UCRL-ID-151771

Scientific Data Management Center Scientific Data Integration

T. Critchlow, L. Lui, C. Pu, A. Gupta, B. Ludaescher, I. Altintas, M. Vouk, D. Bitzer, M. Singh, D. Rosnick

January 31, 2003

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Scientific Data Management Center
Scientific Data Integration

Terence Critchlow (LLNL),
Ling Liu, Calton Pu (Georgia Tech),
Amarnath Gupta, Bertram Ludaescher, Ilkay Altintas (SDSC),
Mladen Vouk, Donald Bitzer, Munindar Singh, David Rosnick (NCSU)

Summary

The Internet is becoming the preferred method for disseminating scientific data from a variety of disciplines. This has resulted in information overload on the part of the scientists, who are unable to query all of the relevant sources, even if they knew where to find them, what they contained, how to interact with them, and how to interpret the results. Thus instead of benefiting from this information rich environment, scientists become experts on a small number of sources and use those sources almost exclusively. Enabling information based scientific advances, in domains such as functional genomics, requires fully utilizing all available information. We are developing an end-to-end solution using leading-edge automatic wrapper generation, mediated query, and agent technology that will allow scientists to interact with more information sources than currently possible. Furthermore, by taking a workflow-based approach to this problem, we allow them to easily adjust the dataflow between the various sources to address their specific research needs.

Over the last year, we have focused on the scientific data management problem from a data mediation and workflow point of view. In addition to working closely as a team, we also collaborated closely with a molecular biologist, Matt Coleman (LLNL). This collaboration led to a better understanding of the actual application domain, which in turn helped to shape and focus our SciDAC goals and research agenda.

We started our collaboration with the implementation of a "promoter identification workflow tool" that chains molecular biology web resources (databases, tools, etc.) together. In a preliminary implementation of this workflow, we have analytically linked Clusfavor, NCBI's Genbank database and Blast tool, and the MatInpector tool which queries the Transfac database. In this initial version, the data and parameters were explicitly linked between steps and/or filled in by the user where necessary, and the decisions were hardwired. (e.g. pick top three matches of the output and feed them as input into the next step.) This version was demonstrated in a tutorial at the NPACI All Hands Meeting 2002.

The first prototype was extended, based on discussions with our domain scientist, to include new resources and allow more realistic decisions. This new prototype also incorporated wrappers generated by Georgia Tech's XWrap Composer program, which was developed as part of this effort, and utilized the NCSU graphical workflow engine and GUI based workflow design tool. This integrated application was demonstrated at the Supercomputing 2002 conference.

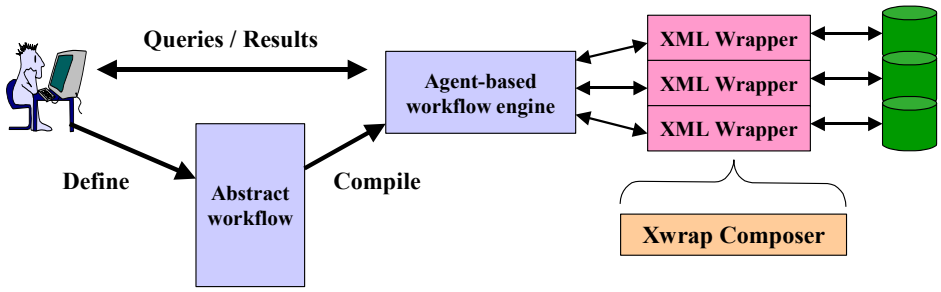

Figure 1 Simple Overview of Approach

As we have developed these prototypes we have given them to our domain scientist. As a result, we have been able to obtain significant feedback on these systems and we have been able to have a direct impact on his work. Matt has used these prototypes as part of his research and this usage has led to two scientific papers being submitted with significantly faster turnaround and less effort on his part than would be possible using any other approach.

Concurrently with the above prototype development, we have undertaken research to determine how the specific workflow example we are using can be generalized to develop a principled approach to workflow design. Our current approach aims at describing the scientist's workflow in an abstract language, such that, given a description and a set of translation rules, a processor of the language would automatically output a specific workflow plan that can be executed for the problem at hand. We researched the necessary properties of the language, and surveyed literature to find other systems with comparable properties.

At this time, we are exploring the two-phase approach shown in Figure 1. In this approach, the domain scientist specifies a general specification of a workflow in the abstract workflow (AWF) language we have developed. It is then compiled into an executable workflow (EWF) definition used by the workflow engine. This translation can be performed in a way similar to the global-as-view translation known from the database mediation area.

The EWF defines the steps to be performed by the workflow engine when the workflow is run. As such, it may be applied to different data sets simply by varying input parameters, without recompiling. We will not be developing an in-house workflow engine. Instead, we have identified an open source workflow engine that meets most of our requirements. Over the coming year, we will extend this system to include all of the basic features required by scientific workflows and use its input language (a community standard) as the initial target for the AWF compiler. We will also extend XWrap Composer to generate wrappers capable of interacting with this workflow.

For further information contact:
Dr. Terence Critchlow
Center for Applied Scientific Computing, LLNL
Phone: (925) 423-5682
Critchlow@llnl.gov

SDM CENTER