2017 NSF BIGDATA grant sponsored by Columbia University builds on Google Cloud Platform to solve problems in climate science

Pangeo, an open source data toolkit, helps researchers store and analyze the huge complex datasets of climate sciences. Using Google Cloud Platform, Pangeo makes it possible for researchers to make progress faster by running parallel computations on a shared infrastructure that is intuitive, scalable, and cost-effective.

In the field of climate science, researchers tackle some of the most pressing problems affecting the world today. With droughts, floods, and hurricanes in the international news, predicting natural disasters and assessing the state of the earth is more and more urgent.

> “
>
> The ability to compute scalably on so much data will allow us to ask questions that we cannot ask now. There are going to be major scientific discoveries enabled by this.
>
> **Ryan Abernathey**, Assistant Professor of Earth and Environmental Sciences, Ocean and Climate Physics, Columbia University

A crisis in data management — and a solution

While some scientists conduct experiments in labs, climate scientists use complex Earth System Modeling (ESM) simulations to understand patterns and make predictions. These simulations draw on the ever-growing accumulation of environmental data from satellites, drones, and sensors monitoring the continents, oceans, and atmosphere.

As the simulations generate ever more data, researchers find themselves in a crisis: the better the models become, the harder they are to use. “We are drowning in data,” says Ryan Abernathey, an oceanography professor at Columbia University’s Lamont-Doherty Earth Observatory. “To do science you need to be able to iterate fast on the data. Try something, get your results back, see if it works. try something new. We’re so far from being able to do that in the old school way of working.”

This problem inspired Pangeo, an open-source data platform based on Python tools that provides a flexible, scalable infrastructure for research in climate science. Existing systems for managing data on high-performance computing clusters were designed for text and tabular data, not the high-resolution three-dimensional datasets that climate scientists analyze. In September 2017 a team from Columbia, the National Center for Atmospheric Research (NCAR), and Anaconda, a private data sciences firm, won a three-year $1.2 million grant through the National Science Foundation (NSF) program Earthcube to develop and pilot Pangeo. Now end users can log onto the system via the web through Jupyter Notebook, design their data analysis through X-Array (a user-friendly wrapper for Common Data Model metadata), and distribute the computation through DASK (an open-source system for parallel computing built by Anaconda). Abernathey, the project’s lead investigator, says he hopes the tools will allow him to work more efficiently with three-dimensional simulations of ocean eddies to predict ocean uptake of heat and carbon, which impacts climate change.

Deploying on Google Cloud Platform

To manage the datasets economically and efficiently, the Pangeo team turned to Google Cloud Platform (GCP), and in particular Google Cloud Storage and Kubernetes. Since GCP has a well-supported, certified, and fully-managed Kubernetes service, Google Kubernetes Engine, transitioning to GCP was an easy choice. With a grant of $100K in GCP credits over the three years of the NSF grant, the team was able to launch a prototype quickly. The project is only a few months old but Abernathey says he is excited by the soft launch so far, with hundreds of users logging on, ten thousand hours of compute, and 100 terabytes of stored data coming soon. As a bonus, Google matches 100% of the energy consumed by its data centers with renewable energy purchases. By running Pangeo on GCP, the net carbon emissions are zero, helping to reduce the relative impact on the environment.

What’s next?

Abernathey and the Pangeo team plan to make the output more easily customizable for each user, optimize the data storage, and finish the conversion from legacy platforms. The next ESM dataset of 100 petabytes is set to be released over the next few years, and Abernathey wants Pangeo to be ready: “Working on the cloud, I do an analysis and I can immediately pass it on to you, you can build on it, extend on it. The ability to compute scalably on so much data will allow us to ask questions that we cannot ask now. There are going to be major scientific discoveries enabled by this.” Previously data scientists had to act as systems administrators for every new proprietary software program; with cloud computing they can focus on their research. The big picture, Abernathey points out, is how the cloud is revolutionizing scientific research: “This is the future of what day-to-day science research computing will look like.”

ORGANIZATION PROFILE

One of the oldest universities in the nation, Columbia University is a leading research and teaching institution based in the Morningside Heights neighborhood of New York City. It serves over 32,000 students in eighteen graduate and undergraduate schools and programs. In 1949 it established the Lamont-Doherty Earth Observatory for interdisciplinary research in the earth sciences.

PRODUCTS USED

Google Cloud Platform, Cloud Storage, Kubernetes