Closing the Analytics Loop:

How Graph Databases complement your data warehouse

A dbInsight white paper for TigerGraph

Trigger

Making analytics operational is becoming table stakes. Most organizations have invested heavily in building data warehouses and data lakes that provide the basis for operational and historical reporting. Yet increasingly, a full understanding of challenges ranging from enhancing customer experiences to understanding drug interactions, preventing fraud, optimizing complex networks and more, requires understanding the interrelationships between the data. The challenge, and opportunity, is uncovering connections within these massive datasets that can provide businesses with competitive advantages.

Our Take

Traditional data warehouses and analytic tools are useful for producing the baseline insights that allow organizations to work more intelligently. For instance, they can pinpoint customer segments responsible for churn, identify candidate drugs to study for harmful or beneficial interactions, uncover patterns of financial fraud, or point to sources of cyberattacks. But consider the following questions: Why do customers defect? Why do certain combinations of drugs aid or abet treatment? How cyberattacks propagate? Answering each of these questions requires understanding the underlying interrelationships because scenarios such as customer churn, drug interactions, or cyberattacks do not happen as the result of any event or condition in isolation.

Graph databases build on the operational or historical analytics foundation established by your data warehouse or data lake, providing the last link in the puzzle to identify how or why desired or undesired events occur. The underlying relationships between people, things, and/or events are often critical to understanding why something has happened, and discovering *similar* relationships can indicate paths to solutions. These are capabilities that add the missing link that build on the discoveries from traditional data warehousing. When evaluating graph platforms, the ability to perform at scale is becoming a key requirement because it provides the big picture that is missing when analyzing only parts of the problem.

The nature of the challenge

The world has always been connected, but is becoming even more so thanks to sensors connecting devices, social networks and messaging connecting people, supply chains/value chains connecting enterprises, and so on. Let’s look at several representative use cases.

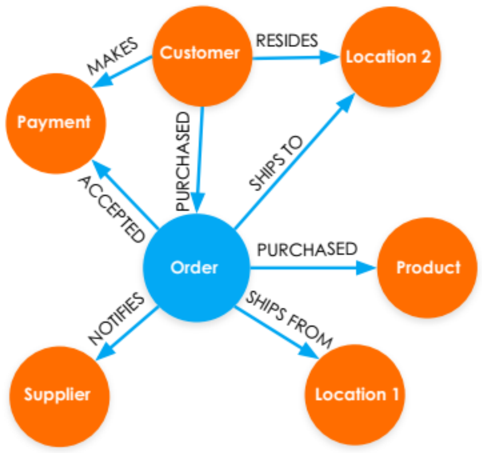
Customer Journey

This is a perennial concern for any mass consumer-facing business. Consumer packaged goods (CPG) manufacturers look for that magic product that answers what customers want or need, while their partners in the supply chain – retailers and distributors – seek to capitalize on that opportunity by ensuring the right product mixes are available and are effectively merchandised.

A data warehouse or data lake can be used for identifying and segmenting customers into demographic cohorts that account for gender, region, income, age group, and other factors. They can then conduct historical analytics such as market basket analyses to get a better idea of their buying preferences. They can conduct classic focus groups to dive deeper into what drives customer purchasing decisions. They can take advantage of recent innovations in artificial intelligence and machine learning to predict, based on historical patterns and current events, what customers will buy. And the learnings that are conducted from understanding customers' journeys can form the basis for building knowledge graphs of customer behavior that yield long-term benefits on the organization's ability to make the right calls.

All of these analytics are valuable in building a base of assumptions. But today, customers are increasingly connected through smart mobile devices, texting each other and participating in social networks. Meanwhile, retailers are employing devices such as beacons that track customer movements when they are inside a brick and mortar store or track navigation when they are ordering online. In today's landscape, classic analytics presents only a partial, static picture of consumer buying habits and brand loyalty.

Figure 1. The multiple relationships involved with customer fulfillment

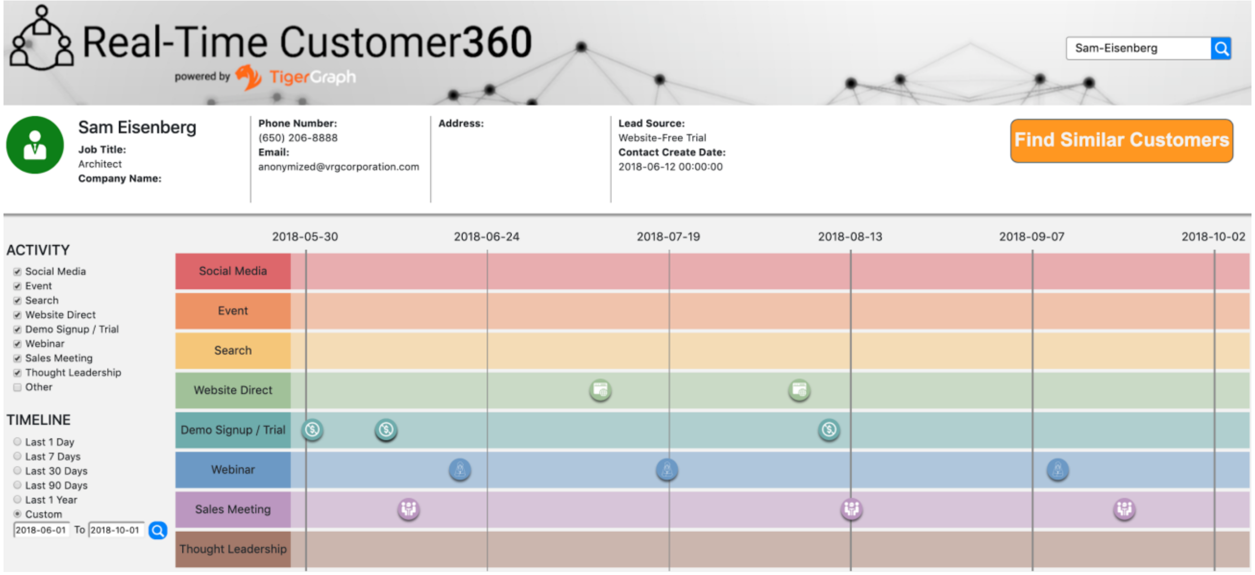


Source: TigerGraph

Graph databases can pick up where data warehouses leave off, conducting similarity matches across millions of customers to identify the most likely customers with similar behavior patterns. Graph databases shed light on the impacts of a variety of factors driving the customer journey, ranging from how they use different channels to interact with CPG brands to the impacts of diverse factors such as ease of navigation using mobile or web apps, weather, traffic conditions, discussions on social networks, and more. Figure 1 (on the previous page) shows a simplified view of the complex relationships that may be involved when analyzing customer buying patterns from the perspective of order fulfillment.

In so doing, CPG companies can identify and take action on customers most likely to churn, or at the other end of the spectrum, are most likely to be the top influencers driving brand loyalty, as shown in Figure 2.

Figure 2. Identifying the customer journey using graph analytics



Sam-Eisenberg

Sam Eisenberg

Job Title: Architect

Company Name:

Phone Number: (650) 206-8888

Email: anonymized@vrgcorporation.com

Address:

Lead Source: Website-Free Trial

Contact Create Date: 2018-06-12 00:00:00

ACTIVITY

☑ Social Media
☑ Event
☑ Search
☑ Website Direct
☑ Demo Signup / Trial
☑ Webinar
☑ Sales Meeting
☑ Thought Leadership
☐ Other

TIMELINE

○ Last 1 Day
○ Last 7 Days
○ Last 30 Days
○ Last 90 Days
○ Last 1 Year
● Custom
2018-06-01 To 2018-10-01

Source: TigerGraph

Drug Interactions

It can cost up to $1 billion to deliver a new drug. Life sciences companies have long used analytics to gain insight from clinical studies and outcomes analysis. Understanding drug interactions has long been a pillar of new drug research because, for most treatments, patients take multiple medications. This is a classic case of a “many-to-many” relationships problem because, across a population of patients, the mixes of medications will vary. Understanding the chain of events when patients take medications is difficult for classic relational data warehouses because of the need to perform hundreds, if not thousands of table joins to identify a critical mass of the possible permutations in the mixes of medications that people take.

Graph complements classic relational analytics by providing that additional context that pulls back the covers on what occurs when new medications are added to the mix. Ultimately, it can identify the medication(s) having the most positive or negative impact on patient care, and the optimum dosages. It can also be used for PageRank algorithms that pinpoint the most influential "members" (e.g., drugs) driving particular outcomes, and Community Detection algorithms that identify groups connecting to those medications driving PageRank. Graph databases can also serve as the springboard for machine learning approaches that can predict outcomes or interactions based on similarity ranks of other drug regimen, or identify features based on data relationships. They are key to the effectiveness of running these models because they can analyze entire datasets at once. By contrast, data warehouses or data lakes are not tailored for these types of analyses.

Fraud Detection

Detecting fraud traditionally has relied on identifying the history of individual people or organizations, which can be captured and analyzed on a data warehouse or data lake. But often, the actions of individuals may be concealed as they apply for multiple fake accounts.

Peer-to-peer payments present a potentially rich attack surface for bad actors. When a new customer without documented history creates a new account by registering a new phone number and email, using traditional analytics, red flags won't be raised. With a graph database, such "new" users can be connected to previous accounts that may have shared the same phone number or mobile device. This analysis can be performed in real time if the platform supports ACID transactions; traverses multiple "hops" (links) to establish the connection; and then immediately computes a trust score based on the findings.

It goes without saying that fraud is a major concern in the banking world, which spends billions annually on prevention. There are numerous examples where graph analytics can provide an extra measure of intelligence in identifying and stopping financial fraud. For instance, a leading US bank is analyzing the connections between known fraudulent accounts and new applications, while a large payment card provider is using graph to supplement its data warehouse to conduct deep pattern analytics on credit card activity. By identifying outliers and anomalies between people, transactions, and institutions, financial services firms can get more complete views to identify and shut down bad players in the act in real-time. Add in machine learning, and the patterns of interactions and underlying relationships can be used to predict when and where fraud is most likely to occur.

Graph Databases provide the solution

Traditionally, historical and operational analytics have been conducted on relational databases. These platforms are useful for well-structured data that can be represented through tables of columns and rows. Relationships are established through foreign key relationships that link specific tables. Relational databases hit the wall handling problems with complex relationships because the solution requires countless joins, presenting a huge computational burden. The alternative was hard coding relationships into an application, an approach that can work for one-off scenarios but requires considerable coding effort to stay in sync each time.

Graph databases are uniquely designed for managing and representing data with complex patterns of interrelationships. They can be used for resolving problems involving complex relationships and forming the foundation for knowledge graphs that organize variably structured data that can be used for feeding AI models. Graphs provide a flexible means for representing the interrelationships between connected data, and unlike relational databases, have flexible schema that can adapt to change.

Where TigerGraph plays

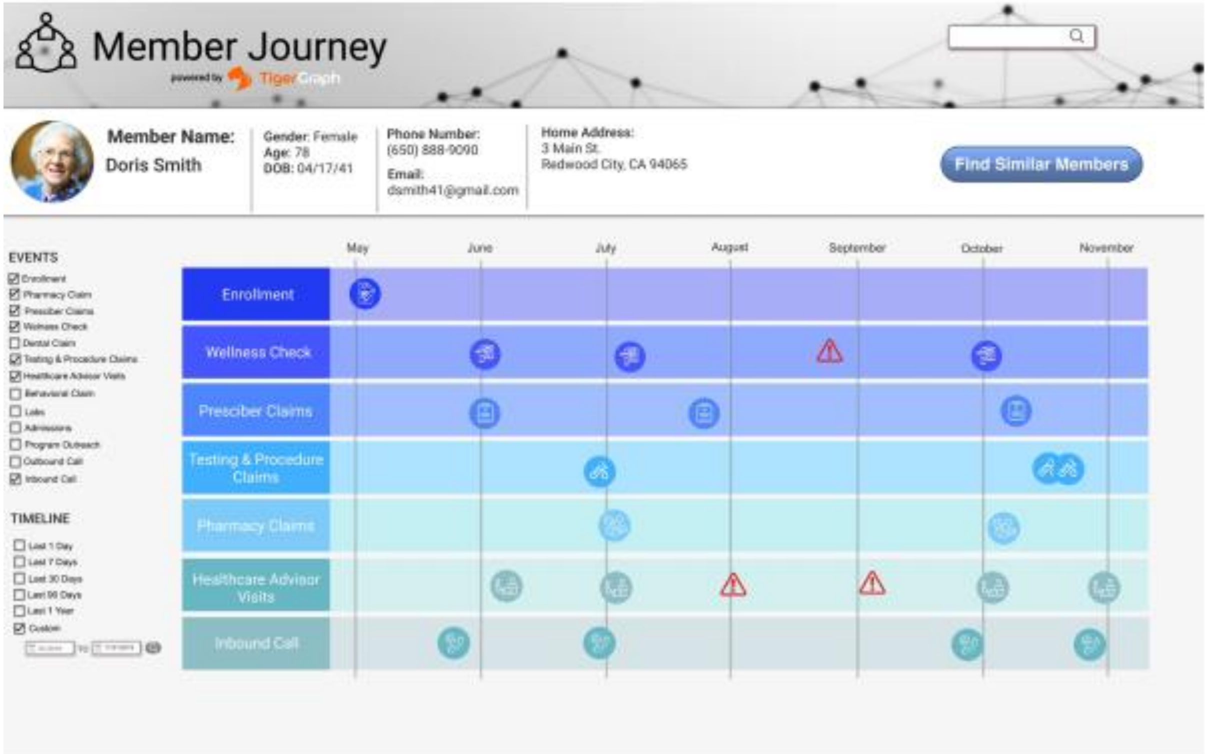
TigerGraph differentiates itself through its scale and speed. Because it was designed for the type of scale-out architectures that cloud data centers have made economical, TigerGraph can handle problems with billions of vertices (each vertex representing an entity such as a payment, claim, order, customer account etc.). It can compute these problems through parallel processing in a single pass. It does so through a series of design features leveraging query compilation; data filtering; data compression; advanced partition measures; and tunable ACID database consistency (through snapshots and other measures) that allows developers to create applications and write queries that address all of the data, while providing the flexibility to balance the demands for currency vs. performance.

Here’s what these advantages mean. TigerGraph’s design supports the ability to generate answers in real time, spanning historical and incoming streaming data. That is where ACID snapshot consistency is critical, because it is essential when working with scenarios where real-time answers are critical. That can encompass uses cases such as nipping fraudulent financial transactions in the bud when someone presents a false or stolen credit card, or identifying a national security, public safety, or health risk when the wrong person crosses a national border or tries to gain entry to a protected area.

Conversely, when conducting analyses to identify similar cases, prioritizing broad access to data over currency becomes important. Figure 3 shows how a large healthcare payer sought to improve the efficacy of care by understanding patient “journeys.” They implemented a

TigerGraph database to paint a complete picture on, not only the prescribed care regimen, but also the patient's actions on whether, or how much of the care plan had been followed, across a membership list numbering in the tens of millions. It involved getting the big picture from data residing in multiple databases ranging from membership to medical history, prescriber, claims, healthcare facility, and other sources, amounting to tens of billions of records.

Figure 3. Identifying the healthcare member journey using graph analytics



Member Name: Doris Smith

Gender: Female
Age: 78
DOB: 04/17/41

Phone Number:
(650) 888-9090

Email:
dsmith41@gmail.com

Home Address:
3 Main St.
Redwood City, CA 94065

EVENTS

☑ Enrollment
☑ Pharmacy Claim
☑ Prescriber Claims
☑ Wellness Check
□ Dental Claim
☑ Testing & Procedure Claims
☑ Healthcare Advisor Visits
□ Behavioral Claim
□ Labs
□ Admissions
□ Program Outreach
□ Outbound Call
☑ Inbound Call

TIMELINE

□ Last 1 Day
□ Last 7 Days
□ Last 30 Days
□ Last 90 Days
□ Last 1 Year
☑ Custom

Source: TigerGraph

Getting a complete picture was essential, not only for customer service representatives fielding calls, but for gaining an understanding of how to improve the customer's experience given their history. It requires the ability to query this data in real-time to provide answers. TigerGraph's similarity query capability allowed the healthcare payer to drill down through tens of millions of patient profiles to identify the patients with similar trajectories, to provide care path recommendations. Its RESTful interfaces allowed to plug these analytics into member tracking and call center applications, so that customers call in, and representatives can work within their existing screens.

Takeaways

Graph databases pick up where data warehouses, the traditional workhorses for analytics, leave off. Both work together in generating the insights for enterprises to make smart decisions. They provide the building blocks for building knowledge graphs that make organizations smarter. Data warehouses generate baseline insights, such as identifying customer preferences or hotspots for fraud. Then, graph databases complement data warehouses, enriching insights with added context driven by underlying relationships in the data. For instance, they can provide the nuance for understanding how and why consumers are motivated to buy specific products. Or they can uncover patterns of financial fraud from individual transactions that in isolation would otherwise appear valid. This is critical in an increasingly connected world, where events and outcomes do not happen in isolation.

When evaluating graph databases, there are several crucial capabilities that stand out. First is the ability to get the complete picture and query *all* of the data; that is where the ability to scale plays a pivotal role. Secondly, is the ability to get answers in real time – this is critical for supporting operational use cases, especially when contending with live streaming data (e.g., IoT, location data) where relevance may have a short shelf life. And finally, the flexibility of prioritizing database consistency vs. data access or performance allows organizations to get the answers that they need.

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

Twitter @TonyBaer

About dbInsight

dbInsight LLC® provides an independent view on the database and analytics technology ecosystem. dbInsight publishes independent research, and from our research, distills insights to help data and analytics technology providers understand their competitive positioning and sharpen their message.

Tony Baer, the founder and principal of dbInsight, is a recognized industry expert on data-driven transformation. *Analytics Insight* named him one of the 2019 Top 100 Artificial Intelligence and Big Data Influencers. His combined expertise in both legacy database technologies and emerging cloud and analytics technologies shapes how technology providers go to market in an industry undergoing significant transformation. His regular ZDnet *"Big on Data"* posts are read 25,000 – 30,000 times monthly.