Microsoft

Azure, Asteroids, and AI: How academic researchers are using Microsoft Azure to transform their research

Universiteit Leiden
The Netherlands

Customer
Leiden University

Products and Services
Azure
Azure Batch

Industry
Higher Education

Organization Size
Large (1,000 - 9,999 employees)

Country
Netherlands

March 24, 2020

Microsoft

After studying physics as an undergraduate at Texas Tech University, John Hefele decided to pursue a Master's degree in astronomy at Leiden University in the Netherlands. Several months into the program, he began an internship with Microsoft in Amsterdam, where he ran a Python-based simulation program in Azure to aid in identifying asteroids that could potentially hit Earth. Using Azure Batch, he was able to generate a million asteroid trajectories, in nine hours, at a cost of $150.

View video:
https://play.vidyard.com/hbzuCxNgs6ELnbwid8Lsnq.jpg

The 6 terabytes of data generated was used to train a neural network to identify known asteroids that match the trajectories of simulated asteroids that pose potential danger for Earth. His neural network identified most asteroids already classified by NASA as dangerous, as well as 11 additional asteroids that pose some considerable risk of impacting Earth.

The quest to identify asteroids that pose a potential danger for planet Earth is no longer science fiction. It is scientific fact. A recent student researcher in the Netherlands was able to shed more light on the process by simulating the trajectories of millions of asteroids to help identify those that have a high probability of hitting Earth—sometime in the next thousand years.

Hefele, a Master's degree candidate in astronomy at Leiden University in the Netherlands, earned his undergraduate degree in physics at Texas Tech. In graduate school, he wanted to explore *exoplanets*—planets that revolve around stars beyond our solar system—and astronomical instrumentation. He chose Leiden University because it offers a strong Master's program in those areas. In addition, Leiden Observatory, the oldest university observatory in operation today, conducts world-class research on a wide range of

astronomical topics, and is helping to build new instruments to find and characterize exoplanets.

In his early months at Leiden, Hefele began his research into using deep learning to determine which asteroids could potentially hit Earth. He started running simulations using the Python-based Astrophysical Multipurpose Software Environment (AMUSE) framework. Designed to run on Linux, AMUSE was created at Leiden Observatory by doctorate and post-doctorate students under the supervision of Simon Portegies Zwart, Professor of Numerical Star Dynamics at Leiden University, and Hefele's former supervisor.

"For this project to work, I needed lots of examples of dangerous asteroids that I knew would collide with Earth," says Hefele. "I needed to create around a million examples, which would take a tremendous amount of time on my laptop or even on a supercomputer." One day early in his tenure at Leiden, Hefele saw a Microsoft ad on LinkedIn looking for students interested in data science internships. Hefele applied and ultimately got the six-month internship, with the blessing of the university.

Exploring the potential of cloud computing

The first step was for Hefele to decide what to focus on during his internship. "Cloud computing seemed like it had a lot of potential for my research, and I knew next to nothing about it at the time," Hefele says. "I did know about running astronomical simulations on Linux-based machines. And I was surprised and excited to learn that I could create Linux virtual machines [VMs] in Azure. That is important because the vast majority of software for astronomy is written for Linux machines."

As he learned more about the cloud, Hefele discovered Microsoft Azure Batch can run batch and high-performance computing (HPC) applications in the cloud, scaling from tens to thousands of virtual machines. Batch is especially useful for running thousands of parallel

operations, which made it ideal for Hefele's asteroid trajectory calculations. "Azure Batch allows you to spin up a bunch of virtual machines, simulate thousands of trajectories at once, and then spin down the machines," says Hefele.

Because he would be using AMUSE to run the simulations and it was written for on-premises use, Hefele had to figure out how to port it to the cloud. "AMUSE was designed to run on any Linux machine so it was really easy to adapt it to run on Linux VMs in Azure," says Hefele.

He read documentation and found an online tutorial that explained how to run Python jobs with Batch. He made a few tweaks to make it work for his project. "I just modified the process in the tutorial and created a script that downloads AMUSE, runs the asteroid trajectory simulations, and then shuts itself off," says Hefele.

Generating a million simulations for $150

It is estimated that there are approximately 850,000 known asteroids in our solar system and potentially millions more, so Hefele wanted to generate up to a million simulated asteroids. "For a neural network to work really well, you need a lot of training data," explains Hefele. The simulated asteroids are launched from hundreds of thousands of points on the Earth's surface and then reversed so that they can be tracked on the journey back down. These simulated trajectories are stored in Microsoft Azure Storage and then fed into a neural network. Artificial intelligence (AI) and deep-learning tools are used to compare the simulated trajectories to the trajectories of known asteroids in the NASA database to train the neural network to identify those that are potentially dangerous.

Hefele ran his first batch of simulations on 20 Linux VMs in Azure and then scaled up to 100, 200, and ultimately 500 VMs. He estimates it took only about 15 minutes to spin up 200 machines.

which generated a million asteroid trajectories in about nine hours, at a cost of only $150.

“A lot of people in academia assume that cloud computing is going to be really expensive, but I was able to generate a million simulations using Azure Batch for just $150,” says Hefele. “When you compare that to the cost of purchasing a large supercomputer, maintaining it, and having it outdated in five years, it makes the cloud seem like a more economic option.”

After he generated the first batch of asteroids, Hefele realized there was a small error. The asteroids were launched in a manner that created an uneven distribution. After changing the launch spacing to be over a few Jupiter periods, this problem was fixed. “It was no problem to correct it and run it again,” says Hefele. "It just took another nine hours. If I wasn’t running it on the cloud, it would have been another two months.”

Hefele got a grant to cover the cloud costs, and the economy of running in the cloud made that grant go a lot further than he originally expected. Thus, he was able to generate a larger volume of training data than initially planned, which greatly improved his results. “If I hadn’t been able to generate this huge amount of training data on the cloud, my neural network wouldn’t have performed as well, and I wouldn’t have been able to achieve the results that I did.”

Saving weeks, months of effort

As low as that cost was, the time saved was even more remarkable. “These calculations require lots of computer cores to run at optimum speed,” notes Hefele. “Normally a supercomputer has at most 100 cores, but in the cloud, you have access to a virtually unlimited number of cores. I used up to 1,000 cores, so I was able to do a lot of things in parallel at once. On my laptop, I would probably have to run simulations for two or three months to get a million examples. If I had full access to the university supercomputer it would probably

take me three to four weeks. Using AMUSE with Azure Batch, I was able to generate a million asteroid trajectories in a day."

The time-savings also contributes to the low cost. "The virtual machines were costing money only while they were running computations," says Hefele. "As soon as they finished computing the simulated trajectories, they shut themselves down, and the charges stopped as well."

Hefele also notes that accessibility is a big advantage of using the cloud. "If you're a student here at Leiden, you have unlimited access to the supercomputer, but you're competing with a lot of other people," says Hefele. "You sometimes log on and find that all the GPUs are being used, so you get put in a queue. With the cloud, you're never put in a queue. I could spin up my virtual machines in Azure whenever and wherever I wanted."

One of the most exciting outcomes of Hefele's research was the identification of new, potentially problematic asteroids, which he is documenting in a paper that will be published in late 2018. "By creating all the training data in the cloud and using it to train my neural network, I was able to identify the vast majority of asteroids that we already know are dangerous through the NASA database;" says Hefele. "Even more importantly, I found about 200 to 300 other asteroids that hadn't been identified as dangerous until now."

The ability to detect potentially dangerous asteroids as much as 100 years ahead of time can literally change the course of history. It gives scientists more time to respond, which means they have more options for trying to slow down or adjust the asteroids' trajectories to steer clear of Earth.