Company Helps Researchers Unravel DNA Mysteries Faster with Google Cloud Storage

At a Glance

What they wanted to do
- Build a secure mirrored version of the NCBI SRA repository
- Develop a web-based user-friendly interface
- Allow for scalability and future data growth

What they did
- Used Google Cloud Storage to host 350 terabytes of DNA sequencing data
- Focused resources on developing the user interface, since Google handles security, scalability and other issues

What they accomplished
- Mirrored a comprehensive cloud-based DNA archive using minimal internal resources
- Created a platform that has received praise from researchers on how easy it is to use
- Help expedite research by allowing scientists to sort through and pinpoint specific DNA sequencing data more easily

Organization

DNAnexus is a Mountain View, Calif.-based company focused on unlocking the potential of DNA-based medicine and biotechnology with a collaborative and scalable data technology platform built on the cloud. One aspect of their work is focused on making large genomic datasets broadly accessible to the research community and coupling them with sophisticated analysis tools. The company recently worked with Google Cloud Storage to develop a mirror of the National Center for Biotechnology Information's Sequence Read Archive (SRA), a public repository of DNA sequencing data from some of the world's leading research institutions. This project provides a complementary way for researchers to freely access these important data and exemplifies how cloud-based technologies are enabling completely new approaches to large-scale data access and analysis.

Challenge

"The SRA is an important resource for the research community and we felt that we could leverage our expertise in developing easy-to-use data analysis tools to help preserve and enhance its usability," says Brigitte Ganter, Ph.D., Director of Product Marketing at DNAnexus.

Along with replicating the SRA data and maintaining free access, DNAnexus was focused on improving the overall experience of using and accessing these data. To accomplish this, they needed a large-scale data storage solution capable of hosting more than 350 terabytes of data and robust enough to handle growth and large downloads. They also wanted to completely re-engineer the way that users interacted with the data, mined results and downloaded datasets of interest.

"Rather than building your own infrastructure and taking time and resources away from your company, you can use Google's infrastructure and know that it's scalable and secure."
—Brigitte Ganter, Director of Product Marketing, DNAnexus

Solution

DNAnexus began this project in June 2011. To host the massive SRA data set, the company looked to Google Cloud Storage, a service that lets companies store their data in Google's cloud. DNAnexus knew the service provided the reliability, security and vast scalability needed to support the mirrored SRA site.

About Google Cloud Storage

Google Cloud Storage allows companies to store and access their data in Google's highly scalable storage and networking infrastructure. Developers can store objects of any size and manage access to their data on an individual or group basis.

For more information visit
www.code.google.com/apis/storage

"We're able to have our site up and running 24 hours a day, seven days a week. We don't have to manage and maintain servers."
—Brigitte Ganter, Director of Product Marketing, DNAnexus

With Google's help, the team downloaded approximately 100,000 files from the NCBI SRA website and converted them into the more popular FASTQ format, which would be easier for researchers to work with than the standard SRA format. They then uploaded both versions of the files to Google Cloud Storage.

"Because Google Cloud Storage is highly scalable, we're able to provide the data in both formats, which provides tremendous value to researchers who are more familiar with using FASTQ files," Ganter says.

With Google hosting the data, the new **SRA site powered by DNAnexus** – which launched in October 2011 – has required very minimal maintenance on DNAnexus' part, Ganter adds. This has freed up the company to focus on enhancing the user interface so researchers can more easily search the database, filter results and pinpoint the specific data they are interested in investigating further. DNAnexus provides an instant online genomics data and analysis center where researchers can upload SRA datasets and access a suite of tools for performing additional analyses.

Results

Google Cloud Storage provided DNAnexus with the robust data capacity it needed to build a mirrored version of the NCBI SRA repository and removed the administrative burden of managing these large datasets.

"There's so much we take for granted with Google Cloud Storage," Ganter says. "We're able to have our site up and running 24 hours a day, seven days a week. We don't have to manage and maintain servers." Ganter says that the DNAnexus team members spend practically no time managing the system, with the exception of overseeing incremental updates.

Researchers around the world are now able to search for and access SRA datasets through an intuitive, web-based user-friendly interface. The mirrored website earned immediate praise from the research community for its ease of use; as word about the site spreads, Ganter is confident that Google Cloud Storage can accommodate any spikes in usage.

"If you have big data, you want to work with someone who truly understands the unique challenges of working on the terabyte scale," she says. "Rather than building your own infrastructure and taking time and resources away from your company, you can use Google's infrastructure and know that it's scalable and secure."

Google™