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
“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)
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.
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
Federated Architecture

- Multiple Job Managers
- Offers Dataset Grouping
- Supports More Parallel Jobs
- Reduced Memory in Engines
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

2014 AGU Fall Meeting – IN24A-07
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
• Services availability
• Ingest and Archive workflow
• Storage Management
• Metrics
DMAS Total Datasets
(exclude Dormant)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1053</td>
<td>1082</td>
<td>1082</td>
<td>1089</td>
<td>1128</td>
<td>1141</td>
<td>1259</td>
<td>1245</td>
<td>1380</td>
<td>1387</td>
<td>1389</td>
<td>1464</td>
<td>1464</td>
<td>1474</td>
</tr>
</tbody>
</table>

Actively Ingesting Datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Legacy</td>
<td>114</td>
<td>114</td>
<td>114</td>
<td>111</td>
<td>103</td>
<td>103</td>
<td>312</td>
<td></td>
<td>110</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>DMAS</td>
<td>125</td>
<td>148</td>
<td>136</td>
<td>132</td>
<td>153</td>
<td>119</td>
<td>172</td>
<td>173</td>
<td>149</td>
<td>165</td>
<td>222</td>
<td>146</td>
<td>169</td>
<td></td>
</tr>
</tbody>
</table>

*Legacy is best estimate;
Spike in May & Aug (DMAS) is due to Aquarius V2.10.1, V3 and V3.1 addition.
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
Two subsystems

- **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.
A system that feeds itself
COMING TO HORIZON
• 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
“Architecture starts when you carefully put two bricks together. There it begins.”

Ludwig Mies van der Rohe (1886 – 1969)
“Premature optimization is the root of all evil.”

Donald Knuth, “The Art of Computer Programming”
THANKS

Questions, and more information

thomas.huang@jpl.nasa.gov