ILLUSTRATION BY THE PROJECT TWINS

ASTRONOMY TAKES TO THE CLOUD

Telescope users reveal six lessons of migrating big data from custom servers to the cloud. **By Charles Q. Choi**

Astronomers typically work by asking observatories for time on a telescope and downloading the resulting data. But as the amount of data that telescopes produce grows, well, astronomically, old methods can't keep pace.

The Vera C. Rubin Observatory in Chile is geared up to collect 20 terabytes per night as part of its 10-year Legacy Survey of Space and Time (LSST), once it becomes operational in 2022. That's as much as the Sloan Digital Sky Survey – which created the most detailed 3D maps of the Universe so far – collected in total between 2000 and 2010. The Square Kilometre Array, which is set to become the world's largest radio telescope, with sites in Australia and South Africa, will generate 100 times that amount – up to 2 petabytes daily – when it goes online in 2028. And the next-generation Very Large Array (ngVLA) will generate hundreds of petabytes annually, roughly 1,000 times more than the VLA does today, says Brian Glendenning, assistant director for data management and software at the National Radio Astronomy Observatory, who is based in Albuquerque, New Mexico.

Such data sets are out of reach for conventional workflows: it isn't feasible to download that much data and store them locally, says Mario Juric, an astronomer at the University of Washington in Seattle. Building and maintaining local computing resources to cope is similarly impractical. William O'Mullane, project manager for LSST data management in Tucson, Arizona, estimates that the cost of developing the computing infrastructure and personnel needed to run the LSST project from scratch could approach US$150 million over 10 years. Instead, they – like much of the rest of the astronomy community – looked to the cloud. Here are six lessons from astronomers' experiences.

Invest in computing power

It's not enough to migrate data to the cloud; researchers need to be able to interact with them. "Instead of the traditional model where astronomers brought their data to their computer, we want them to upload their code to the data" and perform analyses remotely, says Frossie Economou, who manages the Rubin Observatory's science platform.

The LSST, for instance, will provide free online access to its science platform – a collection of Jupyter computational notebooks,

© 2020 Springer Nature Limited. All rights reserved.

web portals and application programming interfaces (APIs) for data analysis, browsing and retrieval, says Leanne Guy, LSST data management scientist at the Rubin Observatory in Tucson. Using a web browser, the LSST's users will be able to write and run code in the Python programming language to analyse the entire LSST data set remotely on servers hosted at the National Center for Supercomputing Applications in Urbana, Illinois, rather than downloading the data to their own computer.

Other disciplines have also found success with this approach. The Pangeo project, for instance, which is a platform for analysing big geosciences data, has partnered with Google Cloud to make petabytes of climate data publicly available and computable – which makes it easier for researchers to collaborate, scale and reproduce their work, says climate scientist Joe Hamman at the National Center for Atmospheric Research in Boulder, Colorado.

No big data? No problem

"There are definitely times when projects involving just mid-size data can see a lot of benefits from cloud computing," says Ivelina Momcheva, a mission scientist at the Space Telescope Science Institute in Baltimore, Maryland. Researchers can access computational resources that greatly exceed those of their laptops for relatively little cost, Momcheva notes. And some cloud providers offer free computing resources for educational purposes.

In 2015, when Momcheva and her colleagues had only an 8-core server available for their 3D-HST project, which analysed data from the Hubble Space Telescope to better understand the forces that shape galaxies in the distant Universe. So they turned to Amazon Web Services (AWS) instead. They ended up renting five 32-core machines. "Our back-of-the-envelope calculations suggested everything we had to do would have taken us three months on our machines," Momcheva says. "With a cloud provider, it took us five days and less than $1,000."

Price isn't everything

Whether commercial cloud services are cheaper than a researcher's local data centre remains debatable. The US Department of Energy's 2011 Magellan report on cloud computing found that the department's computing centres were typically 3–7 times less expensive than commercial cloud providers. The cost difference remains roughly the same today. Yet with code optimizations, researchers can reduce those differences. According to estimates from the University of Washington, for instance, cloud-based processes that cost $43 per experiment cost only $6 after a few months of optimization, Juric says. Executing the same tasks in comparable times using a local data centre would have cost the team roughly $75,000 in hardware, electricity and personnel, he estimates, and the servers would have had to be active 87% of the time over three years. That level of usage is "highly unlikely", he says.

Time savings can influence decision-making. "If it takes nine months to process your analyses at your data centre but one month on the cloud for the same cost, that difference of eight months becomes very interesting," Juric notes.

The choice isn't necessarily binary. Projects can use local data centres for routine storage and computing but supplement those resources through 'cloud bursting' when demand spikes and an extra boost of computing power is needed, says O'Mullane.

> "Cloud computing allows even small institutions to make big discoveries."

In the meantime, funding agencies might be able to help researchers negotiate better rates, says Philip Bourne, dean of data science at the University of Virginia in Charlottesville. The US National Institutes of Health (NIH) does this with its Science and Technology Research Infrastructure for Discovery, Experimentation and Sustainability (STRIDES) Initiative – which uses cloud resources to streamline NIH data.

"With STRIDES, if an institution commits a given amount of their dollars on grants, then the Googles and Microsofts and Amazons of the world can compete with each other so investigators can get the best deals with vendors," Bourne says. Since its inception in 2018, STRIDES has helped researchers with more than 225 projects exceeding 20 million total computing hours, and saved roughly $6 million, says Susan Gregurick, who leads the NIH's strategic plan for data science in Bethesda, Maryland.

Consolidate data

By bringing multiple data sets together, cloud computing can reveal insights that might not be apparent from each data set alone. "Astronomical data becomes exponentially more useful the more there is of it in one place," Momcheva says.

Inspired by the NIH's Data Commons, a pilot project in which researchers store and share biomedical and behavioural data and software, Juric and others have requested funding to build an Astronomy Data Commons to co-locate astronomical data sets and tools in the cloud. The hope is to "eliminate the infrastructure and software barrier to entry for big-data analysis", Juric says. He and his colleagues have already released one data set called the Zwicky Transient Facility, which encompasses 100 billion observations of some 2 billion celestial objects; if they can show that their work yields benefits, other projects might follow suit.

The impact, Juric says, could be akin to that of Google's release of Google Maps and its accompanying API: the company "allowed a whole new ecosystem of apps to develop that we didn't even know were possible".

Train and train again

To set up a cloud project, users need to create an account with a cloud provider, choose from a bewildering array of options and install their software, usually tweaking it so that it can run on many machines at once. Mistakes can be costly, warns Bourne. "Inexperienced grad students have inadvertently burnt thousands of CPU hours with runaway processes," he says, referring to computer tasks that never complete owing to coding errors.

To avoid that, Bruce Berriman, a senior scientist at the Infrared Processing and Analysis Center at the California Institute of Technology in Pasadena, advises users to train beforehand, for instance by running small-scale pilot projects using local machines or academic clouds. In the cloud, he says, "the meter is always running".

Don't neglect security considerations, Juric adds. Although privacy and security in the cloud surpass those of local resources, configuring cloud resources can be tricky, and a mistake by an inexperienced programmer can expose your data to the world. "Private data centres tend to be more 'institutionally locked down'," Juric notes, whereas a commercial provider might allow such a mistake to proceed unimpeded.

Focus on outreach

By providing computing resources for little or no cost, cloud computing allows even small institutions to make big discoveries. "I could set up a notebook in South Africa to run on the LSST Science Platform that had all the same tools as if I was in Princeton," O'Mullane says. "All I'd need is a web browser."

But to do so effectively requires education, says Dara Norman, a research astronomer at the National Optical-Infrared Astronomy Research Laboratory in Tucson. Good starting points include Cloud Computing for Science and Engineering (go.nature.com/338hdpt), which Berriman calls "the best practical guide to getting started on the cloud". There is also the experimental MAST Labs project run by the Mikulski Archive for Space Telescopes (MAST), which has sample notebooks for accessing MAST data on AWS (go.nature.com/314gxyo); and an AWS tutorial using Hubble data (go.nature.com/33bad0a).

Work with researchers at smaller institutions to make sure proposed ideas will actually work for them, Norman says. And encourage networking with their students. "If a lot of really good students that you want in your institution are coming from smaller, underserved places, it's to your benefit to involve them in your research to have them ready to do this work in grad school."

Charles Q. Choi is a freelance writer in New York City.

© 2020 Springer Nature Limited. All rights reserved.

Correction

This Technology feature erroneously affiliated Dara Norman with the National Optical Astronomy Observatory. She is, in fact, at the National Optical-Infrared Astronomy Research Laboratory.

© 2020 Springer Nature Limited. All rights reserved.