Big, fast, weird data

April 8 2014, by Eric Brown

The "Big Data" research that continues to dominate IT agendas has traditionally focused on making sense of the growing volumes of computer data. Yet in recent years, the volume question has given way to the other V's of Big Data: velocity and variety.

"In the past, we've focused on scale, but over the last few years, the big new problems have been about variety," says Sam Madden, professor of electrical engineering at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). "Big Data is no longer just about processing a huge number of bytes, but doing things with data that you couldn't do previously. Increasingly, data is coming at you really fast, and it's much more complex. It's not just tabular data you can easily stick into a spreadsheet or a database."

The velocity and variety challenges stem from faster computers and networks, as well as the proliferation of data types, such as streaming media, social networking trends, and real-time financial transactions. "Suddenly we're trying to process things like graphs of relationships between people in a social network or trying to make sense of a lot of images," says Madden. "We're trying to extract and mine patterns in data."

The problems of speed and diversity have always figured in Madden's database projects. Over a decade ago, his TinyDB project tackled the problem of on-the-fly analysis of networked sensor data. Five years ago, he helped launch CSAIL's CarTel mobile sensor network, which he and other MIT researchers continue to extend as it rolls around Boston on

cars, taxis, and buses. More recently, Madden's research has included Qurk, a system that lets people answer queries via the Mechanical Turk crowdsourcing platform, and Relational Cloud, which investigates architectures for cloud-based data managment. Other recent projects include two we'll examine here: research into social networking analysis (MapD) and scientific databases (SciDB).

Madden is the director of MIT's BigData@CSAIL industry initiative, and he co-directs the more research-focused Intel Science and Technology Center (ISTC) for Big Data. BigData@CSAIL is designed to educate industrial partners on Big Data issues using workshops, white papers, and seminars. Recently, for example, BigData@CSAIL ran a workshop on Big Data and privacy.

With the ITSC, Intel and MIT are attempting to optimize software and tools to exploit the chipmaker's latest hardware and improve the performance of Big Data algorithms. At the same time, "Intel wants to understand how new trends in Big Data are going to change the hardware they ought to be building," says Madden.

The multidisciplinary initiative involves 10 faculty at MIT and about the same number spread around several other universities. "We try to get people from different areas of computer science to think about hard and interesting problems," says Madden. "We might take an architect who understands next-generation flash memory and combine them with someone who's really good with algorithms, and then explore how to gain the most efficiency from a certain piece of hardware."

MapD: Mapping Twitter Trends in Real-Time

Lately, Madden has been working on a SQL-based analytics database of Twitter data called MapD (Massively Parallel Database). MapD was principally designed by Todd Mostak, a Middle East specialist who was

frustrated by the lack of research tools when he was studying Twitter data from Egypt, trying to correlate tweeting trends with census data. Madden helped Mostak mine the growing wealth of location data served up with mobile tweets tagged with "geocodes," or the location of the people sending the tweets.

"Out of the hundreds of millions of tweets sent every day, there are 5 to 7 million geocoded tweets, typically using the position sensors on smartphones," says Madden. "MapD plots them on a map over time and allows users to search them and interact with them using real-time animations."

MapD harvests the processing power of the phones' graphic processing units to parallelize computing tasks such as querying and rendering very large data sets. The data can be used to analyze localized and time-stamped events ranging from political movements to responses to new products to patterns of disease or weather.

"People tweet when it's raining or snowing or if they have the flu or went to the doctor," says Madden. "Those things can be visualized in a very compelling way."

MapD poses challenges in both variety and a velocity, says Madden. "The data is text, not tabular, so you need different ways of pre-processing, compressing and formatting it to make it computable," he says. "You have to convert it into maps that show tweet density visualizations in different regions of the world, or into word clouds or histograms that track the frequency of different words over time. That's a lot of complexity."

SciDB: Data Meets the Matrix

While social networking analysis pushes the envelope of database

management systems on speed and diversity, the problem of analyzing scientific data involves challenges of volume and complexity. Madden is working with DBMS pioneer Michael Stonebraker, the principal developer of INGRES and POSTGRES, on SciDB, a project for developing an optimized DBMS for science research. SciDB integrates a high level query language that enables the computation over matrices and provides an efficient engine to execute those queries on computer clusters.

"We're targeting scientific applications that conventional relational databases aren't very good at," says Madden. "In SciDB we think of data as arrays or matrices, multidimensional structures that you can use to run matrix algorithms and operations."

SciDB is particularly suitable for analyzing large volumes of genomic and patient data. "You can have a matrix of patients on one dimension and a matrix of genes on another, and then do different types of analyses on those matrices," says Madden. "You can find patterns or genes that are correlated with patients that have certain conditions or outcomes."

CarTel and the Boston Transportation Challenge

Madden and other researchers continue to find new applications for MIT's mobile CarTel research platform. Originally, CarTel focused on mapping traffic conditions or building maps of cellular or WiFi connectivity. Now Madden is using the platform to explore the larger picture of the "patterns of movement of people" in an urban environment.

Madden and others continue to fine-tune the wireless and sensor technology deployed on the vehicles, as well as finesse the databases used to analyze their data. Instead of depending on a persistent cellular connection requiring an expensive cellular modem, the CarTel

researchers have explored "opportunistically connecting to WiFi networks as cars drive by or developing multi-hop networks of vehicles that swap data as they pass each other," he says.

Increasingly, CarTel has used smartphones as the mobile computer, giving the researchers a built-in array of sensors. "We've used the phones' accelerometers to build maps of potholes," says Madden. "You have to aggregate this data from many different vehicles and differentiate potholes from things like speed bumps. We can even rank the potholes by size to help determine which to fix first."

One project aims to determine driver safety by analyzing acceleration profiles. This is difficult, however, if the phone in not in a fixed orientation. "We have to recalibrate the orientation of the device as it travels," says Madden. "Now we can tell if an abrupt spike in acceleration is really acceleration or whether the user is just waving the phone around. It poses some interesting signal processing challenges."

Both CarTel and MapD provided some of the inspiration for a recent Boston Transportation Challenge, the first of several challenges to be hosted by BigData@CSAIL. Madden and his team collected a wide variety of transportation data including traffic information, taxi pickups and dropoffs, and public transportation statistics, along with a database of localized tweets and major events in Boston.

"We asked our member participants to use the data to predict taxi ridership at a given time period for which they didn't have data," explains Madden. The challenge forced users to look beyond simply analyzing historical values of baseline information in taxi ridership and explore how variations caused by events and weather affect ridership.

"It's a good proxy for some of the challenges our members might have," says Madden. "A lot of businesses use big data to make predictions, like

how many users are going to show up on my website or click on this ad, or how many widgets I can sell in a month. But there are some fundamental problems in combining all this complex data together. You have location data, textual information about events, traffic jams, rankings of different performers or Red Sox games. Companies have these same sorts of diverse information sets about their customers. Now they can begin to aggregate all this into predictive models."

More information: www.map-d.com/

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology