Microsoft

Unlock insights on any data
Customer Solution Case Study

VirginiaTech

University Transforms Life Sciences Research with Big Data Solution in the Cloud

Overview

Customer: Virginia Polytechnic Institute and State University (Virginia Tech)
Website: www.vt.edu
Customer Size: 7,035 employees
Country: United States
Industry: Education—Universities

Customer Profile

Virginia Tech is one of the country's leading research institutions. The university manages a research portfolio of US$454 million.

Business Situation

As the volume of genome sequencing data has grown exponentially, many researchers lack the storage and processing resources to analyze this data.

Solution

Virginia Tech developed an on-demand, cloud-computing model using the Windows Azure HDInsight Service to reduce costs and improve access to DNA sequencing tools and resources.

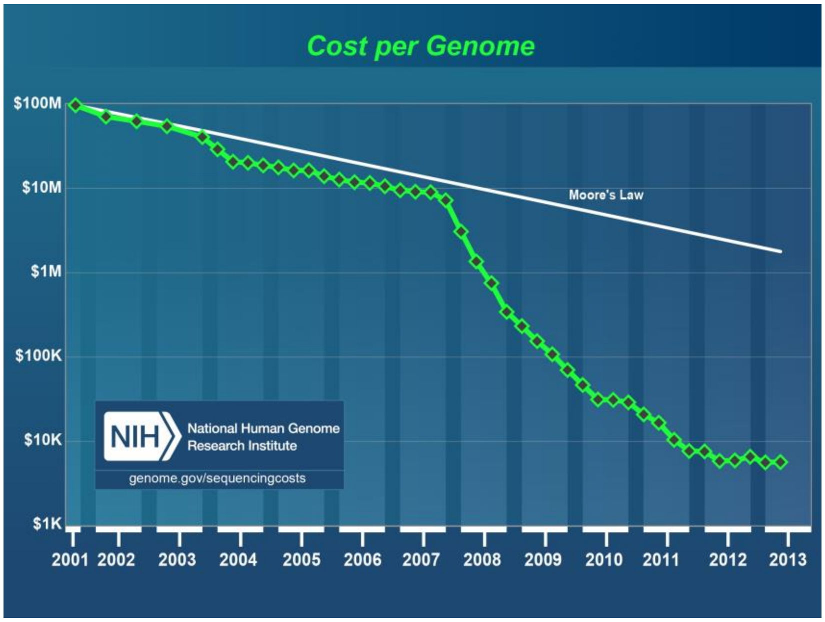
Benefits

- Provides significant cost savings
- Enables collaborative analysis
- Supports research anytime, anywhere

> "Windows Azure is enabling us to keep up with the data deluge in the DNA sequencing space. We're not only analyzing data faster, but analyzing it more intelligently."
>
> Wu Feng, Professor of Computer Science, Virginia Tech

DNA sequencing analysis is a form of life sciences research that has the potential to lead to a wide range of medical and pharmaceutical breakthroughs. However, this type of analysis requires supercomputing resources and Big Data storage that many researchers lack. Working through a grant provided by the National Science Foundation in partnership with Microsoft, a team of computer scientists at Virginia Tech addressed this challenge by developing an on-demand, cloud-computing model using the Windows Azure HDInsight Service. By moving to an on-demand cloud computing model, researchers will now have easier, more cost-effective access to DNA sequencing tools and resources, which could lead to even faster, more exciting advancements in medical research.

Although the cost to compute on a genome continues to drop exponentially, the cost to sequence a genome has dropped even more dramatically—from $95 million in 2001 to $5,700 in 2013. This stunning decline in cost is being driven by the arrival of next-generation sequencers (NGS), which produce 15 petabytes of data for analysis each year.



Situation

The Virginia Bioinformatics Institute and the Department of Computer Science at Virginia Tech began using a network of supercomputers to locate undetected genes in a massive genome database. This and related work by other institutions has the potential to lead to exciting medical breakthroughs, including new cancer therapies and antibiotics used to combat the emergence of drug-resistant bugs.

However, as the size of genome databases grows, so has the challenge of analyzing them. And with the advent of next-generation sequencers (NGS), this growth has been exponential. “Of the estimated 2,000 DNA sequencers worldwide, they are generating 15 petabytes of genome data every year,” explains Wu Feng, Professor of Computer Science at Virginia Tech. Many life sciences institutions simply do not have access to the computational and storage resources required to work with data sets of this size. In other words, says Feng, “We’re generating data faster than we can analyze it.”

The team had already recognized the potential of high-performance cloud computing to address the resource challenge. But now they wanted to develop software that would make it even easier for scientists to take advantage of these cloud resources, which would lead to faster genome analysis. And that’s how Feng was introduced to the potential of the Windows Azure HDInsight Service running on the Windows Azure platform.

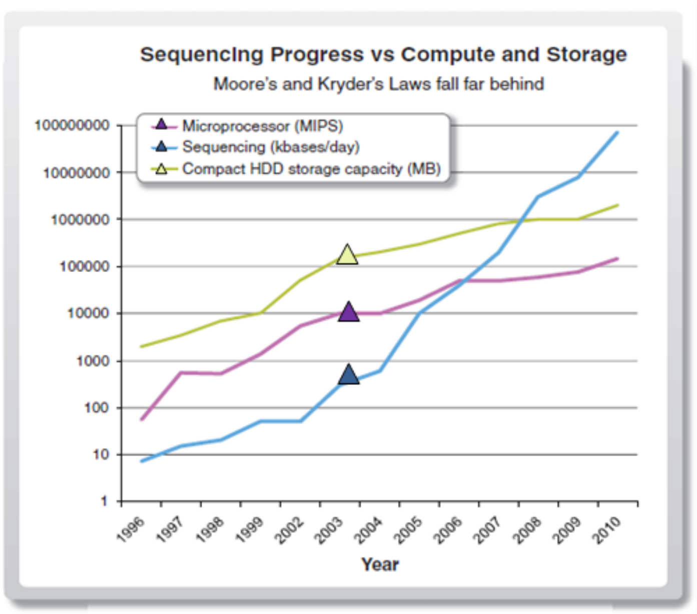
Feng’s team was one of only 13 from across the country selected by a research program called Computing in the Cloud. Run by the National Science Foundation in partnership with Microsoft, the program was designed to accelerate access to cloud computing for research discovery, data analysis, and multidisciplinary collaboration. Based on the potential of their proposal, Feng’s team was awarded both a grant that covered the cost of using the Windows Azure platform and its supporting technical resources.

Feng had looked at an alternative cloud service on which to do the work but found that it did not meet the requirements needed for the team’s development efforts. This included resource and support levels that simply weren’t as robust as the Microsoft offering. For Feng and his team, Windows Azure provided an ideal combination of infrastructure and technical support “to conduct the research and development necessary to facilitate personalized genomics for the broader research community.”

Solution

Since being awarded the grant, Feng and his team have developed two software artifacts: SeqInCloud, a popular genetic variant pipeline called the Genome Analysis Toolkit (GATK), and CloudFlow, a workflow management framework that uses both client and cloud resources.

SeqInCloud (short for "sequencing in the cloud") is based on the Broad Institute's Genome Analysis Toolkit (GATK), a toolkit for analyzing next-generation sequencing



The rate at which genome sequences are being generated is increasing faster than researchers' ability to analyze and store them. (Source: Scott D. Kahn, "On the Future of Genomic Data," Science, 331, 728 (2011)

data, with the main focus on variant discovery and genotyping.

SeqInCloud seamlessly generalizes the GATK pipeline, allowing it to run in the cloud using HDInsight and Windows Azure in order to maximize portability. The SeqInCloud application also features a novel design strategy for data partitioning, data transfer, and storage optimization on Windows Azure. The result is more efficient use of Azure cloud resources and better performance overall.

CloudFlow is a workflow management framework that can be installed on a researcher's PC to help interactions with the Windows Azure HDInsight Service. As Feng explains, "It allows us to compose flexible MapReduce pipelines that simultaneously utilize both client and cloud resources for running the pipeline and automating data transfers. This is where the HDInsight resource has been particularly useful." To run large tasks, researchers can automatically provision HDInsight clusters on demand.

The CloudFlow framework delivers unique features that are not offered by existing MapReduce-based workflow managers, including enabling the simultaneous use of client and cloud resources, automatic data-dependency handling between client and cloud resources, and the flexibility of implementing user-defined plugins for data transformations.

Benefits

By taking advantage of the Windows Azure platform, Feng and his team showed how well the Windows Azure HDInsight Service can be used seamlessly to deliver cloud applications with advanced capabilities.

By making the Windows Azure HDInsight Service more effective and accessible for DNA sequencing researchers, the project has produced several key benefits.

Provides Significant Cost Savings

The cloud computing solution developed by Feng's team can address growing resource issues that come with analysis of genome sequencing data. As Feng notes, "Life scientists and their institutions no longer have to find millions of dollars to establish their own supercomputing center. Rather than incur the cost of housing their own data center resources and create their own provisioning and scheduling policies, this is done for them through the Windows Azure ecosystem."

Because the data persists in Azure blob stores independently of HDInsight, additional costs savings are realized by only paying for the compute power of the HDInsight clusters for the duration of their actual use, all without losing data.

Supports Collaborative Analysis Anytime, Anywhere

Feng also notes the value of the Azure cloud platform as an effective collaborative tool. "The model enables the easy sharing of public data sets and helps to facilitate large-scale collaborative research."

For More Information

For more information about Microsoft products and services, call the Microsoft Sales Information Center at (800) 426-9400. In Canada, call the Microsoft Canada Information Centre at (877) 568-2495. Customers in the United States and Canada who are deaf or hard-of-hearing can reach Microsoft text telephone (TTY/TDD) services at (800) 892-5234. Outside the 50 United States and Canada, please contact your local Microsoft subsidiary. To access information using the World Wide Web, go to:
www.microsoft.com

For more information about Virginia Tech, visit the website at:
www.vt.edu

To learn more about the DNA sequencing program described in this case study, please refer to: N. Mohamed, H. Lin, and W. Feng, "Accelerating Data-Intensive Genome Analysis in the Cloud," in Proceedings of the 5th International Conference on Bioinformatics and Computational Biology (Honolulu, Hawaii, March 2013), pp. 297–304
http://synergy.cs.vt.edu/pubs/papers/nabeel-bicob13-genome-analysis-cloud.pdf

This case study is for informational purposes only.

And because the applications can be accessed from virtually anywhere, including on mobile devices, Feng sees an opportunity not far in the future when researchers will be able to engage in genome analysis outside the laboratory, "say at a hospital, which could lead to faster, prescribed treatments."

Even in these early stages of development, the benefits of the solution are quickly being recognized by other top research institutions across the country. For example, says Feng, "The solution has already generated interest from the University of California at Berkeley and here at the Virginia Bioinformatics Institute."

At the same time, Feng's team continue to expand their research using the Windows Azure HDInsight Service. "Windows Azure is enabling us to keep up with the data deluge in the DNA sequencing space," says Feng. "We're not only analyzing data faster, but analyzing it more intelligently."

Unlock insights on any data

Microsoft business intelligence (BI) solutions simplify access to virtually any type of data, whether it resides in the business or the cloud. Powered by Microsoft SQL Server, and built into familiar programs such as Microsoft Excel, BI tools speed insight into data from multiple sources, including business applications, blogs, and sensors.

For more information about unlocking insights on any data, go to:
http://www.microsoft.com/en-us/server-cloud/cloud-os/data-insights.aspx#fbid=I-JoQIGZin5

Software and Services

- Windows Azure platform
  - Windows Azure
  - Windows Azure HDInsight Service