Democratizing Data

How a new generation of cloud technologies gives more students and institutions access to high-performance computing

Introduction

Clemson University researcher Dr. F. Alex Feltus calls his computational biology lab "an effort to propel genomics research from the Excel-scale toward the exascale."

His lab's genomics workflows break down the building blocks of life by analyzing DNA sequences that range from "a few dozen to 20,000 datasets that can sum into the petabyte range," says Feltus, a professor in Clemson's genetics and biochemistry department.

As researchers have moved from spreadsheets to these massive stores of data, the kind of high-performance computing required to generate insights and advance science was once only available to top-tier research institutions using proprietary tools. Today, Feltus and his group publish their workflows as open source projects that other labs can adopt freely, and next-generation cloud technology provides the scalable computing resources required to make them work. The result, he says, is "democratizing" access to data and analytics.

> “ With open source tools, any user at any institution can have the power of an elite research lab as long as they have access to a cloud system.
>
> *Dr. F. Alex Feltus, Professor, Clemson University*

"With open source tools, any user at any institution can have the power of an elite research lab as long as they have access to a cloud system," Feltus says. "In pilot projects, we are training brilliant people in high schools, community colleges, colleges and universities how to access and run these workflows on democratized systems."

This paper from the Center for Digital Education and Cisco focuses on how high-performance cloud computing can help institutions of all sizes meet the evolving needs of their students, researchers and communities.

The Next Generation of Research

Much of today's most promising research is happening in the cloud.

At Clemson, for example, Feltus and his lab sequenced the DNA of a tumor of a friend suffering from a rare form of kidney cancer. But that was just the beginning. The Clemson lab also held a hackathon to bring together researchers at other universities and Silicon Valley companies. The purpose was to compare the genetic makeup of the tumor to thousands of others the National Institutes of Health (NIH) has studied and uploaded to a database.

"We all worked together … trying to understand how his tumor had mutated during the tumor progression process," Feltus says.

The results helped identify treatment options, and the patient ultimately went into remission. This kind of precision research – zeroing in on specific details hidden within mountains of data – relies on two massive trends in science and technology.

Computational science has given researchers the power of high-performance computers to tackle complex real-world problems in a range of fields – from medicine and biology to engineering and social sciences. Researchers develop complex algorithms and simulations to create an environment where large amounts of data are analyzed to identify best practices or personalize treatments.

However, doing so requires a lot of computing power – the kind that only universities with high-performance computing clusters once could marshal. Researchers used to "wait for time in the lab," says Mike Shepherd, business development manager in Cisco's U.S. Public Sector Organization. Today, the sheer size and scope of these projects tax even those high-performance clusters.

"Every lab is going to be generating terabytes and gigabytes of data. Ten or 20 years from now it will be exabytes of data, and people can barely process things at giga scale right now," Feltus says. "It's really important the computer systems match what the biologists need so they can do their work."

Enter the flexibility of the cloud, which allows researchers to obtain the right amount of processing power to meet the needs of a specific project.

> “What the cloud ultimately allows researchers to do is determine where the best place is to run a job based on their needs – cost, speed and infrastructure.
>
> *Mike Shepherd, Business Development Manager,*
> *Cisco's U.S. Public Sector Organization*

"The ability to scale just-in-time using the infinite resources of [cloud providers] has led many researchers to pivot to the cloud," Shepherd says.

Today, many research projects are run in hybrid environments, depending on the scope and specific needs of each task.

"What the cloud ultimately allows researchers to do is determine where the best place is to run a job based on their needs – cost, speed and infrastructure. If they need massive computational horsepower, they might run that at the high-performance computing center. If it needs to be dynamic, they might pick the cloud," Shepherd says. "They can figure out the work they need to do, what the best way to do it is and what will yield the fastest results."

But that's just the beginning. When the complex algorithms and computing power of computational science are married with the cloud, opportunities for collaboration like the Clemson hackathon emerge. And one component of cloud computing in particular is dramatically simplifying the process.

Containers and Collaboration

Institutions have collaborated on research in the cloud since at least 2007, when the first multi-university project, the Academic Cloud Computing Initiative (ACCI), was announced. In the years that have followed, the National Science Foundation and other federal agencies devoted to research have provided cloud testbeds, datasets and other resources. As the cloud has matured, open source and software-as-a-service (SaaS) applications have become available for discipline-specific research in a range of areas.

"However, it is important to recognize the difference between technology being ready and the disciplines being ready," state the authors of a National Science Foundation-funded report.[1]

Researchers in many disciplines are aware that bringing peer institutions and their resources together is the best way to solve the complex challenges the world faces, and the cloud can improve collaboration. But replicating complex research projects across different institutions and cloud platforms can be difficult.

“One huge flaw in computational science is reproducibility,” says Cole McKnight, a cloud architect at Clemson. “It’s really hard to reproduce these experiments because the configuration needed is very exact.”

Enter containers – which essentially do what their name suggests by allowing researchers to take a complex computational environment and standardize it so it can be run by anyone on any cloud platform.

“Imagine having to build a spreadsheet from scratch every time you want to do an analysis,” Shepherd says. “Think of Excel as a template. That reproducibility, that kind of standard approach enables you to scale and be far more efficient.”

Containers have become a key technology driving the adoption and flexibility of cloud systems. Nearly 8 in 10 enterprise managers consider technology for deploying and managing containers a priority, according to one study.[2]

For researchers, containers can allow more time to focus on science, not setup. Feltus says that 20 percent of the time it takes to prepare the computing environment for a specific research project involves configuration. With containers, he says, researchers can spend more of their time collecting and analyzing data.

“We can now spend our time troubleshooting the scientific discovery through cross-institutional workflow engineering in a central repository and data exceptions instead of troubleshooting software installation and compatibility issues,” Feltus says.

Containers also allow researchers to choose which cloud provider offers the right combination of power and pricing for a specific project. One of today’s most popular approaches to containerization, called kubernetes or K8s, is a cloud-native open source project, meaning it is freely usable across providers and platforms.[3]

Kubernetes and other container technologies provide opportunities to plug in solutions that bring artificial intelligence and machine learning to bear on complex problems and datasets. “These are baby steps into flattening the research curve,” says Shepherd.

Container solutions also allow researchers at other institutions to easily create an identical computing environment to work on common problems.

“Using containers allows you to contain something that you can reproduce in other labs and platforms,” McKnight says.

Factors to Consider

When evaluating cloud solutions for research, institutions should focus on critical capabilities to best manage the computing environment and containers, including:

✓ **Scalability.** Because of the size and scope of modern research projects, the cloud environment must easily scale storage, memory and processing power. Containers should be able to grow in size or be duplicated to meet these needs.

✓ **Elasticity.** Different researchers or institutions may experiment with variations of analytical models in shared projects to identify the most effective approach to addressing a research question. Containers provide the ability to swap in and out new models with the same dataset without having to start rom scratch.

✓ **Reliability.** An array of “good enough” tools and limited standards can make deploying, monitoring and managing containers challenging. As a result, creating and managing containers can produce errors that lead to an inconsistent computing environment. Automation and unified management tools can help simplify the process and reduce mistakes.

✓ **Shareability.** The ability to allow other researchers or institutions to replicate a computing environment to work on a common problem is critical. Containers also can allow institutions with their own on-premises computing resources to move projects to and from the cloud as needed.

Technology that allows researchers to manage multiple containers across hybrid systems and share them with peers allows researchers at colleges and universities to teach real-world skills their students will take with them to future projects and careers.

“In my lab, I have trained a score of undergraduate and graduate students how to run containerized workflows in cloud systems to discover knowledge in human and crop genomics fields,” Feltus says. “As they graduate, they are doing science in this new way, and they are getting great jobs in life science and other industries.”

Democratizing Data

Perhaps the greatest opportunity of cloud-based computational research is increased access.

“There’s a real thirst to enable all students to be able to tap in and contribute,” Shepherd says. “Think of social media, but social research. That’s a big deal, and it represents orders of magnitude.”

To that end, Clemson’s Feltus is ensuring his lab’s work can be replicated elsewhere.

“All of our released workflows are open source, can be found in GitHub and are accessible by anyone,” he says. Because the workflows are described in ways that ensure compatibility, “all we have to do is point a collaborator or interested party to the GitHub repository and they will run software the same way, ensuring reproducibility.”

But the potential of democratizing data and analytics goes beyond college research. Feltus and others envision using the power of the cloud to draw in younger students and citizen scientists to help solve complex problems.

“I am especially interested in training students in high school career centers and technical colleges as they can fill critical jobs and start innovative new companies in the cloud without the expense of traditional university degree programs,” Feltus says.

> “I am especially interested in training students in high school career centers and technical colleges as they can fill critical jobs and start innovative new companies in the cloud without the expense of traditional university degree programs.
>
> *Dr. F Alex Feltus, Professor, Clemson University*

To that end, he has held workshops and is developing a cloud learning platform called Praxis AI to provide training that can be applied in a range of domains, including biotechnology, energy, geographic information systems, finance and the social sciences. Over spring break, he used the platform to teach high school students from the Porter-Gaud school in Charleston how to sequence the DNA genome of the coronavirus – which has nearly 30,000 “letters” in its sequence – and translate it into protein structures.[4]

“If I can get the genius kid in high school who is bored out of his or her mind in the classroom to plug into a kubernetes cluster and do some cancer research in the back of the room, that’s what I want. To enable citizens to do this work even if they don’t have a PhDs. I want to be able to access those people’s brains through these compute systems.”

This piece was developed and written by the Center for Digital Education Content Studio, with information and input from Cisco Systems.

1. http://dsc.soic.indiana.edu/publications/Cloud%20Forward%20Final%20Report%205.3%20CLEAN%20.pdf
2. https://www.cisco.com/c/dam/en/us/products/collateral/cloud-systems-management/container-platform/container-management-idc-white-paper.pdf
3. https://kubernetes.io/
4. prxai.com

Produced by:

CENTER FOR DIGITAL EDUCATION

The Center for Digital Education is a national research and advisory institute specializing in K-12 and higher education technology trends, policy and funding. The Center provides education and industry leaders with decision support and actionable insight to help effectively incorporate new technologies in the 21st century. **www.centerdigitaled.com**

For:

CISCO

Digital education is making it possible for students to learn more, in new ways, in new places, with new connections to resources around the globe. Cisco is leading this new digital world in education, including with solutions for Safer Schools, which support students as they learn without limits. **www.cisco.com/go/education**