Earth Science Mining Web Services

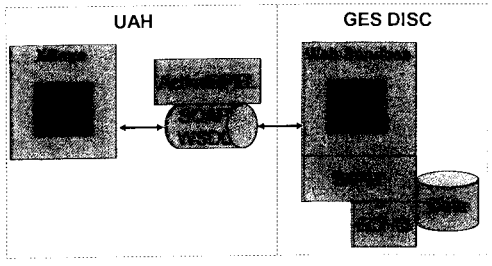
< *Infusion of Diverse Technologies via Web Services* >

[1]Long Pham, [1]Chris Lynnes, [1]Mahabaleshwa Hegde, [2]Sara Graves, [2]Rahul Ramachandran, [2]Manil Maskey, [2]Ken Keiser ([1]NASA GSFC, [2]UAH)

NASA Goddard Earth Sciences (GES) Data & Information Services Center (DISC)
Distributed Active Archive Center (DAAC)
Code 610.2
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771 USA
long.b.pham@nasa.gov

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 pre-existing technologies: the Simple Scalable Script-based Science Processor for Measurements (S4PM), 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).

XBaya is used to construct an analysis workflow at UAH using ADaM components, which are also installed remotely at the GES DISC, wrapped as Web Services. The S4PM processing engine searches ECHO for data using space-time criteria, staging them to cache, allowing the ActiveBPEL engine to remotely orchestrates the processing workflow within S4PM. As mining is completed, the output is placed in an FTP holding area for the end user. The goals are to give users control over the data they want to process, while mining data at the data source using the server's resources rather than transferring the full volume over the internet. These diverse technologies have been infused into a functioning, distributed system with only minor changes to the underlying technologies. The key to this infusion is the loosely coupled, Web- Services based architecture: All of the participating components are accessible (one way or another) through (Simple Object Access Protocol) SOAP-based Web Services.



Web Services enables the infusion of diverse technologies

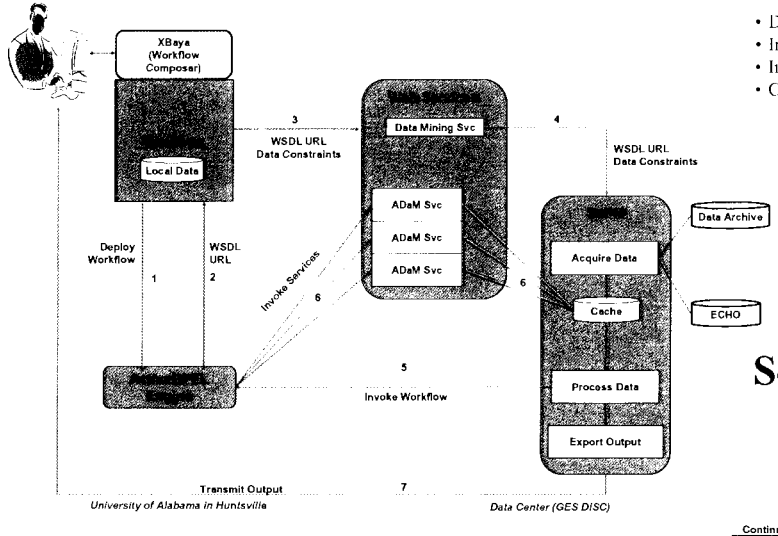
XBaya: *mining workflow composer*
- User authors workflow and deploys to ActiveBPEL engine.

ActiveBPEL: *workflow orchestration engine*
- Exposes a URL pointing to the WSDL for that workflow
- Workflow URL is sent to the GES DISC Data Mining Services via SOAP

EOS Clearinghouse (ECHO)
- Discovery service for data files

Simple, Scalable, Script-Based Science Processor for Measurements (S4PM) :
- Processing engine searches ECHO for data, and caches locally
- Invokes ActiveBPEL to execute ADaM via web service according to workflow
- Stages output to an FTP area for pickup by external user.

Mining Web Services Architecture



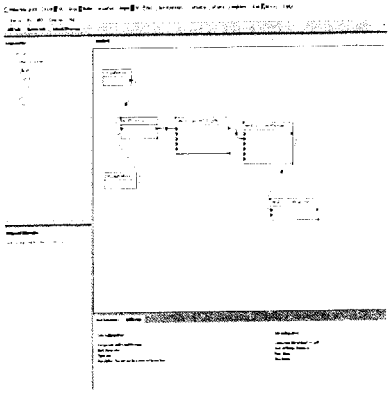
ADaM // Command-line data mining algorithm from UAH wrapped as Web Services

- 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

Pattern Recognition
- Classification Techniques
- Bayes Classifier
- Naïve Bayes Classifier
- Bayes Network Classifier
- Classifier
- And more…

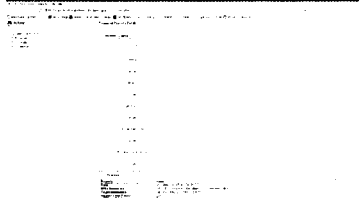
Image Processing
- Arithmetic Operations(+-*/)
- Collaging
- Cropping
- Image Difference
- Image Normalization
- And more…

XBaya // Web Service workflow authoring tools from University of Indiana with modifications from UAH



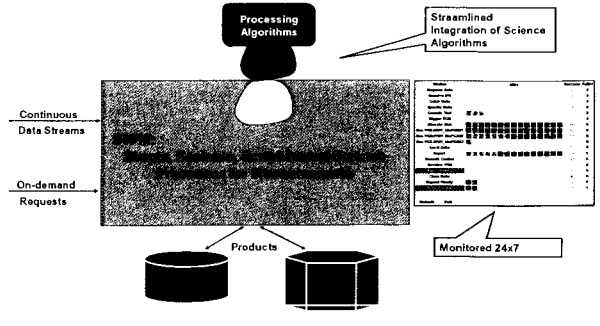
- Java based client-side GUI
- Compose / monitor workflows for Web Services
- Hides complexities of Business Process Execution Language (BPEL)
- Decouples workflow execution from composition
- Deploys WSDL workflows to different BPEL workflow engines (e.g. ActiveBPEL)
- Save workflow to invoke later

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

S4PM // Perl data processing engine from GES DISC triggers Web Service workflow via ActiveBPEL engine



- 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.

AGU Fall 2008 – SF, CA - IN31A-1129