kinetica

Building Real-time Location Applications on Massive Datasets

Vectorization opens the door for fast analysis of large geospatial datasets

Inside

WHITE PAPER

1 / Introduction

Gartner predicts that by 2022, 30% of customer interactions will be influenced by real-time location analysis, up from 4%. This shouldn't come as a surprise given the explosion of smart devices and sensors and the amount of customer data that's being captured — most of it with location and time information. But this also means increasing pressure to improve the speed and accuracy of business processes and deliver contextually relevant experiences by analyzing massive data sets, visualizing the results and delivering highly customized location-aware analytics applications.

Businesses that generate location data can create new revenue streams by building data services for other companies and industries. Telecom companies, for example, generate a huge amount of location data with high value that can be sold to retailers and other businesses. The same is true for payment processing vendors. Location data is valuable, even when anonymized to maintain user privacy.

Location intelligence plays an important role in improving operational efficiency across telecom, logistics and utility organizations. Location analysis on network usage can help identify periods of heavy network usage, forecast network capacity, and plan for potential network outages or short-term surges. Location is starting to play a part in fraud detection as well.

Financial services use a combination of location and time data to prevent fraud. Using location for fraud detection is also applicable to any kind of product or drug fulfillment service such as prescription medication pickups at pharmacies.

However, existing geospatial intelligence systems (GIS) as well as traditional relational databases that support spatial capabilities are really struggling to meet the requirements of business to build next-generation location apps. This paper highlights the existing challenges businesses face, how data level parallelism can dramatically accelerate processing, introduces Kinetica–a single engine that combines location intelligence, visualization, advanced analytics and the power of AI–and highlights a range of real-world use cases that are changing the game for businesses.

2 / The Challenge in Building Real-time Analytical Applications on Large Data Sets

As more and more location data becomes available from sensors, from customers, and from transactions, there are increasing demands to analyze these data sets through applications and visualize the results on maps.

But today's geospatial technologies are hardly up to the task. Today, there are three approaches to leveraging location data:

1. Using business reporting tools like Tableau that supports some capabilities to visualize location information and analyze the basics
2. Building a custom location-aware application on relational databases like Microsoft SQL Server or Oracle that provides some basic spatial capabilities
3. Leveraging traditional GIS systems that provide advanced spatial analysis, rendering and visualization capabilities

However, all these approaches are starting to falter today as location data is increasing in volume, speed and complexity, and is now more important to be analyzed. Reporting tools are starting to have capability and speed challenges, and need a faster engine underneath them to perform better. Relational databases are hitting limits in performance, features and functionality, and scale. And GIS systems that have provided businesses with a foundation to store, manage and analyze spatial data, in reality use a relational database to store data and are so are also hitting a performance and scale boundary, especially on large, fast-moving datasets.

Moreover, traditional GIS systems struggle with large compute workloads because operations such as spatial joins, geometry functions and spatial aggregations are workload intensive. For a traditional system based on task parallelism, scaling the analysis of big data with high demands for time-series and geo-location is costly and slow. This puts real-time location-based insights out of reach.

Spatial databases weren't designed for a world where IoT systems might be tracking millions of sensors generating frequent updates. If you want to do analysis on large datasets with any sort of interactivity, if you want to visualize millions or billions of records—and then filter them and see results based on different groupings—then you're going to need to solve a couple fundamental challenges.

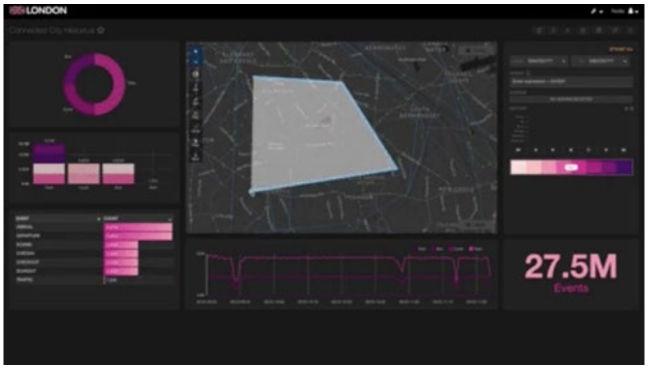
2 / The Challenge in Building Real-time Analytical Applications on Large Data Sets (continued)

The first challenge: Scale

The first challenge is that most databases are simply not designed to perform large-scale geospatial analytics in a reasonable amount of time. Imagine trying to analyze millions of customer purchases and aggregating those based on ad-hoc proximity to retail stores. The types of polygon intersection calculations needed to run this are expensive from a compute and I/O perspective. Multiply that by millions or billions of records, and your query may leave you waiting all week. Not a problem if you only need to do it once, but if business users want the ability to make ad-hoc queries on data, this is not acceptable.

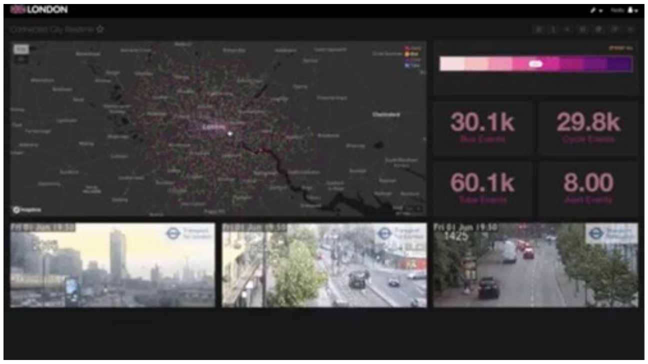
The second challenge: Interactive Visualization at Scale

The second challenge is how to visualize large geospatial datasets with any sort of interactivity. Web browsers on the client-side of an application struggle to handle more than a few thousand features, and it takes time to send large volumes of data over the wire. Send more than several thousand points or a thousand polygons to a browser, and you will see your browser slow to a crawl. Eventually there is a threshold where it’s not practical to send all your data across the wire for the client to sort out. Visuals need to be rendered on the server-side.



The third challenge: Streaming Data

The third challenge is analyzing and visualizing streaming data in context with other data sets. As IoT devices increase in number, advanced location analysis on streaming data is a key requirement of modern location intelligence applications. However, existing systems find this quite challenging after hitting both I/O and compute bottlenecks.



3 / Leveraging Vectorization for Ultra-Fast Spatial Computation

So how do you solve for the performance challenges of more complex analytics with large geospatial datasets?

The founders of Kinetica faced this challenge when working with the US Army Intelligence community to tackle real-time mapping of security threats. Existing database options weren't designed to handle streaming data with spatial and time series attributes, and they were unable to provide the geospatial capabilities or performance for this type of data to be useful in real time.

Graphics Processing Units (GPUs) presented an opportunity to solve the computational challenge. GPUs were originally designed to speed up video games, but their parallelized processing capabilities are also ideal for the types of vector and matrix calculations needed for geospatial analysis. The Kinetica founders coupled this with the high performance of in-memory computing, creating a converged engine for advanced analytics and complex geospatial analysis and rendering.

Recognizing the performance improvements, CPU vendors began to incorporate vectorization capabilities into their architectures. With roots in vector and matrix calculations, Kinetica engineers embraced this complimentary form of data parallelism within CPUs. Kinetica can gracefully switch between GPUs or CPUs to deliver data level parallelism (aka vectorized parallelism) to produce performance improvements in excess of 16X compared to environments that only do task level parallelism.

Almost any database can be used to store geospatial data: coordinates can be stored as floats, and shapes can be converted into a WKT format and stored in a text column. But while a database can store data this way, it isn't readily available for query. A separate geospatial system needs to retrieve these records, convert them into a geometry objects, and evaluate the match—one record at a time. What is needed is a spatially aware database that has a geometry engine built in, that can work natively with geospatial data and compute relationships between shapes and objects within a single system. Kinetica was designed with such spatial capabilities.

Native Geospatial Objects

Native support for geospatial objects such as points, lines, polygons, tracks, and labels, and storing and interpreting these objects as OGC-compliant WKT, makes it easy to ingest from, or export to, other systems.

3 / Leveraging Vectorization for Ultra-Fast Spatial Computation (continued)

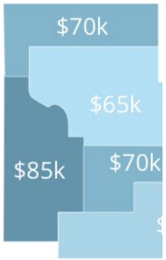
Native Spatial Operators

A suite of more than 80+ geospatial functions that run natively within the database, many of them data level parallelized, makes it possible to get fast results on queries such as the following:

SPATIAL QUERIES

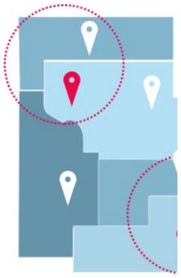
Find all households within 10mi of a store

SPATIAL JOINS



Calculate average income per zip code

ADVANCED SPATIAL ANALYTICS



Identify retail locations where over 50% of households earn more than $50k, and where there are over 5,000 households within within 10mi of a store.

Geospatial queries are frequently compute intensive. As the volume of data increases, performance becomes an ever more critical issue. Kinetica is designed to harness the power of vectorization for exceptionally fast query response, even across the largest datasets. Spreading spatial computations across thousands of nodes, across multiple cards, and multiple machines is an exceptional solution for the types of brute-force calculations needed for advanced analysis of large and streaming geospatial-temporal datasets. In addition, the data is managed and served from system memory to create high-performance applications.

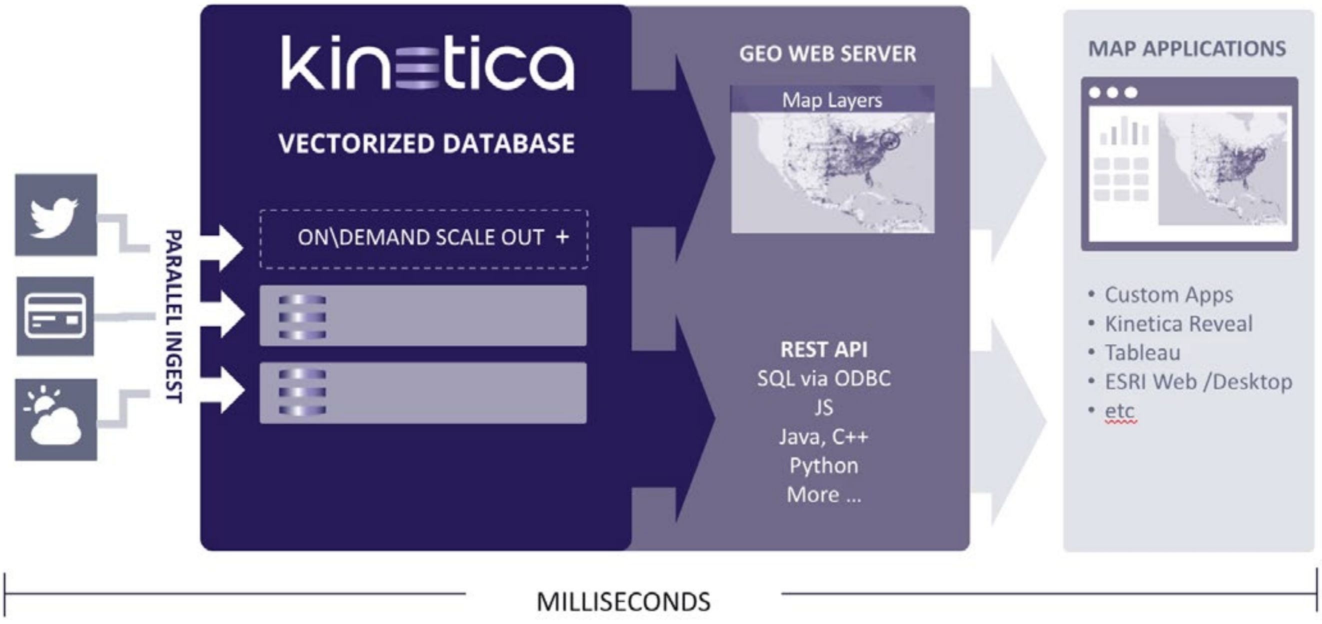
And with the rise of IoT data – such as social media feeds, moving vehicles and sensors – a modern geospatial database must also be able to handle high velocity streaming data. Multi-head ingest design enables each node within a cluster to share the work of absorbing streaming data. And since vectorization offers such tremendous performance improvements on query, less indexing is required, and data can be made available to query the moment it arrives. This opens up tremendous new opportunities for real-time and predictive analytics.

4 / A Geospatial Visualization Pipeline for Large Datasets

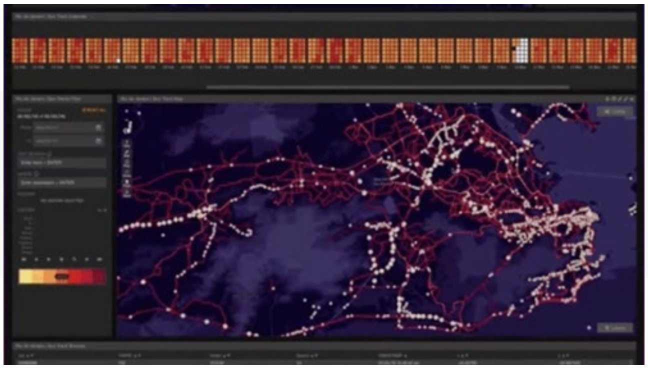
So, the challenge of real-time geospatial analytics at scale can be met with a vectorized parallel database, but that still leaves the challenge of how to visualize large datasets with any sort of interactivity. Remember, if you're outputting more than a few thousand points or polygons across a wire to a mapping client, things are liable to grind to a crawl.

Kinetica addresses this with a native visualization pipeline capable of leveraging parallelized vector processing to quickly render geospatial visualizations on-the-fly.

The Kinetica Visualization API also comes with the tools necessary to interact with those maps, drill into, and explore individual points and shapes on that map. These can be overlaid on top of base-maps from ESRI, Mapbox, etc.

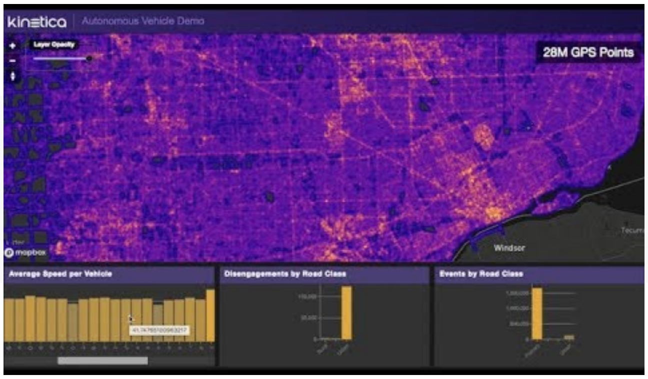


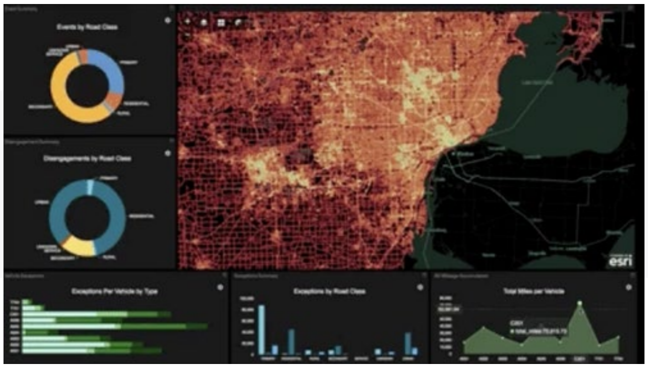
4 / A Geospatial Visualization Pipeline for Large Datasets (continued)



Smart city data filtered and analyzed in sub-second time. No indexes, no caching.

Autonomous vehicle test drive data using Kickbox, Kinetica's Mapbox integration





Autonomous vehicle test drive data using Kinetica & ESRI

5 / Advanced Visualizations and Analytics

Advanced Analytics and Machine Learning with Geospatial Data

Geospatial analysis can be further extended through the User-Defined Functions (UDF) API – an interface that makes it possible for custom code to run from within the database. Through UDFs, almost any type of analysis is possible.

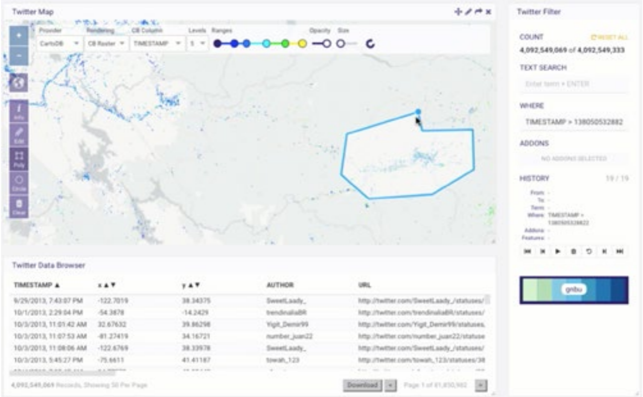
Even for highly customized geospatial operations, the dataset does not need to be extracted into a separate system for analysis. Instead, models can be brought to the data, to be run 'in-database.'

This opens a world of possibilities—custom code can even call out to machine learning libraries, such as TensorFlow, for advanced geospatial predictions.

This might make it possible for deliveries to be flagged when they are unlikely to arrive on time—based on traffic, weather, or other indicators. Insurance companies could better analyze drivers that are most likely to be involved in an accident based on driving behavior, or they could calculate risk for assets from weather models.

Advanced Visualization Options

The visualization layer also includes some more advanced functionality, including heat maps and color-coded filtering. These dynamic map visualizations can be integrated with any OGC-compliant web mapping API to allow for interaction with the visualization layer's features. The below visualization of querying a large Twitter dataset illustrates some of these capabilities:



COUNT
4,092,549,069 of 4,092,549,333

TEXT SEARCH
Enter term + ENTER

WHERE
TIMESTAMP > 138050532882

ADDONS
NO ADDONS SELECTED

HISTORY 19 / 19
From: -
To: -
Term: -
Where: TIMESTAMP > 138050532882
Addons: -
Features: -

Twitter Data Browser

TIMESTAMP ▲	x ▲▼	y ▲▼	AUTHOR	URL
9/29/2013, 7:43:07 PM	-122.7019	38.34375	SweetLaady_	http://twitter.com/SweetLaady_/statuses/
10/1/2013, 2:29:04 PM	-54.3878	-14.2429	trendinaliaBR	http://twitter.com/trendinaliaBR/statuses/
10/3/2013, 11:01:42 AM	32.67632	39.86298	Yigit_Demir99	http://twitter.com/Yigit_Demir99/statuses.
10/3/2013, 11:07:53 AM	-81.27419	34.16721	number_juan22	http://twitter.com/number_juan22/statuse
10/3/2013, 11:08:06 AM	-122.6769	38.33978	SweetLaady_	http://twitter.com/SweetLaady_/statuses/
10/3/2013, 5:45:27 PM	-75.6611	41.41187	towah_123	http://twitter.com/towah_123/statuses/38

4,092,549,069 Records, Showing 50 Per Page

Download

Page 1 of 81,850,982

6 / Case Studies

These are real-world examples from businesses across industries that are beginning to understand what is possible with the new location analytics technology on the market, and what is achievable once they leverage the power of data level parallelism.

1 / **Oil and gas.** A leading international oil and gas company has a newfound ability to visualize huge datasets. These data visualizations improve the way they extract hydrocarbons, leading to reduced environmental impact and resource consumption. Leaders gain visibility to drive exploration as they integrate real-time data about wells, pipelines, and land ownership. Analytics on extreme data empowers them to focus on the most promising new oil and gas fields. This is only possible because they are able to use all their data, including geospatial data, coupled with machine learning and predictive modeling, to assess which assets currently provide value and which do not.

2 / **Telecom.** Telecom companies face exponential and even suffocating growth in data from mobile devices and the IoT. They need a platform to ingest and analyze all of this data. Consider network efficiency and performance: Telecoms need to assess where signal strength is optimal or poor and layer it in with information about dropped calls to ensure high-quality service. This type of wide-scale, granular, spatial-temporal analytics can lead directly to network improvements and to marketing in undersaturated areas, mixing in customer experience data to target, attract, and retain high-value clients. Telecoms can capitalize on increasing real-time data volume to build new revenue streams.

3 / **Logistics and inventory.** A data-powered system enables machine learning to drive real-time inventory decisioning. If demand for a certain product spikes, operational algorithms send out more of that product to stores. Route optimization is performed at scale with a more efficient infrastructure on a dramatically reduced technology footprint. New business models and services powered by real-time data, such as 15-minute delivery windows, support additional revenue streams.

4 / **Media and broadcasting.** Up until now, though set-top boxes generate tremendous amounts of data about customer usage, communications service providers have not been able to leverage it. But with data-powered tools that ingest 300 million to a billion events a day, companies now answer questions about the specific reach of an advertiser on a program, down to the time and location of viewers, as well as how users are viewing media, whether on TV or on mobile devices. This empowers communications companies to improve advertising and target programs more specifically to certain demographics, and, in some cases, develop the kind of programming that appeals most to particular audiences.

7 / Conclusion

The combination of real-time streaming query, native geospatial operators, and advanced map-based visualizations opens opportunities for businesses to perform analyses that were previously difficult or impossible.

Any company with large volumes of geo-temporal data will have their own demands and opportunities from that data. For many, the solution lies in using Kinetica to map infrastructure, for logistics, for customer research, and more. Find out more about location-based analytics with Kinetica. Have a discussion with one of our solutions engineers, or contact us for a demo.

Get Demo

About Kinetica

Kinetica is the insight engine for the Extreme Data Economy. The Kinetica engine combines artificial intelligence and machine learning, data visualization and location-based analytics, and the accelerated computing power of a vectorized parallel database across healthcare, energy, telecommunications, retail, and financial services. Kinetica has a rich partner ecosystem, including NVIDIA, Dell, HP, and IBM, and is privately held, backed by leading global venture capital firms Canvas Ventures, Citi Ventures, GreatPoint Ventures, and Meritech Capital Partners.For more information and trial downloads, visit kinetica.com or follow us on LinkedIn and Twitter.

kinetica

Headquarters

101 California St, Suite 4560
San Francisco, CA 94111

(888) 504-7832
+1 415 604-3444

To learn more

visit us at
kinetica.com

or
contact us