MapD

Somewhere, something incredible is waiting to be known.

Carl Sagan, astronomer

Introduction to MapD
The world's fastest platform for data exploration

Summer 2015

What could you do if your data was 1,000x faster?

As technology advancement creates an explosion of human data—millions of tweets, hours of video, conversions, and patterns every day—a new universe is emerging. Software limitations have left data scientists stuck on the shore of this new "data ocean"—only able to provide a coarse view even with clusters of hundreds of servers.

MapD's breakthrough system is changing this, enabling data queries 1,000x faster than anything ever developed. Using Graphics Processing Units (GPUs) to make supercomputing mainstream, MapD creates an immersive, weightless experience for data analysts to explore patterns at the speed of discovery.

With MapD, it is possible to query and visualize billions of records in tens of milliseconds. This enables the creation of hyper-interactive dashboards in which dozens of attributes can be correlated and cross-filtered without lag. Using MapD, analysts can query data at rates approaching three terabytes per second on a single server.

MapD delivers the world's fastest data exploration on a single server by harnessing the massive parallelism of GPUs. Originally designed to render video games, GPUs have evolved into general purpose computational engines that excel at performing repetitive tasks in parallel. Buttressed by high-speed memory, GPUs can perform thousands of calculations simultaneously, making them an excellent fit for data queries. The GPUs' prodigious compute capabilities also allow them to excel at many machine learning algorithms, while the graphics pipeline of the cards means they can be used for rendering large datasets in milliseconds.

This white paper begins with an overview of MapD and the key reasons for MapD's extraordinary performance. It then highlights the paradigm shift MapD's solution is enabling in today's dense data analytics and visualization space. The paper proceeds to overview how early commercial pilots, with diverse big data needs, are utilizing MapD's platform and outlines next steps for interested data explorers.

Founded in 2013
Based in San Francisco, CA

Index

MapD.com
info@mapd.com
@datarefined

or follow us on

> "The power of imagination makes us infinite."
>
> John Muir, naturalist

Mission: Immersive data exploration

MapD was founded by scientists, engineers, and data analysts on a mission to make data exploration an immersive experience. CEO Todd Mostak first developed a prototype of MapD while waiting hours and sometimes days for a single query to process patterns in hundreds of millions of tweets for his Harvard thesis on the Arab Spring. Frustrated that he couldn't access a cluster of computers to perform his computations, he created his own solution by pairing off-the-shelf video game GPU cards with a new design for parallel databases. Todd pursued this technology as a researcher at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) before launching MapD in 2013.

MapD is backed by Google Ventures and NVIDIA. The company was awarded the $100,000 grand prize as winner of the 2014 GPU Technology Conference's Early Stage Challenge and is currently in beta pilots with a select group of Fortune 100 companies and academic research organizations.

MapD's focus is to not just make queries faster, but to create a fluid and immersive data exploration experience that removes the disconnect between analyst and data. To take data from a position of "mining" and "visualization" to a sense of weightless exploration. Queries at the speed of imagination for understanding the expanding data universe.

Read more about MapD's story in VentureBeat

MapD's team holds degrees in computer science, anthropology, economics, Middle Eastern studies, and business administration from the following universities:

HARVARD UNIVERSITY

Berkeley UNIVERSITY OF CALIFORNIA

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

TEXAS A&M UNIVERSITY

"Anyone can build a fast CPU The trick is to build a fast system."

Seymour Cray, inventor

Putting supercomputing into context

Organizations are finding themselves awash with bigger and bigger datasets with greater need to extract actionable insight from them in real time. While hardware has become significantly faster over the last several years, data analytics tools have not kept up. The status quo solution has been to compensate for lack of software performance with more hardware, running clusters of hundreds of servers to query datasets not even a terabyte in size.

Away from the mainstream, the 50-year history of supercomputers is full of massive, warehouse-sized machines only accessible to the most elite scientists and researchers. Only a few hundred supercomputers are in operation in the U.S. today, being used for work on nuclear weaponry, weather modeling, and astrophysics at high-security facilities such as Lawrence Livermore National Laboratory and Argonne National Laboratory.

More recently, there has been migration away from supercomputers toward large clusters of commodity servers, particularly for data analytics. While these clusters are typically less expensive and less exotic than traditional supercomputers, they still require significant amounts of rack space, electricity, and trained technicians and continue to exhibit latencies that prevent true interactivity.

Whether by using traditional supercomputers or massive Hadoop clusters, big compute resources have enabled organizations to extract valuable and actionable insight from their data. However, many companies cannot afford the hardware, rack space, and personnel required to host such systems, and for those that can, the costs are tremendous. While the advent of cloud computing lessens the barrier to entry to obtaining heavy compute resources, oftentimes datasets must remain behind the company's firewall due to privacy issues or to protect a competitive advantage embodied in the data.

In short, it is still difficult for organizations to marshall the necessary computing resources to extract the insight hidden in their data. What if the work of racks of computers could be done by a single box, deployable by large and small organizations alike?

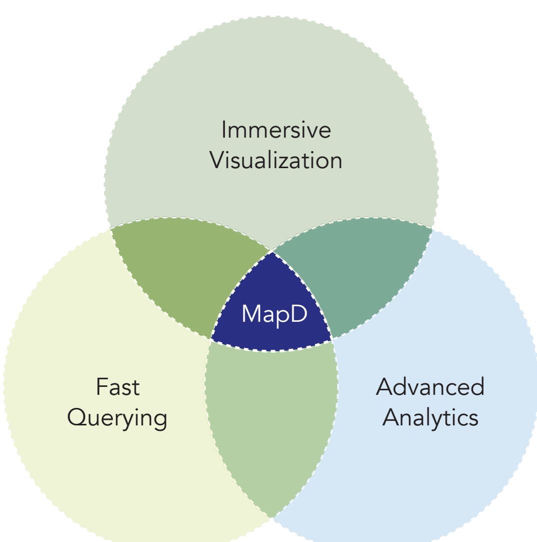
MapD's name is an acronym for "Massively Parallel Database."

Supercomputer in a server, supercomputer on a laptop

MapD is packaged in two versions, MapD Server and MapD Desktop. MapD Server is designed to run in headless server environments and supports multiple GPUs (up to 16 per server), allowing for analysis of multi-billion-row datasets by multiple simultaneous users.

MapD Desktop is for single-user deployment on laptops, desktops, and workstations. It can pull subsets of data from MapD Server for local analysis. By running simultaneously on the CPU and GPU of these platforms, MapD Desktop allows for analysis of hundreds of millions of rows of data at speeds far exceeding what other platforms achieve on dedicated servers. Analysts running MapD locally on a laptop or desktop can "check out" data from a MapD server, running a filter on potentially terabytes of data server-side to extract and download a subset that they can then analyze locally without network latencies or having to share resources with other users.

Both MapD Server and MapD Desktop support a subset of SQL and can be queried both in the console and via bindings to major programming languages. They also support User Defined Functions (UDFs) so that complex analytics such as regression and deep learning can be performed directly on GPU query results without copying the data.



Both versions of MapD are packaged with a high-performance web-based frontend - MapD Explorer - that enables interactive visual analysis of the data in MapD. MapD Explorer can generate a wide array of chart and geographic visualizations in-browser as well as request GPU-rendered visualizations from the MapD backend when result sets are large. The frontend is built around the cross-filter model, in that a filter applied to one data attribute in the dashboard is applied to all other attributes as well, allowing for easy drill-down and correlation analysis. The MapD is an end-to-end big data analytics and visualization platform that leverages the massive parallelism and memory bandwidth of GPUs to execute queries 1,000x faster than competing solutions. The unprecedented speed of the system allows for hyper-interactive exploration of multi-billion-row datasets by many simultaneous users, allowing analysts to generate and test hypotheses unencumbered by the latency associated with CPU-only platforms. MapD takes advantage of the three distinguishing traits of GPUs: their computational parallelism, their high-speed memory, and their graphics pipeline to form an end-to-end system that allows for high-speed querying, analytics, and visualization on big datasets.

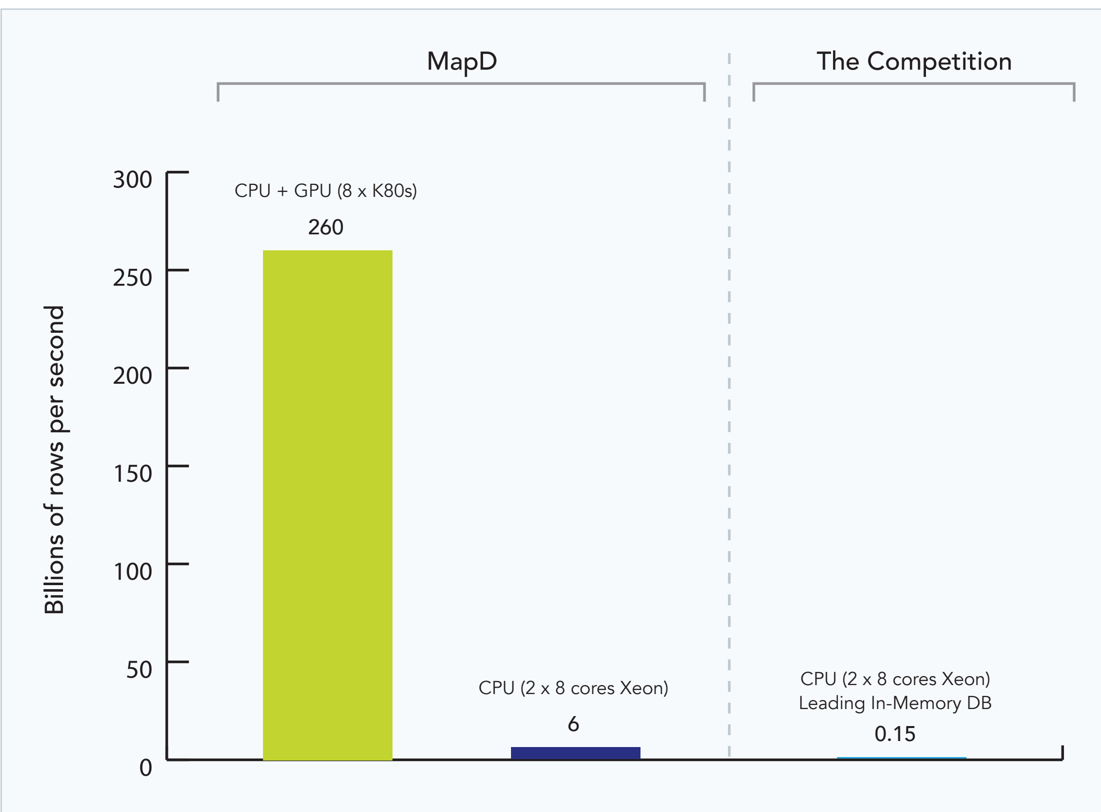
MapD outpaces competitors running on dedicated servers even when running on a laptop.

MapD's technology framework

While GPU performance and memory capacity are still rapidly increasing, in the current generation, up to eight GPU cards with 192GB total GPU memory can be installed in a single server, allowing data to be queried at rates approaching three terabytes per second by almost 40,000 cores. Furthermore, using the sophisticated graphics pipeline on each GPU, the data can be visualized in situ and sent to the client without the need for costly transfers to other platforms.

Although MapD was designed to deliver weightless data exploration straight to the browser, it is equally capable of processing programmatic SQL queries with blistering speed (at a rate over a trillion rows per second per server for certain workloads).

MapD performance compared to leading in-memory database on 2-socket, 8-GPU system



The MapD platform uniquely:
Caches the hot data in GPU memory such that no time is lost moving the data across the PCIe bus onto the GPU when a query is executed.

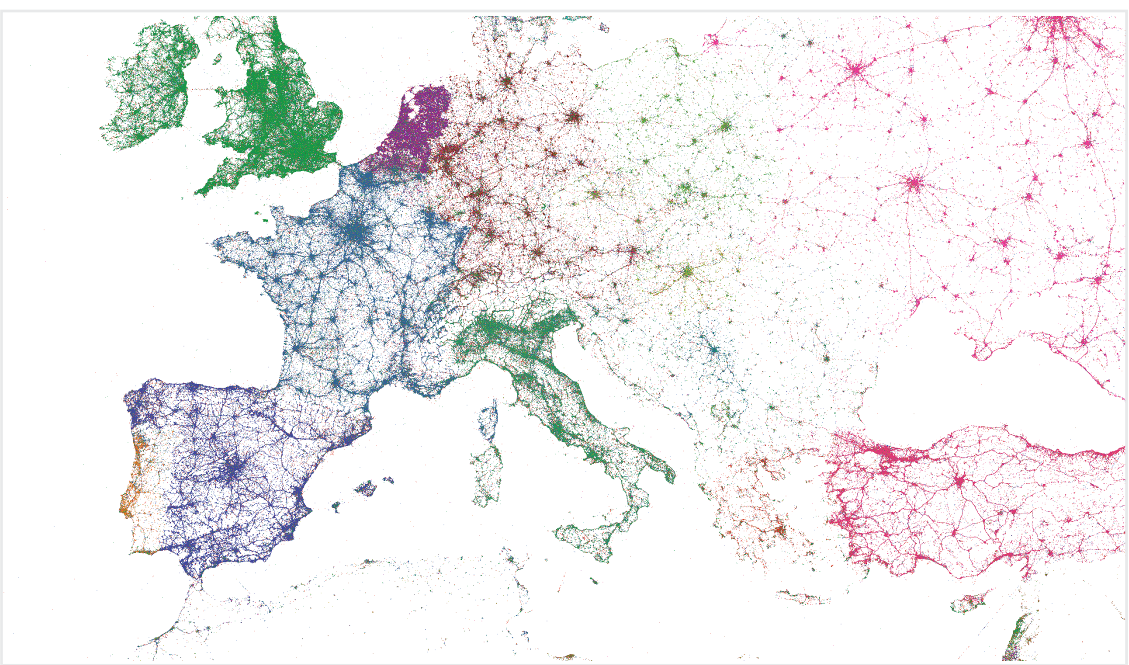
- Vectorizes execution of queries whenever possible. Vectorized code allows the compute resources of a processor to process multiple data items simultaneously. This is a must to achieve good performance on GPUs, which can comprise thousands of execution units. Additionally, optimizing vectorized execution also translates well to CPUs, which increasingly have "wide" execution units capable of processing multiple data items at once.

- Employs highly-optimized kernels for common database operations. Although GPUs are extremely fast, GPU code often requires more optimization work than comparable CPU code to achieve the best performance. MapD is the product of thousands of hours of benchmarking and testing to extract the maximum vectorized performance from both GPU and CPU.

- Compiles queries on the fly for maximum execution speed.

- Executes queries simultaneously on both CPU and GPU, or entirely on CPU if the working dataset cannot fit in GPU memory.

- Visualizes and performs complicated analytics on data in situ. Since the relevant data is already cached on the GPUs, MapD does not need to copy the query result set before rendering it (using the GPUs) or using it as input to a follow-on machine learning algorithm.



A MapD-rendered visualization of 500 million tweets in Europe colored by language used

> "Exploration is the engine that drives innovation. Innovation drives economic growth."
>
> Edith Widder, oceanographer

Case studies

Speed of understanding is one of the most significant business advantages available in today's market. The ability to see patterns and optimize for them has become a requirement for any large-scale operation. Early pilots of MapD's data supercomputing system have explored commercial applications for the technology with some of the largest companies in the country. Business names not disclosed under pilot agreement terms:

Finding patterns in social advertising
A social network with $12 billion in annual advertising revenue launched a pilot with MapD to explore real-time effectiveness of ads across different demographics and geographic regions. MapD's built-in text mining features enable the network's analysts to quickly determine key brand associations for their top accounts. Marketers at the organization are exploring terabytes of data across many simultaneous dimensions without lag. This knowledge allows the company to grow revenue by demonstrating complex conversion lift performance across mobile and web to advertising clients.

Troubleshooting telecom issues in real time
MapD is being tested for real-time analysis on streaming call records and server logs with one of the largest telecommunications companies in the U.S. The company's analysts use MapD to visually correlate call records with server performance data to determine in real time how network traffic is affecting load on the company's servers. They are able to instantly drill down to an individual cell tower, quickly determining if there are any malfunctions or if a device update for a particular handset is causing abnormal load on the network. Faster identification of issues helps the company reduce customer outages and more efficiently deploy technicians.

Finding hyper-regional fashion trends
A team of hundreds of analysts at a Fortune 500 apparel company are using MapD to interactively analyze historical sales transaction records to assess product demand and to help determine future inventory needs. With MapD's platform, they can query several billions of data rows in milliseconds, a significant improvement over the several minutes required by their previous OLAP tools. The company has used MapD's lightning-fast query processing and visualization capabilities to discover store-level fashion preferences and optimize their shipments with new efficiency.

Next steps

MapD combines hyper-optimized software with the fastest available hardware to create the world's fastest solution for data exploration and analytics, 1,000x faster than existing solutions. While such performance increases can be instrumental for achieving a desired level of performance within a smaller cost, space, and energy footprint, we hope we have also made it clear that the speed of such a system enables a paradigm shift in how analysts can explore the expanding data universe.

MapD plans to continue early commercial pilots through Q2 2015 before expanding its private beta to new organizations. Please select a next step to explore below:

Experience data exploration for yourself.
Find patterns in a half-billion tweets over a three-year period with MapD's live demo.

What could you do if your data was 1,000x faster?
Apply to MapD's pilot program >

MapD.com
info@mapd.com
@datarefined

or follow us on