Chasing science as a service

Eurekalert!

These days, we do everything on the web. We bank, shop, flirt… so why not conduct science on the web, too?

Web-based science has been an aspiration of researchers and research funding organizations for more than two decades. Currently, it is a driving force behind the National Science Foundation's (NSF) $121 million investment in cyberinfrastructure through the Extreme Science and Engineering Discovery Environment (XSEDE), and a leading factor in the creation of the iPlant Collaborative, a $50 million effort to facilitate research in plant genetics.

For the thousands of scientists who use web portals and gateways to perform simulated experiments ranging from gene sequence alignments to complex datasets analysis, web-based science is already a reality.

Yet the field has a long way to go.

According to a 2009 study by researchers at the University of Toronto, 84 percent of scientists said developing scientific software is important for their own research. Scientists also said they spend, on average, 30 percent of their work time developing scientific software.

This means scientists are devoting a lot of their effort to writing code when they should be (and want to be) doing science.

The AGAVE Advanced Programming Interface (API) is a new tool created by staff at the Texas Advanced Computing Center (TACC) to make scientific computing on the web more functional and intuitive. In the broadest sense, it aims to make launching a computational experiment using supercomputers as easy as buying a book from Amazon.

"When services have been built to that level, research starts moving really fast," said Rion Dooley, a research associate at TACC and one of the creators of the API. "You can start leveraging manpower and focus exclusively on the science rather than the computation and technology needed to accomplish that science."

Some software development cannot be avoided. Specific tools are often required to answer specific scientific questions. However, the work researchers do can be abridged and improved with a platform that provides common functionality.

The AGAVE API does just that.

An API is the set of rules and specifications that software uses to communicate with other software. It serves as an interface between programs and facilitates their

interaction. AGAVE (A Grid and Virtualization Environment) is flexible, web-friendly, and easy to integrate with. It allows researchers with little programming experience to add big time functionality to their scientific computing software.

"If we can give thousands of researchers a few percent of their time back, that's a win," Dooley said.

The AGAVE API helps both the developers and the users of scientific software. For developers, it means user-friendly tools are easy to add. Offload data management and experiment execution with a few lines of code and the power of an entire supercomputing center can be available to users of your scientific software.

For users, the AGAVE API provides science-as-a-service. This changes the paradigm from a manual, command line-driven experience to one that can be done entirely online. For many, it is the difference between walking up to a teller and banking online.

Log in, drag and drop, select and copy

We take for granted all the wonderful web-based technologies we have at our fingertips until we step into a laboratory. So many of the online tools that we rely on in our daily routines simply do not exist in a common, recognizable way inside a scientific environment.

"Millions of people each year interact with Google Maps or Dropbox or Paypal," Dooley said. "There's no learning curve with these tools; they just make sense. That's what APIs do. They provide a platform where people don't think about what's going on underneath."

What's going on underneath the AGAVE API is pretty amazing. AGAVE gives scientists and developers the ability to access and use some of the nation's most powerful supercomputers to accelerate their research.

In its current implemented state within the iPlant project, AGAVE leverages supercomputing resources at Pittsburg Supercomputing Center (PSC), San Diego Supercomputing Center (SDSC), and TACC through the XSEDE project, as well as several smaller, but equally valuable HPC and cloud resources through FutureGrid and the University of Arizona.

Access to these resources brings with it the powerful parallel software tools and libraries required to tackle big problems, supersized storage systems for archiving and sharing massive datasets, and the high speed network necessary to pull it all off.

Developers are already leveraging the AGAVE API to meet their needs. For example, the developers of BioExtract Server used AGAVE to address the large storage and intensive computing needs of bioinformatics researchers.

Created by researchers at Iowa State and the University of South Dakota (USD),

BioExtract Server harnesses online informatics tools and databases to let scientists create and customize workflows for web-based analyses of genomes. Users search online sequence data, analyze the data using informatics tools, create and share custom workflows for repeated analysis, and save the resulting workflows in standardized reports.

The NSF supported the development of BioExtract Server and recently funded Dr. Carol Lushbough, a computer science professor at USD, to integrate the tool into iPlant. The tool accelerates the work of hundreds of academic and industry researchers each year, but it had drawbacks.

"BioExtract Server couldn't handle large datasets and it was really hard for our servers to execute analytic tools that are very CPU intensive," Lushbough said. "We didn't have the horsepower."

iPlant, however, can support both of those functions, and the AGAVE API allowed Lushbough to link together her tool and the iPlant resources to add these capabilities to the BioExtract tool.

"Through their API, we're able to execute iPlant tools, deploy the iPlant environment, store our data at iPlant, use the data at iPlant as input into analytic tools, and then store the output on their environment," Lushbough said. "The API gives us access to resources that we don't have. That's the biggest advantage."

The ultimate goal of AGAVE is to extend the nation's advanced computing resources to a far larger audience across the sciences.

"When a service organization exposes its resources through APIs like iPlant's Foundation API and our AGAVE platform, we can impact thousands of scientists, cutting across disciplines and demographics," Dooley said. "We can do for computational-driven scientific discovery what the printing press did for literacy."

Some of the functionalities AGAVE provides are simple, like the ability to create profiles, authorize the use of software, or move data among resources. Other tools like metadata creation, job monitoring, and auditing are more sophisticated.

The API is preparing for its second major release this summer. The new version is primarily focused on adding new types of systems, such as public and private clouds, that will give users faster turnaround times on their experiments, Dooley said.

Even now, the API is seeing a surge in use. As of July 2012, more than 3,000 researchers were using the API; 75 applications were available for use; and the API was being accessed upwards of 50,000 times per month.

To help the project gain traction, TACC staff provide a developers' console built by Apigee that allows individuals to test the interface and exercise the API without having to write any code. They also provided libraries and wrote a demo application that makes use of the entire API.

According to Karen Cranston, bioinformatics project manager at the National Evolutionary Synthesis Center, the project represents an important step for community software development.

"The promise of easily executable scientific workflows has yet to be realized. The Agave platform represents an important component, allowing developers to access common services and computational resources through a language-independent programming interface," she said.

"As large data sets, such as those from next-generation sequencing, become the norm, scientists will need access to methods, storage and computational power through services that don't require them to become software engineers."

Using AGAVE, scientists hoping to do research on the web will have fewer obstacles and far greater computing power in support of their projects.

"We're not reinventing the wheel," Dooley said. "What we're trying to do is develop a service platform that takes the spotlight off of technology and puts it back onto science."

Source URL (retrieved on *12/08/2013 - 1:46pm*):
http://www.ecnmag.com/news/2012/07/chasing-science-service