Whitepaper

GPU-accelerated Analytics

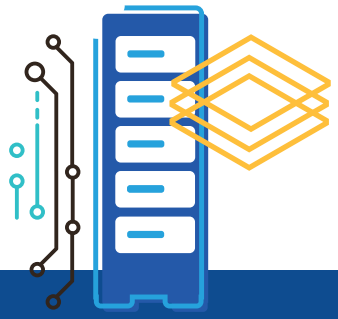
Big Data Analytics at Speed and Scale

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

—Carl Sagan, astronomer

Table of Contents

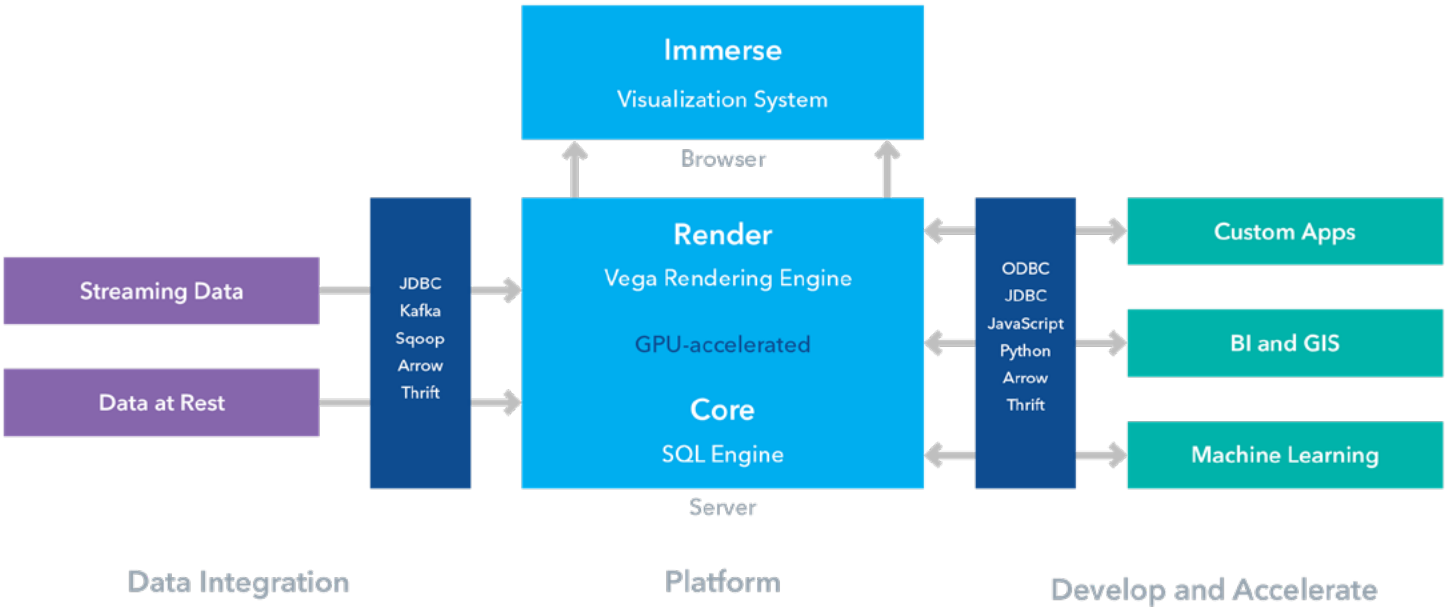
What would be possible if you could explore your data 100x faster?



Executive Summary

We live in a world of exponentially growing data. Billions of social and internet interactions, mobile usage logs, sensor streams, and transactions remake our information universe on a daily basis. Mainstream hardware and analytics software, designed for a past era, are too slow to be effective with the scale of today's data.

Businesses and governments respond to this speed-at-scale challenge with ever-expanding hardware footprints and armies of data wranglers. Yet lag times and poor visual interaction with the data still frustrate analysts and data scientists and harm their productivity. Decision-makers receive insights that are outdated, incomplete, and lacking the detail that they need to move forward with precision.



OmniSci: a vertically-integrated platform for querying, rendering and visualizing big data

The OmniSci platform, overcomes these challenges of speed, scale and interactivity. Born out of research at MIT, OmniSci is a breakthrough analytics technology, the first to harness the massive parallel processing and visual rendering power of graphics processing units (GPUs) which have evolved from their video game origins to power the latest innovation in artificial intelligence and general purpose analytics.

OmniSci delivers an immersive, instantaneous and interactive way to explore massive datasets in real-time. Using the OmniSci platform, teams can query and visualize datasets with many billions of rows and answer their analytical questions in milliseconds. This transforms the way customers conduct operational and geospatial analytics, data science and big data exploration.

OmniSci users can create interactive dashboards displaying dozens of distinctive attributes that can be correlated and cross-filtered in real-time. When one filter is adjusted, all related charts, graphs and maps immediately redraw within that context.

With data throughput approaching six terabytes per second on a single server, users do not need to wait on the technology before asking their next question. The OmniSci platform removes the technological performance tax that used to constrain their curiosity. For this reason, a growing set of companies is standardizing around the open source OmniSci Core SQL engine—turning to OmniSci to accelerate their existing analytical applications.

This whitepaper begins with a technical overview of the OmniSci platform and the three components that make it so fast at scale: the OmniSci Core SQL engine, the OmniSci rendering engine and the OmniSci Immerse visualization system. It then describes solutions that OmniSci enables, with specific examples of how OmniSci fundamentally transforms the work of going from data to insight to action.

We describe the OmniSci partner ecosystem, the importance of OmniSci Core as an open source engine and our leadership in creating a common data framework for end-to-end analytics on the GPU. Finally, we share some real-life examples of how our customers transform their organizations with the OmniSci platform.

This white paper begins with a technical overview of the OmniSci platform and the three components that make it so fast at scale: the OmniSci Core SQL engine, the OmniSci rendering engine and the OmniSci Immerse visualization system. It then describes solutions that OmniSci enables, with specific examples of how OmniSci fundamentally transforms the work of going from data to insight to action.

OmniSci Technology

OmniSci Core: The World's Fastest Open Source SQL Engine

Omnisci designed its open source OmniSci Core SQL engine so that analysts could query big data with millisecond results.

OmniSci designed its open source OmniSci Core SQL engine so that analysts could query big data with millisecond results. With OmniSci Core, they can ask and answer at least one hundred times more questions in the same amount of time as it would take for one query to complete on a legacy, CPU-only platform.

OmniSci Core operates in single or multi-node configurations. The single node configuration supports eight NVIDIA Tesla P40 cards, with a total of 192GB of GPU RAM. It can cache a dataset of 1 to 3 raw terabytes within GPU RAM. Larger datasets of 10 to 15 terabytes can still be processed using CPU RAM, with far faster results than other CPU-based analytic tools. With distributed configurations, query speeds scale linearly by adding more GPU nodes as the data accumulates.

The Technical Drivers of Query Performance at Speed

Native SQL

OmniSci uses standard SQL, but returns query results one hundred times faster than CPU-based analytic platforms. Analysts and data scientists can still rely on their existing SQL knowledge, querying data using industry-standard SQL, via any of these methods:

- interaction with the OmniSci Immerse visualization system,
- from the native Command-Line Interface (CLI) via bindings to major programming languages, including a DB API-compliant Python client, or
- with Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers

Rapid Query Compilation

OmniSci was architected for the GPU. In particular, our founders focused development to enable common SQL analytic operations such as filtering (WHERE), segmenting (GROUP BY) and joining (JOIN) to run as fast as possible with native GPU speed. OmniSci engineers also had the benefit of developing the code in a new big data era—they had the benefit of accumulated experience with past CPU-based systems. For this reason, OmniSci code also performs well on CPUs.

One significant innovation in this regard involves a Just-In-Time (JIT) compilation frame-work built on Low-level Virtual Machine (LLVM). LLVM allows OmniSci to transform query plans into an architecture-independent intermediate representation code (LLVM IR). It uses any of the LLVM architecture-specific "back-ends" to compile the IR code for the needed target, such as NVIDIA GPUs, x64, POWER and ARM CPUs. This approach to query compilation is far more efficient for both memory bandwidth and cache space. By pre-generating compiled code for the query, OmniSci avoids many of the inefficiencies of traditional virtual machine or transpiler approaches.

Using LLVM, compilation times are much quicker—generally under 30 milliseconds for entirely new queries. The system can also cache templated versions of compiled query plans for reuse. This is important in situations where users are leveraging OmniSci Immerse to cross-filter billions of rows over multiple correlated visualizations.

Advanced Memory Management with Three-tier Caching

In addition to its innovative use of LLVM, OmniSci Core also optimizes the memory and compute layers to deliver unprecedented performance. OmniSci architects designed OmniSci Core to keep hot data in GPU memory for the fastest access possible. Other GPU systems have taken the approach of storing the data in CPU memory, only moving it to GPU at query time.

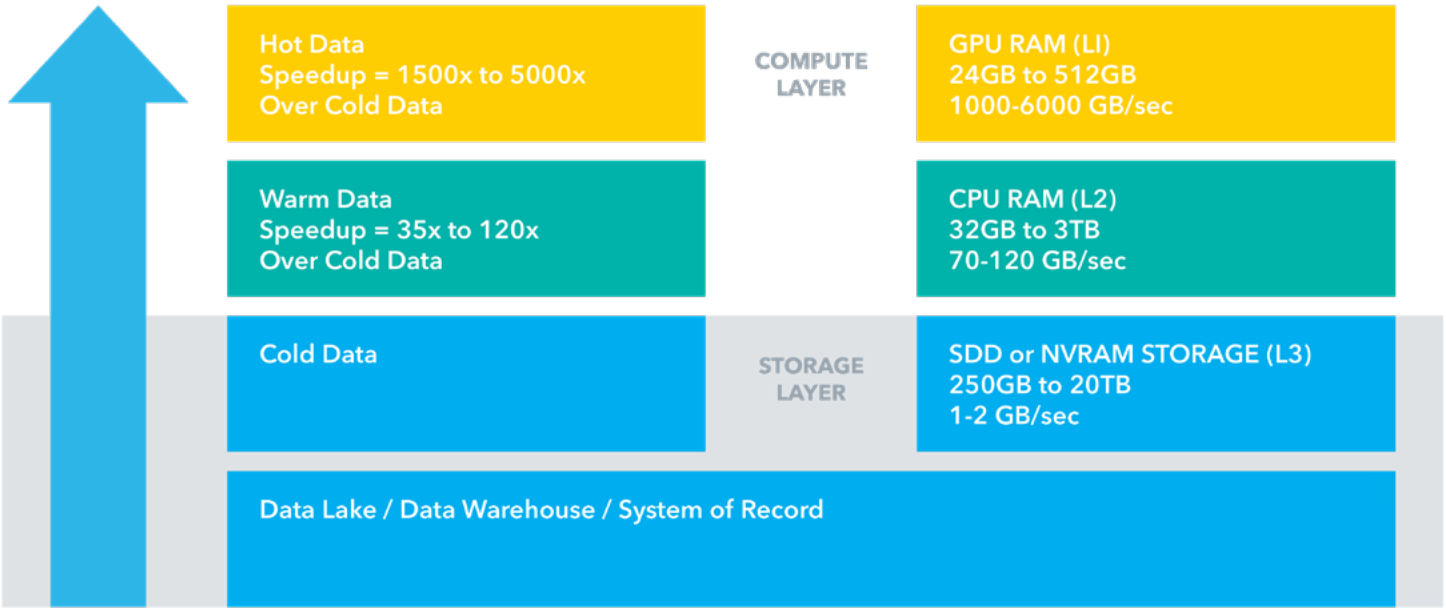
Unfortunately, this means that those other systems give back gains from GPU parallel velocity through the transfer overhead of moving data over the PCIe bus. OmniSci Core avoids this transfer inefficiency by caching up to 512GB of the most recently touched data in the GPU's ultra-fast video RAM. OmniSci Core can also leverage the NVIDIA® NVLink™ technology to accelerate CPU-to-GPU data transfers up to 2.5 times faster than systems without NVLink, now available on IBM OpenPOWER LC servers.

OmniSci designed its open source OmniSci Core SQL engine so that multiple analysts could query big data with millisecond results.

Query Vectorization and Hybrid Execution

Another important component of performance is vectorization (that is, parallelization) of query execution. Vectorized code allows a processor to compute multiple data items simultaneously. This is necessary to achieve optimal performance on GPUs, which can com-prise thousands of execution units. Optimizing vectorized execution also translates well to CPUs, which increasingly have "wide" execution units capable of processing multiple data items at once.

OmniSci Core parallelizes computation across multiple GPUs and CPUs. The SQL engine can also execute purely on CPUs, delivering superior performance even on CPU-only systems.



OmniSci optimizes memory and compute layers with three-tier caching

Native Geo Data Types for Geospatial Processing at Scale

OmniSci Core can store and query data using native Open Geospatial Consortium (OGC) types, including POINT, LINESTRING, POLYGON, and MULTIPOLYGON. With native geo type support, analysts can query geo data at scale using a growing number of geospatial functions. This opens up a wide range of use cases for geospatial analysts, who use the full power of GPU parallel processing to quickly and interactively calculate distances between two points or analyze characteristics of geometries within billions of polygons.

High Availability (HA)

The goal of HA is to meet an organization's service level agreements for performance and uptime. High Availability configurations allow a set of OmniSci servers that are running together in a High Availability Group to be synchronized in a guaranteed way.

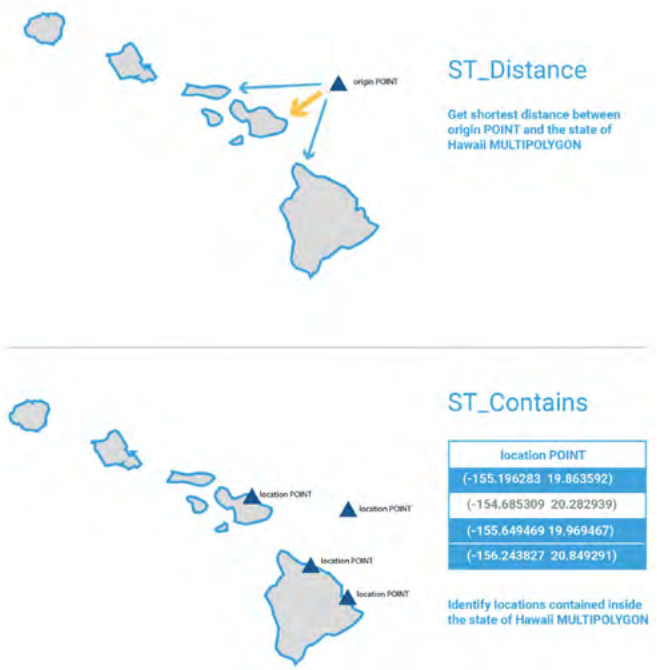
In conjunction with a load balancer, OmniSci ensures that service will remain available in the event of a server failure. As HA group members receive updates, back-end synchronization orchestrates and manages replication, then updates the OmniSci servers in the HA group using Kakfa topics as a distributed resilient logging system.

If a OmniSci server becomes unavailable, the load balancer redirects traffic to a different available OmniSci server, preserving availability even with a reduction in capacity.

While multiple servers are active in an HA group, average response times tend to improve, due to the efficient distribution of query load across the members. A load balancer distributes users across the available OmniSci Servers, improving concurrency and throughput as more servers are added.

Distributed Scale-out

The OmniSci distributed scale-out configuration allows single queries to span more than one



physical host when data is too large to fit on a single machine. Within one node, OmniSci uses a shared-nothing architecture between GPUs. When a query is launched, each GPU processes a slice of data independently from other GPUs. Even though multiple GPUs reside within a single machine, data is fanned out from CPU to multiple GPUs and then gathered back together onto the CPU.

OmniSci Enterprise edition includes a distributed setup that efficiently transports data between the leaf nodes and aggregator node of a distributed cluster. While architecting OmniSci for scale-out, our engineers were conscious of preserving the high levels of performance characteristic of the OmniSci platform by minimizing overhead in sending data over the wire between nodes. Rather than fully converting data into idiomatic Thrift (our binary protocol of choice), we instead broadcast compressed binary data, virtually eliminating any conversion cost.

In addition to efficiently transporting query result data, OmniSci also efficiently transports image data between nodes, by compressing the image's color channel information before sending over the wire for final compositing on the aggregator node.

The OmniSci Core SQL engine, however, is just one third of the story. The OmniSci platform also includes the OmniSci rendering engine and the OmniSci Immerse visualization system. Each platform component was architected from the beginning to take full advantage of the GPU's inherent strengths for parallel data processing and visualization.

A distributed configuration also provides faster data load times. Import times speed up linearly with the number of nodes since OmniSci Core has no centralized index to maintain. Reads from disk also benefit from similar acceleration in a distributed configuration.

An End to Indexing, Downsampling, and Pre-aggregation

Analytic platforms designed for a previous era did not need to account for big data or the pace of today's decision-making. As those mainstream analytics tools began to collapse under the weight of large datasets, traditional data warehouses resorted to techniques such as indexing, downsampling or pre-aggregation for a modicum of performance at scale.

By comparison, OmniSci Core leverages the massive performance gains achievable with GPUs, coupled with its best-in-breed execution engine. Working together, they deliver instantaneous query execution without the need to index, downsample or pre-aggregate beforehand. The result is that OmniSci queries are predictably fast, with results in tens to hundreds of milliseconds even over many billions of records.

This approach has two major upsides. First, organizations deploying OmniSci do not need to spend significant time and resources modeling the data, they simply load it and immediately enjoy fast SQL. Second, the lack of any significant preprocessing of data means that OmniSci can typically load data much more quickly than other systems, a particular boon for streaming and high-ingest-rate scenarios.

Query Benchmark Comparisons with CPU-based Platforms

The innovations detailed above make OmniSci Core the world's fastest open source SQL engine. In November 2017, noted Big Data consultant Mark Litwintschik published benchmark comparisons on "The Taxi Dataset"—1.2 billion individual taxi trips made available by the NYC Taxi and Limousine Commission (TLC). The following table shows a summary of the benchmark results.

Litwintschik tested OmniSci under a variety of hardware configurations, from consumer grade Titan X cards to enterprise grade Tesla K80s. To summarize, Mark found that OmniSci running on eight NVIDIA Tesla K80s was up to 41,000 times faster for the first query than other large CPU clusters he tested. Those CPU-based systems included: Amazon Redshift, BigQuery, Elastic, PostgreSQL, Presto and Apache Spark.

Of course, this taxi trip data is simply an illustration of how OmniSci performs at large scale. In actual practice, OmniSci Core helps businesses and government agencies analyze the data that relates to their objectives. With the same unprecedented speed to insight, they can improve productivity, reduce risk, foster creativity and make fast analytics a sustainable competitive advantage.

The OmniSci Core SQL engine, however, is just one third of the story. The OmniSci platform also includes the OmniSci rendering engine and the OmniSci Immerse visualization system. Each component of the platform was architected from the beginning to take full advantage of the GPU's inherent strengths for parallel data processing and visualization.

Setup	Query 1	Query 2	Query 3	Query 4
OmniSci & 8 NVIDIA Pascal Titan Xs	0.021	0.053	0.165	0.510
OmniSci & 8 NVIDIA Tesla K80s	0.027	0.083	0.163	0.891
OmniSci & 4-node g2.8xlarge cluster	0.028	0.2	0.237	0.578
OmniSci & 2-node p2.8xlarge cluster	0.034	0.061	0.178	0.498
OmniSci & 4 NVIDIA Titan Xs	0.036	0.131	0.439	0.964
kdb+/q & 4 Intel Xeon Phi 7210 CPUs	0.051	0.146	0.047	0.794
ClickHouse, Intel Core i5 4670K	1.034	3.058	5.354	12.748
Redshift, 6-node ds2.8xlarge cluster	1.56	1.25	2.25	2.97
BigQuery	2	2	1	3
Presto, 50-node n1-standard-4 cluster	4	4	10	21
Amazon Athena	6.41	6.19	6.09	6.63
Elasticsearch (heavily tuned)	8.1	18.18	n/a	n/a
Spark 2.1, 11 x m3.xlarge cluster w/ HDFS	10.19	8.134	19.624	85.942
Presto, 10-node n1-standard-4 cluster	11	10	21	31
Vertica, Intel Core i5 4670K	14.389	32.148	33.448	67.312
Elasticsearch (lightly tuned)	34.48	63.3	n/a	n/a
Presto, 5-node m3.xlarge cluster w/ HDFS	35	39	64	81
Presto, 50-node m3.xlarge cluster w/ S3	43	45	27	44
PostgreSQL 9.5 & cstore_fdw	152	175	235	368
Spark 1.6, 5-node m3.xlarge cluster w/ S3	264	313	620	961
Spark 2.2, 3-node Raspberry Pi cluster	1103	1198	2278	6446

Source: http://tech.marksblogg.com/benchmarks.html (Response times above are in seconds)

OmniSci Render

Delivering Query Results to the Front-end for Complex Visualizations

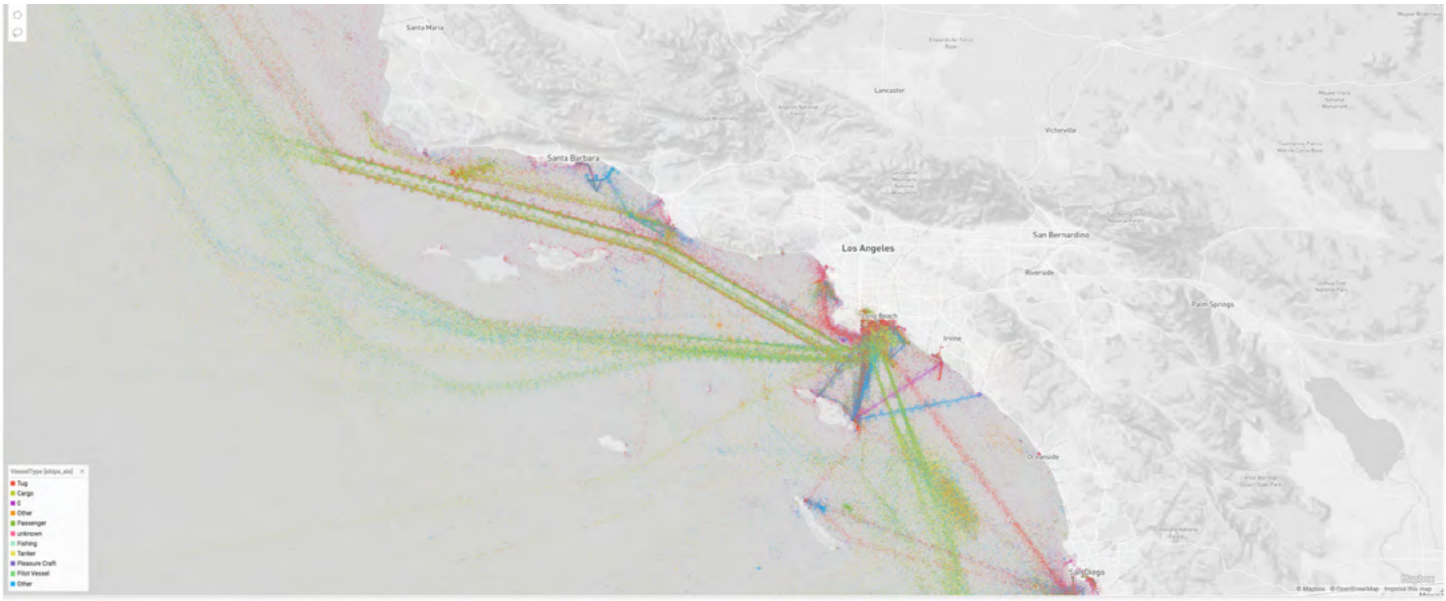
The OmniSci Render rendering engine connects the extreme speed of OmniSci Core SQL queries to complex front-end visualizations in OmniSci Immerse and custom applications. It superpowers the interactivity of OmniSci geopoint maps, geo heat maps and scatter plots.

The OmniSci Render rendering engine connects the extreme speed of OmniSci Core SQL queries to complex front-end visualizations in OmniSci Immerse and custom applications. It superpowers the interactivity of OmniSci geopoint maps, geo heat maps and scatter plots.

Server-side Rendering

OmniSci leverages the full graphics pipeline of multiple GPUs for rendering visualizations. Network bandwidth is a bottleneck for the uncompressed data required by complex charts, so OmniSci developed server-side technology to accelerate the rendering of granular data. This distinguishes OmniSci Immerse from other technologies that execute queries on the GPU but then render them on CPUs.

OmniSci not only executes the underlying SQL query that drives a visualization on the server-side GPUs, but also renders and rasterizes the query results directly on the GPU. This avoids any slowdowns associated with transferring the results from GPU to CPU and then over the network to the client.



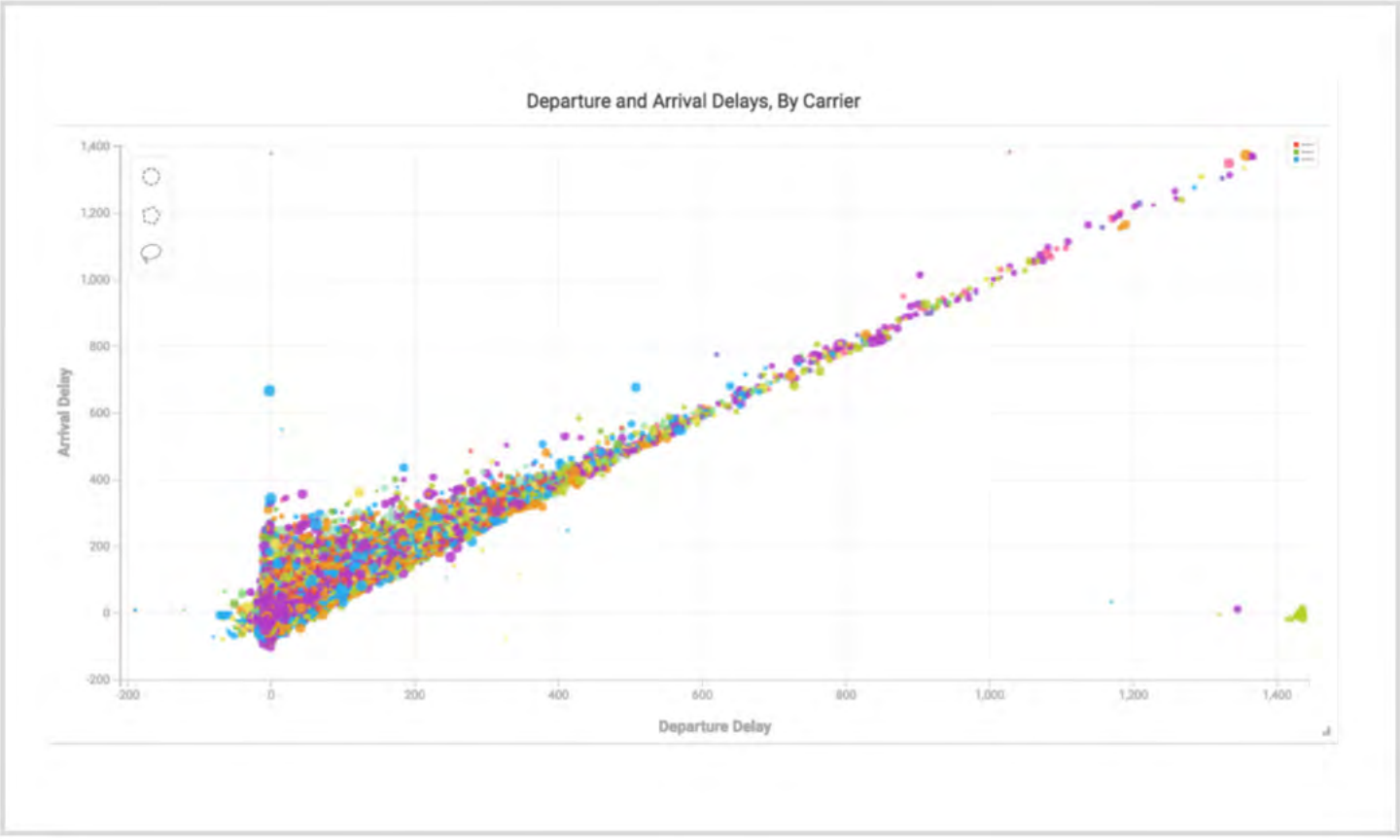
Geo chart showing millions of ship movement records

Server-side rendering works for both data points and polygons. Users can visually interact with billions of points and millions of polygons on the same geo chart, to understand granular or summary detail on events that occurred within a given geographical area.

Visualization with Vega

Complex server-side visualizations are specified via an adaptation of the Vega Visualization Grammar, a declarative API developed by the creators of D3. OmniSci Immerse generates such rendering specifications using this grammar behind the scenes, however OmniSci users can also generate custom visualizations and dashboards using the same API.

The result is an innovative visualization system that combines the agility of a lightweight front-end with the parallel power and rendering capabilities of a GPU engine. Since the relevant data is already cached on the GPU, OmniSci does not need to copy the query result set before rendering it or using it as input to a follow-on machine learning algorithm. For geospatial visualization, the Vega rendering engine API allows users to visualize billions of points, lines, and polygons on a geo-chart. It can also render geographical heatmaps. Geospatial visualizations can be overlaid on a Mapbox basemap. Users can dynamically lasso data at locations-of-interest to cross-filter just on those hand-chosen areas.



Scatter plot of airplane arrivals and departures, utilizing Vega Visualization Grammar

OmniSci Immerse

Interactive Data Visualization at Extreme Scale

While a lightning-fast SQL engine is valuable in its own right, OmniSci Immerse allows organizations to bring that speed all the way to the end user. The visualization system transforms the work of data scientists and analysts, who can now interactively explore large datasets that they've always had to approach from the command line or via incomplete and slow visualizations. OmniSci Immerse leverages the speed and rendering capabilities of the OmniSci Core SQL engine to power both standard visualizations (such as line graphs, bar charts, and pie charts) and also complex data visualizations rendered in geo-point maps, geo heat maps, choropleths, and scatter plots. All of these visualizations can be dynamically filtered together, and they refresh in milliseconds.

OmniSci Immerse leverages the speed and rendering capabilities of the OmniSci Core SQL engine to power both standard visualizations to complex data visualizations.

Polygons are displayed on a geo chart in OmniSci Immerse via the rendering API

OmniSci users can place charts and complex visualizations alongside one another within a single dashboard, providing multi-dimensional insights into large datasets. Standard charts are rendered in the web browser, while complex renderings of large datasets are fetched from OmniSci Core.

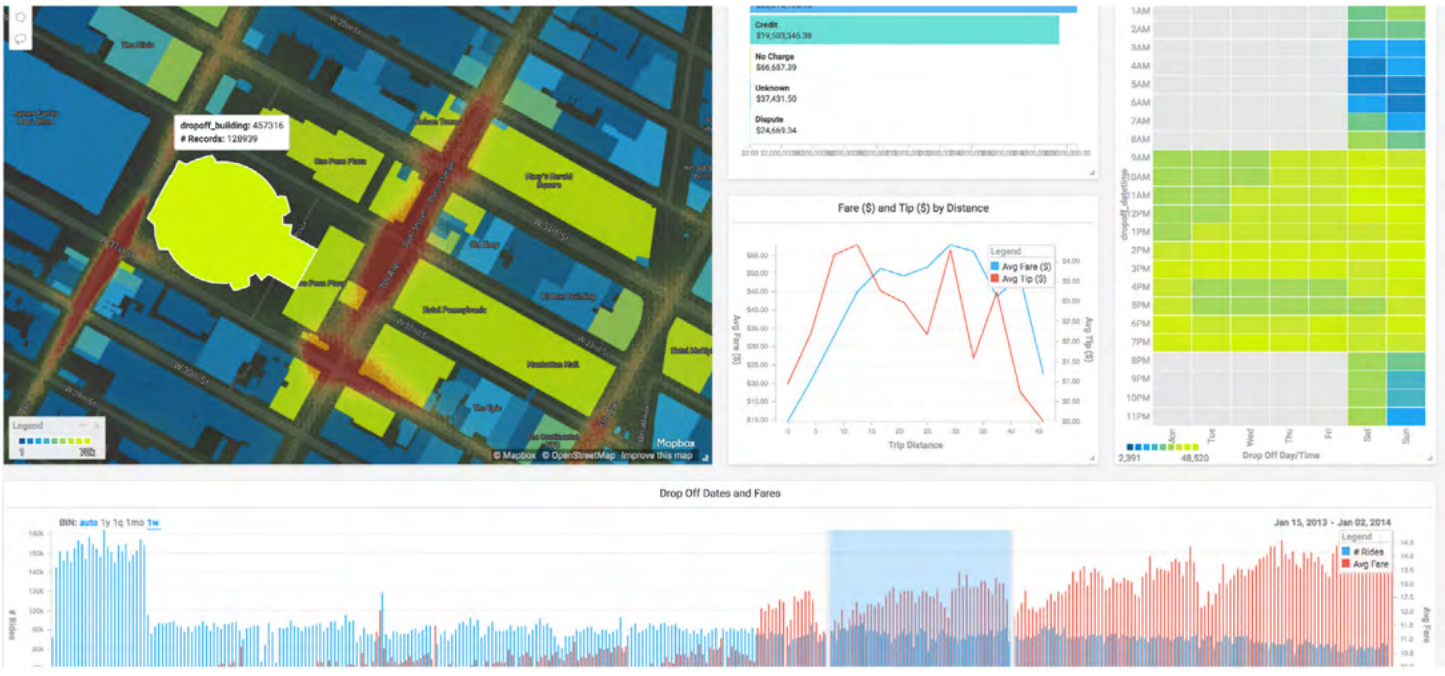
Native Cross-Filtering with Geospatial Context

In today's data-driven world, multiple stakeholders within an organization are demanding access to a large and growing range of datasets. More critically, those stakeholders want to interrogate the data and get immediate answers to their own important business, defense or intelligence questions.

The cross-filter paradigm of OmniSci Immerse meets this need for self-service data discovery. When an OmniSci user clicks on any dimension in a chart or graph, Immerse simultaneously redraws all other visualizations in a dashboard to reflect that new context. This is a transformative way to quickly find correlations and outliers in data.

The cross-filter paradigm of OmniSci Immerse meets this need for self-service data discovery. When an OmniSci user clicks on any dimension in a chart or graph, Immerse simultaneously redraws all other visualizations in a dashboard to reflect that new context.

Some OmniSci customers have multiple simultaneous users cross-filtering over datasets in the tens of billions of records, something they could have never dreamed of with conventional tools. This is especially true if interactive geo charts are shown next to more conventional bar charts or line graphs. When an analyst moves around a geo chart in OmniSci, those visualizations refresh immediately for that new location. Users can see what matters to them, in the location where it matters most.



OmniSci Immerse combines elements of GIS and BI for a new breed of interactive geospatial exploration

Dashboard Auto-refresh

Users appreciate OmniSci Immerse because its extreme speed helps them monitor and respond to real-time changes. The term "real-time" may have many different meanings for different audiences and use cases, and so the dashboard auto-refresh capabilities gives teams the flexibility to set the refresh intervals according to their own needs related to the rate of streaming data ingest.

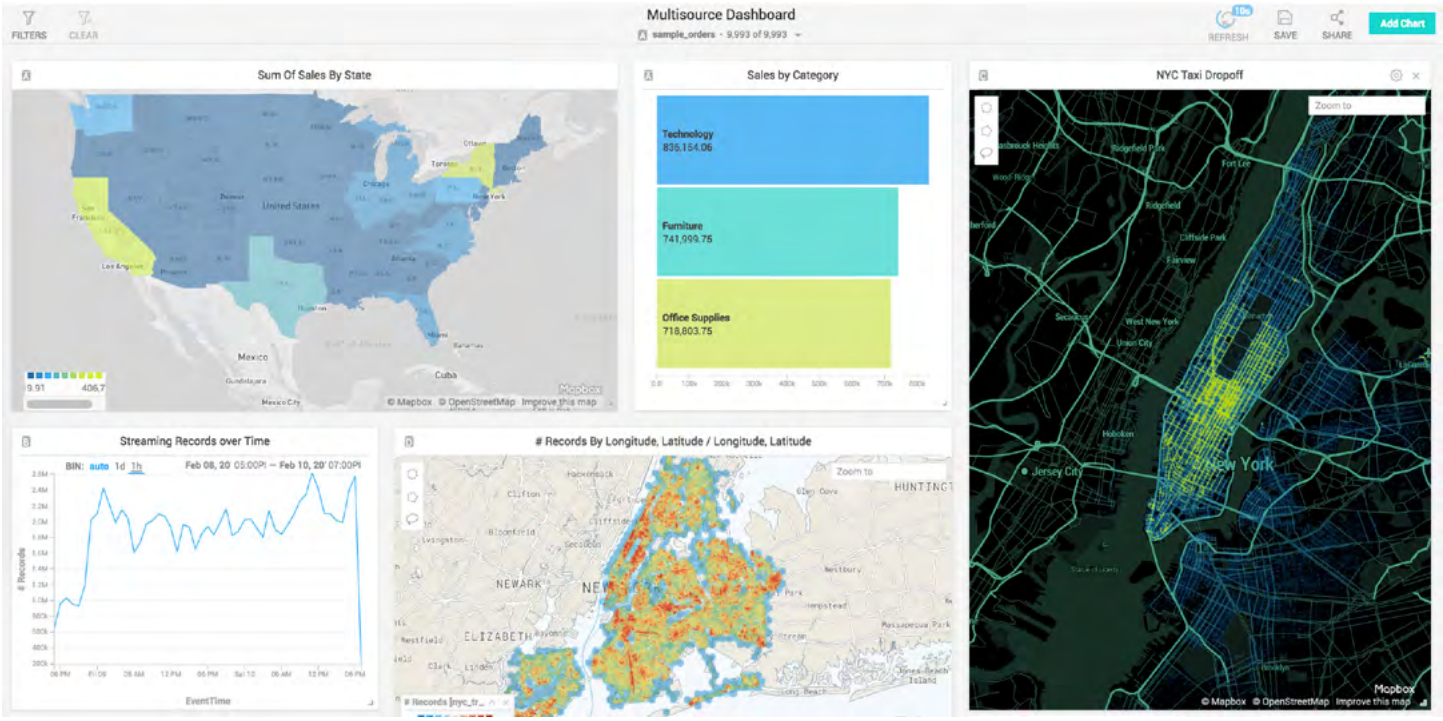
Dashboards can be set to automatically refresh every 10 seconds, 30 seconds, or at 1, 2 or 5-minute intervals.

That means millisecond queries and renderings will show a business or government agency how its operations are changing from moment to moment.

Multi-source Dashboards

OmniSci Immerse can visually display dozens of distinct datasets in the same dashboard, without having to join underlying tables. This saves data preparation time and uncovers surprising multi-factor relationships that an analyst might not think to look for in a visualization system that can handle only one data source with fewer records.

Each chart (or groups of charts) in a dashboard can now point to a different table, and filters are applied at the dataset level. Multisource dashboards expand an analyst's ability to compare across datasets, without having to merge the underlying tables (which can be time consuming).



Multi-source dashboards let analysts compare disparate data, such as Tweets and taxi trips

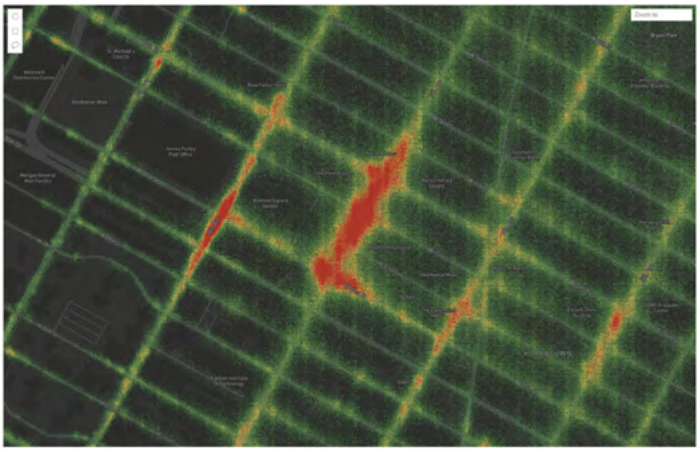
Multi-layer Geo Charts

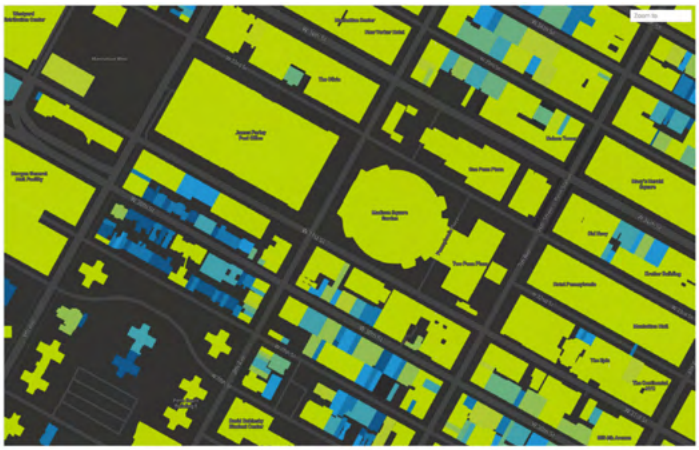
OmniSci users can create geo charts with multiple layers of data, in order to visualize the relationship between factors within a geographic area. Each layer represents a distinct metric overlaid on the same map. Those different metrics can come from the same or a different underlying dataset.

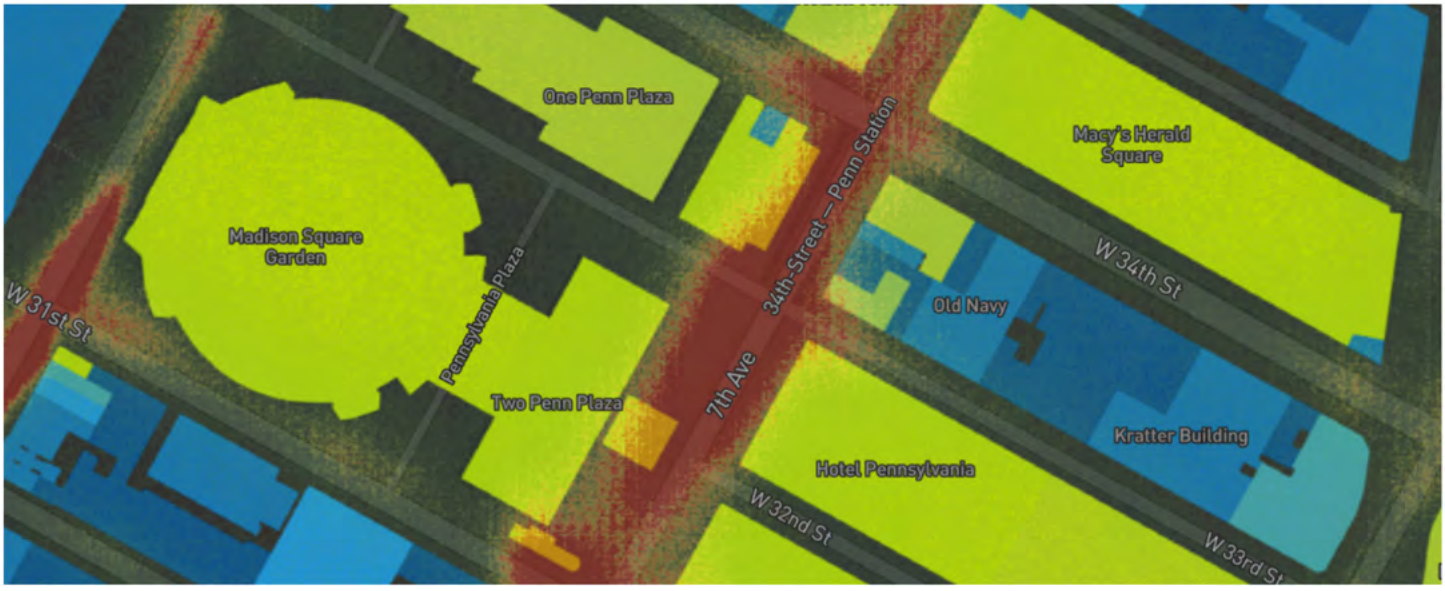
Analysts can compose multiple layers, reorder layers, choose to show or hide layers, or adjust the opacity of each layer. They can also toggle on or off the legends for any layer or turn off the legends completely.

Collaboration Across Teams

An OmniSci user can create a dashboard and share it with colleagues. If those who receive that dashboard do not have permission to see the data tables that it includes, they cannot view the dashboard. Since dashboard permissions are decoupled from data table permissions, users can share their work freely, without having to worry that someone might see sensitive information beyond their authority.







Multi-layer geo charts to compare different datasets in one map

OmniSci Solutions

Giving Business and Government the Fastest Operational Analytics, Geospatial Analytics and Data Science Capabilities

Traditional CPU-based data warehouses are often adequate for generating static reports. It used to be "good enough" to create an overnight report from the previous day's data and have it ready for a CEO or commanding officer in the morning.

But today, this is unacceptable. Decision-makers are increasingly demanding up-to-the-minute information and insights.

With OmniSci, users can immediately answer questions like these:

- How much lift is the organization getting from its new ad campaign, broken down by channel, country and demographic?
- Did a software update to a telecommunication company's cell towers lead to an increase in dropped calls, or was it a bad firmware refresh on a particular brand of handset?
- Are our troops in the right locations before launching a mission?

Traditional CPU-based data warehouses are often adequate for generating static reports. It used to be "good enough" to create an overnight report from the previous day's data and have it ready for a CEO or commanding officer in the morning.

Operational Analytics For Ongoing, Instant Decisions

OmniSci eliminates the technological barriers to achieving high-frequency decision-making on very large streams of operational data. The platform redefines operational analytics by giving analysts the power to query and visually explore multi-billion row, high-velocity datasets on their own. With immediate query responses and visualization, teams dramatically improve their situational awareness and decision-making. They no longer need to wait for tomorrow's static report before making important decisions.

Analytics used to be a two-person sport: decision-makers asked questions and analysts found answers. Now, the velocity and usability of OmniSci dashboards lets business leaders discover their own answers. That means more people can watch operations, notice anomalies, and take data-driven actions at a moment's notice.

Examples of use cases include media or advertising companies that uncover hidden insights quickly enough to improve campaign results. Risk teams at financial services companies can visually explore billions of records to spot hidden patterns of fraud they can address directly and then use to train machine learning models that automatically spot similar patterns in the future. Security Operations Centers (SOCs) can give their security analysts OmniSci to immediately respond to zero-day cyber incidents and telecom Network Operations Centers (NOCs) quickly identify network anomalies and resolve them before too many customers are affected.

Geospatial Analytics that Combines BI and GIS in One Dashboard

Today, most data has been enriched with a location element that mainstream BI tools cannot manage gracefully. GIS systems can plot points on a map, but only up to a few hundred thousand records, and usually outside the context of important nongeographical information. Geo location datasets are now big enough to overwhelm traditional GIS and BI platforms.

While OmniSci is a general-purpose analytics platform, many of our customers derive particular value out of its ability to visualize and explore large geospatial datasets interactively, at the granular level, and in the context of other charts and graphs. By virtue of OmniSci query speeds and rendering capabilities, users can visualize and cross-filter location-specific data from cell phones, self-driving cars, weather satellites or oil wells without lag.

OmniSci geospatial analytics let a telecom team query and explore CDRs with geo charts. In the military, logistics officers can visualize the location of soldiers, vehicles, and weapons and give their commanders a comprehensive, real-time view into every facet of a mission. Dynamic weather maps can show first responders, airlines, construction businesses and insurance companies how to protect people and property from imminent weather damage.

Data Science with Greater Speed and Transparency

Front-line analysts using OmniSci immediately understand and appreciate the value of its performance for ad hoc exploration of big data. In addition, more sophisticated data scientists use the power of rapid iteration to construct machine learning (ML) models. Technology partners of OmniSci may execute those ML models, but OmniSci helps build them far more efficiently.

One of the most tedious and time-consuming parts of ML modeling is feature engineering. Feature engineering is the process of building domain knowledge of a particular dataset to choose the features that make machine learning algorithms work—choosing the variables likely to be most predictive.

Data scientists who use OmniSci to explore datasets with hundreds of columns and billions of rows, have the power to regain the hours or days they used to spend feature engineering, even before getting to train the first iteration of their model.

Fast visual exploration in OmniSci helps shrink the feature engineering workflow from days or hours to minutes, and OmniSci has partnered with Anaconda and H2O.ai as founders of the GPU Open Analytics Initiative (GOAI) to build a common framework for engineering, training, running and tuning ML models.

The first project within this open source framework for GPU-accelerated analytics is the GPU Data Frame (GDF). Leveraging the GDF, a data scientist can pass the results of a OmniSci query to Anaconda's NumPy-like Python API for data manipulation with zero overhead. Then the user can build a predictive model from H2O.ai's suite of machine learning algorithms. The output of the H2O model can be imported back into OmniSci for interactive querying and visualization. If a model becomes less predictive, the data scientist can see that in OmniSci Immerse and do further feature engineering to tune it again.

OmniSci visualization also gives data science teams a much needed tool for transparency. Machine learning algorithms make predictions with minimal (or no) human involvement. The resulting algorithms may be efficient and accurate, but opaque models are often mistrusted by the executives who approve them and regulators that certify them. With OmniSci, anyone can visually explore the underlying data and data scientists can explain how they trained their models.

The "black box" is still black, but a broader audience can see and understand what data went in and compare that to the recommendations that come out.

Big Data Research

With the ability to consume large amounts of structured data from Big Data repositories, and query and visualize that data at the speed of thought, OmniSci is enabling analysts to derive knowledge from that data. Companies have invested in distributed storage platforms based on approaches like Apache Hadoop. Now many of those platforms contain petabytes of data. However, relatively slow query performance and the lack of visualization tools for interacting with that data at scale has slowed the value that those organizations realize from those investments. Large amounts of data that users can't rapidly query and visualize is not particularly useful. Now with extreme speed and granular interaction, enterprises derive more value from the data already sitting in storage.

The OmniSci Partner Ecosystem

The Importance of the Transformative Power of Open Source Software

The Importance of Open Source

On May 8, 2017, the OmniSci Core SQL engine was released as an open source project. The code is available on Github under an Apache 2.0 license. The Github repository has everything needed to build a fully-functional installation of OmniSci Core. The key components including its tiered caching system and LLVM query compilation engine, are included in the open source release.

OmniSci believes in the transformative power of open source software, both to accelerate innovation and also to encourage users to try first-hand the power of the OmniSci platform, without the fear of becoming locked in to a proprietary approach. We contributed OmniSci Core to the open source community so that our customers could integrate OmniSci into their existing data and analytics ecosystems, with an ease just not possible with a proprietary platform.

Since open sourcing OmniSci Core, our company has seen proof of that community momentum. The GOAI (mentioned above) is a perfect example. The GOAI collaboration has already created a GPU Data Frame (GDF), which makes it possible for complex, multi-party workflows to run entirely on GPUs.

Custom Applications Supercharged with OmniSci

Most organizations persist their data in data warehouses, data lakes, flat files, or systems like Apache Spark™. OmniSci provides a variety of connectors to move data from these persistent stores into the OmniSci analytics platform.

These connectors let you pull high-value data from systems as diverse as Oracle, Spark, or Hadoop, then cache it on the GPU for interactive visual analysis and machine learning. As Spark has brought orders-of-magnitude acceleration to Hadoop, OmniSci can can bring similar speed-ups to Spark and other existing CPU-based analytics systems, such as:

- Apache Hadoop: Pull data from your Hadoop data lake with a Sqoop-based connector.
- Apache Spark: Supercharge your persistent Spark data stores by loading your information into OmniSci Core to take advantage of GPU processing speed.
- ODBC and JDBC: Load records in controlled batches using the OmniSci Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC) connectors.
- Flat Files: Use the COPY FROM command to create tables in OmniSci Core from standard CSV or TSV-delimited files.
- GPU Data Frame (GDF): OmniSci Core can output the results of a SQL query into the GDF. Users can then manipulate the data with Continuum Analytics' Anaconda NumPy-like Python API, or use the data as input to the H2O.ai suite of machine learning algorithms, without the overhead of copying data to CPU. In early tests, this approach exhibited order-of-magnitude improvements in processing times compared to passing the data between applications on a CPU.
- Kafka: Use a Kafka Consumer to stream data directly to OmniSci Core for real-time analysis of data as it becomes available from any Kafka Broker.

Case Studies

Rapid time to insight is one of the most critical advantages for leaders at business, government, defense or intelligence organizations. The ability to see patterns in real-time and take action accordingly is a basic requirement for any modern, large-scale endeavor. Clients of OmniSci have implemented commercial and government applications for the technology across a number of different use cases, industries and solution sets. Here are just three examples of those successful applications.

A leading telco provider resolves network issues in real-time

A leading telecommunications provider in the US leverages the breakthrough operational analytics capabilities of OmniSci for real-time analysis of its streaming network logs. The company's commercial data systems (CDS) team interacts with over 10 billion rows of new data every five days. Even at that scale the CDS team uses high-speed visual analysis to improve service availability, increase personnel efficiency, and provide industry-leading network reliability. They are able to instantly drill down to an individual cell tower, quickly determine if there are any malfunctions or if a device update for a particular handset is causing abnormal load on the network. Faster issue identification and resolution helps the company proactively triage potential customer outages and efficiently deploy technicians.

Simulmedia interactively explores ad performance with customers on the phone

Simulmedia uses data to optimize advertising on national television. The company has deployed OmniSci across its organization to accelerate insight discovery from its massive proprietary database of television viewing behavior, demographics, spending behavior and ratings data. The ability to query the data at will, and to interact with it in team meetings, encourages confident engagement with clients. Before OmniSci, Simulmedia sales consultants could answer client questions about ad performance, but only after running analytics and calling the customer back at a later time. Now, they can look up the answers in real-time and answer new questions with the customer on the line. This opens more opportunities to win business.

A hedge fund validates investment strategies with unprecedented insight

A leading quantitative hedge fund adopted OmniSci to speed the development and validation of investment theses. The firm assembled a massive store of proprietary data on everything from smart-phone application usage to the economic output of China. Before, the company felt constrained by its traditional CPU-bound architecture that often labored for minutes to return results to only moder-ately complex queries. Using OmniSci, the company improved the performance of its analytic models by 120X over a 20-node Impala cluster. The firm parlayed that speedup into much faster analysis, which boosting analyst performance and hypothesis validation.

OmniSci in the Cloud

OmniSci offers a range of cloud-based options. These allow a broad range of business leaders and data scientists to become familiar with our solutions, even before they themselves have articulated a strategy for GPU-enabled analytics. Here are current cloud offerings on major public cloud providers:

- **AWS Marketplace – We offer an AMI for OmniSci Enterprise on AWS. Users can easily spin up OmniSci running on an AWS P2 instance.**
- **Microsoft Azure Cloud – Users can spin up OmniSci running on a variety of N-Series configurations.**
- **Google Cloud Platform – We offer our Community and Enterprise editions for K80, Pascal and Volta GPU instances in the GCP Marketplace.**

Closing Thoughts

Today, data is created at a pace that overwhelms the capability of CPU-based solutions. Those who fall behind won't survive. As a result, a solution is needed for use cases that are too big, important and fast for platforms architected in an earlier era.

This compute and visualization inflection point has broad implications for operational analytics, geospatial analytics, data science and research and discovery on existing big datasets. In-house developers and software vendors are leveraging the OmniSci open source SQL engine—the fastest ever invented—to accelerate existing analytics for creating tomorrow's applications.

While the compute platforms of the future will run with GPUs, this fast hardware requires software designed specifically for the GPU's unique advantages. OmniSci has taken this GPU-native approach from the beginning. As a result, our software harnesses the hardware attributes that deliver analytic performance at speed, even when running on hardware with both GPUs and CPUs.

Experience for yourself the transformative power brought by this new paradigm in interactive analytics. Interact with 11 billion shipping location signals or 1.2 billion taxi rides. Explore 400 million tweets. Soar through 30 years of flight data and visualize every US political donation since 2001. Analyze every play from every NBA game over 11 seasons. All of the datasets are at your fingertips and directly in your line of sight via the OmniSci collection of online demos. These demos will show you the power of GPU-accelerated analytics.

For more information about how OmniSci can help you obtain a competitive advantage, please schedule your personal demo.

Alternatively, try a 14-day free trial of OmniSci Cloud. With a dedicated single node pre-loaded with multiple dashboards, you can practice using OmniSci and easily import your own datasets. You can also download the OmniSci Core open source code and become part of the OmniSci developer community.