Asynchronous Data Mining Tools at the GES-DISC

Long B. Pham, Stephen W. Berrick, Christopher S. Lynnes and Eunice K. Eng
NASA – Goddard Space Flight Center
Distributed Active Archive Center

Introduction

At the NASA Goddard Earth Science Data and Information Services Center (GES-DISC), terabytes of data are received daily from the NASA Earth Observing System (EOS). As the volume of data increases, the all-encompassing management of data becomes more complex. Data access, data mining, data analysis and data retrieval may also become cumbersome processes for users, especially when data files routinely exceed 200 MB.

Two solutions to minimize the complexities of data management and to assist users with mining for data closer to the data source are the Simple Scalable Script-based Science Processor for Measurements – Data Mining Edition (S4PM-DME) and the GES-DISC Interactive Online Visualization ANd aNalysis Infrastructure (Giovanni.) The first solution, S4PM-DME, assists users with mining for data within the GES-DISC using the users' own algorithms uploaded to the S4PM-DME system. Users can test and mine for data through the GES-DISC's website. After the mining is completed, the output is placed in an FTP holding area. A daily email is sent to the user with information about the pickup directory and the number of files (Figure 1). S4PM-DME's goals are to give users control over the data they want to process, to allow users easy access over the Internet and to mine for data at the data source using the server's resources rather than the user's local system resources. Thus, the data volume transferred over the Internet will be minimized.

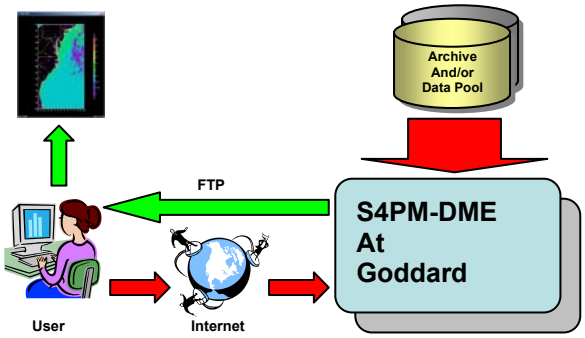

Figure 1: S4PM-DME

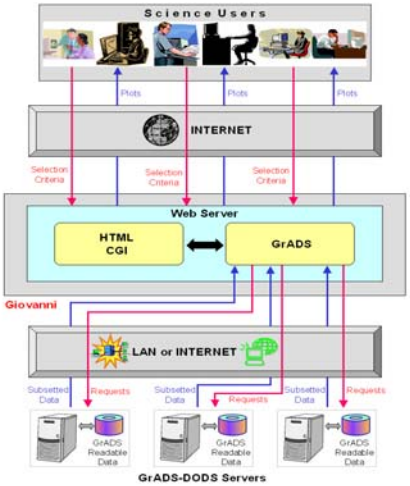

Figure 2: Current Giovanni

The second solution, Giovanni, allows users the capabilities to discover and explore GES-DISC gridded data through sophisticated analyses and visualizations. Users can access multiple data formats, apply server-side temporal or spatial subsetting and view multiple plot types including area, time, Hovmoller, and image animation. Along with these functions, users can also

intercompare multiple parameters from multiple instruments. Giovanni's goal is to off-load the data processing workload onto the machines hosting the data and to minimize the data transfer. The current Giovanni architecture (Figure 2) supports only synchronous processing in which the results are provided to the user in the current Web session. The new Giovanni architecture will be asynchronous in nature. For basic plots or images over relatively small amounts of data, the user will get results quickly within the current session. For requests that either require heavy duty processing or processing of large amounts of data, users will be notified asynchronously when the results are ready. Based on a services oriented architecture, the new Giovanni will be able to handle generic processing requests for any type of data reduction as well as those that produce visualizations.

S4PM-DME

The S4PM-DME system is based on an earlier data mining system developed for data from the Tropical Rainfall Measuring Mission (TRMM) (Lynnes and Mack, 2001). Both systems are based on the underlying Simple, Scalable Script-based Science Processor, also known as S4P (Lynnes et al. 2001). The S4P concept is similar to a factory assembly line where pieces of the product are assembled at each station until the product is completed.

S4PM-DME Typical Mining Scenario

1. First-time user registers with GES-DISC.
2. Users then log onto website: http://g0dup05u.ecs.nasa.gov/S4PM_DM/index.html
3. The user enters the requested information. New user should expect a phone call for user name and password. An established user signs in with her username and password.
4. For a first time user, an algorithm will need to be uploaded to the S4PM-DME system, using the development tools.
5. Once uploaded, S4PM-DME scans, compiles and installs the algorithm. If errors or security flaws are detected a message is displayed to the user.
6. The user can run the algorithm interactively or the algorithm can run automatically by initiating the algorithm for data subscription.
7. If the algorithm has been set up for automatic execution, the user can expect daily email notification of the previous day's output data.

S4PM-DME Future Direction

To further utilize S4PM-DME system and to allow science users more data mining capabilities, GES-DISC will be collaborating with the University of Alabama in Huntsville to utilize their ADaM system (http://datamining.itsc.uah.edu/adam/index.html). ADaM is a mining and image processing toolkits with components that can be configured in a variety of ways to create customized mining processes (i.e. classification techniques, pattern recognition utilities, image processing, filtering and more.) With these added enhancements to S4PM-DME, scientists and data miners will be able to advance future research and effective analyze the vast NASA's Earth science data collection.

Giovanni

Giovanni started out as a data exploration tool through which TRMM data could be explored via sophisticated visualizations and analyses without having to download the data. Because of the popularity of this approach, Giovanni was extended to supporting other user communities and the data sets of most interest to them, typically gridded Level 3 data.

In addition to providing visualizations and analyses of individual data sets, Giovanni also allows users to visually intercompare one or more parameters from multiple instruments. This powerful capability allows users to hone in on the data that will provide the most benefit to their investigations before data are downloaded.

Giovanni currently supports many visualizations: latitude-longitude area maps, time series plots, Hovmoler diagrams (latitude or longitude versus time), and area map animations. For atmospheric profile data, Giovanni offers pressure versus parameter (e.g. species) plots. In almost all cases, the user may opt for ASCII rather than an image output. For multi-parameter intercomparisons, Giovanni provides area maps of time-averaged parameters, time series plots of area-averaged parameters, parameter difference maps, scatter plots, and parameter correlation plots. The GES-DISC is responsive to data user needs and is continually feeding community requests for new data sets, features, and enhancements back into Giovanni.

Current Giovanni Architecture

The principal design goal for the Giovanni architecture was to provide a quick and simple interactive interface for users to generate various visualizations of parameters measured by multiple instruments and draw conclusions, all without downloading data. Alternatively, Giovanni would provide a means to ask relevant what-if questions and get back answers that would stimulate further investigations.

Giovanni consists of HTML templates, CGI scripts written in Perl and in Grid Analysis and Display System (GrADS) language. In addition, there is an image map Java applet through which a user can select a bounding box area to process. Access to data is via one or more GrADS Data Servers (GDS) running on remote machines that have GrADS readable data.

Via the Giovanni Web interface, the user selects one or more data sets, the spatial area, the time range, and the type of output. CGI scripts process the parameters submitted by the user. GrADS scripts are then used to formulate GDS data requests to the appropriate servers with the resulting subsetted data sent back to the client (Giovanni). For local data, the data are read directly. The subsetted data are processed by the CGI scripts into an image or into ASCII data that the user can download. For plots or graphs, the results will be displayed in the Web browser.

New Giovanni Architecture and Future Directions

The new architecture, currently being tested, will be based upon a service oriented architecture supporting a generalized workflow engine. Processing, whether to produce subsetted data, visualizations, or other operations, will be handled via Web services. Unlike the current

Giovanni which only can support processing that can be completed within a user's Web session, the new Giovanni will be inheritently asynchronous and able to support processing beyond this limitation. Users will choose to be notified of completion via a Web page with status information or via a RSS feed.

Data Available

Currently these products are available to S4PM-DME.

1. Atmospheric Infrared Sounder (AIRS) data
2. Solar Radiation and Climate Experiment (SORCE) data
3. Tropical Rain Measurement Mission (TRMM)
4. Ozone Monitoring Instrument (OMI)
5. Microwave Limb Sounder (MLS)
6. Upper Atmosphere Research Satellite (UARS)
7. Future products will include model output from the Global Modeling ad Assimilation Office

Currently these Giovanni interfaces are available. See http://giovanni.gsfc.nasa.gov.

1. Agricultural Online Visualization and Analysis System
2. AIRS Online Visualization and Analysis System
3. Aura MLS Online Visualization and Analysis System
4. MODIS Online Visualization and Analysis System (MOVAS)
5. Ocean Color Time-Series Project
6. OMI Online Visualization and Analysis System
7. TOMS Online Visualization and Analysis System
8. TRMM Online Visualization and Analysis System (TOVAS)
9. UARS HALOE Online Visualization and Analysis System

Conclusion

The S4PM-DME and Giovanni systems can greatly benefit users who dislike transferring and managing huge amounts of data on their own system. Giovanni allows users to explore data for interesting regional or global phenomena and S4PM-DME allows the information gained to be applied to data mining on existing data holdings as well as on new data as they are continually produced. Investigations are underway to develop an architecture that would make Giovanni and S4PM-DME interoperable. This would allow S4PM-DME users to take advantage of Giovanni interactive and visualization capabilities. At the same time, this new architecture would allow Giovanni to access S4PM-DME's user delivered algorithm and S4PM-DME's subscription capability.

To assist users in getting started with S4PM-DME an online tour can be found at http://g0dup05u.ecs.nasa.gov/S4PM_DM/S4PM-DMEQuickGuide.htm.

References

Lynnes, C. and R. Mack, 2001. KDD Services at the Goddard Earth Sciences Distributed Active Archive Center, in Grossman RL, Kamath C, Kegelmeyer P, Kumar V, Namburu RR ed., *Data Mining for Scientific and Engineering Applications*, Kluwer, Dordrecht, pp 165-182.

Lynnes, C., B. Vollmer, S. Berrick, R. Mack, L. Pham, and B. Zhou, 2001. Simple Scalable, Script-based Science Processor (S4P), International Geoscience and Remote Sensing Symposium, Sydney, Australia.

Berrick, S., G. Leptoukh, Z. Liu, H. Rui, S. Shen, W. Teng, and T. Zhu. 2005. Online Interactive Data Analysis of Multi-Sensor Data Using Giovanni, AGU Fall Meeting, San Francisco, CA, December 5-9