In Brief

Although researchers are able to collect, store, and analyze tremendous volumes of data, technological and storage limitations can severely impact the speed at which these data can be analyzed. A new algorithm developed by Microsoft Research, FaST-LMM, runs on Windows Azure in the cloud and expedites analysis time—reducing processing periods from years to just days or hours. An early application of FaST-LMM and Windows Azure helps researchers analyze data for the genetic causes of common diseases.

David Heckerman
Distinguished Scientist
Microsoft Research

Robert Davidson
Principal Software Architect
Microsoft Research,
eScience

Jeff Baxter
Development Lead
Windows HPC, Microsoft

Jennifer Listgarten
Researcher
Microsoft Research
Connections

Christoph Lippert
Researcher
Microsoft Research
Connections

Websites:

research.microsoft.com/
cloudresearch

windowsazure.com

FaST-LMM and Windows Azure Put Genetics Research on Faster Track

The ability to store and analyze tremendous volumes of data is critical to researchers around the world. Although technological advances enable researchers to store and process ever-larger data stores, there are still challenges. CPU and storage limitations can hinder researchers, particularly those dealing with massive volumes of information, or "big data."

"With this new, huge amount of data that's coming online, we're now able to find connections between our DNA and who we are that we could never find before."

David Heckerman
Distinguished Scientist
Microsoft Research

Scientists typically have used supercomputers to handle big data research, but CPU and storage limitations have restricted the potential speed and breadth of data processing that can be done at any given time. Microsoft Research has developed an alternative method for processing big data: the FaST-LMM (Factored Spectrally Transformed Linear Mixed Models) algorithm running on Windows Azure.

FaST-LMM is a high-speed algorithm that is capable of analyzing data at an accelerated rate (when compared to supercomputers). It is designed to process big data projects that are comprised of massive volumes of data, such as those available to scientists conducting genome-wide association studies.

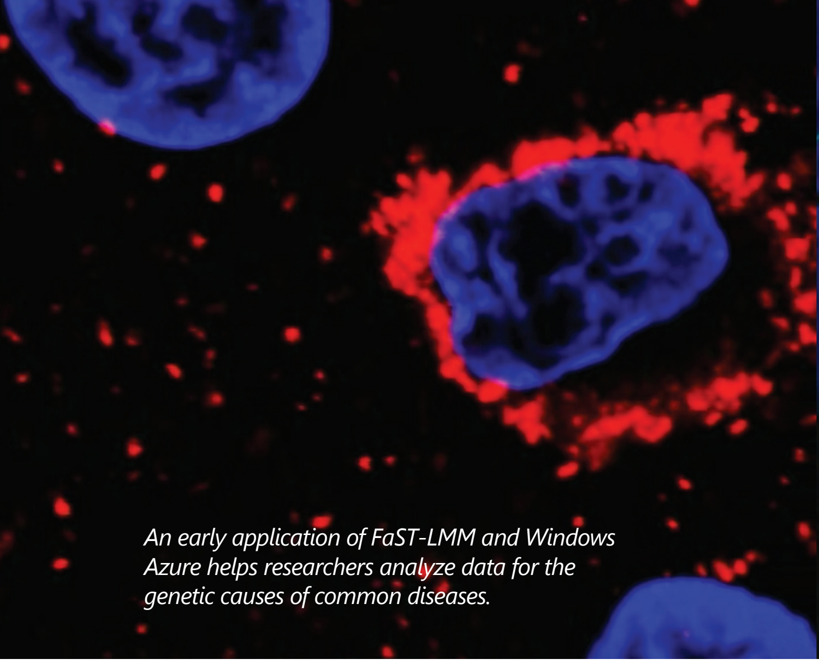
"When I hear [the term] 'big data,' I think of hundreds of thousands of individuals—and the DNA for those individuals and the intricate algorithms we need to process all that data," remarks David Heckerman, distinguished scientist at Microsoft Research. Today, the process of measuring the sequence of an individual's genome is inexpensive and relatively simple, yet analyzing the data is arduous and costly. FaST-LMM was developed to assist with that part of the process.

The FaST-LMM algorithm is capable of analyzing massive volumes of data in less time than existing alternatives. Microsoft Research also has the infrastructure required to perform the computations, Heckerman states. Specifically, Microsoft Research has access to server farms with thousands of compute nodes. With more CPUs dedicated to a job, the computations complete more quickly. Now, computations that would ordinarily take years to finish can be completed in just hours.

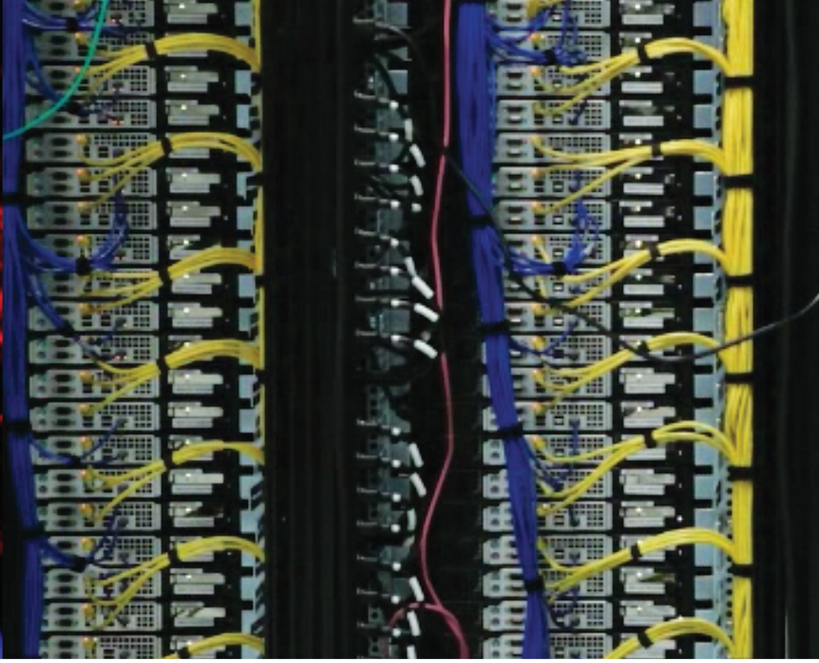
Microsoft Research's role does not end with the development of algorithms, Heckerman explains. The team is also applying those algorithms to various data sets provided by collaborators—including the Wellcome Trust in Cambridge, England.

SEARCHING FOR DNA CLUES IN SINGLE NUCLEOTIDE POLYMOPHISM PAIRS

The Wellcome Trust database contains genetic data from 2,000 individuals and a shared set of approximately 13,000 controls for each of seven major diseases: bipolar disease, coronary artery disease, hypertension, inflammatory bowel disease (Crohn's disease), rheumatoid arthritis, and types I and II diabetes. "We're



An early application of FaST-LMM and Windows Azure helps researchers analyze data for the genetic causes of common diseases.



looking at the genetic causes of common diseases such as diabetes and hypertension," says Jennifer Listgarten, researcher, Microsoft Research Connections. The project involves searching for combinations of genomic information that interact to make a person more or less susceptible to a certain phenotype.

"We look at SNP pairs—a SNP being a single nucleotide polymorphism," explains Robert Davidson, principal software architect in the Microsoft Research eScience group. Each SNP is a genetic marker residing on a specific chromosome within an individual's genome. The Wellcome Trust data also indicates if the individual has a given disease. The Microsoft Research team believes that they will gain better insight into an individual's propensity to develop common diseases by reviewing the SNPs in pairs, rather than individually.

The SNPs are being stored in the Windows Azure cloud instead of traditional hardware storage—a profound shift in how big data are stored. "We are taking on the challenge of taking what would be traditional high-performance computing, one of the hardest workloads to move to the cloud, and moving to the cloud," observes Jeff Baxter, development lead in the Windows HPC team at Microsoft. "There's a variety of both technical and business challenges, which makes it exciting and interesting."

EXPLORING THE POWER OF THE CLOUD

Windows Azure is typically used for website storage—a mere fraction of the potential storage that is required for a big data project. The product team had been seeking ways to demonstrate the true power of cloud storage: "We were looking for a suitable workload as an exemplar," Baxter says. "After looking around the company, [the Microsoft Research project] jumped out."

Resource management is one of the primary issues associated with big data: not only determining how many resources are required for the project, but also identifying the right type of resources—within the available budget. For example, running a large project on fewer machines might save on hardware costs but result in substantial project delays. Researchers must find a balance that will keep their project on track while working with available resources.

For the Wellcome Trust project, the team's available resources for the project included a combination of Windows HPC Server, Windows Azure, and the FaST-LMM algorithm. The team knew that they had a powerful set of technologies. The question was, could it achieve the results required in the desired timeline?

"For this project, we would need to do about 125 compute years of work. We wanted to get that work done in about three days," explains Baxter. By running FaST-LMM on Windows Azure, the team had access to tens of thousands of computer cores and an improved algorithm that was able to expedite the work. "You're still doing hundreds of compute years of work," he explains, "but with these resources, we can actually do hundreds of compute years in a couple of days."

"We are taking on the challenge of taking what would be traditional high performance computing, one of the hardest workloads to move to the cloud, and moving it to the cloud."

Jeff Baxter
Development Lead
Windows HPC, Microsoft

While the results were impressive, there was something that had an even bigger impact. "The most impressive thing was how quickly we could take this project from inception to actually completing it and generating new science," Baxter notes. "This is stuff that, without both the improvements in the algorithms that the Microsoft Research guys had come up with and the ability for us to provide the tens to hundreds of thousands of cores, would have been infeasible."

The Wellcome Trust project is just the beginning of what could be a major shift in how research is stored and analyzed. "With this new, huge amount of data that's coming online, we're now able to find connections between our DNA and who we are that we could never find before," Heckerman says. The ability to analyze that data more quickly, and with greater depth, could result in *faster* breakthroughs in genetic research—and breakthroughs in *critical* genetic research.

"Just about every day, you turn on the radio and you hear about some new discovery of a gene that's associated with some disease, or a gene that's associated with some favorable reaction to a drug," Christoph Lippert, researcher, Microsoft Research Connections remarks. The FaST-LMM algorithm running on Windows Azure is helping to accelerate just such discoveries.