PHYS.ORG

Harnessing the petabyte at Rensselaer Polytechnic Institute

September 9 2013

The petabyte—a quantity of digital information 12 orders of magnitude greater than the lowly kilobyte—looms large as a future standard for data. To glean knowledge from this deluge of data, a team of researchers at the Data Science Research Center at Rensselaer Polytechnic Institute is combining the reach of cloud computing with the precision of supercomputers in a new approach to Big Data analysis.

"Advances in technology for medical imaging devices, sensors, and in powerful scientific simulations are producing data that we must be able to access and mine," said Bulent Yener, founding director of the DSRC, a professor of computer science within the Rensselaer School of Science, and a member of the research team. "The trend is heading toward petabyte data and we need to develop algorithms and methods that can help us understand the knowledge contained within it."

The team, led by Petros Drineas, associate professor of computer science at Rensselaer, has been awarded a four-year, $1 million grant from the National Science Foundation Division of Information & Intelligent Systems to explore the new strategies for mining petabyte data. The project will enlist key faculty from across the Institute including Drineas and Yener; Christopher Carothers, director of the Rensselaer supercomputing center, the Computational Center for Nanotechnology Innovations (CCNI), and professor of computer science; Mohammed Zaki, professor of computer science; and Angel Garcia, head of the Department of Physics, Applied Physics, and Astronomy and senior chaired professor in the Biocomputation and

Bioinformatics Constellation.

Drineas said the team proposes a novel two-stage approach to harnessing the petabyte.

"This is a new paradigm in dealing with massive amounts of data," Drineas said. "In the first stage, we will use cloud computing—which is cheap and easily accessible—to create a sketch or a statistical summary of the data. In the second stage, we feed those sketches to a more precise—but also more expensive—computational system, like those in the Rensselaer supercomputing center, to mine the data for information."

The problem, according to Yener, is that data on the petabyte scale is so large, scientists do not yet have a means to extract knowledge from the bounty.

"Scientifically, it is difficult to manage a petabyte of data," said Yener. "It's an enormous amount of data. If, for example, you wanted to transfer a petabyte of data from California to New York, you would need to hire an entire fleet of trucks to carry the disks. What we are trying to do is establish methods for mining and for extracting knowledge from this much data."

Although petabyte data is still uncommon and not easily obtained (for this particular research project Angel Garcia will generate and provide a petabyte simulation of atomic-level movements), it is a visible frontier, and standard approaches to data analysis will be too costly, too time-consuming, and not sufficiently powerful to do the job given current computing power.

"Having a supercomputer process a petabyte of data is not a feasible model, but cloud computing cannot do the job alone either," Yener said. "In this way, we do some pre-processing with the cloud, and then we do more precise computing with CCNI. So it is finding this balance between how much you are going to execute, and how accurately you can execute it."

The work will include developing the techniques for pre-processing and precision processing, such as sampling, rank reduction, and search techniques. In one simplistic example, Yener said the cloud may calculate some simple statistics for the data—mean, maximum, average—which could be used to reduce the data into a "sketch" that could be further analyzed by a supercomputer.

Balancing between the two stages is critical, said Drineas.

"How do you execute these two stages? There are some steps, some algorithms, some techniques that we will be developing," Drineas said. "The steps in cloud computing will all be directed to pre-processing, and the steps in supercomputing will all be directed to more exact, expensive, and precise calculations to mine the data."

Established in 2010, the DSRC is focused on fostering research and development to address today's most pressing data-centric and data-intensive research challenges, utilizing the unique resources available at Rensselaer. Recently, the DSRC welcomed General Dynamics Advanced Information Systems and General Electric as its first two corporate members.

Big Data, broad data, high performance computing, data analytics, and Web science are creating a significant transformation globally in the way we make connections, make discoveries, make decisions, make products, and ultimately make progress. The DSRC is a component of Rensselaer's university-wide effort to maximize the capabilities of these tools and technologies for the purpose of expediting scientific discovery and innovation, developing the next generation of these digital enablers, and preparing our students to succeed and lead in this new data-driven world.

Provided by Rensselaer Polytechnic Institute

Citation: Harnessing the petabyte at Rensselaer Polytechnic Institute (2013, September 9) retrieved 25 April 2024 from https://phys.org/news/2013-09-harnessing-petabyte-rensselaer-polytechnic.html