Whitepaper

GEOSPATIAL BIG DATA: CHALLENGES AND OPPORTUNITIES

A GUIDE FOR GEOSPATIAL LEADERS

omni·sci

Table of Contents

Since its emergence almost 200 years ago, geospatial data has always been a means for humans to make sense of their world and solve the most daunting problems. In fact, the seeds for what we know of geospatial technology today were planted in 1832 during a cholera outbreak in Paris. French cartographer Charles Picquet created one of the first heat maps to show where the incidents of illness were concentrated- a breakthrough that has been expanded on ever since. Indeed, similar techniques are being applied today for COVID-19. The data volumes and analysis techniques have changed of course, but the ability to represent spatial and temporal patterns and convey that to others remains central.

Today we rely more than ever on geospatial data--and geospatial analysts--to solve major societal, economic and business problems. From minimizing the impact of natural disasters and planning public health initiatives, to optimizing infrastructure investments and predicting customer demand, geospatial data is playing an increasingly vital role.

As a manager of a geospatial data team, you're likely being asked to handle more data, deliver faster insights and support new applications of that data. You're also likely partnering with other analytics groups across the organization who are recognizing the value of geospatial data for their own functions. This increased demand brings challenges and, in this paper, we'll explore the most common problems encountered when utilizing geospatial data and show you how to overcome them, resulting in a robust geospatial data strategy that delivers real impact to your organization and beyond.

Figure 1

The Growth of Geospatial Data

But first, let's explore what's driving an explosive growth in both geospatial data and analytical needs. There are a number of technological trends behind the recent surge. Chief among them is the proliferation of intelligent sensors, in everything from point of sale systems to vehicles to machinery to wildlife and even consumer weather stations, which generate location-specific and timestamped transaction data. This sensor data has become a valuable asset to companies, helping them understand local market conditions or conduct preventive maintenance. On top of this, machine learning techniques are automating map generation from primary data, providing a geospatial lens where it didn't exist before.

Another trend is that microsatellites and drones are dramatically lowering the cost of geoimage acquisition, which multiplies the volume of data available. These systems are imaging the entire planet daily, and parts of it in incredible detail. Open public data available today is imaging the planet weekly at 10-60m resolution. And, as

competition steps up between satellite, aerial and drone imagery providers, we'll likely see increased resolutions, band depths and image frequency--all of which means more data.

And, when we add temporal data to geospatial, we multiply data volumes further. Looking at geospatial data over a period of several years can quickly add millions of records to the dataset. For example, a single Sentinel-2 image covering 100x100km constitutes 100m records, requiring 800MB of storage. But Sentinel-2 provides repeat coverage on average every 5 days, so a single year's worth of data for one area is 58.4GB. An area the size of California requires 2.3TB per year.

So, geospatial data is now easier to come by. But what's driving organizations to invest in their own--and incorporate external--geospatial data? In short, the answer is they can't afford to ignore it. If a commercial business fails to consider location-specific trends, its competitors will seize the advantage. Consider the revenue opportunities from delivering services that meet specific needs of customers in a particular location. And consider the cost savings from continuously monitoring assets in the field, perhaps through a digital twin strategy, and identifying problems before they happen. In government, failure to embrace the best available data to analyze local conditions can (and has) resulted in stalled projects and major crises, at great fiscal and human cost.

If you're seeing a growing demand for geospatial data reflected in your own organization, now is the time to ensure you're positioned to take full advantage.

Overcoming Geospatial Data Challenges

So what are the challenges in bringing robust geospatial analysis and insight into organizations that are clearly demanding it? Let's lay out four primary challenges, then dive deeper into how to overcome them:

> *1. Delivery: With the growing availability of geospatial data, most organizations recognize the need to ensure their architecture can handle significant data volumes.*
>
> *2. Transformation: Organizations often have to contend with complex schemas in unsupported formats, especially when bringing in disparate geospatial data sets.*
>
> *3. Collaboration: Sharing geospatial data with non-technical users presents challenges in communication and access, as well as a need for ease of use.*
>
> *4. Visualization: New geospatial models must integrate interactive visualization into traditional maps and render changes in real time.*

These are not insignificant obstacles. However, with some planning and the right solutions, organizations are able to overcome them and realize the tremendous potential in geospatial data.

Challenge One: Delivery

As described above, geospatial data volumes have grown by orders of magnitude, resulting in increased capacity and processing requirements

for most organizations. But as data volumes grow, it makes sense to get much more selective about exactly what data are transferred. For example, the satellite imagery from Sentinel-2 is available from AWS cloud storage. If you can extract only the spatial data you need, and only the specific bands you need, you can vastly decrease both transfer times and volumes. Taking this one step further are "analysis-ready data" (ARD) providers such as Google Earth Engine, Planet or Tesselo.

These providers stitch together mosaics of the best available imagery, perform atmospheric corrections, and even classify imagery with machine learning. An organization only needs to ingest data ready for geospatial analysis. Much of the earth is cloudy on a given day, so why bother downloading imagery you cannot use? Most of the forest is hopefully not on fire, so doing your "change detection" near the data source can save you significantly on transfer costs and downstream storage.

To plan hardware capacity, you need to calculate data volumes at rest. Then consider data update frequency. Exactly how much of this do you need to store in transitory fashion, and how much must be archived? For example, you might keep a rolling one month, or rolling one year of data in live storage. As noted earlier, typical open satellite data--and daily-updated commercial data--are now available on demand. Weather data updates every six minutes. Often organizations fail to implement change detection, or they use static data, because the volumes involved with up-to-date data are just too vast to handle. But this obviously results in inaccurate insights so it's worth taking the time to plan hardware needs properly. Another important consideration for many organizations is maintaining an appropriate "record of decision." This implies near-permanent archiving of the data used to make decisions. In many cases, it is sufficient to use only the fraction of the data actually used, and screen-resolution can substitute for full-resolution data. However neglecting this important aspect can leave you with insufficient hardware, or even put you in legal jeopardy.

Once you have identified your data volume and update frequency, you can calculate the compute times for different hardware scenarios. The image below (figure 2) shows how this calculation might look in practice.

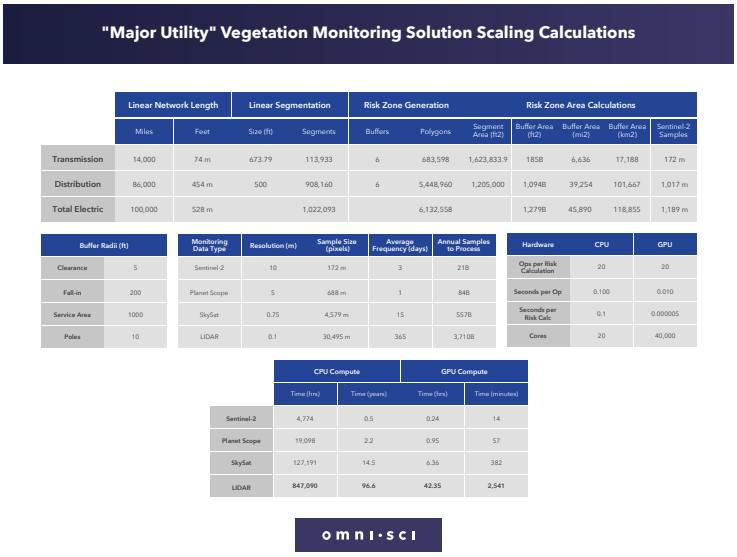

Figure 2

Here we're looking at possible data sources used by a major utility to manage vegetation near hundreds of thousands of utility lines. We consider the resolution sample size and average frequency for different data sources, such as Sentinel-2 or LiDAR, then calculate the compute time for both a CPU-based platform and a GPU-based alternative using ops per calculation, seconds per op, seconds per calculation, and cores. Note that the processing

requirements increase with the square of spatial resolution, and linearly with time sample intervals. Thus Planet Scope daily 5m imagery has 4x the pixels of Sentinel-2 but 12x the data volume once update times are considered. The results show the stark difference in compute time between CPU and GPU hardware: for instance processing the required volume of Sentinel-2 data would require six months with a CPU option and just 14 minutes if using GPUs.

Next you need to consider query response times, i.e. how many users will be accessing the data and at what frequency? Where will these users be located? Look at the example of an energy company which required its analysts to download extremely large files every time they wanted to explore the data. While this may have been feasible on-site, as employees shifted to working from home, they could no longer handle large downloads over a VPN. They decided to move data into the cloud for exploration and require downloads only when an analyst needed to manipulate the data, significantly reducing the burden on the VPN. In general, this strategy is known as "moving compute to the data" (as opposed to moving data to the compute). In many big data contexts, there is considerable advantage to doing as much processing in the cloud as possible, returning to the user just for feedback and final results. Conventional software licensing can be a barrier here, so be sure to check with vendors and understand how their licensing works relative to remote server deployments. But, overall, considering which data needs to be easily accessible and which can be processed more slowly in the background can save you meaningful amounts in infrastructure upgrades.

There are many solutions for small datasets, but given how easily geospatial data can reach millions, or billions, of records, it's more critical than with any other type of data to be prepared with a platform that offers the compute power and memory you need to scale.

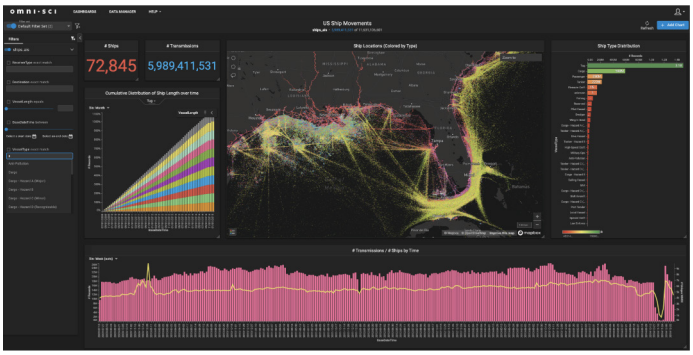

Figure 3

Challenge Two: Transformation

If your geospatial data is mostly clean and in common supported formats, you'll have a much easier time reading it. However, as with all data analytics, it is far more likely that you're dealing with multiple schema and diverse data types. With geospatial in particular, since much of the data typically comes from external sources, not only data formats but also coordinate systems and schema can vary widely.

There are a number of ways to create standardized processing steps across a range of data formats. A priority is being able to explore large datasets, with cross-filtering, so that you can understand the data before preparing it for more rigorous use. OmniSci Immerse enables this level of exploration with zero-latency, saving analysts from tedious prep time. OmniSci also integrates with Python Jupyter Notebook, where analysts can synthesize datasets and clean up inconsistencies. For the

most challenging scenarios, there are commercial tools such as Airflow, Prefect or Safe Software's FME, that allow you to orchestrate data at scale.

As described above, the relatively recent emergence of ARD (Analysis Ready Data) platforms, such as Google Earth Engine or Tesselo provides organizations with a second way to minimize data transformation challenges. The proliferation of cloud-based data providers allows organizations to outsource data cleanup, and enables analysts to more quickly incorporate powerful geospatial data into their own business.

Geospatial data consultancy, EcoAcumen, used ARD from Microsoft to study structure-level flood risk in communities. Microsoft's data applies machine learning to airphoto data to extract footprints for every building in the U.S.. EcoAcumen used this data to compute a "Height Above Nearest Drainage" (HAND) model, showing the height of every building in a town relative to the nearest water feature. The model allowed the team to create a risk scale based on likelihood to flood, providing more granular detail than current standard flood maps, as seen in figure 4 below. The model was powerful, but relatively easy to create thanks to the reduced data transformation required.

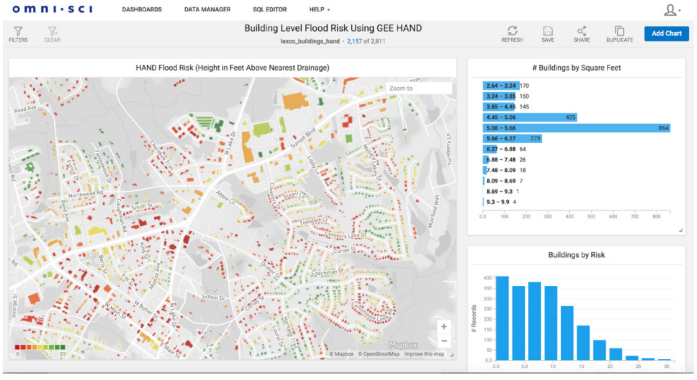

Figure 4

In order to reap the benefits of geospatial data, complexity is inevitable. Combining multiple data sources is not only possible, but necessary. The temptation is to over-simplify in order to avoid the transformation and orchestration challenges, as well as the hardware pressures. But that can lead to poor outcomes. Take another flood data example: the standard approach when creating flood maps is to use average annual soil moisture when calculating risk. Why don't these maps factor in actual moisture levels? We have access to actual soil moisture levels via satellite data. And, given how profound an impact on drainage soil moisture can have (if you've ever tried to water your garden with a hose in a drought, you'll know why), it seems obvious we'd want to include this data in our calculations. But we don't because of the complexity. Now it is becoming faster and easier to incorporate multiple data sources into our models, this must change.

Challenge Three: Collaboration

You've overcome volume and transformation challenges in order to deliver usable data. Now the question looms: who gets access to the data? And how can we foster collaboration among teams with different interests and varying levels of exposure to geospatial information?

A key trend seen across industries is the growth in use of geospatial data among non-geospatial professionals. As the value of geospatial data for a whole host of business decisions (from product development to sales to finance) is being recognized, other teams are requesting access to geo data. Geospatial team managers report that requests for business applications based on location data are on the rise. Collaboration with other analytics teams is now an everyday occurrence.

Sharing sounds simple--but as many an IT lead knows, the access and privileging of large, sensitive datasets across an organization is a big task. One collaboration-enabling technique that has seen widespread success is object-level permissions. By creating object-level permission to geospatial data, you've opened up invaluable analytical resources to both technical and business teams, while ensuring the integrity of your data.

Beyond granting access, organizations must consider ease-of-use. For example, in the military, battlespace awareness is critical to shortening decision cycles and to making better decisions than one's adversaries. Government intelligence teams must rapidly sense, collect, process, analyze, evaluate, and exploit streaming data. And that analysis must also be accessible to commanders and soldiers in the field. This demands an interface that can be used by those with relatively little data analytics training.

By implementing an accelerated analytics platform, intelligence analysts and commanders can query and interact with big geolocation datasets with zero perceptible latency. Then an intuitive, easy-to-use dashboard, powered by advanced analytics engines, gives them visual exploration at the speed of thought.

Self-service dashboards are becoming a must-have for most geospatial teams. They allow them to respond to internal demand for geospatial insights and create a much faster feedback loop of inquiry and analysis across teams. Geospatial data experts, non-geospatial counterparts, and team members with just a passing familiarity with the insights geospatial data affords find themselves working collaboratively on collective organizational goals.

Challenge Four: Visualizing and Sharing Insights

There is an enormous demand for geospatial insights, and for many good reasons. But until internal teams can truly harness the insights that come from geospatial datasets, many organizations will find themselves unable to apply it to ongoing decision support. That's where visualization and dashboard sharing come in.

Game-changing visualization begins with compute power, which occurs with massively parallel processing and GPU-accelerated operational analytics. As data volumes are exploding, increased interactivity is needed to keep pace. Instead of a map changing once every year, you might have one changing hourly, while accumulating years of history. Essentially, users now need to be able to scan through both time and space, and rendering interactive visualizations is technically much harder than classic semi-static maps. But that's just the initial step--supporting individual insight. In order to have business value, this insight typically needs to trigger one or more downstream processes, and that requires sharing. At an early stage, perhaps you need to share an observation with a single colleague for confirmation. In other cases, you might need to distribute tasks across the enterprise. Either way, the challenge is that you need a simple way to share a view of a complex system. Perhaps you could email a screen capture, but more often than not you'll find yourself wanting to share your interactive dashboard, perhaps with an annotation. With appropriate software architecture, you can do so while maintaining appropriate access control and avoiding sending enormous copies of data around. Teams from across the organization

explore up to millions of polygons and billions of mapped points, placing their own particular lens over the dataset to find answers to their particular set of questions.

Strong visualization in geospatial data also has an emphasis on temporality. Achieving real-time or near real-time visualization of geospatial data allows for critical, uninterrupted decision-making that can make a fundamental difference.

The Department of Housing and Urban Development (HUD) is leveraging geospatial data visualization as a solution to the ongoing problem of massive affordable housing shortages in many major metropolitan areas throughout the U.S.. A paradox within this crisis is that, while there are so many homeless, there are vast tracts of land that are suitable for development and construction. Zoning laws, city and county ordinances and recordkeeping often hinder what seems like an obvious solution to a devastating problem. Enter geospatial data. By providing developers, builders, and other commercial interests with a way to visualize housing and urban development from multiple sources, companies can clearly see paths to viable, profitable construction projects.

Delivering useful visualizations, of course, is not simple. It requires processing multiple disparate datasets at speed and scale, enabling relational searches across those datasets, and then rendering highly granular data in real-time.

Pactriglo is a market intelligence platform for real-estate developers that was confronted with exactly this issue. The Pactriglo platform leverages geospatial data to identify suitable development opportunities for affordable housing in Los Angeles. Identifying suitable development opportunities usually requires sifting through data from multiple government agencies, ranging from federal to county level to identify any possible obstacles such as zoning restrictions, neighborhood plans, or preservation overlay zones. Any of these so-called spot zones could cause a 12-18 month administrative review and delay in development. Pactriglo was able to use OmniSci's massively parallel processing to handle large real-estate datasets and render them in a way that gave developers an immediate view of viable development opportunities, as illustrated in figure 5.

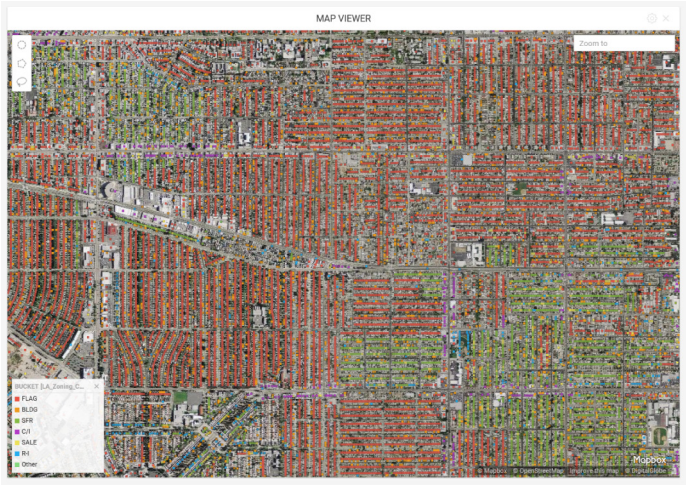

Figure 5

Pactriglo conducts processing of 100 million data points and renders them in real-time in response to developer queries. Red buildings on the map indicate properties or land with restrictions that would slow development, while green buildings represent viable opportunities. As a result, developers are able to accelerate decisions and prioritize investments. In this instance, intuitive visualization speeds up the delivery of a desperately needed resource such as new dwelling units.

Putting It All Together: Solving Challenges with Geospatial Data to Deliver Crucial Insights

If you can solve the issues of delivery, transformation, collaboration and visualization, geospatial data insights can play an essential role in major investment decisions.

Specifically, today's telcos face an interesting challenge with 5G investments. 5G technology allows much higher data speeds than prior generations of technology, but requires approximately ten times as many antennas. Fortunately, these antennas no longer need to be located on large towers. Instead, they can fit on a lamp post or innocuously on a rooftop. This allows antennas to be put closer to the users they serve, but also requires that these locations be more precisely located and targeted.

Traditional CPU-based tools for mapping radio frequency simply aren't up to this job. It can take weeks to generate a coverage analysis for a single market area. For each tower, thousands of individual rays must be created and then intersected with buildings, vegetation and terrain in order to compute RF signal strength. On a large modern CPU server, this work can be done partially in parallel, but only on 16-20 cores. This then in practice requires spreading the load across many dozens to hundreds of servers. This "horizontal scaling" requires a great deal of time and significant resources to move data to and from all these servers.

Meanwhile, a single modern GPU can boast over 10,000 cores, each of which has access to local data. The difference in practical terms: RF mapping for a major metro can be done in seconds instead of weeks.

Figure 6

Then, by mapping RF signal strength to the individual building level, telcos can focus more directly on the business value generated by adding infrastructure. By correlating with customer data, as in the display above, they can see the number of potential customers and their lifetime commercial value (top left), and are able to model the potential return on investment. Moreover, these visualizations can be shared and collaborated on with different stakeholders across the business, to inform decision-making. Sitting behind the map is a GPU dataframe supporting direct access to this data from machine learning and optimization tools critical for downstream applications.

What OmniSci Offers

A Forbes article declared that "66% of enterprises rank location intelligence as either critical or very important to revenue growth strategies." Yet, mainstream Geographical Information System (GIS) platforms struggle to analyze more than a few hundred thousand data points, and then only with latency that makes visual exploration of the data incomplete, slow and uncomfortable.

OmniSci addresses all the common challenges of geospatial data analysis by leveraging the power of CPUs and GPUs to accelerate existing analytics solutions and render interactive visualization of massive geospatial datasets, all with millisecond latency:

Delivery - The OmniSciDB SQL engine natively stores geographic and geometric data types (POINT, LINESTRING, POLYGON, and MULTIPOLYGON), and can not only query this data in real-time, but can render it interactively using the built-in graphics capabilities of GPUs. This enables organizations to run geo calculations in real-time and interactively explore millions of geometries on a chart in a web browser. Speed and power come from OmniSci Core, the world's fastest open source SQL engine, which harnesses the power of GPUsto allow multiple analysts to query big data with millisecond results. OmniSci Core features an innovative Just-In-Time query compilation framework, which is more efficient for memory bandwidth and cache space and delivers much faster compilation times--generally under 30 milliseconds for entirely new queries. OmniSci Core also includes advanced memory management and query vectorization to further query vectorization to further optimize performance, as well as High Availability and Distributed Scale-Out.

Transformation - To minimize data prep time, the OmniSci analytics platform speeds data exploration with extraordinarily fast SQL queries, rendering and visualization. Analysts can import native geo data into OmniSci, explore it visually, and clean it quickly, so it's ready for analysis. With OmniSci Immerse, a browser-based, interactive data visualization client that works seamlessly with the OmniSci server-side technologies, analysts can also 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).

Collaboration - OmniSci's visualization system makes it easy to share geo charts and geospatial calculations with others. 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.

Visualization and Sharing - Exceptionally easy to use, 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 and refresh them in milliseconds. Immerse even enables multiple stakeholders with an organization to interrogate data and get immediate answers to their questions thanks to its cross-filter paradigm: 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. Some OmniSci customers have multiple simultaneous users cross-filtering over datasets in the tens of billions of records. OmniSci users can also 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.

And while many organizations are being introduced to geospatial data for the first time, OmniSci has forged partnerships with some of the biggest data providers over the years. These established working relationships can greatly reduce training time for analysts while dramatically accelerating the pace of their queries.

To illustrate the power of OmniSci, we can use an example that harks back to the very first use case of geospatial data--to track infection incidents. Leveraging the parallel processing power of GPUs, OmniSci was able to cut through over 16 billion location records from 6 million mobile devices to investigate outbreaks of COVID-19 at meat processing plants. By ingesting massive volumes of geotemporal data at speed and visualizing them, we were able to show how the communities around these plants experienced disproportionately higher COVID-19 infection rates compared to nearby counties. See how this was achieved here or visit our live demo.

Geospatial data is undoubtedly helping to solve many of the most pressing questions faced by businesses and governments today. From managing natural disasters and public health crises, to making crucial business decisions, it's clear that location intelligence must be part of any organization's strategy. Are you ready to lead the charge in your own organization?

About Us

OmniSci is the pioneer in accelerated analytics. The OmniSci platform is used in business and government to find insights in data beyond the limits of mainstream analytics tools. Harnessing the massive parallelism of modern CPU and GPU hardware, the platform is available in the cloud and on-premise. OmniSci originated from research at Harvard and MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). OmniSci is funded by GV, In-Q-Tel, New Enterprise Associates (NEA), NVIDIA, Tiger Global Management, Vanedge Capital and Verizon Ventures. The company is headquartered in San Francisco. Learn more about OmniSci at www.omnisci.com.

omni sci

Contact Us

www.omnisci.com/contact
sales@omnisci.com