Facing a Downpour of Data, Scientists Look to the Cloud

February 3, 2020 • *Physics* 13, 14

To improve access to large data sets, scientists are looking to cloud-based solutions for data management.

iStock.com/JuSun

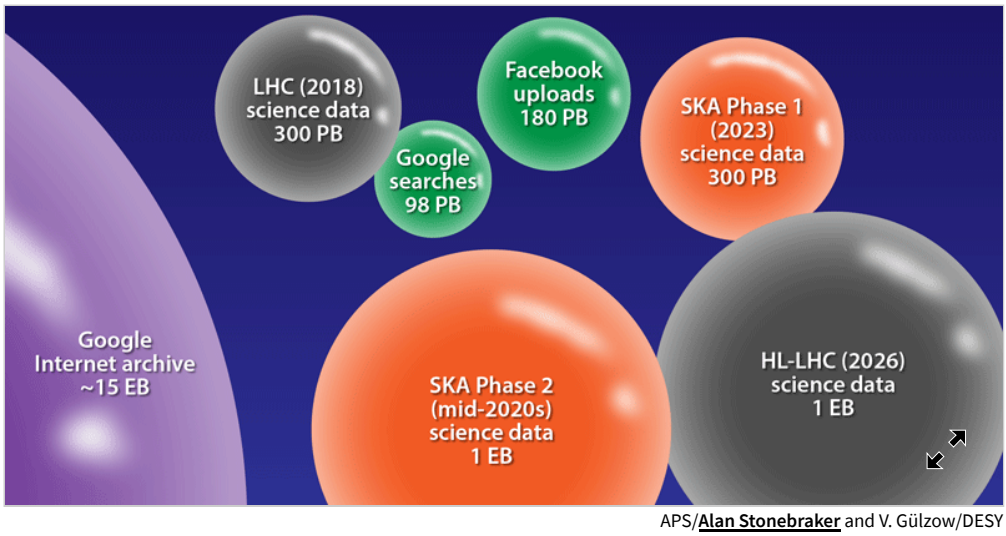
Storing experimental data in a “science cloud” has some advantages, such as making information more accessible to a wider scientific community.

In the coming decade, big projects like the Large Hadron Collider (LHC) and the Square Kilometre Array (SKA) are each expected to produce an exabyte of data yearly, which is about 20 times the digital content of all the written works throughout human history. This information overload requires new thinking about data management, which is why scientists have begun to look to the “cloud.” In such a scenario, data would be stored and analyzed remotely, with the advantage that information would become more accessible to a wider scientific community. Efforts are underway to create “science clouds,” but disagreements remain over their structure and implementation. To discuss these details, around 60 scientists came together for **The Science Cloud** meeting in Bad Honnef, Germany. The attendees shared lessons from past and ongoing projects in the hope of building the groundwork for a future scientific computing infrastructure.

The three-day meeting, held in a 19th century building of the **German Physical Society**, brought together physicists, astronomers, and computer scientists for the purpose of identifying “a common set of needs,” says meeting organizer Karl Mannheim, an astrophysicist from the Julius Maximilian University of Würzburg, Germany. He believes these sorts of discussions will help steer scientific data projects at the national and international level. In Europe, for example, in 2018 the European Union launched the European Open Science Cloud, which aims to be a one-stop portal for multidisciplinary research.

Many questions face the developers of science clouds, such as what fraction of computer hardware should be devoted to flexible, all-purpose central processing units (CPUs) versus faster, more specialized components like graphics processing units and neuromorphic chips. There are also software decisions, like choosing which platform will connect users to the cloud. These issues could be handled by a commercial cloud provider, like Google or Amazon. But many scientists fret over giving up too much control over their data and becoming “addicted” to a company’s software interface, says Benedikt Riedel, a researcher from the University of Wisconsin–Madison who works at the IceCube Neutrino Observatory.

Several of the meeting participants presented their particular computing challenges in a kind of group therapy session for big data users. Michiel van Haarlem from the Netherlands Institute for Radio Astronomy described the IT setup for the SKA telescope, which is scheduled to start taking data in the mid-2020s. SKA’s low-frequency arrays will produce 157 terabytes per second of raw data, which is 5 times the internet traffic in 2015, Haarlem says. To avoid that storage nightmare, the SKA collaboration plans to reduce the data—through processing—by about a factor of 1000. Some users may not feel comfortable with this pre-packaged data product, but “do you really want to see the raw data?” van Haarlem asks.



APS/**Alan Stonebraker** and V. Gülzow/DESY

A comparison of the yearly data volumes of current and future projects, where PB stands for petabyte (10^{15} bytes) and EB stands for exabyte (10^{18} bytes).

The LHC Collaboration has been dealing with big data since its inception, but the problem is about to balloon in size. In 2026, the LHC plans to boost its luminosity by a factor of 10, which will mean tripling the amount of data stored each year. But that's only part of the problem, explained Volker Gülzow, the IT director at DESY in Hamburg, Germany. To "reconstruct" what happens in a collision event, LHC researchers currently run large-scale computer simulations on a **grid** of 15,000 computer servers from around the world. With HL-LHC—as the luminosity boost is called—the number of computing tasks is expected to jump by a factor of 50 to 100, which is beyond the current capacity of the LHC grid. "It's very likely that HL-LHC will use cloud technology for almost all of its simulations," Gülzow says.

One of the advantages of using a cloud is that it can be "elastic." If on one day your experiment needs 50% more computing power than average, a cloud can accommodate that spike without you having to invest in extra computers. Another perk is that a cloud can help streamline the way data is accessed. This sort of user interface can be modeled after previous science projects, like the Sloan Digital Sky Survey. When the Sloan project started in 2000, the data management team—led by Alex Szalay from Johns Hopkins University, Maryland—developed a kind of "virtual observatory" for downloading data. Szalay compares it to iTunes, in that scientists can select a specific data set (like an MP3 file) rather than taking home a complete hard copy (like a CD). Since its debut, the Sloan survey has had seven million users, making it the world's most-used astronomy facility today. But Szalay admits the challenges are evolving. "The data is coming faster than we can consume it," he says.

The tools for managing all this data are evolving as well. Frossie Economou from the Legacy Survey of Space and Time project in Chile made a plea that researchers stop writing their own computer codes for handling data. Rather than "re-invent the wheel," she says, scientists should collaborate with data specialists working at the cutting-edge of IT development. But many young scientists at the meeting complained that data management experience is not valued in the academic job market. Meeting organizer Andreas Haungs, an astroparticle physicist from the Karlsruhe Institute of Technology in Germany, agrees that this is an important issue. "We all have to work on better recognition and visibility for people working on the interface between information technology and science," he says.

The value of good data management should be factored into the price of science, argues Szalay. "We are spending billions of dollars on big experiments, yet we have no plan for long-term data," he says. This issue was less of a problem when experiments had a high turnover rate and new data was always being produced. However, experiments have become so expensive that the time between them may stretch to decades, which means their data need

to have a long shelf life. To ensure that data archives remain active, Szalay proposes that projects include in their budgets a 5% “cost overrun” that would cover data archive management for 20 years after an experiment has shut down.

Call it saving for a cloudy day.

–Michael Schirber

Michael Schirber is a Corresponding Editor for ***Physics*** based in Lyon, France.

Subject Areas

Particles and Fields | Astrophysics

Print