NIH Prioritizes Cloud Migration in 2020 IT Ecosystem Plan

The agency plans to tackle research-data roadblocks through its STRIDES Initiative.

Faith Ryan
Mon, 12/30/2019 - 15:02

Photo credit: ojogabonitoo/iStock

The National Institutes of Health is aggregating and integrating scientific data across various enterprise data platforms to advance biomedical research. With that effort, researchers are encountering challenges that cloud-service technologies could address in the coming year.

Firstly, it is particularly difficult for researchers to link patient data for those who've participated in multiple research studies. The same NIH research subjects are often involved in other studies that could be used for correlational research. Researchers studying patients with chronic obstructive pulmonary disease, for instance, could

connect relevant data from the same participants or research studies.

"There are genetic and dietary data that are available that can be used to study the effects of this disease," noted the agency's Office of Data Science Strategy Director Susan Gregurick at the agency's 119th Advisory Committee to the Director meeting this month, but it is timely and cumbersome for researchers to search data-rich platforms for relevantly linked patient information.

Another challenge is studying rare diseases like pediatric cancer can be negatively impacted by disjointed data queries across multiple resources, Gregurick said. Given that time is of the essence for certain diseases, incorporating external research groups that have integrable data resources could further enhance NIH's research capabilities, she added.

Finally, integrating data from multiple repositories and platforms is simply challenging. Greguick explained that NIH's research data is housed across various data platforms that are incapable of talking to one another. Moreover, being able to access data from multiple platforms easily would allow for better analyses and insights, coupled with implementing data analytics tools and user-friendly interfaces.

By harmonizing data, correlations between certain diseases could be discovered. Any associations between cardiovascular disease and aging-related dementias, for example, could be analyzed using interoperable data platforms, such as NHLBI's Framingham Heart Study and NIA's Alzheimer's health data.

"This is the holy grail. This is what we're going for," she said.

To overcome these data challenges, NIH is currently enhancing its cloud-computing capabilities, specifically through its Science and Technology Research Infrastructure for Discovery, Experimentation and Sustainability (STRIDES) Initiative. Last year, NIH and NIH institutions partnered with cloud-service providers Google and Amazon Web Services to support its large-scale dataset migration, which saved the agency over $5 million so far.

NIH has moved over 30 petabytes of data into the cloud since. (Gregurick mentioned that one petabyte is equivalent to storing about 4,000 photos per day over an entire person's lifetime.) Additionally, the NHLBI's Framingham Heart Study and NIA's Alzheimer's health data have already both been moved to the cloud.

"This could possibly represent the largest amount of biomedical data available for research," Gregurick noted, "yet it's still very hard to aggregate access across our platforms."

In 2020, NIH hopes to move 50 petabytes of data. This includes data from large health-science research datasets, such as the [All of Us Research Program](#) and the NCI Cancer Research Data Commons.

"I suspect that when the All of Us Research Program is reaching its full stride, we can actually see zettabytes of data in the cloud, particularly if we start integrating things from sensor data. That's an unprecedented scale of data," Gregurick said. "There aren't computational algorithms I think today that can really work at this scale."

In order to expand the access to the wide-spanning impact of research data, NIH is working to facilitate STRIDES to lower the costs for NIH-funded institutions, increase the training available for cloud computing, add STRIDES industry partners, and create common controls and protection standards.

The ultimate end goals, as Gregurick outlined, are to create seamless access to STRIDES accounts through universities and institutions; increase research data science skills; foster greater analytics and data management capabilities for NIH-funded programs; and provide assurance in confidentiality, integrity, and availability of data.

Lastly, NIH wants to leverage cloud for biomedical research because it is computing on data at an unprecedented scale with a need for increased data access, sharing and reuse of data with other researchers using a community-driven approach.

"That's the problem I really want to solve," Gregurick said. "I want to provide for you, our research community, a greater data platform that leverages all the excellent data capabilities we're already standing up."

[View printer friendly version](#)
[data](#)
[cloud migration](#)
[data science](#)
[NIH](#)
[cloud computing](#)

[Health IT Standard](#)