Free tool kit to assist big-data scientists

19 July 2011, By Rob Knies

Two years ago, during a cyberinfrastructure meeting convened by the U.S. National Science Foundation, principal investigators from across the country found their scientific concerns begin to converge.

"All around the table," recalls Roger Barga, an architect in Microsoft Research's eXtreme Computing Group (XCG), "people were saying, 'I need the means to analyze data,' or 'I need a library of analytics that scale out over large data.'"

Barga and his colleagues took note, and in Redmond, Wash., on July 18, the opening day of the 12th annual Microsoft Research Faculty Summit, they provided their response to the scientists' plea: a platform code-named "Daytona," designed to expand the tool set for scientists who require large-scale data computation.

"'Daytona' gives scientists more ways to use the cloud without being tied to one computer or needing detailed knowledge of cloud programming-ultimately letting scientists be scientists," says Dan Reed, corporate vice president of the Technology Policy Group at Microsoft. "We're very excited to empower the research community with this enhanced tool kit that will, hopefully, lead to greater scientific insights as a result of large-scale data-analytics capabilities."

That offering is capturing plenty of attention during the Faculty Summit, a three-day event at the Microsoft Conference Center in which more than 300 leading computer scientists, academics, educators, and governmental officials consult with Microsoft researchers on challenges and trends in computing. The theme of this year's event is Future World, and, accordingly, attendees will explore new advances in natural user interfaces, cloud computing, and machine learning that reflect Microsoft Research's significant collaboration with computer scientists and academic researchers to advance the state of the art in computing.

"Daytona," Barga explains, represents the next step in that long-term commitment.

"'Daytona' has a very simple, easy-to-use programming interface for developers to write machine-learning and data-analytics algorithms," he says. "They don't have to know too much about distributed computing or how they're going to spread the computation out, and they don't need to know the specifics of Windows Azure."

From the outset of the project, Barga and his XCG colleagues had one goal in mind: to make it easier for scientists and researchers to explore today's burgeoning data sets effectively.

"Increasingly," he says, "we have seen in our cloud-engagement program scientists and researchers wrestling with large data collections. They want to extract insight from these collections, and they know which algorithms to use. But they need them to scale out the algorithm to process data volumes that they've never tackled before: terabytes of data, not just megabytes or gigabytes.

"These are the people we see using 'Daytona' to run algorithms over large-scale data collections to extract insights: find patterns, find clusters, find outliers, and build training models so they can classify incoming data based on the data they've captured and classified to date."

Not only are new releases of "Daytona" scheduled monthly, to incorporate the latest advances and feedback from the scientific and research communities, but it also is being distributed for free.

"We wanted something that we could freely distribute as soon as possible," Barga explains. "We now have something that accumulates all the knowledge our team has gathered on how to work with Azure, build applications on Azure, and scale out data on Azure. What we're releasing now has been rigorously tested and well documented. We have code samples and programming guides-a full kit. People can now build on top of that."

Barga says success for "Daytona" will be measured in a couple of ways.

"One," he says, "is that groups both inside Microsoft and outside Microsoft are installing it on Azure and they're building libraries on top of it that they're basing their research on. Two, that we're getting active feedback from the community on how to enhance it.

"We'll be getting suggestions from the community on how to make it better, and we can respond to that, in addition to putting our own ideas in. If we have an active dialogue, what more could you ask?"

Provided by Microsoft Corporation