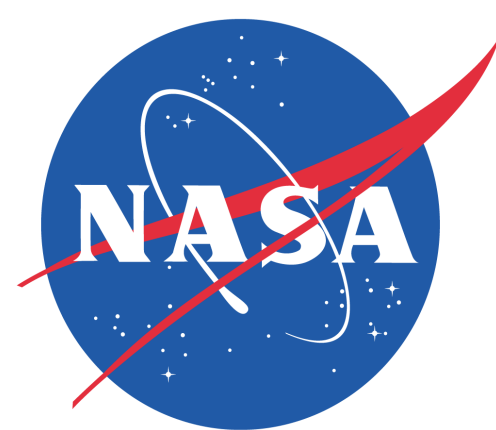




New Era, New Opportunity, Is GES DISC Ready for Big Data Challenge?



NH41B-0924

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

A. Li¹, C. Shie^{1, 2}, J. Wei¹, L. Pham¹ and D. Meyer¹

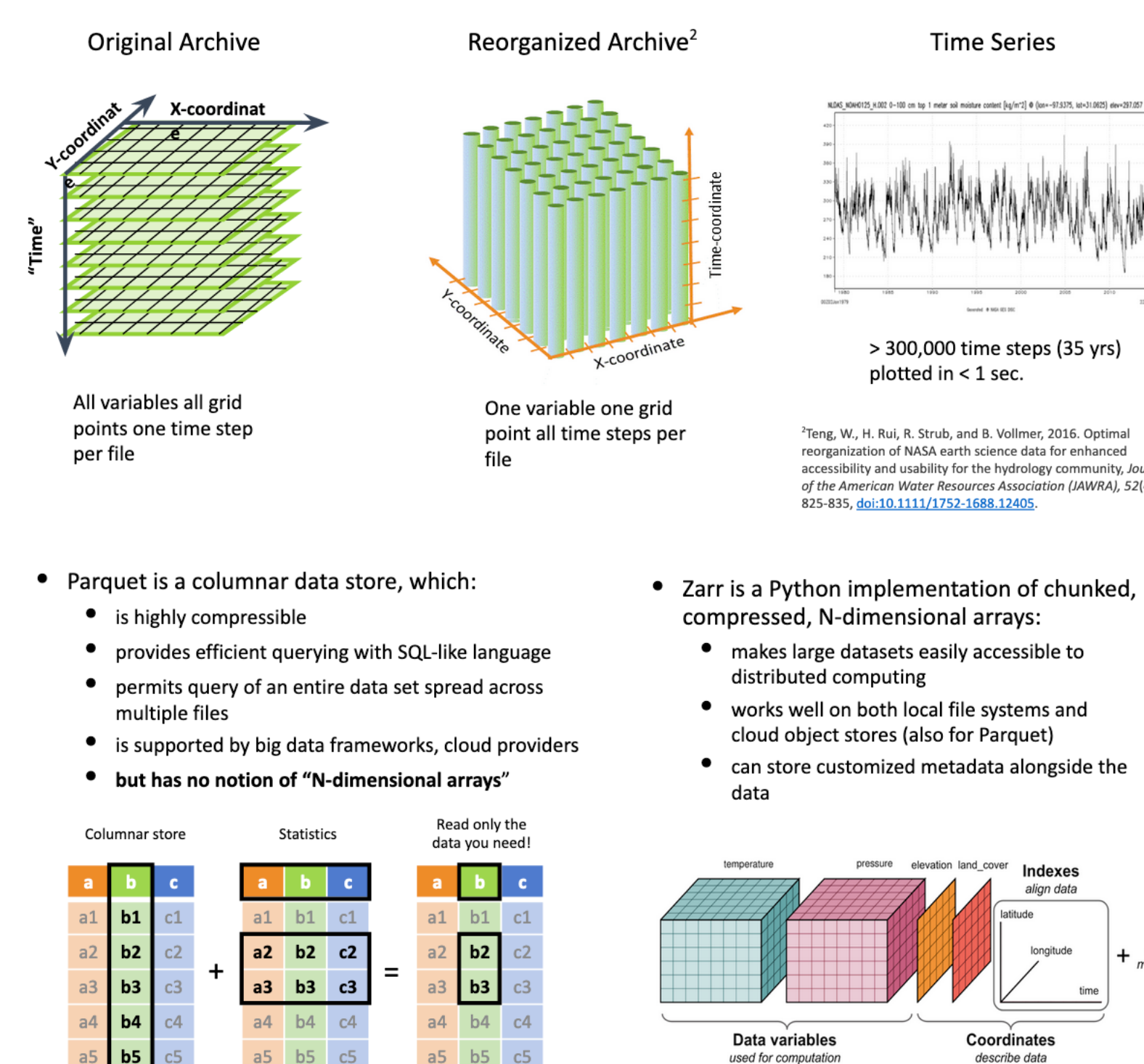
Abstract

The new era of Big Data has opened doors for many new opportunities, as well as new challenges, for both Earth science research/application and data communities. As one of the twelve NASA data centers - Goddard Earth Sciences Data and Information Services Center (GES DISC), one of our great challenges has been how to help research/application community efficiently (quickly and properly) accessing, visualizing and analyzing the massive and diverse data in natural hazard research, management, or even prediction. GES DISC has archived over 2000 TB data on premises and distributed over 23,000 TB of data since 2010. Our data has been widely used in every phase of natural hazard management and research, i.e. long term risk assessment and reduction, forecasting and predicting, monitoring and detection, early warning, damage assessment and response.

The big data challenge is not just about data storage, but also about data discoverability and accessibility, and even more, about data migration/mirroring in the cloud. This paper is going to demonstrate GES DISC's efforts and approaches of evolving our overall Web services and powerful Giovanni (Geospatial Interactive Online Visualization AND aNalysis Infrastructure) tool into further improving data discoverability and accessibility. Prototype works will also be presented.

Analysis Ready/Optimized Data Storage

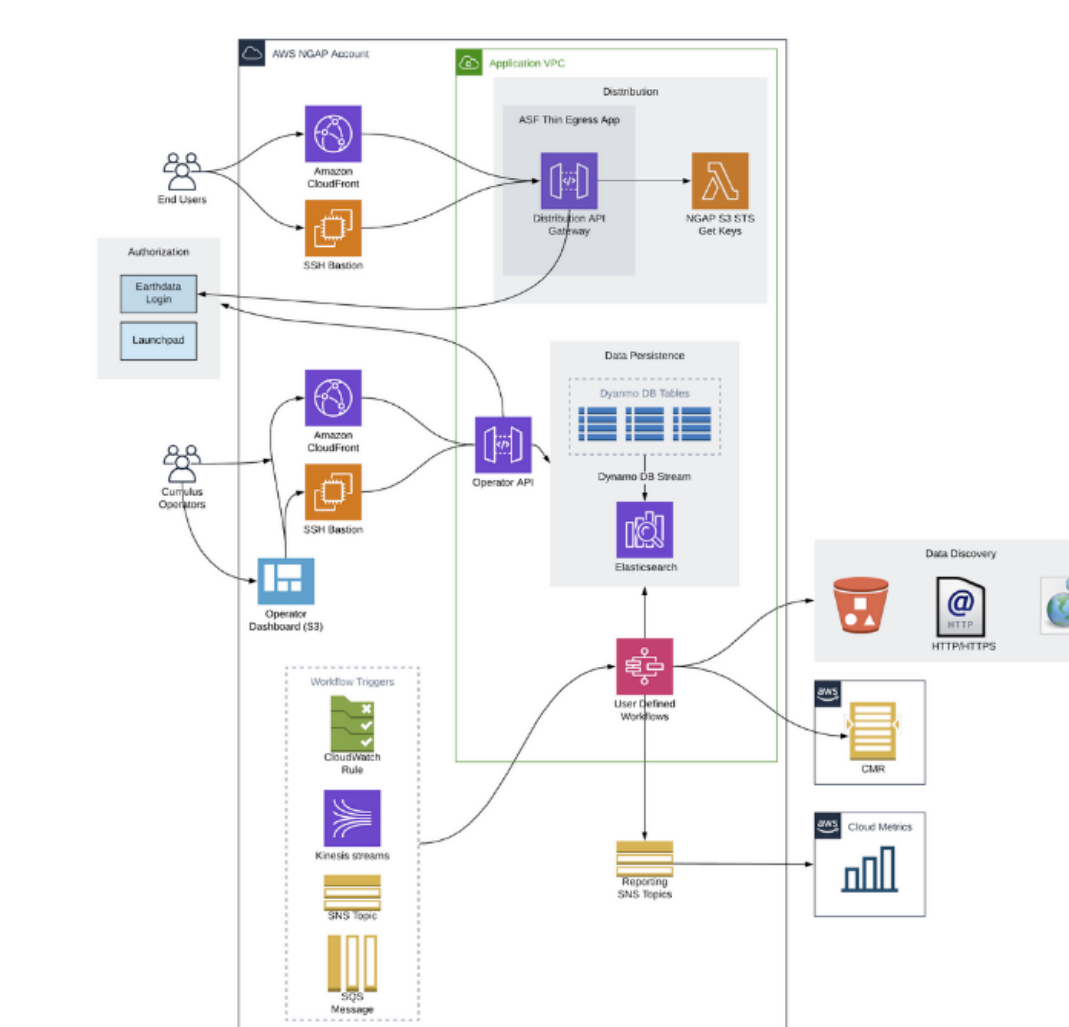
Data stores optimized for temporal analysis - data rods



- Data rods is in on-prem production, will be migrated to Cloud in two years
- Both Parquet and Zarr are prototyped in Cloud Giovanni; one of them will be selected in Cloud production next year
- Both Parquet and Zarr can be deployed on-prem too

Cloud Data Migration - Cumulus

Cumulus Deployment Architecture



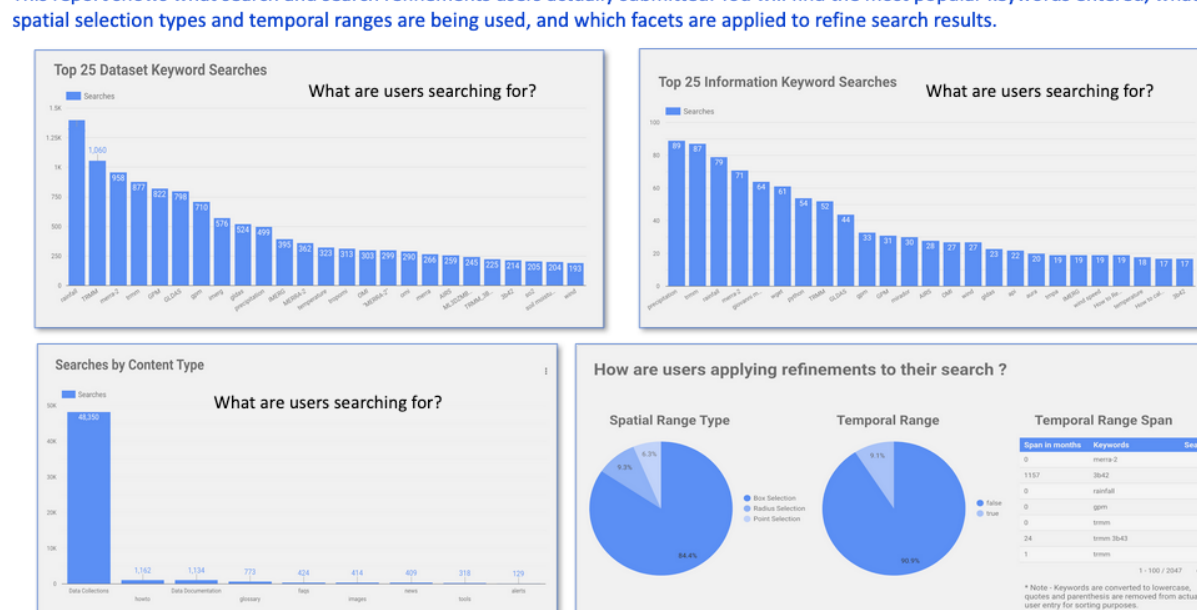
- Cumulus – ESDIS project for “native” cloud-based data ingest, archive, distribution and management system
- Benefit:
 - Scalable performance
 - Co-location of data to facility data fusion
- GES DISC is migrating two datasets (IMERG and MERRA2) to Earthdata cloud using Cumulus
- Expect to achieve parallel operation (in cloud and on-prem) by September 2020
- Future dataset migration includes:
 - AIRS L2
 - OCO 2/3 L2 data comparison
 - OMI and TROPOMI L2 Comparison
 - Selected L3 Data Rods

Deep Study User Metrics

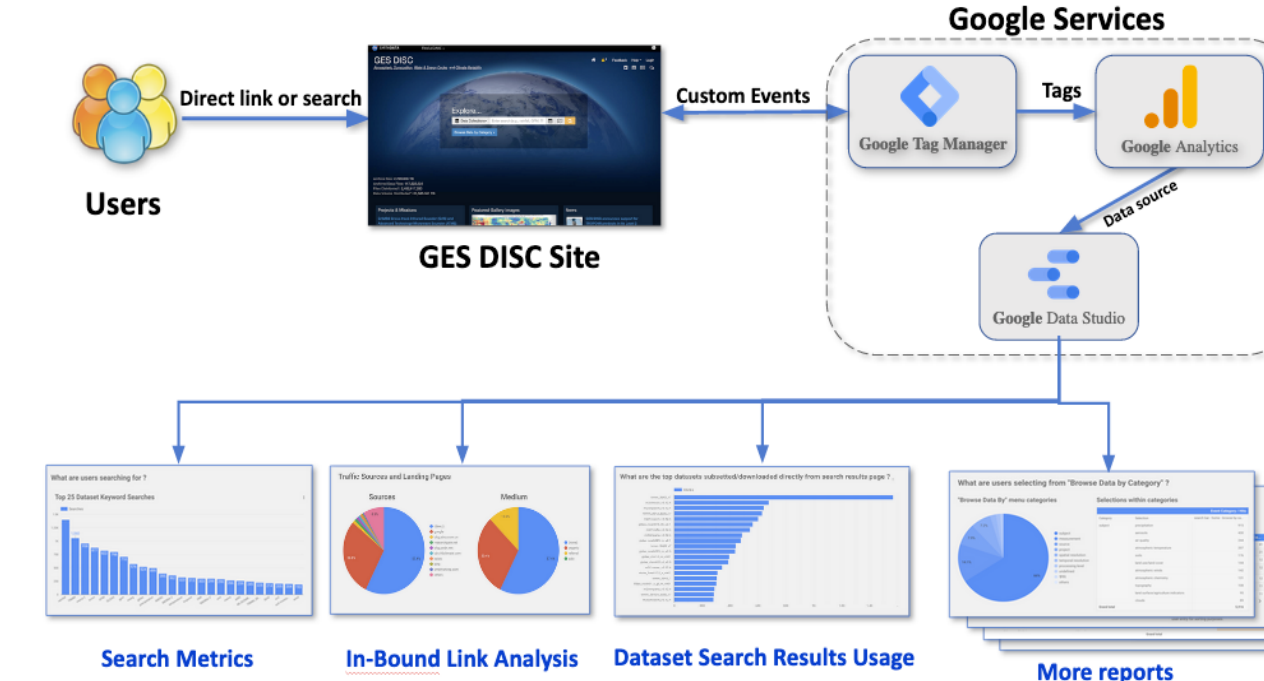
Standard “Out-of-the-box” Metrics



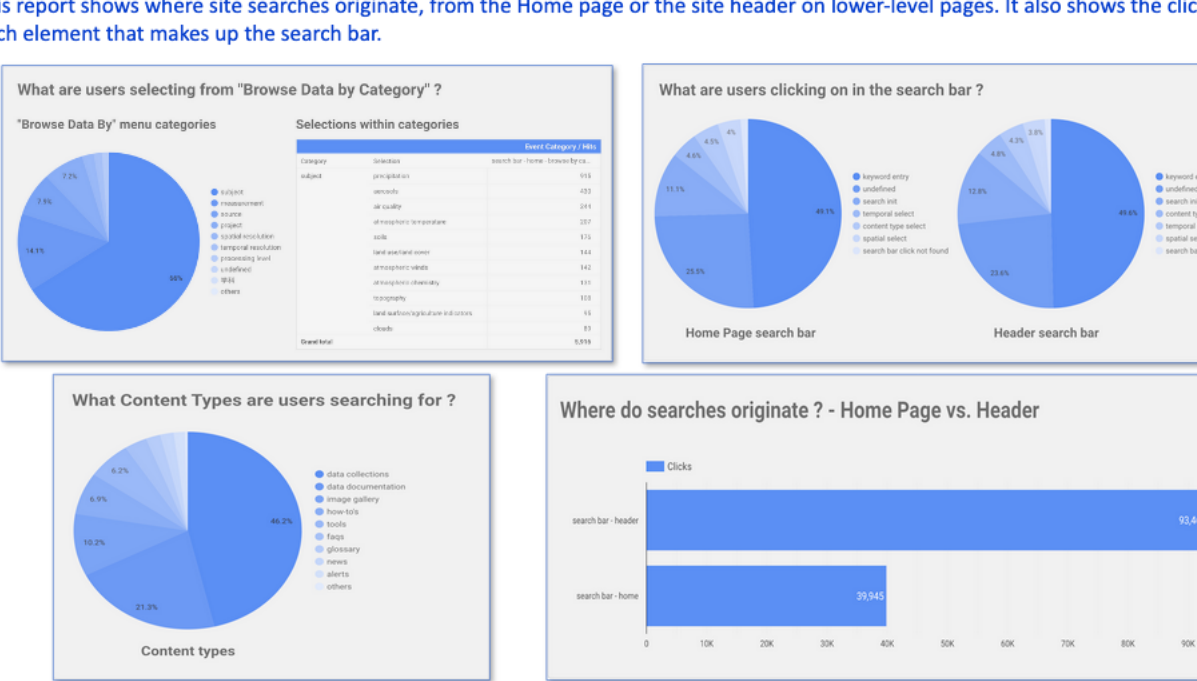
Search Metrics / Search Refinement Report (4/1/19 – 9/30/19)



Custom Metrics Reports



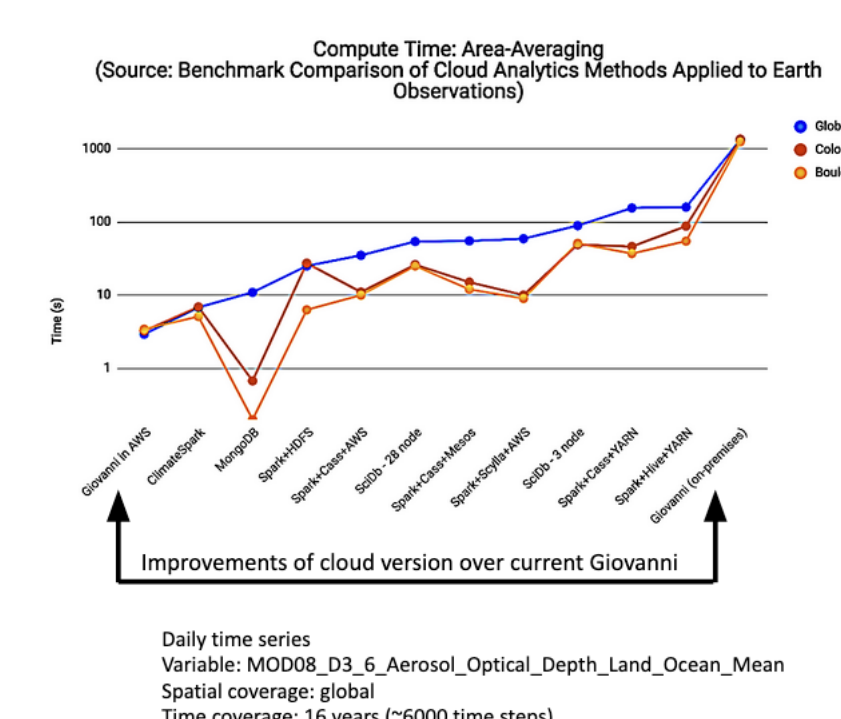
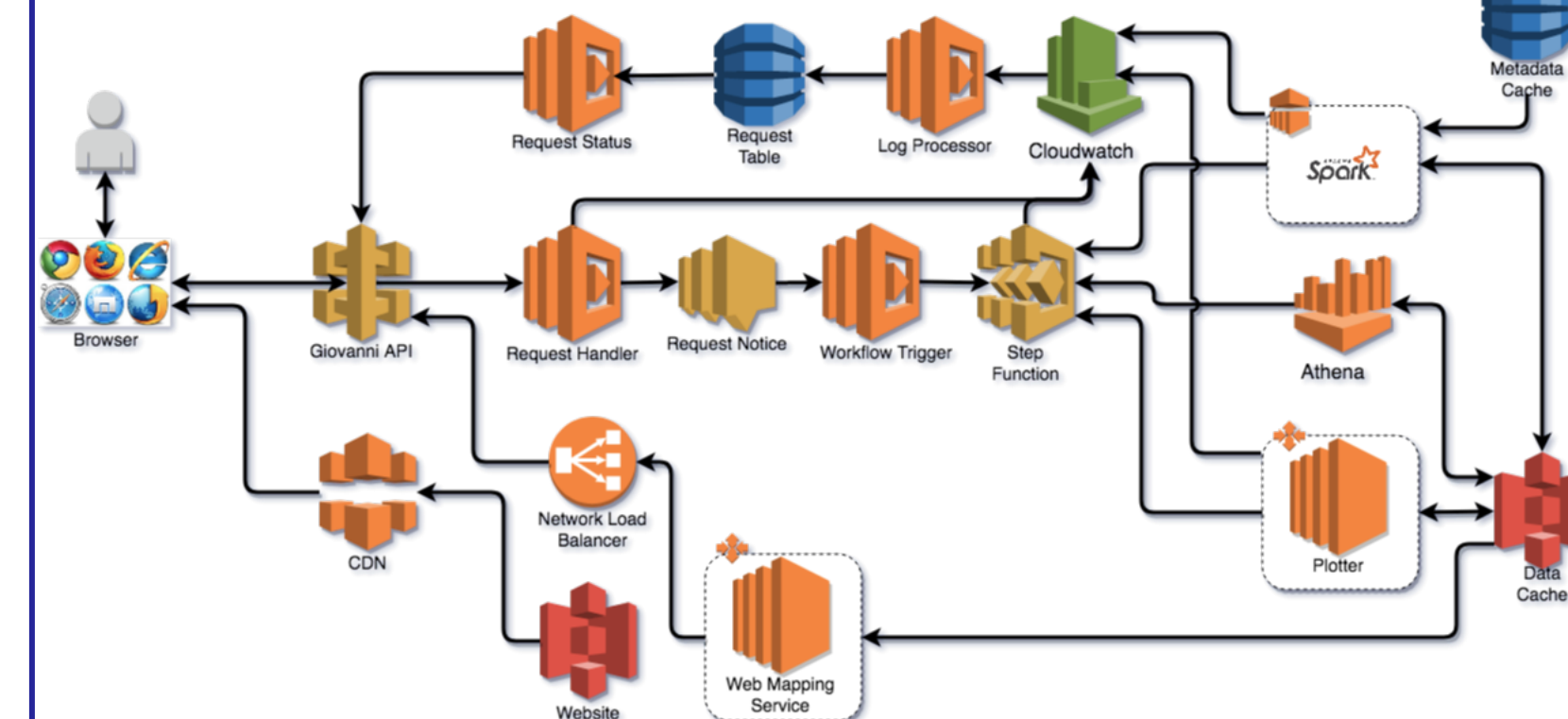
Search Bar Usage Report (4/1/19 – 9/30/19)



- Better understanding of user behavior
- Determine data service needs
- Improve search relevancy
- Future – use AI/ML to study user metrics

Cloud Giovanni

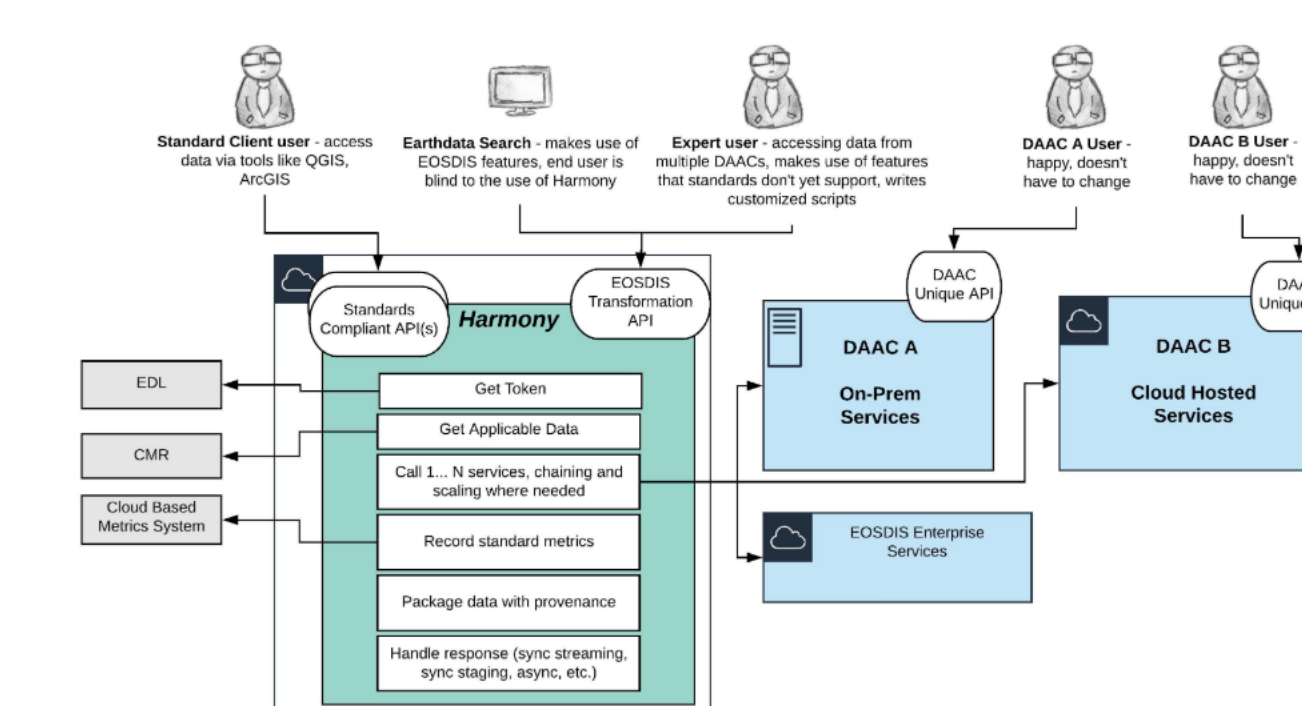
Cloud Giovanni Reference Architecture



- Achieved closed to fixed-cost (per request) both in clock time and dollar amount
- Extreme reduction in compute cost: < \$5 per month for a service which operationally is a cause of frequent server overloading
- Portable Cloud-based data store
- Certain computationally intensive services saw a 500x increase in performance (from 1.5 hours to 11 seconds)
- Expected in Earthdata Cloud operationally next year.

Cloud Data Services - Harmony

Harmony Architecture



- Increase usage and ease of use of EOSDIS data
- Focus on opportunities when multiple DAAC's data all exist in the Earthdata Cloud
- Users can work seamlessly with data from different DAACs in ways previously not possible
- GES DISC is evaluating services to migrate to Harmony and expect to have some services in operation later in 2020

Conclusions

- GES DISC is moving high-value data into AWS to test core archival functions and cloud analytics.
- For more information on GES DISC cloud efforts, please contact the author.

Acronyms

ESDIS - Earth Science Data and Information System
MERRA2 – Modern-Era retrospective analysis for Research and Applications, Version 2
OCO – Orbiting Carbon Observatory

DAAC - Data Active Archive Center
AIRS – Atmospheric Infrared Sounder
TROPOMI – Tropospheric Monitoring Instrument

IMERG – Integrated Multi-satellite Retrievals for GPM
OMI – Ozone Monitoring Instrument
EOSDIS – Earth Observing System Data and Information System

Related Poster

- IN12B-06 - Metrics Learning at NASA GES DISC
- IN12B-14 - Integrated Analysis of Multiple User Metrics - A “Sequel”; and Introducing the Google Analytics
- IN13B-0708 - Cloud Giovanni: Reining in Costs and Improving Performance with Analytical Data Stores using Scalable Serverless Architecture

Authors

¹ NASA Goddard Space Flight Center
² University of Maryland, Baltimore County



<https://disc.gsfc.nasa.gov>