Cloud Giovanni: Reining in Costs and Improving Performance with Analytical Data Stores using Scalable Serverless Architecture

AGU100 ADVANCING EARTH AND SPACE SCIENCE

IN13B-0708

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

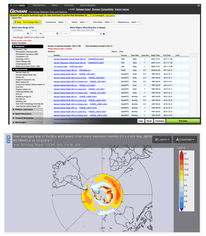
Hailiang Zhang[1,2], Mahabal Hegde[1,2], Christine Smit[1,3], Maksym Petrenko[1,2] and Long Pham[1]

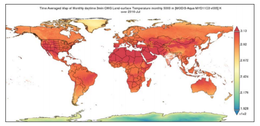
[1] NASA Goddard Space Flight Center, [2] ADNET Systems Inc., [3] Telophase Corp

Background

Giovanni (https://giovanni.gsfc.nasa.gov) is the Geospatial Interactive Online Visualization ANd aNalysis Infrastructure developed at NASA's GES DISC which provides a simple and intuitive way to visualize, analyze, and access vast amounts of Earth science data:

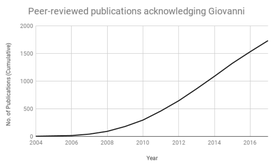
- Twenty-two (22) analysis and visualization services at the click of a button
- Access to over 1477 data variables
- Persistent URLs for sharing data and visualizations

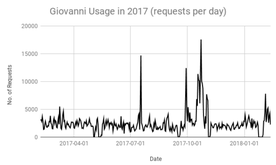




Challenges

- Emphasis on feature set rather than reliability and performance, the two key pillars of a well architected framework
- Unable to meet spikes in demand during training, and "seasonal" events such as conferences and end of academic terms
- Increased demand on resources due to higher resolution data and user demand for data statistics





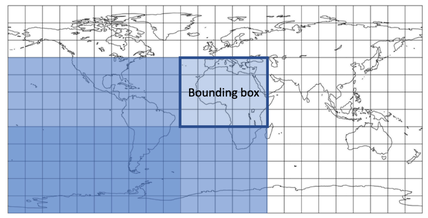
Algorithm

Multi-dimensional global averaging

$$\frac{\iiint variable(\,time,lat,lon\ldots)\cdot weight(\,time,lat,lon\ldots)\,dV}{\iiint weight(\,time,lat,lon\ldots)\,dV}$$

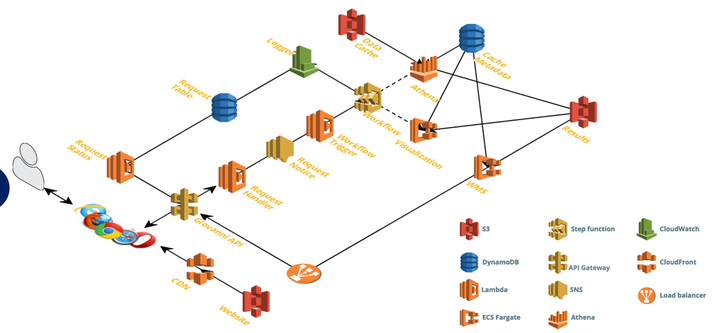
2-dimensional averaging (time series)

$$\frac{\sum_0^{lat2}\sum_0^{lon2} v\cdot\cos(lat)-\sum_0^{lat2}\sum_0^{lon1} v\cdot\cos(lat)-\sum_0^{lat1}\sum_0^{lon2} v\cdot\cos(lat)+\sum_0^{lat1}\sum_0^{lon1} v\cdot\cos(lat)}{\sum_0^{lat2}\sum_0^{lon2}\cos(lat)-\sum_0^{lat2}\sum_0^{lon1}\cos(lat)-\sum_0^{lat1}\sum_0^{lon2}\cos(lat)+\sum_0^{lat1}\sum_0^{lon1}\cos(lat)}$$



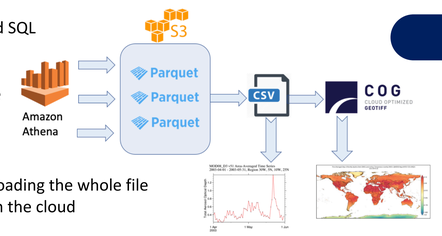
Architecture

AWS stack



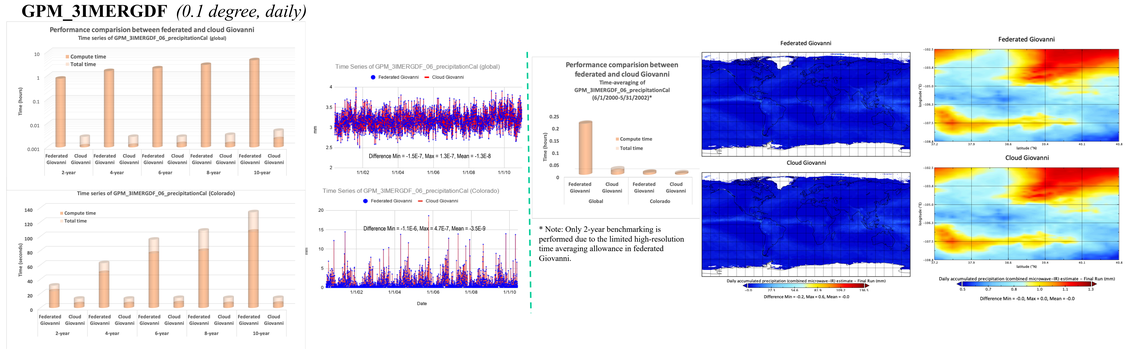
Data analysis frameworks

- AWS Athena
  - Efficient query service using standard SQL
  - High performance
  - Cost efficient: $5/TB
- Parquet-based cloud object storage
  - Columnar data store
  - Highly compressible
- Cloud optimized Geotiff
  - Streaming GeoTIFF instead of downloading the whole file
  - Enabling more efficient workflows on the cloud

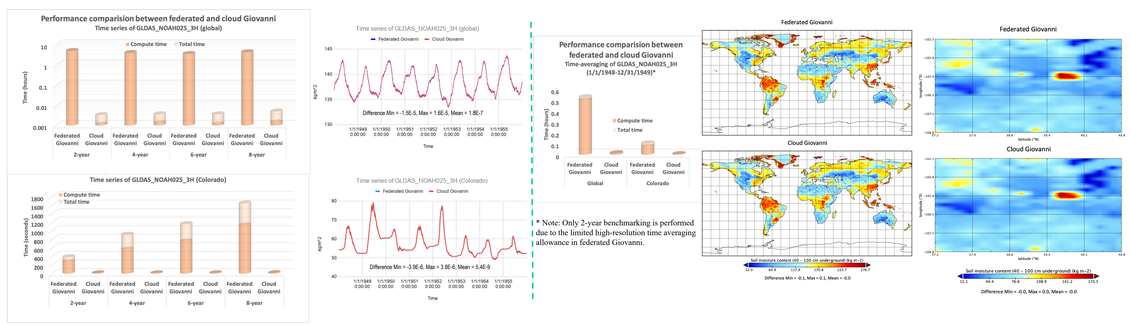


Results and Performance

GPM_3IMERGDF *(0.1 degree, daily)*



GLDAS_NOAH025_3H *(0.25 degree, 3-hourly)*



Cost Analysis

GPM_3IMERGDF
(0.1 degree, daily)

Service type	Duration	Total cost ($) /1000 requests	
		Global	Colorado
Area-averaged time series	2-year	0.4	0.58
	4-year	0.58	0.94
	6-year	0.75	1.28
	8-year	0.92	1.63
Time-averaged map	2-year	12.12	0.52

AWS total cost for multi-dimensional accumulation-based cloud Giovanni

GLDAS_NOAH025_3H *(0.25 degree, 3-hourly)*

Service type	Duration	Total cost ($) /1000 requests	
		Global	Colorado
Area-averaged time series	2-year	0.79	1.22
	4-year	1.33	2.18
	6-year	1.85	3.11
	8-year	2.35	4.02
Time-averaged map	2-year	10.07	0.57

Conclusion

- Multi-dimensional accumulation-based cloud Giovanni significantly improved the performance of Giovanni services including area-averaged time series and time-averaged map. It can improve the performance by >1000 times for time series of variables with high spatial and temporal resolution.
- Multi-dimensional accumulation-based cloud Giovanni provides nearly identical results compared to federated Giovanni.
- Cloud Giovanni incurs reasonable AWS cost for highly intensive averaging services.
- Work is currently underway to further improve cloud Giovanni performance and reduce AWS costs.