Solving the Problem of Managing Big Genomic Data

Researchers at Nationwide Children's Hospital complete a first-of-its-kind project to evaluate a large-scale genomic data management system on the scale of up to one million genomes.

The influx of genomics data resulting from the increasing affordability of whole exome/genome sequencing and President Obama's Precision Medicine Initiative requires a novel technological solution to data storage, communication with other clinical decision support systems and health information exchange. Any solutions must also enable the use of the data in secondary research studies.

Researchers at Nationwide Children's Hospital may have found the solution: Apache Hadoop, the open-source Big Data ecosystem employed by Facebook and Google to handle extremely high volumes of transactions and computational data, provides a secure, highly scalable and inexpensive method of managing massive genomics datasets.

Researchers at Nationwide Children's are the first to complete a project to evaluate a large-scale genomic data management system versus conventional methods.

"In the medical domain, we find very few applications using Hadoop," says Simon Lin, MD, MBA, chief information officer of The Research Institute at Nationwide Children's and senior researcher on the project. "We are on the frontier of utilizing this tool for genomics data and have found it to be an ideal candidate."

The project, which will be presented as a poster at the American Society of Human Genetics (ASHG) Conference on October 18, simulated 750,000 genomic records to test the ability of the system to handle such large-scale data.

"We need to look beyond current needs and think about the millions of patients' genomic data that we will need to handle in five years," says Yungui Huang, PhD, research and development director in The Research Institute's Research Information Solutions and Innovation (RISI) department at Nationwide Children's and member of the research team. "With this project, we also pushed the envelope with precision medicine to look at the usefulness of the system for both the research and clinical domains."

Using the Hadoop ecosystem, the team designed an open-source Genome Archiving and Communication System (GACS) for clinical genomics, which is able to securely interface with the medical records systems (such as EPIC), much like the system used for radiology – Picture Archiving and Communication System (PACS).

"Hadoop's suite of authentication, authorization and auditing schemes enables our system to be HIPAA compliant," says Dr. Huang. "Supporting the security of genomics data is an important consideration for us."

When fully developed, GACS will allow clinicians to query the genomic variants a patient may have beyond those reported in the PDF clinical summary from the sequencing lab.

“Commercial medical record systems are not yet capable of handling the large data sets associated with genomics,” says Dr. Lin. “Now, a physician might find a static PDF of the geneticist’s summary following a genomics study. We want that physician to have real time access to the summary and the granular data, and Hadoop enables us to do that.”

That access would enable clinicians to search a patient’s genome for a specific variant and have an answer in milliseconds. The opportunities for application are infinite.

From selecting the right medication to treat a patient’s cancer to looking at variants associated with chronic diseases, managing genomics’ Big Data in a way that is secure and accessible is the key to the promise of precision medicine coming to fruition.

A recent white paper released by the Workgroup for Electronic Data Interchange (WEDI), the nation’s leading authority on the use of health information technology to create efficiencies in health care information exchange, noted that “Genomic medicine offers the potential to greatly improve medical practice by tailoring the preventive, diagnostic and therapeutic care available to each patient,” but that this potential “depends on high-quality data that can be readily accessed and applied in the patient care setting.” It called on the health care community to work aggressively toward better access and seamless integration of genomic data for the benefit of the patient. The GACS implemented at Nationwide Children’s is an important step toward that goal.

“We are preparing for a future — one that is getting closer every day — where each patient will have a complete genetic profile linked to their medical health record,” says Grant M. Wood of Intermountain Healthcare’s Clinical Genetics Institute and chair of the WEDI Genomics Work Group. “We need to access these very large datasets not only for continued research and discovery but also for use by intelligent clinical systems, which will help guide healthcare providers with the accurate interpretation of the genomic data for targeted care for their patients. The work by Nationwide Children’s will make this possible by solving the data storage scalability issue.”

The project at Nationwide Children's is supported by institutional investment in Big Data and through a gift from Alliance Data to support the development and application of Big Data architecture.

For more information about the project and the GACS, please visit http://www.nationwidechildrens.org/GACS.

Citation:

Swaminathan R, Huang Y, Yu E, Fitch J, Lintner K, White P, Lin S. A Scalable and Secure Genome Archiving at Communication System for the Clinical Enterprise. Abstract presented at ASHG, 18 Oct 2016.