The University of Alabama in Huntsville: Data Mining in the Modeling Environment for Atmospheric Discovery

February 2005
NSF Middleware Initiative (NMI) Integration Testbed Case Study Series

Series contact: Mary Fran Yafchak, Southeastern Universities Research Association, maryfran@sura.org.

The NMI Integration Testbed Program provided practical evaluation of NMI components within the context of real projects and application scenarios from June 2002 through November 2004. During that time, NMI Testbed sites collectively submitted over 220 evaluation reports to middleware component developers as direct feedback into the NMI development cycle. Site representatives also actively inspired, promoted and facilitated the integration of middleware throughout their institutions.

The NMI Integration Testbed Case Study Series documents the most significant outcomes and influences of NMI Testbed sites’ middleware integration efforts, highlighting intersections with established projects, application contexts and influences, drivers for innovation, decision points and challenges. Through this documentation, the work of these pioneering institutions is captured to provide a breadth of insight and approaches for others to use towards successful middleware development and deployment.

This NMI Integration Testbed Case Study Series is sponsored by the National Science Foundation Middleware Initiative-Enterprise and Desktop Integration Technologies (NMI-EDIT) Consortium of EDUCAUSE, Internet2, and SURA. Additional support was provided by the National Science Foundation Cooperative Agreement NSF 02-028, ANI-0123937.

Copyright © 2005 University of Alabama in Huntsville. The University of Alabama in Huntsville permits use of this content for noncommercial purposes with proper attribution. All rights reserved.
Executive Summary

Researchers in the Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville (UAH) are performing research in information technologies, such as data mining, to make data more usable. The ITSC-developed Algorithm Development and Mining System (ADaM) is one such tool that can be used to apply data mining technologies to remote sensed and other scientific data.

In addition to their work in data mining, ITSC is also active in leading edge computational environments, such as grid computing. As part of the process of bringing the benefits of grid technology to their research projects, ITSC applied for and was selected to be an NMI (NSF Middleware Initiative) Integration Testbed site. ITSC selected a data mining algorithm, developed with the ADaM Toolkit, as its first application to grid-enable in the Testbed. ITSC's participation in the Modeling Environment for Atmospheric Discovery (MEAD) project provided a real-life scenario in which they could deploy and evaluate NMI software releases. Since the ability to perform data mining in a grid environment was critical to the data analysis requirements of the MEAD Expedition, ITSC was able to leverage their MEAD and NMI Integration Testbed collaborations to grid-enable ADaM applications.

The ADaM data mining toolkit allows researchers to carry out computing tasks critical to atmospheric modeling. The ADaM system mines large scientific data sets for geophysical phenomena detection and feature extraction. ITSC’s Sara Graves led the MEAD Data Mining and Management team, which was focused on adapting data mining tools, such as ADaM, for use on grids. ITSC used the GRIDS Center’s Globus and Condor-G to adapt a Feature Subset Selection application, developed with ADaM, for use on grids.

ITSC’s work with ADaM was one of the earliest deployments of NMI components within the NMI Integration Testbed. Following their work in the Testbed and the MEAD Expedition, ITSC now offers grid-enabled ADaM applications for mining meteorological data. Grid-enabled ADaM applications are also being used in other grid-based environments and ITSC expects they will be used for mining data in other disciplines as well.

For more information about data mining on the grid at UAH, contact Sandra Redman at sredman@itsc.uah.edu.
NMI Components Highlighted in this Case Study

The NMI components discussed in this case study series encompass NMI-Releases 1 through 4. Information about NMI-Releases can be found at http://nsf-middleware.org/.

**Condor-G**
The GRIDS Center’s Condor-G is a computation management agent for the grid. Condor-G is the marriage of technologies from the Condor project and the Globus project. Home site: http://www.cs.wisc.edu/condor/; Globus (see below)

**Globus**
The GRIDS Center’s Globus Toolkit is an open-source collection of modular technologies that simplifies collaboration across dynamic, multi-institutional virtual organizations. It includes tools for authentication, scheduling, file transfer and resource description. Home site: http://www-unix.globus.org/toolkit/
Data mining is focused on techniques that allow knowledge to be extracted from large amounts of data (1). The large volumes of satellite imagery data collected by NASA, for instance, are excellent candidates for knowledge discovery and data mining techniques. Weather researchers use satellite and other types of data to better understand phenomena occurring in weather. Without data mining technology, it would be an insurmountable task to extract useful information from the huge volumes of data available for research.

Researchers in the Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville (UAH) are performing research in information technologies to make data more usable. The ITSC’s Data Mining Solutions Center (DMSC), the focal point for data mining research, development and services at UAH, has researched and developed a number of technologies for researchers and others that need to conduct data mining. The ITSC-developed Algorithm Development and Mining System (ADaM) is one such tool that can be used to apply data mining technologies to remotely sensed and other scientific data.

In addition to their work in data mining, ITSC is also active in leading edge computational environments, such as grid computing, that have the potential to enhance data mining and scientific research. To enrich its experience with grid computing, ITSC became a participant in the NSF (National Science Foundation) Middleware Initiative (NMI) Integration Testbed Program\(^1\). ADaM applications were grid-enabled by ITSC in conjunction with their work in the NMI Integration Testbed and other projects such as Modeling Environment for Atmospheric Discovery (MEAD) project, an NSF Partnerships for Advanced Computational Infrastructure (PACI) National Computational Science Alliance (Alliance) Expedition (2). The Alliance program is a partnership of organizations working together to build and prototype a grid infrastructure with applications and tools for specific scientific disciplines. Prior to their work in the NMI Integration Testbed, ITSC researchers had worked with NASA researchers to port earlier versions of ADaM applications to the NASA Information Power Grid (IPG) (3).

This article will examine how and why ADaM applications were grid-enabled, paying

\(^1\) NMI Integration Testbed information: http://www1.sura.org/3000/NMI-Testbed.html
particular attention to the influence of ITSC’s participation in the NMI Integration Testbed and to the NMI middleware components that made the work possible.

The NMI Integration Testbed
NMI is an NSF-funded program intended to promote the development, deployment and support of middleware and grid computing technologies through the collaborative work of members of the advanced networking community (4). The NMI Integration Testbed program was created to provide practical evaluation of NMI middleware as a tool for scientists and educators to share resources across the Internet. Managed by the Southeastern Universities Research Association (SURA) on behalf of the NMI-Enterprise and Desktop Integration Technologies (EDIT) Consortium, the Testbed consisted of eight universities that participated in a closely coordinated effort to deploy and evaluate grid middleware technologies.

As part of the process of bringing the benefits of grid technology to their research projects, ITSC applied for and was selected to be an NMI Integration Testbed site. By participating in the Testbed, ITSC researchers were able to work with NMI components and to share their experiences with their Testbed collaborators, while learning through the experiences of others in the Testbed. ITSC’s participation in the NMI Integration Testbed coincided with their participation in other projects, providing ITSC researchers with “real life” scenarios in which they could deploy and evaluate NMI middleware components. ITSC’s work with ADaM was one of the earliest deployments of NMI components within the NMI Integration Testbed. This collaboration resulted in successfully grid-enabling ADaM applications and provided valuable feedback for their NMI Integration Testbed collaborators on NMI middleware components, specifications, and services.

ITSC at UAH: Experts in Data Mining
One of the primary missions of ITSC is to develop and transfer innovative technologies and knowledge into real-world areas. ITSC conducts multidisciplinary research in many areas of data intensive computing, including data mining, data management, grid technology, collaborative computing, and high-performance networking. ITSC researchers are especially focused on problems involving large heterogeneous spatial data sets, such as data collected from remote sensors. Scientists and researchers work with data collected by many types of remotely located sensing instruments that are often operated by various agencies or institutions. Not only is this data different in its inherent nature, the same type of data is often represented in different data formats creating difficulties in performing analysis across data sets.

2 NMI-EDIT http://www.nmi-edit.org/index.cfm
ITSC has developed numerous tools to assist in the research of weather and other domains and has engaged in a number of collaborative projects to develop and test these technologies. Technologies developed at ITSC facilitate knowledge discovery through data mining technologies such as machine learning, and phenomena detection. For example, the EnVironmEnt for Onboard Processing (EVE) provides an extensible and highly flexible processing framework that can be used to both define and control the interactions of multiple processes working on many heterogeneous platforms (5). Using ADaM, the on-board processing environment is capable of detecting complex phenomena based upon information from multiple sensors working together. The system is operable within a distributed heterogeneous array of satellite, airborne and ground-based sensor assets, and is interoperable with on-line processing systems. In addition, ITSC has developed the Earth Science Markup Language (ESML), an interchange technology to facilitate the use of Earth Science data of different formats. ESML enables both structural and semantic data interoperability with applications without enforcing a standard format within the Earth science community (6).

The Algorithm Development and Mining (ADaM) Toolkit

ITSC’s ADaM data mining toolkit allows researchers to carry out computing tasks critical to atmospheric modeling. The ADaM system mines large scientific data sets for geophysical phenomena detection and feature extraction. ADaM components can be classified into those for pattern recognition (component techniques for classification, clustering, and optimization, plus pattern recognition utilities, and association rules) and those for image processing (components for basic image operations, segmentation/edge and shape detection, filtering, and texture features).

The ADaM toolkit’s interoperable components can be linked together in various ways that allow researchers to define and perform data mining operations on scientific data (7). ADaM can be used, for instance, to help researchers working with satellite sensor data to discern lightning patterns by letting them remove the extraneous light from cities and moonlight that the satellite sensors detect (8). ADaM components can also be put together in customized ways and even combined with other specialized software modules. The use of ADaM to detect tropical cyclones and estimate their maximum sustained winds is an example. In this type of research, a combination of general-purpose image analysis modules and special purpose modules developed specifically for the problem are being used (9).

Grid-Enabling ADaM

Why Grid-Enable ADaM?

The knowledge discovery and data mining techniques used in ADaM are well-suited for use with large volumes of satellite imagery
and other types of data that are typically highly computationally intensive. A single high performance computing facility is one option for providing the necessary computational power, but grid technology brings the potential of combining distributed resources to offer higher levels of computational processing power than might otherwise be available. In addition to increased computational resources, grids offer a framework for a standards-based distributed collaborative environment for research and operations.

ITSC selected ADaM as its first application to grid-enable in the NMI Integration Testbed because ADaM’s flexible framework design has a lightweight architecture that can take advantage of the distributed computational and storage resources of a grid. ITSC’s ultimate goal for ADaM is to allow scientists and researchers the flexibility to build analysis systems using the framework or use individual components from the framework for their analysis.

**ADaM within MEAD**

The goal of the MEAD Expedition is to develop and adapt a grid-enabled cyberinfrastructure for mesoscale storm and hurricane research and education, allowing a MEAD user to configure and integrate model simulations, manage resulting model and derived data, and analyze, mine, and visualize large model data suites in a research (not predictive) context. Such a system will allow researchers to work with a broad array of meteorological information, independent of data format and physical location. A particular focus in MEAD is to enable the retrospective analysis of hurricanes and severe storms on the TeraGrid, an NSF-funded collaborative endeavor to provide remote access to a rich set of tightly integrated computing resources for scientific research, including supercomputers, data mining tools, sensors and instruments, data storage, and visualization tools (10,11). The resources of the TeraGrid make it an ideal platform for mining large volumes of computationally intensive observational data such as those from Doppler radar and satellite imagery.

UAH’s Sara Graves led the MEAD Data Mining and Management team, one of two subgroups of the Alliance’s Data Mining and Machine Learning (DMML) working group. The DMML team’s work was focused on adapting data mining tools, such as ADaM, for use on grids. ADaM’s data mining operations were migrated into modular and programmatically scriptable components in order for ADaM operations to be easily integrated into the grid operational environment, and to allow other applications within MEAD to call these operations through a variety of mechanisms (e.g., Python scripts, command line calls), as individual components.

The ability to use grid-enabled ADaM applications was critical to the data analysis requirements of the MEAD Expedition. ITSC

---

*Sara Graves is the Director of The Information Technology and Systems Center at UAH, a Professor in the UAH Computer Science Department. She is also the Director of The Information Technology Research Center at the National Space Science and Technology Center.*
developed select data mining algorithms targeting MEAD’s datasets and problems, and enhanced the selected algorithms for the TeraGrid’s parallel Linux environment. For this effort, researchers developed parallel versions of algorithms for neural net image classification, and studied the performance improvements of the parallelized versions vs. the previous, un-parallelized versions. The Alliance DMML team’s second subgroup, the Workflow, Metadata and Data Management subgroup created a “NCSA grid desktop” which uses the standard grid protocols GridFTP and GRAM for moving data and submitting jobs. ITSC researchers worked with others in this subgroup to take a script utilizing data mining operations and incorporate it into a MEAD workflow.

**ADaM within NMI**

ITSC’s Sandra Redman served as Integration Lead for UAH’s participation in the NMI Integration Testbed. As an NMI Integration Testbed site, ITSC received sequential releases of NMI middleware for integration and evaluation within “real life” projects and applications. ITSC leveraged participation in MEAD and the grid-enabling of ADaM applications to provide a context for NMI evaluation.

UAH began integrating NMI grid components for use with ADaM applications during their evaluation of NMI-Release 2 (NMI-R2), completed in April 2003. ADaM’s Feature Subset Selection application was used as the focus for the evaluation of this release. Supervised pattern classification is a problem that is important in many domains, including medicine, defense, business and the geosciences. It is often possible to improve both the runtime and accuracy of a supervised pattern classifier by eliminating noisy, irrelevant or redundant attributes or features from the data set. Feature subset selection is the process of choosing a subset of the features from the original data set in order to maximize classifier accuracy. Wrapper methods for feature subset selection use a trainable classifier to help choose the features. Wrapper methods work by evaluating how well the classifier performs for a number of different feature subsets, then choosing the subset that exhibits the greatest performance. It is possible to evaluate the performance of a classifier on a selected feature subset by training a classifier with a portion of the available samples, then use an independent set of samples to evaluate the classifier. However, for problems with a large number of features, this is not feasible. In such cases, it is necessary to use a search or optimization procedure to drive the feature subset selection process, which is both processor and data-intensive.

ITSC used the GRIDS Center’s Globus component from NMI-R2 to grid-enable

---

4 Sandra Redman is a Research Scientist in the ITSC at UAH. She performs research in high-performance networking, grid computing, security and video technologies.

ADaM’s Feature Subset Selection application:

- The Globus Toolkit was installed on seven systems in the UAH ITSC lab. Client and server host certificates were obtained from the Globus Certificate Authority. Each system ran RedHat Linux 7.3, with the Globus software installed under /usr/grid.
- The Feature Subset Selection application was successfully ported to work in a grid environment by:
  - Porting the application to Linux
  - Downloading and testing a Linux SVM (Support Vector Machine) implementation
  - Application scripts were developed in Python and tested on Linux, and then modified for the Globus environment by writing a simple Globus RSL file.
  - The application scripts were modified to run each combination of tools on a different node on the grid. Globus was used to execute the jobs on different machines. Only minor modifications to the scripts were required, mostly to change paths so that executables and data files were visible.

Experiments were conducted using both real and synthetic data to determine the extent to which classification accuracy achieved with SVMs can be improved by using optimized feature subset selection. A widely used public domain data set was obtained from the University of California at Irvine (UCI) for testing. This data set involved breast cancer data and is commonly used for testing classifiers. A synthetic dataset was also used to determine whether the Feature Subset Selection application can overcome the susceptibility of SVMs to noisy or irrelevant attributes.

During NMI-Release 3 (NMI-R3) testing, completed in July 2003, ITSC continued their work to grid-enable ADaM applications. The GRIDS Center’s NMI-R3 Condor-G and Globus components were used directly with ADaM. UAH performed application development and testing with Globus from NMI-R3 using an enhanced version of the Feature Subset Selection application that was originally ported to work in a grid environment for NMI-R2. For NMI-R3 testing, classification based on texture features and edge density was used for image analysis. Cumulus cloud fields have a very characteristic texture signature in GOES visible imagery, so detection of cumulus cloud fields in GOES can be accomplished by using texture features or edge detectors. Cumulus clouds are an important area of study since it is thought that man-made changes to land use cause changes in weather patterns, especially cumulus clouds. Comparisons were performed based on accuracy and the amount of time required to classify. Though the Feature Subset Selection Application was updated significantly since NMI-R2, no problems were encountered in running the enhanced version with the NMI-R3 Globus toolkit. UAH used Condor-G to submit and

---

6 Condor-G information: [http://www.cs.wisc.edu/condor](http://www.cs.wisc.edu/condor)
queue the Feature Subset Selection application for execution, with no problems encountered.

**ADaM – Present & Future**

Following their work in the NMI Integration Testbed and the MEAD Expedition, ITSC now offers grid-enabled ADaM applications for mining meteorological data. This offering provides researchers with the capability to use geographically distributed resources in new ways to achieve greater results than previously possible (12).

Grid-enabled ADaM applications are being developed for other grid-based environments, such as NASA’s system for performing remote payload operations of International Space Station (ISS) experiments (13). This system uses a grid infrastructure to provide International Space Station (ISS) remote Principal Investigators with required ground-based services such as payload planning, telemetry data acquisition and processing, spacecraft voice and video distribution, and commanding, while also meeting security requirements and providing the ability to integrate custom applications and collaboration services.

UAH researchers are also extending ADaM capabilities with the development of Linked Environments for Atmospheric Discovery (LEAD)\(^7\), an NSF Information Technology Research (ITR) program complementary to the MEAD program. LEAD is creating an integrated, scalable cyberinfrastructure in which scientists can more effectively use distributed and heterogeneous data for mesoscale meteorology research and education (14). It is expected that grid-enabled ADaM applications will be used for mining data in other disciplines as well.

**More Information**

For more information about data mining on the grid at UAH, contact Sandra Redman at sredman@itsc.uah.edu.

**References**

1. [http://datamining.itsc.uah.edu/adam/tutorials/adam_tut_02_overview_01.html](http://datamining.itsc.uah.edu/adam/tutorials/adam_tut_02_overview_01.html)
2. [www.nsca.uiuc.edu/expeditions/MEAD](http://www.nsca.uiuc.edu/expeditions/MEAD)
5. [http://eve.itsc.uah.edu/](http://eve.itsc.uah.edu/)
6. [http://esml.itsc.uah.edu/index.jsp](http://esml.itsc.uah.edu/index.jsp)
11. [http://lead.ou.edu/](http://lead.ou.edu/)
13. [http://lead.ou.edu/](http://lead.ou.edu/)

\(^7\) This project is detailed in the NMI Integration Testbed Case Study “UAH Grid Integration with LEAD”, available at: [http://www1.sura.org/3000/NMI-Testbed/UAH-GridInLEAD.pdf](http://www1.sura.org/3000/NMI-Testbed/UAH-GridInLEAD.pdf)
Links of Interest

The University of Alabama in Huntsville http://www.uah.edu/
ADaM http://datamining.itsc.uah.edu/adam/
GRIDS Center http://www.grids-center.org/
ITSC http://www.itsc.uah.edu
MEAD at ITSC www.itsc.uah.edu/meal
MEAD at NCSA http://www.ncsa.uiuc.edu/expeditions/MEAD/
NMI-EDIT http://www.nmi-edit.org/index.cfm
NMI Integration Testbed Program http://www1.sura.org/3000/NMI-Testbed.html
TeraGrid http://www.teragrid.org/
LEAD http://lead.ou.edu/