OCEAN OBSERVATORIES INITIATIVE

Panel: Guiding Principles and Best Practices for large-scale Data Management and Dissemination: Challenges and Opportunities

Jeffrey Glatstein
Senior Manager of Cyberinfrastructure

OCEANOBSERVATORIES.ORG

NSF

OOI Vision

- Real-time data from more than 800 instruments to enable research and education in Earth & Ocean sciences
- Marine arrays at three scales served by a common cyberinfrastructure
- Data freely available online
- 25-year lifetime
- Operated and maintained by WHOI, UW and OSU

- Sponsored by National Science Foundation:
  - Section Heads: Bauke Houtman & Lisa Clough
  - Division of Acquisition & Cooperative Support: Anna-Lee Misiano
  - Large Facilities Office: Ryszard Kaczmarek

OCEAN OBSERVATORIES INITIATIVE

NSF

OOI Data By The Numbers

- Data Collected
  - Cassandra 28 Nodes totaling 25TB
  - Postgres 350GB
  - 119 billion rows of numerical data to date with new data ingress every second
  - Raw data 900TB to date with expected growth rate of doubling every 3 years
  - HD video, digital still pictures, bio-acoustic sonar, Hydrophone acoustic sampling
- Data Delivered
  - 25 million data requests per month deliver terabytes of data
  - These include 5.6 million data requests with 1000 rows or greater of data returned
  - External real-time systems interrogate OOI API every second to every 30 seconds

Background Information

My Background

- Joined OOI in 2018 as the Program Management Office (PMO) Data Delivery Manager
- Started as a Developer on massively parallel computing environments
- Datawarehouse Architect
- Have held many management positions across the technology landscape
- Not a scientist

OOI Cyberinfrastructure

- Current PMO is the second OOI PMO (O&M objective) and not the ‘authors’ of the data system
- Created position dedicated to managing the cyberinfrastructure – First best practice applied here

My definition of “Best Practices”

- Best practices are those process and procedures that work best for your organization. With some exceptions (e.g. backing up data), not all best practices apply to all organizations.
- Taking approach of using ‘issues’ and their ‘solutions’ to communicate ‘best practices’

Challenges

- Particular to OOI
  - Poor performance on both data ingestion and delivery
  - Unfavorable perception of data quality
  - Perception that the system was unsalvageable
  - Technical debt – older technologies, aging versions and/or questionable decisions
- Universal
  - Data discoverability - particularly with large diverse data sets
  - Equipment costs associated with processing and storage of data
  - Disaster recovery and long-term archiving
  - Constant growth requires persistent adaptation for storage, maintenance and delivery
  - Budget perspective

Solutions for OOI Specific Challenges

- Poor performance on both data ingestion and delivery
  - Improve transparency of ingestions by providing mechanisms for communicating status
  - Implement a query governor that prevents one large request from tying up the system
  - Removed worse case scenario data request as the default request
  - Black box syndrome - many of the issues were perception due to lack of system information, codes changes to resolve physical issues were minor
- Unfavorable perception of data quality
  - Transparency – recognize gaps, document current quality procedures and present plan for adoption of QARTOD standards
- Perception that the system was unsalvageable
  - Transparency – discuss perceived system gaps and root causes
  - Performed an internal Self Evaluation followed by an Analysis of Alternative solutions
- Technical debt – older technologies, aging versions and questionable decisions
  - Analyze and document issues in order of impact and complexity
  - Huge opportunity here for modernization and correction of any identified gaps – for OOI that meant a re-assessment of Cassandra

Solutions for Universal Challenges

- Data discoverability - particularly with large diverse data sets
  - Analysis of Alternatives identified a user interface that presented and performed better than the current interface
  - Moved to a user driven model for features and functionality
  - Moved to a pre-calculated data set
- Equipment costs associated with processing and storage of data
  - Use policy to help set resource levels for storage and processing (e.g. store 5 years of data)
  - Archive data to cheaper storage making sure users can still find it
  - Build software rules that limit a single user from over allocating compute resources
- Disaster recovery and long-term archiving
  - Tabletop recovery exercises
  - Back-up procedures (tape and cloud) on top of hardware redundancy
  - Database replication to offsite storage – future solution
- Constant growth requires persistent adaptation for storage, maintenance and delivery (e.g. Cloud and compute in place)
- Budget perspective – tie $$ to physical results (e.g. an increase in sampling rate equals $600k)

Summary of Best Practices Used

- Transparency
- Good communications internally and externally
- Approach issues with open mind
- Involve your user community
- Constant evaluation of technology and data processing methods
- Use code changes to support end goals
- Data backup/archive strategy – short and long term
- Data quality and maintenance procedures
- Keep system reasonably up to date

OCEAN
OBSERVATORIES
INITIATIVE

OCEANOBSERVATORIES.ORG

NSF