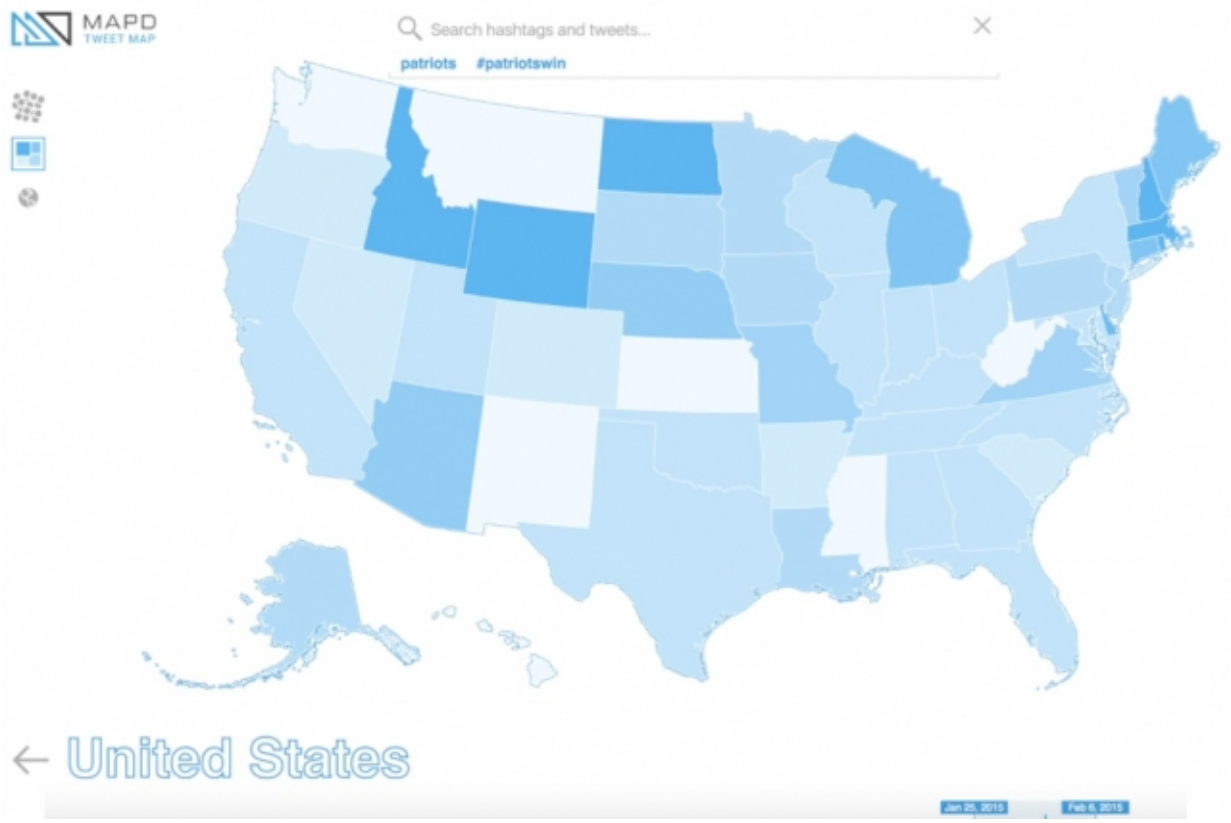
New type of database-analytics platform queries and maps billions of data points in milliseconds

January 11 2017, by Rob Matheson



MapD is a type of commonly used database-management system modified to run on GPUs instead of CPUs, which power most traditional database-management systems. MapD can process billions of data points in milliseconds, making it 100 times faster than those traditional systems, and visualizes that data near instantaneously. Credit: MapD Technologies

People generally associate graphic processing units (GPUs) with imaging processing. Developed for video games in the 1990s, modern GPUs are specialized circuits with thousands of small, efficient processing units, or "cores," that work simultaneously to rapidly render graphics on screen.

But for the better part of a decade, GPUs have also found general computing applications. Because of their incredible parallel-computing speeds and high-performance memory, GPUs are today used for advanced lab simulations and deep-learning programming, among other things.

Now, Todd Mostak, a former researcher at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), is using GPUs to develop an analytic database and visualization platform called MapD, which is the fastest of its kind in the world, according to Mostak.

MapD is essentially a form of a commonly used database-management system that's modified to run on GPUs instead of the central processing units (CPUs) that power most traditional database-management systems. By doing so, MapD can process billions of data points in milliseconds, making it 100 times faster than traditional systems. Moreover, MapD visualizes all processed data points nearly instantaneously—such as, say, plotting tweets on a world map—and parameters can be modified on the fly to adjust the visualized display.

With its first product launched last March, MapD's clients already include Verizon and other big-name telecommunications companies, a social media giant, and financial and advertising firms. In October, the investment arm of the U.S. Central Intelligence Agency, In-Q-Tel, announced that it had invested in MapD's latest funding round to accelerate the development of certain features for the U.S. intelligence community.

"[The CIA has] a lot of geospatial data, and they need to be able to form, visualize, and query that data in real-time. It's a real need across the intelligence community," Mostak says.

"Making GPUs first-class citizens"

GPUs are designed specifically for parallel computing, with thousands of energy-efficient cores that can, for example, simultaneously determine the color of each pixel on a computer screen to render an image. GPUs also use high bandwidth memory, a form of random access memory (RAM) that's about an order of magnitude faster than CPUs.

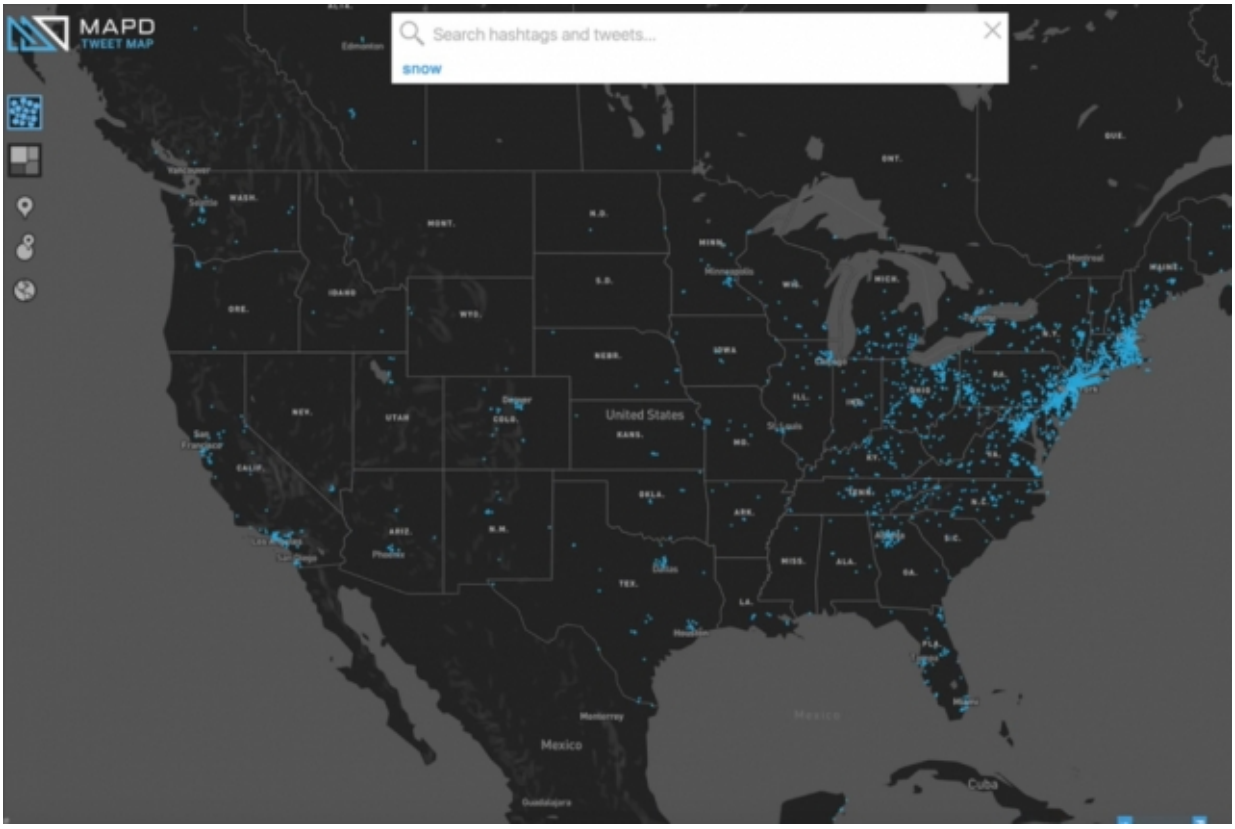
Today, some databases are being powered by GPUs. But these systems suffer from a major design flaw, Mostak says: "In most implementations, the data is initially stored on a CPU, moved to the GPU for a query, and results are moved back to the CPU for storage. Even if you speed up the computation time of a query [by using a GPU], you lose most of the speed by transferring from CPU to GPU and back."

But with MapD, Mostak says, the goal "is making GPUs first-class citizens."

Instead of storing the data on CPUs, MapD caches as much data as possible on multiple GPUs, so there's no moving back and forth between the different circuits and pulling from the hard drive, which saves a lot of time.

The trick, Mostak says, is giving each GPU its own buffer pool—portions of a database memory that temporarily caches the most recent data pulled from the hard. If a database then needs to query the same data point over and over, which is quite common, it accesses that data point in the GPU's ultrafast RAM, instead of pulling from the CPU or hard drive.

By carefully managing the memory on the GPU, MapD can deliver performance that is two to three orders of magnitude faster than CPU-powered database systems, Mostak says.



MapD's live, geolocated "Tweetmap" lets users search for individual Twitter hashtags and see those hashtags appear, in real time, across a world map. Credit: MapD Technologies

Taxis, tweets, telecommunications

In one example of what MapD can do, the system analyzed a dataset that's considered the benchmark for large-scale analytics—a 1.2 billion-record New York City taxi dataset. In a test by an independent big-data

consultant, MapD ran 74 times faster than numerous advanced CPU database systems, completing several queries in milliseconds.

In other examples, MapD's live, geolocated "Tweetmap" lets users search for individual Twitter hashtags and see those hashtags appear, in real time, across a world map. Another map, of the United States, shows every political donation since 2001, color-coded for Republican (red) and Democratic (blue) candidates.

As for MapD clients, financial services agencies and hedge funds can use the system to monitor fraud and make investment decisions; advertising agencies can use it to measure reaction to ads; and social media companies can track usage across the planet.

Verizon used MapD to analyze the activity of updating SIM cards on each of its 85 million subscribers' phones on a weekly basis. With other database systems, the query would take hours to run and hours to evaluate, so the company only did so periodically. Using MapD, Verizon found a glitch in its system that led to SIM card updates upward of a million times per year, which used a lot of server power and was a nuisance for subscribers.

"So that's a big money savings for them, and probably good for the customer, as it's probably not good to have your SIM card updated so frequently," Mostak says.

Putting MapD on the map

The idea for MapD came to Mostak when he was at Harvard University in 2012, writing his political-science master's thesis on the Arab Spring, and analyzing hundreds of millions of Egyptian tweets sent out during the uprisings.

Using CPU-based database-management systems to analyze the data was a time-waster. Often he would run queries overnight and wake up to find an error, meaning the long process would need to be repeated. "It was a frustrating experience," Mostak says.

At the time, Mostak was also taking a CSAIL database course taught by the co-directors of the MIT Database Group: Michael Stonebraker, an adjunct professor in computer science who founded the pioneering database-management company Vertica; and Sam Madden, a professor of electrical engineering and computer science who serves as a MapD advisor.

As a personal project to speed up his thesis research, Mostak invented an early MapD prototype. The professors were impressed. After Mostak completed his thesis, they asked him to join CSAIL as a researcher and build out the prototype, which he did in 2013.

With Madden's encouragement, Mostak also began showcasing the speedy system around MIT's Industrial Liaison Program (ILP), which connects MIT community members with corporations around the world. Companies started asking Mostak where they could buy it. "At the time, I said it was purely an academic project," Mostak says. "But it got me thinking that this was a widespread problem—getting real-time insights out of big data."

In January 2014, Mostak officially launched MapD. Joining ILP's Startup Exchange, an online community for MIT-affiliated startups to connect with each other and with other companies, "put [MapD] on the map with commercial entities," Mostak says.

From there, the startup, then headquartered in Cambridge, Massachusetts, hit the ground running. In March 2014, it won a $100,000 prize from an early startup contest put on by Nvidia, a

prominent GPU manufacturer and current MapD partner. That fall, the startup landed $2 million in seed funding from Nvidia and Google, followed by a $10 million Series A funding round the following year.

Today, MapD is expanding in its new San Francisco headquarters. It's also looking to capitalize on an increased user base, as more companies start launching GPU programming platforms in the cloud. "That'll give us more access to customers," Mostak says, adding, "I feel like we're just getting started."

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