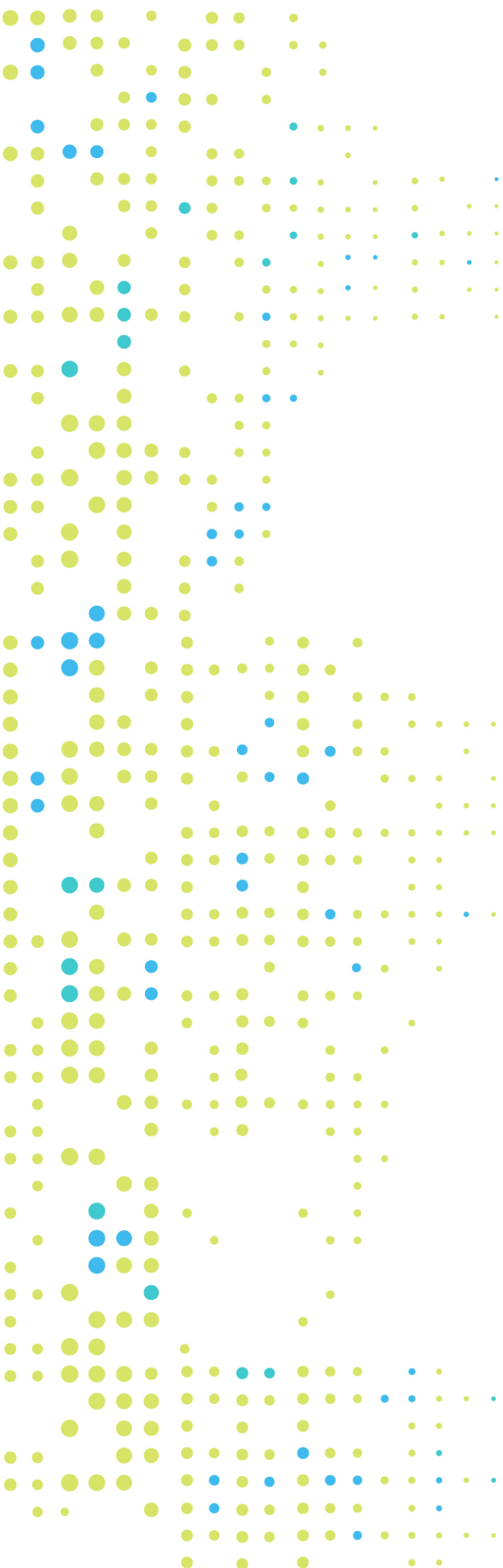


MAPD

MapD Technical Whitepaper

The world's fastest platform for data exploration

Summer 2016

"Somewhere, something incredible is waiting to be known."

Carl Sagan, astronomer

What would be possible if you could explore your data 100x faster?

As technology advancement creates an explosion of data—billions of tweets, mobile usage statistics, sensor data, financial information — a new universe is emerging.

Hardware and software limitations have left data scientists stuck on the shore of this new "data ocean"—unable to find signal in the noise, even with clusters of hundreds of servers.

MapD's breakthrough system is changing this, enabling data queries up to 100x faster than anything previously developed. By leveraging the parallel power of modern graphics processing units (GPUs), MapD offers an immersive, lag-free experience for analysts to explore large datasets interactively and in real-time.

Originally designed to render video games, GPUs have evolved into general purpose computational engines that excel at performing 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.

With MapD's software and GPU compute power, 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. With data throughput approaching three terabytes per second on a single server, analysts and data scientists are now free to interact with multi-billion row datasets more fluidly, more creatively and more productively.

This whitepaper 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 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 then outlines next steps for interested data explorers.

Founded in 2013
Based in San Francisco, CA

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

Seymour Cray, inventor

Lightning fast data visualization and a pioneering SQL database

As Seymour Cray notes, the whole system needs to be fast. In the modern world of analytics, that means two things: query speed and visualization. As a result, MapD packages two primary capabilities: the MapD analytical database and the MapD Immerse visualization platform to deliver, what we believe to be, the world's fastest data discovery and exploration solution.

Database

The MapD database 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. The database supports SQL and can be queried in the console, via bindings to major programming languages, or via Open Database Connectivity (ODBC).

The database can operate in single or multi-node configurations. The single node configuration will support eight NVidia K80 cards, with a total of 192GB of GPU RAM. It can typically handle a dataset of one to three TB in raw size within GPU RAM, and larger datasets of 10 - 15TB can be handled by multi CPU RAM at still-impressive speeds.

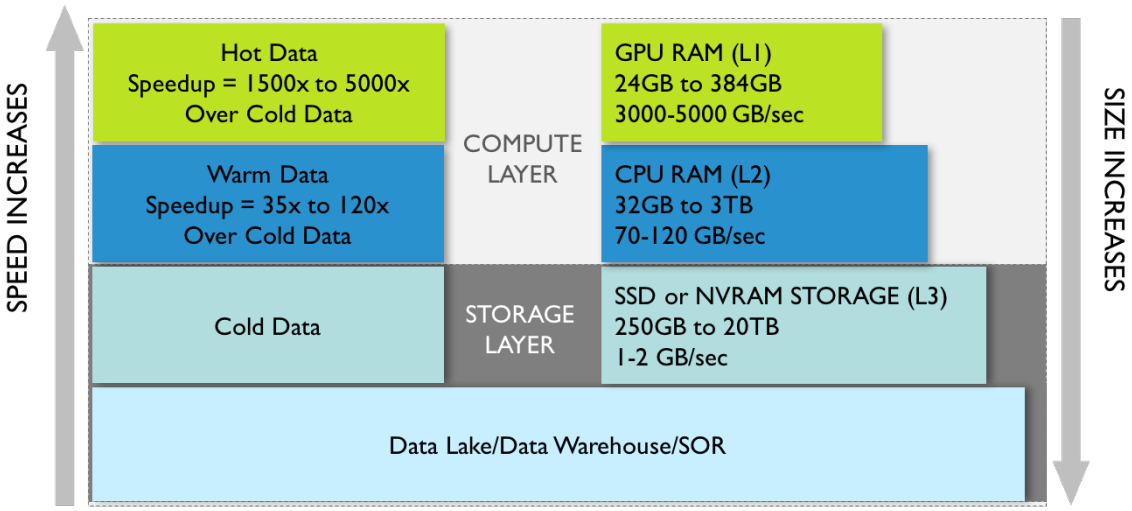
In order to extract the maximum from this powerful hardware platform, MapD has invested heavily into optimizing our code such that a wide range of analytic workloads run optimally on GPUs. In particular, we have focused on enabling common SQL analytic operations, such as filtering (WHERE) and segmenting (GROUP BY) to run as fast as possible. One significant innovation in this regard involves a JIT (Just-In-Time) compilation framework built on LLVM. LLVM allows MapD to transform query plans into architecture-independent intermediate code (LLVM IR) and then use any of the LLVM architecture-specific "backends" to compile that IR code for the needed target, such as NVIDIA GPUs, x64 CPUs, and ARM CPUs.

This approach to query compilation is far more efficient from both a memory bandwidth and/or cache space perspective. By pre-generating compiled code for the query, MapD avoids many of the inefficiencies of traditional virtual machine or transpiler approaches.

The result is that compilation times are much quicker using LLVM –generally under 30 milliseconds for entirely new queries. Furthermore, the system can cache templated

versions of compiled query plans for reuse. This is important in situations where our visualization layer is asked to animate billions of rows over multiple correlated charts.

While MapD's innovative use of LLVM is notable, we have also optimized the memory and compute layers to deliver unprecedented performance.



A key pillar of the MapD architecture is to keep the hot data in GPU memory as much as possible. Other GPU systems have taken the approach of storing the data in the CPU memory, only moving it to GPU at query time. Unfortunately, any gains such systems get by executing on the GPU is often counteracted by the transfer overhead of moving the data over the PCIe bus. MapD instead caches the most recently touched data in the ultra-fast video RAM of all available GPUs, which can amount to up to a quarter terabyte per server.

Another important component of MapD's performance is how it vectorizes (i.e. parallelizes) query execution 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.

The net result of these, and other innovations, is one of the world's fastest databases.

The impact to organizations is considerable. From driving productivity to eliminating risk, enhancing creativity and encouraging exploration, faster databases are a critical part of that equation. The backend, however, is just part of the story.

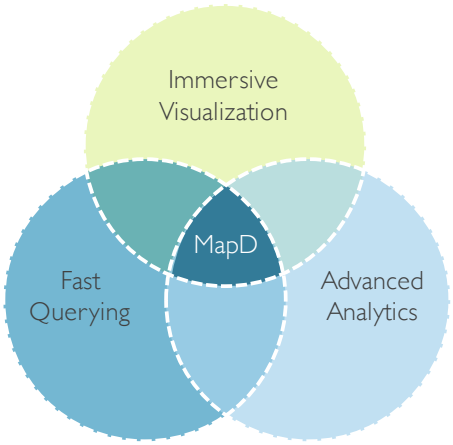
Visualization

MapD's Immerse visualization platform uniquely leverages the power of the MapD database to provide both complex data visualizations (GIS representations, scatterplots, data animations, and more) and standard reporting charts (line, bar, and pie charts, among others) in-browser. Complex, data-rich visualizations and simpler charts can be placed alongside one another within a single dashboard, providing multi-dimensional insights into large datasets. Simpler charts are rendered in the web browser, while complex renderings of large datasets are fetched from the MapD backend, where they can be generated in milliseconds.

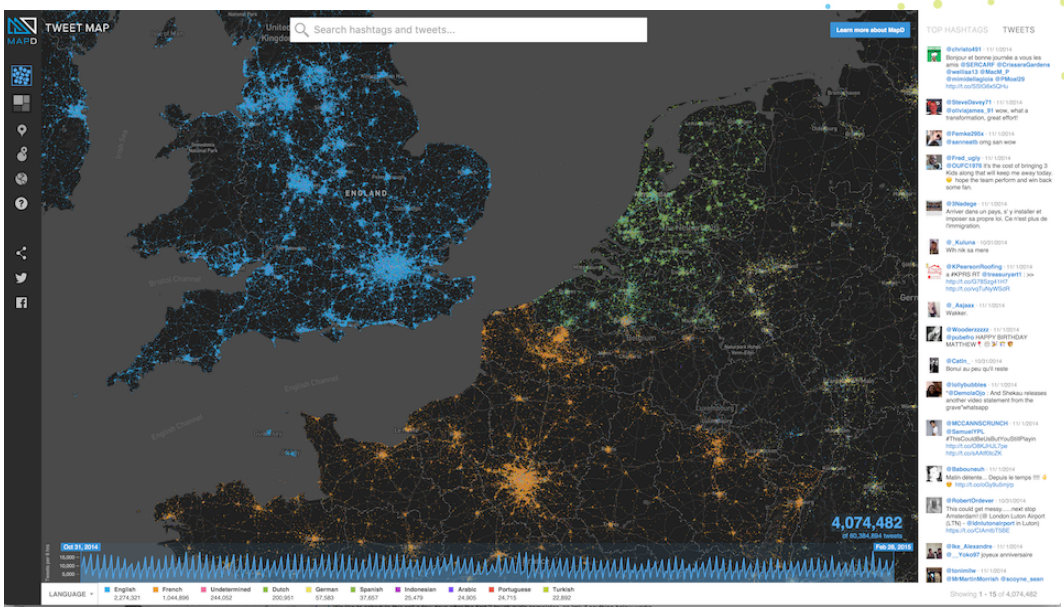
The capacity to leverage the full graphics pipeline of multiple GPUs for rendering visualizations is an area where MapD has developed considerable intellectual property, distinguishing us from other technologies that rely on CPU-driven visualizations.

Network bandwidth is a bottleneck for uncompressed data that might be required for certain visualizations, impacting our goal of providing real-time interactivity. As a result, MapD has developed certain techniques around server-side GPU rendering that have several distinct advantages, many of which are enhanced by our use of GPUs.

First of all, large datasets can be rendered and then rasterized on the server, reducing the data needing to be sent over the network to common compressed images/video streams. Even when rendering only a few million points, sending the raw data over the network to the client for the x, y, color and size channels can consume hundreds of megabytes, while a compressed PNG might weigh in at less than a hundred kilobytes.



Secondly, and even more importantly, the native graphics pipelines of the GPU enable the use of the CUDA/GL interoperability API to map GL buffers to CUDA space for an optimal query-to-render pipeline. Eliminating the copy requirement is necessary to deliver high frame rate experiences and these buffers fit into MapD's query system so naturally that results can be rendered directly without any additional copy or reduction operations.



Since it doesn't always make sense to render all charts on the server, we adopted a hybrid approach, leveraging functionality from both the frontend and the backend. For example, basic chart visualizations that require minimal data are rendered on the frontend with D3, whereas, the aforementioned complex charts are rendered server-side on the GPUs.

On the other hand, MapD employs OpenGL on the server-side to render visualizations, such as, point maps which can involve millions of points, passing the result to the frontend client as a compressed PNG. MapD can then overlay these rendered images on a base map such as MapBox GL, which offers a low-latency, resolution-independent, vector tile format that keeps network demands to a minimum and delivers near-instant rendering performance.

The result is an innovative, hybrid system that combines the agility of a lightweight and efficient frontend with the parallel power and rendering capabilities of a GPU database backend.

Additionally, the MapD Immerse frontend is built around the cross-filter model, where applying a filter to one data attribute in the dashboard applies it all other attributes as well, allowing for intuitive and seamless drill-down and correlation analysis. The frontend also allows for export of subsets of the data in CSV form, for import into other software packages like R or Excel.

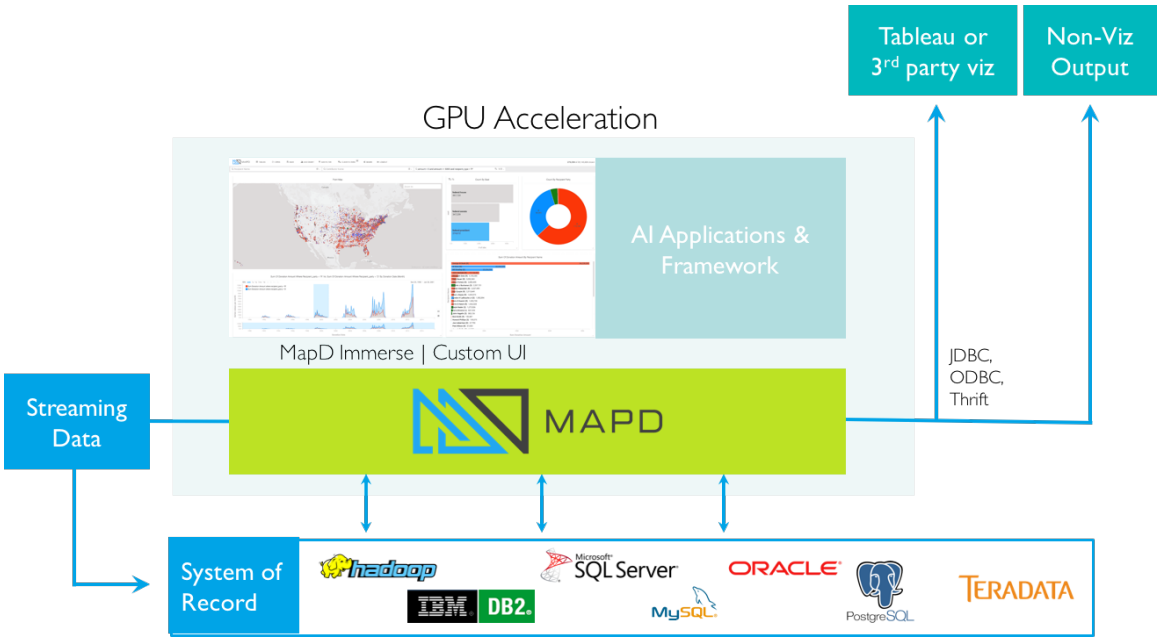
An additional benefit of performing and visualizing complicated analytics on data *in situ* is that since the relevant data is already cached on the GPUs, MapD does not need to copy the query result set before rendering it (using the GPUs) or using it as input to a follow-on machine learning algorithm. Also, while MapD's database and Immerse frontend offer exceptional performance, we also support other frontends via industry standard connectors.

MapD's technology framework

MapD data analytics and visualization platform is designed to operate in a number of different configurations and environments from cloud deployments to on premise deployments.

MapD is rarely deployed as the System of Record, rather, MapD is deployed alongside or on top of existing infrastructure; be it Hadoop, Teradata, HANA or other datastores. From this vantage point, MapD will ingest and operate on the data at GPU speeds.

MapD can also be deployed in a streaming configuration whereby it receives some time-bound portion of data (two weeks, a month, etc.) and makes it available to the organization for exploration. This option is often employed in IOT, social media, financial services and energy/utility applications.



Benchmarks

Benchmarking offers the most transparent way to determine the validity of performance claims. MapD has performed our own benchmarks in addition to independently executed benchmarks.

One of those independent benchmarks has been performed by the well-known blogger and database consultant, Mark Litwintschik. Mark has tested MapD under a variety of hardware configurations, from consumer grade Titan X cards, to enterprise grade Tesla K80s. Mark uses a well-known dataset in all of his benchmarks. It is often referred to as "the taxi dataset," and is a collection of 1.2 billion individual taxi, limo + Uber trips from January 2009 through June 2015 that was compiled and released by the New York City Taxi & Limousine Commission.

His process and results can be found on his blog (tech.marksblogg.com) but to summarize, Mark found that MapD, running on eight Nvidia Tesla K80s was 55x faster than other large CPU clusters tested including Amazon Redshift, BigQuery, Elastic, PostGreSQL and Presto. Additionally, Mark's testing found that even consumer grade GPU hardware such as four Nvidia Titan X cards outperformed CPU-constrained solutions by 43x.

MapD has run its own benchmarks using multiple well-known dataset of US flights. The configuration and queries used were as follows:

Data source:	10x copy of flights dataset (1.2B rows) at http://stat-computing.org/dataexpo/2009/the-data.html
Query 1	`select carrier_name, avg(arrdelay) from flights group by carrier_name`
Query 2	`select origin_name, dest_name, avg (arrdelay) from flights group by origin_name, dest_name`
Query 3	`select date_trunc(month,dep_timestamp) as ym, avg (arrdelay) as del from flights group by ym`
Query 4	`select dest_name, extract(month from dep_timestamp) as m, extract(year from dep_timestamp) as y, avg (arrdelay) as del from flights group by dest_name,y,m`
Query 5	`select count(*) from flights where origin_name='Lambert-St Louis International' and dest_name = 'Lincoln Municipal'`

System configurations:

MapD:	1 machine (8 core, 384GB Ram, 2 x 2TB SSD, 8 Nvidia K40)
In-memory DB 1:	10 machines (16 core, 64GB Ram, EBS storage, m4, 4xlarge)
In-memory DB 2:	3 machines (32 core, 244GB Ram, 2 x 320GB SSD, r3.8xlarge)
Hadoop OLAP:	10 machines (16 core, 64GB Ram, EBS storage, m4, 4xlarge)

We actively encourage our clients and prospects to benchmark their solutions with data and queries that they expect to use.

"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. MapD's clients have explored commercial applications for the technology across a number of different use cases and solution sets:

Finding patterns in social advertising

A social network with billions in annual advertising revenue employs 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 deployed 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 micro-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.

Developing the perfect plan

The leading targeted television advertiser has deployed MapD across its organization to accelerate the speed at which it uncovers insights from its massive proprietary database of television viewing behavior, demographics, spending behavior and ratings data. The ability to query in real time and to display those results within and across DMAs encourages real time collaboration and discussions with customers – resulting in more opportunities to win business.

Developing informational advantage

A leading quantitative hedge fund has adopted MapD to speed the development and validation of investment theses. The firm, which has assembled a massive store of proprietary data on everything from smartphone application usage to the economic output of China was constrained by traditional CPU-bound solutions that often labored for tens of minutes to return even moderately complex queries. Using MapD, the firm was able to achieve performance increases of 120x over a 20 node Impala cluster resulting in speed of thought data exploration –boosting analyst performance, idea generation and hypothesis validation.

Conclusion

MapD combines hyper-optimized software with the fastest available hardware to create the world's fastest (up to 100x faster than existing solutions) and most efficient solution for data exploration and analytics. Such improvements yield unprecedented performance within a smaller cost, space, and energy footprint, and represent a paradigm shift in an organization's ability to rapidly derive insights from their data.

Experience the new paradigm in data exploration for yourself.

Explore 180M tweets, interact with 30 years of flight data and visualize every political donation since 1990 with MapD's collection of live demos.

Contact us to learn more about MapD as a solution for your business, at info@mapd.com.

MapD Technologies, Inc.
One Front Street Suite 2650
San Francisco, CA 94111