Giovanni in the Cloud: Earth Science Data Exploration in Amazon Web Services

Maksym Petrenko1,2, Mahabal Hegde1,2, Christine Smit1,3, Haiang Zhang1,3, Paul Pilone1, Andrey Zasorin1,3, Long Pham1

1 NASA Goddard Space Flight Center, 2 ADNET Systems Inc., 3 Telophase Corp, 4 Element84

Giovanni is an online tool for exploration of geo-spatial data with:
- Twenty-two (22) analysis and visualization services at the click of a button
- Access to over 1600 data variables
- Persistent URLs for sharing data and visualizations

What is Giovanni?

Leveraging Cloud

- Server-less architecture: AWS-managed solutions for services where possible.
  - AWS API Gateway for service endpoints
  - AWS Lambda, Simple Queueing Service (SQS), Simple Notification Service (SNS) for triggering request processing
  - AWS Simple Storage Service (S3) for webhosting and data storage
  - AWS Elastic MapReduce (EMR) for cluster computing
  - AWS Elastic Compute Cloud (EC2) for general computing (Example: Web Mapping Service)
- Micro services: each operation is an independent service, making chaining of services feasible
- OpenAPI based service specifications: enables language agnostic service definition
- Auto-scaling: to meet demand spikes and compute-intensive services
- Use of Apache Parquet, a columnar and open-source agnostic service definition
- Use of Apache Parquet, a columnar and open-source agnostic service definition
- Use of Apache Parquet, a columnar and open-source agnostic service definition

Notable Features

- Single page application for specifying service parameters, navigating and manipulating results
- Rapid exploration of geo-spatial data in time and space
- Serves broad spectrum of users from students to subject matter experts

Pain Points

- Emphasis on feature set rather than reliability and performance, the two key pillars of a well architected framework
- Victim of its own success; 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

See also

IN31A-0068: Use of Schema on Read in Earth Science Data Archives
IN41B-0039: The Value of Data and Metadata Standardization for Interoperability in Giovanni

Prototyped Features

- Services
  - Time-averaged map: most popular service in Giovanni
  - Area-averaged time series: most resource intensive service in Giovanni
- Plot-centric instead of variable-centric user interface: users add data to plots simplifying user experience

Performance Analysis

<table>
<thead>
<tr>
<th>Service</th>
<th>Data</th>
<th>Time HH:MM:SS (On-premises)</th>
<th>Time HH:MM:SS (Cloud)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area-Averaged Time Series</td>
<td>37+ years of Global 0.5° x 0.67° hourly-data</td>
<td>Not Available</td>
<td>00:22:58</td>
</tr>
<tr>
<td>Time-Averaged Map</td>
<td>14 years of 20 x 20 grid 0.25° daily data</td>
<td>00:04:52</td>
<td>00:02:52</td>
</tr>
<tr>
<td>Time-Averaged Map</td>
<td>14 years of Global 0.25° daily data</td>
<td>00:09:03</td>
<td>00:04:00</td>
</tr>
</tbody>
</table>

Cost Analysis

<table>
<thead>
<tr>
<th>Service</th>
<th>Data</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage</td>
<td>25TB</td>
<td>$600 per month</td>
</tr>
<tr>
<td>Compute cost</td>
<td>2500 requests/day</td>
<td>$1600 per month</td>
</tr>
</tbody>
</table>