UC San Diego News Center

April 03, 2014 | By Jan Zverina

SDSC Establishes Data Science Workflows 'Center of Excellence'

Supercomputer Center to provide expertise to researchers across all domains

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, has formally established a new 'center of excellence' to assist researchers in creating workflows to better manage the tremendous amount of data being generated across a wide range of scientific disciplines, from natural sciences to marketing research.

Called the WorDS Center (or Workflows for Data Science Center of Excellence), the initiative leverages more than a decade of experience within SDSC's Scientific Workflow Automation Technologies Laboratory in developing and validating scientific workflows for computational science, data science, and engineering at the intersection of distributed and parallel computing, big data analysis, and reproducible science, while fostering a collaborative working culture.

A full overview of WorDS services and potential applications can be viewed at the WorDS website.

A data science workflow is the process of combining data and processes into a configurable, structured set of steps that lead to automated computational solutions of an application. The workflows contain a full range of capabilities such as execution management, provenance tracking and reporting tools, integration of distributed computational and data management technologies, and data streaming interfaces. Creating such data science workflows, however, is not without technological challenges.

"The WorDS Center's purpose is to allow scientists to focus on their specific areas of research rather than having to solve workflow issues, or the computational challenges that arise as data analysis progresses from task to task," said Ilkay Altintas, SDSC's deputy coordinator for

research and director of SDSC's Scientific Workflow Automation Technologies Laboratory, and director of the new WorDS Center. "The amount of potentially valuable information buried in what is commonly known as 'Big Data' is of interest to numerous data science applications, and big data workflows have been an active area of research ever since the introduction of scientific workflows in the early 2000s."

Specifically, the expertise and services in the center will include:

World-class researchers and developers well-versed in data science and scientific computing technologies.

Research on workflow management technologies that resulted in the collaborative development of the popular Kepler Scientific Workflow System.

Development of data science workflow applications through a combination of tools, technologies, and best practices.

Hands-on consulting on workflow technologies for big data and cloud systems, i.e., MapReduce, Hadoop, Yarn, Cascading.

Technology briefings and classes on end-to-end support for data science.

The WorDS Center will be funded by a combination of sponsored agreements and recharge services. In addition to Altintas, the center's key personnel include the following SDSC researchers: Dr. Jianwu Wang as Assistant Director for Research; Dr. Daniel Crawl as Assistant Director for Development; and Shweta Purawat as New User Applications Specialist.

"We view WorDS as an excellent opportunity to teach a much larger number of researchers how to create efficient and effective workflows," said Altintas, one of the founders of the Kepler workflow collaboration, which provides researchers with the means to access, arrange, and share data and workflows via a common interface. "The technology behind such systems has matured to the point where this is now an opportune time to establish such a center."

"The WorDS Center is a natural addition to SDSC's other centers of excellence, which are part of SDSC's larger strategic focus to help researchers across all domains, including those who are relatively new to computational science, manage the challenges posted by massive data sets or numerous smaller ones," said SDSC Director Michael Norman. "The age of data-enabled science is upon us, and it's here to stay."

SDSC Centers of Excellence

The WorDS Center joins four other SDSC centers of excellence that are focused on big data management across multiple disciplines, as well as Internet topologies.

- The Center for Large-scale Data Systems research (CLDS) was established in 2012 as an industry-university partnership to study the technology and management aspects of big data, with a related goal of developing a set of benchmarks for providing objective measures of the effectiveness of hardware and software systems dealing with big data applications.

- The Predictive Analytics Center of Excellence (PACE), also announced in 2012, was started to foster collaboration and education among industry, government, and academia to provide a multi-level curriculum that gives business and science enterprises the critical skills to design, build, verify, and test predictive data models. Both CLDS and PACE were recognized at a 2013 White House Office of Science and Technology Policy (OSTP) meeting for projects focused on accelerating collaborations in data-enabled science.

- The Cooperative Association for Internet Data Analysis (CAIDA), formed in 1997, is a collaborative undertaking among organizations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure.

- Sherlock is an SDSC center of excellence focused on information technology and data services for healthcare and government that includes cloud computing, cyber security, data management and mining, application development, HPC, big data, and visualization. Sherlock offers four major products: Sherlock Analytics, Sherlock Case Management, Sherlock Cloud, and Sherlock Data Lab, along with an array of services and consulting expertise.