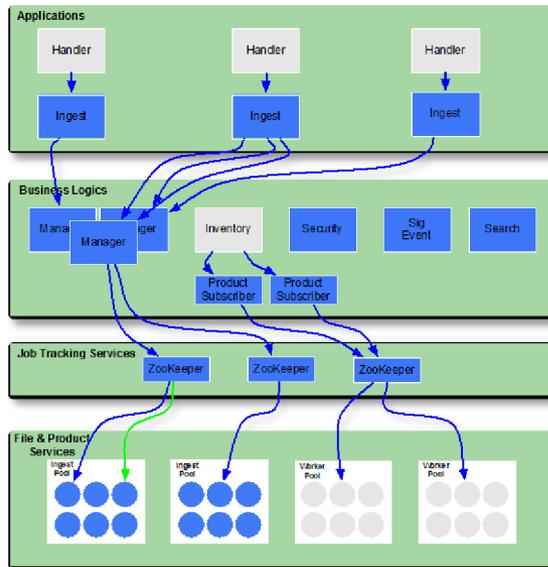
- Two subsystems
 - **The Imagery Exchange (TIE)** – Data Management and Product Generation Workflow Service an extension of the **Data Management and Archive System (DMAS)**
 - **OnEarth** – Image transformation and extremely efficient fast image serving capabilities





The Imagery Exchange (TIE)

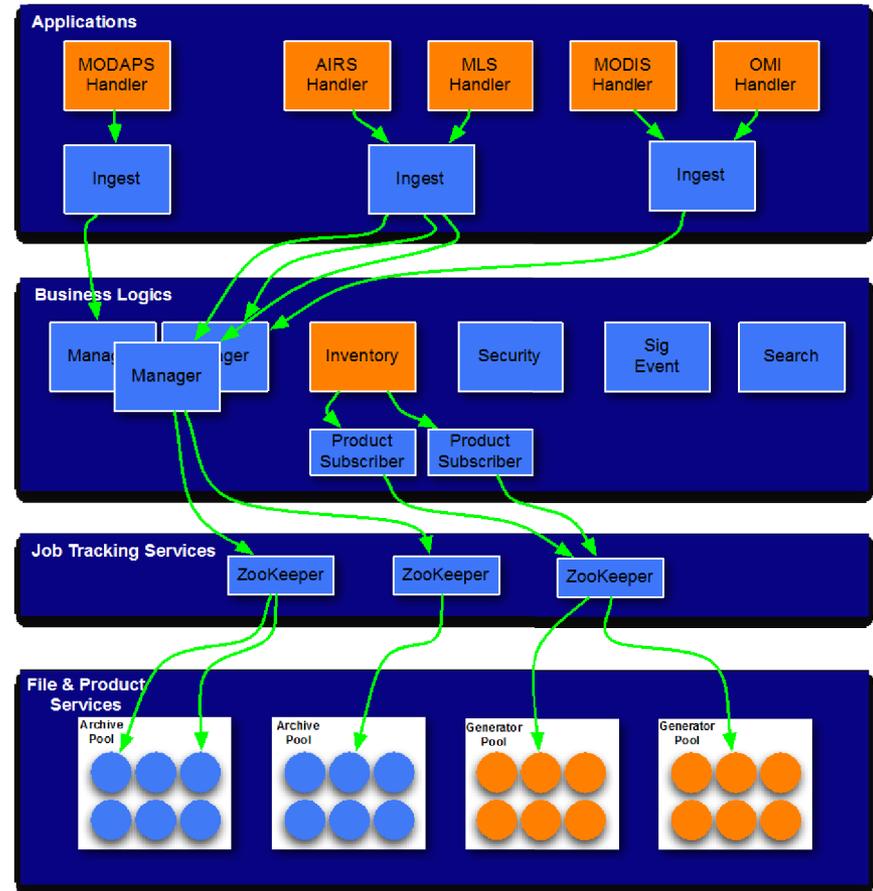
- GIBS needs an imagery archival with product generation workflow capabilities that is scalable to handle the increasing amount of science data.
- Responsibilities
 - Ingest/Archive source imagery
 - Catalog metadata and provenance information
 - Automate and manage the generation of Meta-Raster Format (MRF) products to be served by the GIBS OnEarth subsystem.



HORIZON

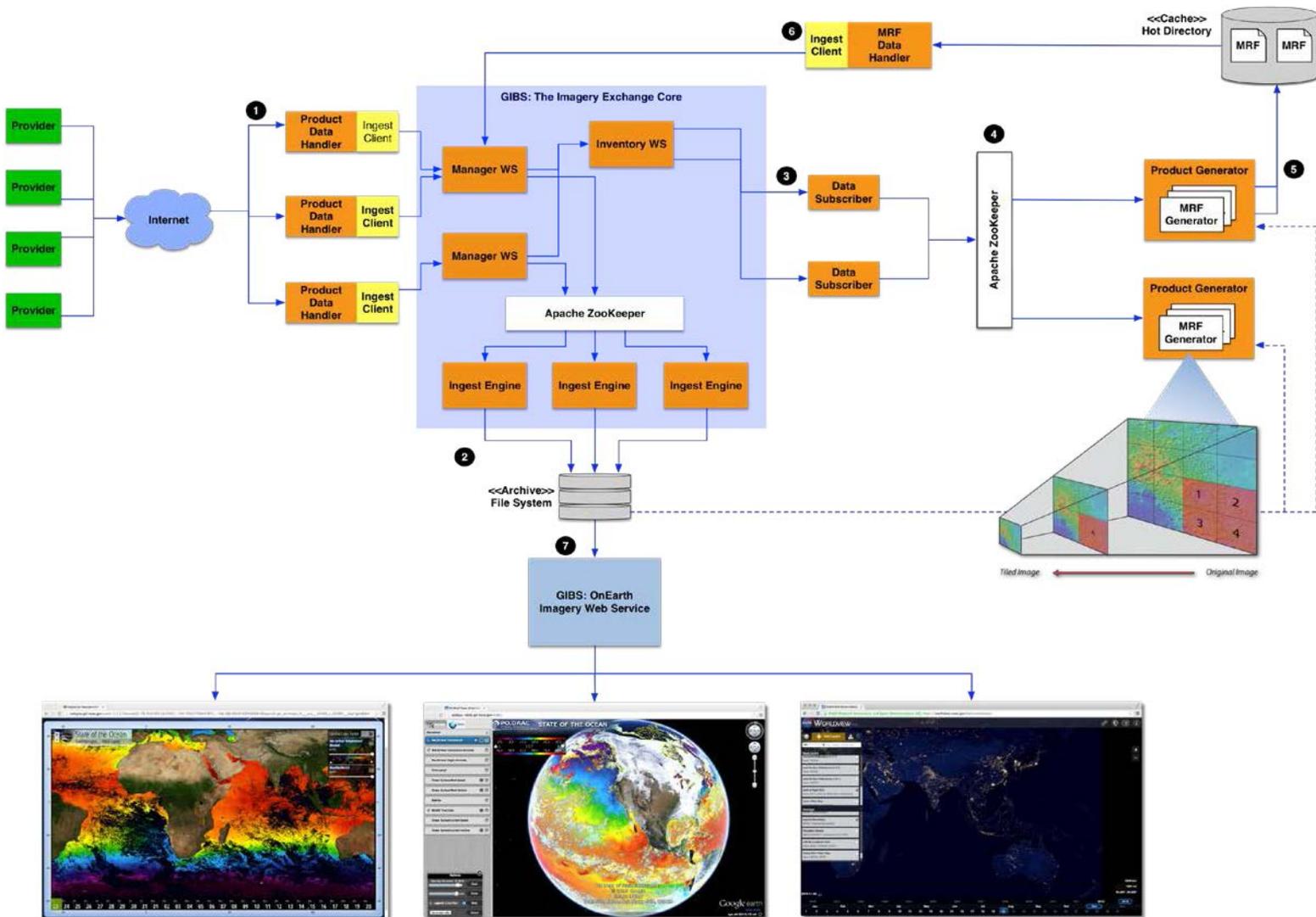
Data Management and Workflow Framework

The Imagery Exchange





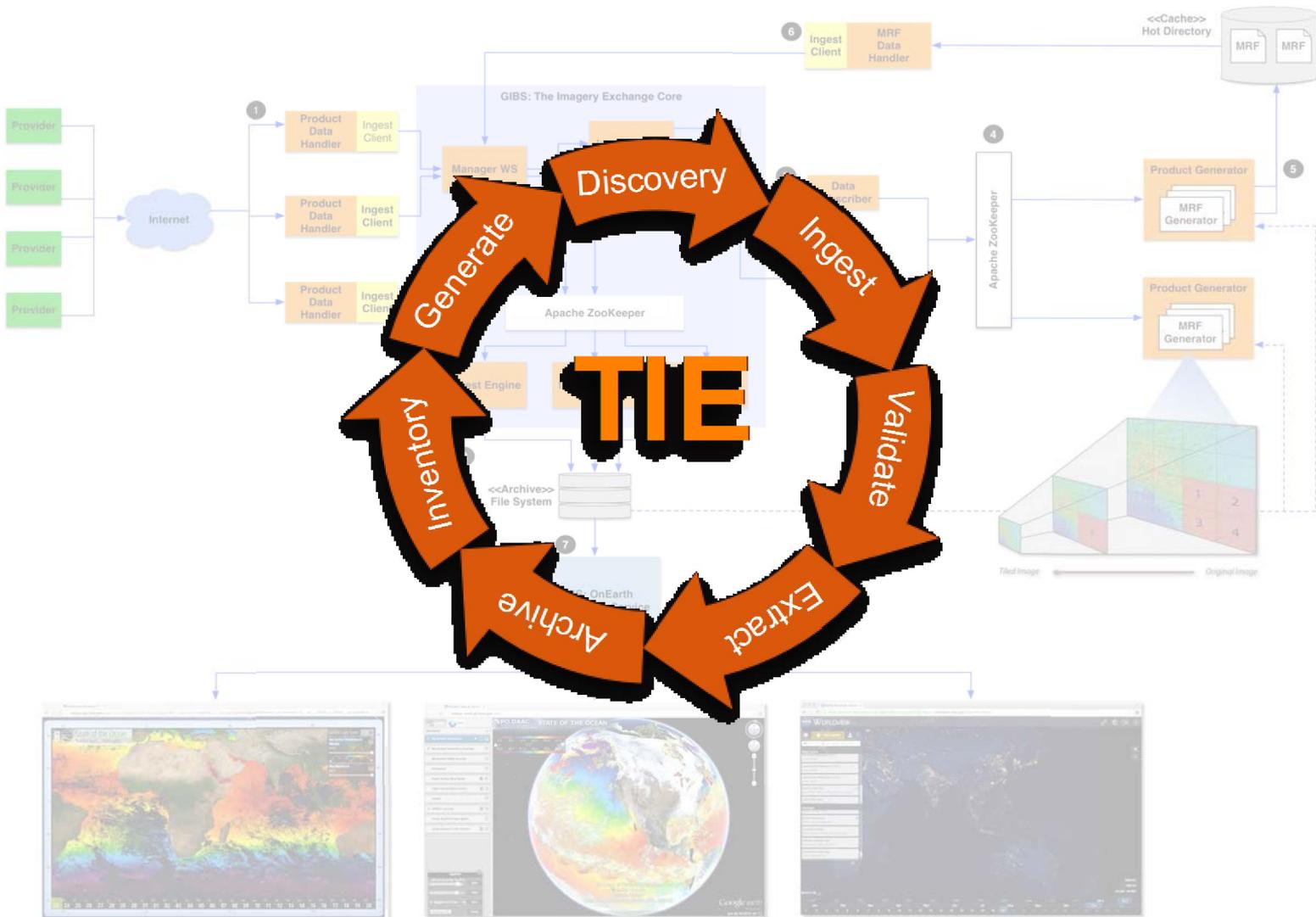
Automation



A system that feeds itself



Automation



A system that feeds itself



**National Aeronautics and
Space Administration**

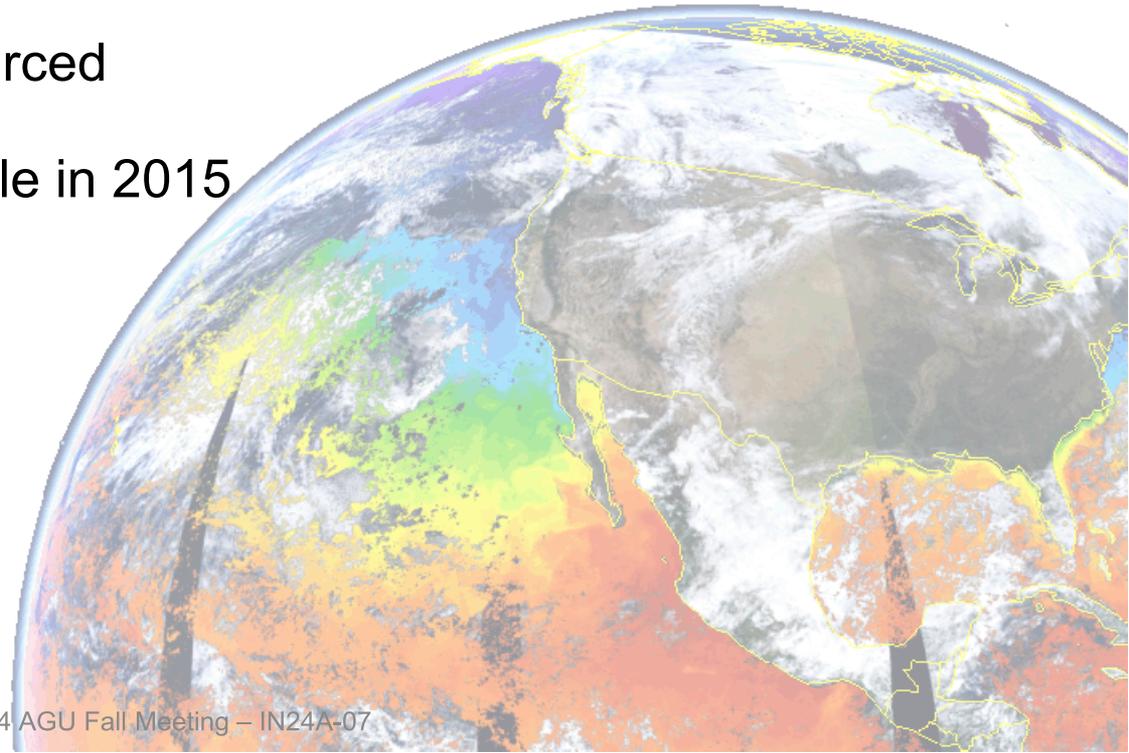
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

COMING TO HORIZON



Horizon, DMAS, and TIE Open Sourcing

- Growing interest and demand for Horizon – our motivation for making these solutions open sourced
- For over a decade we have been benefiting from the open source communities
- We are committed to contribute back to the communities
- In process to be open sourced
- We plan to have it available in 2015





Horizon for Big Earth Data

- Typical analysis drilldown
 - Dataset discovery – keyword, parameter, instrument, satellite, processing level, temporal, spatial, etc.
 - Dataset metadata
 - Granule Search – temporal, spatial
 - Granule Metadata
 - Granule Download
 - Apply Local Processing Tool: time series, subset, re-gridding, etc.
- There has been various tools and services developed over the years close the archives
 - Limited in performance and scalability – typically constrained by I/O and data movement
- The Horizon team is tackling Big Earth Data by developing a new Data Architecture that leverages
 - Private and Multi-Cloud Computing
 - Cloud DB
 - High Performance Data Access
 - Scalable Geospatial Search
 - New Data Processing Framework
- Integrates with the Horizon data management workflow
- Multi-project collaboration effort – goals: fast subsetting and climatology
- Initial release: Spring 2015



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

“Architecture starts when you carefully put two bricks together. There it begins.”

Ludwig Mies van der Rohe (1886 – 1969)



**National Aeronautics and
Space Administration**

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

“Premature optimization is the root of all evil.”

Donald Knuth, “The Art of Computer Programming”



**National Aeronautics and
Space Administration**

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

THANKS

Questions, and more information

thomas.huang@jpl.nasa.gov