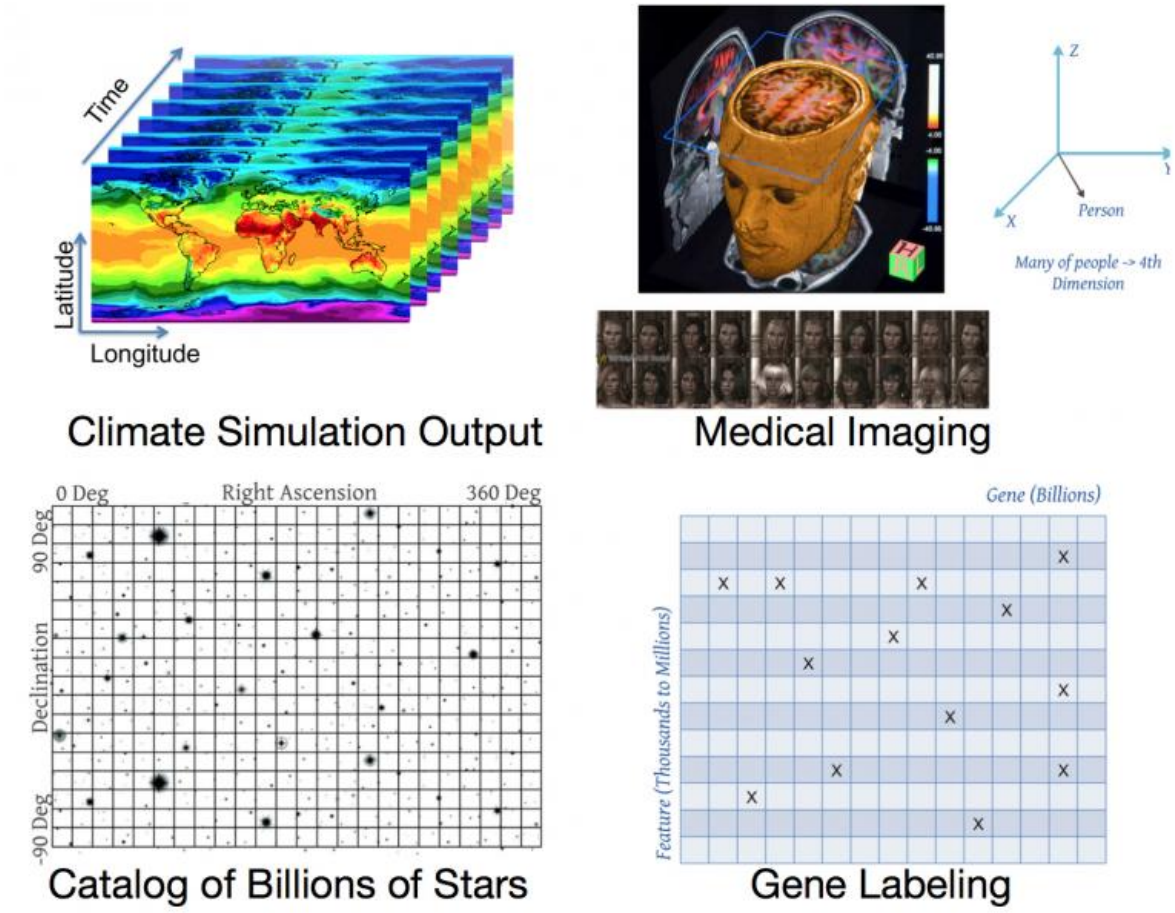
System designed to store and analyze extremely large array-structured data

September 1 2015, by Linda Vu



SciDB harnesses parallel architectures for fast analysis of terabyte (TBs) arrays of scientific data. This collage illustrates some of the scientific areas that have benefited from NERSC's implementation of SciDB, including astronomy, biology and climate. Credit: Yushu Yao, Berkeley Lab

Science is swimming in data. And, the already daunting task of managing and analyzing this information will only become more difficult as scientific instruments—especially those capable of delivering more than a petabyte (that's a quadrillion bytes) of information per day—come online.

Tackling these extreme data challenges will require a system that is easy enough for any scientist to use, that can effectively harness the power of ever more powerful supercomputers, and that is unified and extendable. This is where the Department of Energy's (DOE) National Energy Research Scientific Computing Center's (NERSC's) implementation of SciDB comes in.

"SciDB is an open source database system designed to store and analyze extremely large array-structured data—like pictures from light sources and telescopes, time-series data collected from sensors, spectral data produced by spectrometers and spectrographs, and graph-like structures that illustrate relationships between entities," says Yushu Yao, of NERSC's Analytics Group.

He notes that the advantage for science is that the database can scale on hundreds of nodes, can easily be deployed on commodity hardware or standing DOE supercomputers, has efficient parallel input/output (I/O), includes a large variety of built-in generic analysis tools and that it is relatively easy to integrate new algorithms that can transparently access efficient I/O. Back in 2013, NERSC set up a SciDB 20-node cluster and set out help scientists use SciDB on real science data problems. To date, NERSC has initiated more than 10 such partnerships across a broad range of science topics—from astronomy and climate to biology.

"Our aim was to help the scientists build SciDB into their normal science workflows, with the assumption that the lessons learned from each case study would provide insight into how to create new technologies and

environments for other data-intensive computing projects at NERSC," says Yao.

"This is a really useful and exciting tool if you have a large array of data," says Lisa Gerhardt, a physicist and NERSC User Consultant, who helped evaluate the usefulness of SciDB for high-energy physics analyses.

As a case study for high energy physics, Gerhardt worked with the LUX team to load the raw data collected from the instrument's inaugural run in 2013. The LUX instrument was built to directly detect the faint interactions from galactic dark matter in the form of Weakly Interacting Massive Particles (WIMPS). In its initial data run, LUX collected 83 million events, containing 600 million pulses and wrote 32TB of raw data. Of this data, only 160 events survived basic analysis cuts for potential WIMP candidates. The data rate for the new dark matter run, which started in 2014, exceeds 250 TB/year.

"Typically, analysis of this kind of data requires a researcher to search through about 1 million 10 MB files within a 10 TB dataset. If a WIMP candidate is spotted the researcher would save the candidate to another file. This process is slow and cumbersome, it took about a day at best to do this work and the analyses steps are difficult to share," says Gerhardt. "But with the SciDB testbed at NERSC, the same search took 1.5 minutes from start to finish. This is a tremendous breakthrough because it allows researchers to ask more questions and spend more time doing science."

Gerhardt notes that one of the greatest benefits of SciDB is that it makes parallel computing transparent to users. Once the SciDB interface is set up, researchers do not need to know anything about its configuration to run analysis. And since SciDB has a robust python and R interface—two programming languages that are widely used in scientific data

analysis—this significantly lowers the threshold for using this tool.

Of the 10 SciDB use-case partnerships that the NERSC has initiated so far, there has been so much interest in continuing this work that NERSC deployed a dedicated set of nodes for this kind of analysis. These nodes currently serve as the backbone for the Metabolite Atlas web portal and will be one of the main analysis platforms for the LUX project.

"We are in the process of putting together a broad software stack for enabling big data workloads, and scalable database technologies, like SciDB, are key for enhancing the productivity of our scientific users," says Prabhat, who leads NERSC's Data & Analytics Services Group. "Our partnership with Paradigm4 has been crucial to the evaluation and deployment of SciDB at NERSC."

More information: Any researchers interested in exploring SciDB can get more information and request access here: www.nersc.gov/users/data-analy … ent/databases/scidb/

Provided by Lawrence Berkeley National Laboratory

Citation: System designed to store and analyze extremely large array-structured data (2015, September 1) retrieved 24 April 2024 from https://phys.org/news/2015-09-extremely-large-array-structured.html