The New York Times

Technology

Next Article in Technology (1 of 16) »

Training to Climb an Everest of Digital Data

By ASHLEE VANCE
Published: October 11, 2009

SIGN IN TO RECOMMEND
TWITTER
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

MOUNTAIN VIEW, Calif. — It is a rare criticism of elite American university students that they do not think big enough. But that is exactly the complaint from some of the largest technology companies and the federal government.

Enlarge This Image

Steve Ruark for The New York Times

Jimmy Lin is an associate professor at the University of Maryland.

At the heart of this criticism is data. Researchers and workers in fields as diverse as bio-technology, astronomy and computer science will soon find themselves overwhelmed with information. Better telescopes and genome sequencers are as much to blame for this data glut as are faster computers and bigger hard drives.

While consumers are just starting to comprehend the idea of buying external hard drives for the home capable of storing a terabyte of data, computer scientists need to grapple with data sets thousands of times as large and growing ever larger. (A single terabyte equals 1,000 gigabytes and could store about 1,000 copies of the Encyclopedia Britannica.)

The next generation of computer scientists has to think in terms of what could be described as Internet scale. Facebook, for example, uses more than 1 petabyte of storage space to manage its users' 40 billion photos. (A petabyte is about 1,000 times as large as a terabyte, and could store about 500 billion pages of text.)

It was not long ago that the notion of one company having anything close to 40 billion photos would have seemed tough to fathom. Google, meanwhile, churns through 20 times that amount of information every single day just running data analysis jobs. In short order, DNA sequencing systems too will generate many petabytes of information a year.

"It sounds like science fiction, but soon enough, you'll hand a machine a strand of hair, and a DNA sequence will come out the other side," said Jimmy Lin, an associate professor at the University of Maryland, during a technology conference held here last week.

The big question is whether the person on the other side of that machine will have the

Personal Tech E-Mail

Sign up for David Pogue's exclusive column, sent every Thursday.
Sign Up
See Sample | Privacy Policy

Subscribe to Technology RSS Feeds

Technology News
Internet
Start-Ups
Business Computing
Companies
Bits Blog
Personal Tech
Pogue's Posts

MOST POPULAR - TECHNOLOGY

E-MAILED | BLOGGED

1. Training to Climb an Everest of Digital Data
2. Closing the Deal at the Virtual Checkout Counter
3. 5 Easy Steps to Stay Safe (and Private!) on Facebook
4. Bits: Coming Soon to a Theater Near You, If You Demand It
5. Gadgetwise: App of the Week: A Nostalgic Time Killer
6. Some Users May Lose Data on a T-Mobile Smartphone
7. Parental Guidance on Web Video for Children
8. Bits: Start-Ups Aim to Transform Visits to the Doctor
9. Seeing the Future in 3-D Television
10. Phone Smart: In Choosing a New Phone, Online Research Goes Only So Far

Go to Complete List »

wherewithal to do something interesting with an almost limitless supply of genetic information.

At the moment, companies like I.B.M. and Google have their doubts.

For the most part, university students have used rather modest computing systems to support their studies. They are learning to collect and manipulate information on personal computers or what are known as clusters, where computer servers are cabled together to form a larger computer. But even these machines fail to churn through enough data to really challenge and train a young mind meant to ponder the mega-scale problems of tomorrow.

"If they imprint on these small systems, that becomes their frame of reference and what they're always thinking about," said Jim Spohrer, a director at I.B.M.'s Almaden Research Center.

Two years ago, I.B.M. and Google set out to change the mindset at universities by giving students broad access to some of the largest computers on the planet. The companies then outfitted the computers with software that Internet companies use to tackle their toughest data analysis jobs.

And, rather than building a big computer at each university, the companies created a system that let students and researchers tap into giant computers over the Internet.

This year, the National Science Foundation, a federal government agency, issued a vote of confidence for the project by splitting $5 million among 14 universities that want to teach their students how to grapple with big data questions.

The types of projects the 14 universities have already tackled veer into the mind-bending. For example, Andrew J. Connolly, an associate professor at the University of Washington, has turned to the high-powered computers to aid his work on the evolution of galaxies. Mr. Connolly works with data gathered by large telescopes that inch their way across the sky taking pictures of various objects.

The largest public database of such images available today comes from the Sloan Digital Sky Survey, which has about 80 terabytes of data, according to Mr. Connolly. A new system called the Large Synoptic Survey Telescope is set to take more detailed images of larger chunks of the sky and produce about 30 terabytes of data each night. Mr. Connolly's graduate students have been set to work trying to figure out ways of coping with this much information.

Purdue, meanwhile, looks to carry out techniques used to map the interactions between people in social networks into the biological realm. Researchers are creating complex diagrams that illuminate the links between chemical reactions taking place in cells.

A similar effort at the University of California, Santa Barbara, centers on making a simple software interface — akin to the Google search bar — that will let researchers examine huge biological data sets for answers to specific queries.

Mr. Lin has encouraged his students to illuminate data with the help of Hadoop, an open-source software package that companies like Facebook and Yahoo use to split vast amounts of information into more manageable chunks.

One of these projects included a deep dive into the reams of documents released after the government's probe into Enron, to create an analysis system that could identify how one employee's internal communications had been connected to those from other employees and who had originated a specific decision.

Mr. Lin shares the opinion of numerous other researchers that learning these types of analysis techniques will be vital for students in the coming years.

Go to Complete List »

“Science these days has basically turned into a data-management problem,” Mr. Lin said.

By donating their computing wares to the universities, Google and I.B.M. hope to train a new breed of engineers and scientists to think in Internet scale. Of course, it’s not all good will backing these gestures. I.B.M. is looking for big data experts who can complement its consulting in areas like health care and financial services. It has already started working with customers to put together analytics systems built on top of Hadoop. Meanwhile, Google promotes just about anything that creates more information to index and search.

Nonetheless, the universities and the government benefit from I.B.M. and Google providing access to big data sets for experiments, simpler software and their computing wares.

“Historically, it has been tough to get the type of data these researchers need out of industry,” said James C. French, a research director at the National Science Foundation. “But we’re at this point where a biologist needs to see these types of volumes of information to begin to think about what is possible in terms of commercial applications.”

Sign in to Recommend

A version of this article appeared in print on October 12, 2009, on page B1 of the New York edition.

Next Article in Technology (1 of 16) »

SIGN IN TO E-MAIL

PRINT

REPRINTS

Related Articles

FROM THE NEW YORK TIMES

IN THE HUNT; Finding the Path to Success by Changing Directions
(September 10, 2009)

A Small-Business Guide to Disaster Recovery
(September 10, 2009)

As Banks Retreat, Valley Financier Looks to Fill a Gap
(August 3, 2009)

THE MEDIUM; The Great Crash
(June 14, 2009)

INSIDE NYTIMES.COM

MUSIC »

Instead of Retiring, Carly Simon Is Suing Her Label

BUSINESS »

Mutual Funds Report: Third Quarter

OPINION »

Op-Ed: Health Care Reform ... for the N.F.L.

WEEK IN REVIEW »

Audio Slide Show: Irving Penn at the Getty

OPINION Â»

The High Cost of Empty Prisons

By downsizing its prison system, New York can help point criminal justice in a more constructive direction.

N.Y. / REGION »

A Quest to Read a Book a Day for 365 Days