Earth Science Mining Web Services

< Infusion of Diverse Technologies via Web Services >

Abstract To allow scientists further capabilities in the area of data mining and web services, the Goddard Earth Sciences Data and Information Services Center (GES DISC) and researchers at the University of Alabama in Huntsville (UAH) have developed a system to mine data at the source without the need of network transfers. The system has been constructed by linking together several existing technologies: the Simple Scalable Script-based Science Processor for Measurements (4SPM), a processing engine at the GES DISC; the Algorithm Development and Mining (ADaM) system, a data mining toolkit from UAH that can be configured in a variety of ways to create customized mining processes; ActiveBPEL, a workflow execution engine based on BPEL (Business Process Execution Language); XBaya, a graphical workflow composer, and the EOS Clearinghouse (ECHO).

Mining Web Services Architecture

ADaM // Command-line data mining algorithm from UAH wrapped as Web Services
- Data mining toolkit developed by UAH
- Includes image processing, pattern recognition and other complex algorithms
- Includes over 100 scientific utilities
- Customizable as well as traditional data mining capabilities

S4PM // GES DISC triggers Web Service workflow via ActiveBPEL engine

ActiveBPEL // Remotely hosted Web Service orchestration
- Open source Java based implementation of the BPEL engine
- Reads WSDL file from XBaya
- Orchestrates processes from initial stage to execution
- Manages flow control, alarms and other executions
- Flexible Perl based processing engine for Mining Web Services
- Used heavily in all GES DISC processing applications
- Robust and reliable tool for process automation
- Capable of accessing large online data collection via ECHO search
- Customizable to meet the needs of most data mining applications
- Open source

Conclusion

Earth Science Mining Web Services, created from an infusion of well-known technologies, have shown promising results to the data mining/scientific community. With an abundance of algorithms available, users can create and execute their data mining workflows without any data transfer. In turn this gives user control over the data they want to process at the server's source. The next phase will be the Smart Assistant for Earth Science Data Mining (SAM). SAM will provide data type/mining ontologies to aid in workflow composition, expansion of existing workflow composer tool and deployment of existing mining services in additional environments.