Introducing Connection Analytics

An Ovum white paper for Teradata

SUMMARY

Catalyst

SQL relational data warehouses have long track records of success, generating insights from data originating from internal enterprise applications. Big Data technologies are opening the floodgates to new data types and new analytics approaches that complement SQL-based querying. Ovum has found that enterprises are using Big Data analytics to complement traditional SQL queries in answering very familiar questions, such as customer retention, marketing attribution, risk mitigation, and operational efficiency. One of the most promising new Big Data approaches supplements traditional analytics by uncovering interdependencies between people and/or things to answer questions such as:

- How do relationships between people in various groups impact the likelihood that demand for a product goes viral?
- What are the ripple effects that could happen when a broken outflow pipe contaminates a water supply, or the occurrence of illness triggers a mass outbreak?
- What will happen when a merchant, auctioneer, or bidder drops their price for a product?
- Changes in bodily metabolism impact the effects and interactions between different drugs?

Until now, answering such questions required enormous compute power, time-consuming data management and the need for learning highly specialized programming and query languages.

Ovum view

Connection Analytics provides a new way of looking at people, products, physical phenomena, or events. It provides insights by dissecting the types of relationships between entities to determine causation and can be used for generating predictive intelligence based on the patterns of interactions. Connection Analytics can address queries such as identifying influencers, the groups that they influence, and where promotions or other forms of marketing are best directed. It can be utilized for product affinity analysis by taking a bottom up look at how the decisions to buy different items are linked. Likewise, this approach can help analyze networks by patterns of activity, and fraud and money laundering through the actions (rather than identities) of involved actors. It can help segment customers based on behavior patterns like past purchase behavior or reviews vs. traditional segmentation techniques like income & demographics. Graph analytics is one of the most promising approaches to performing Connection Analytics. Teradata is the first analytics data platform provider to make graph computing accessible to the existing base of data scientists, database developers and business analysts by introducing a SQL-friendly approach. Underneath the hood, Teradata Aster is using a compute approach that allows the data to leverage the power and performance of massively parallel analytic processing engines and pre-built algorithms.

Key messages

- SQL-based analytics remains as relevant to enterprises today as ever, but new approaches are needed for addressing queries that would be difficult to answer using SQL.
- Connection Analytics is an emerging discipline that provides answers to persistent business questions such as identification and influence of thought leaders, impact of external events or players on financial risk, or analysis of network performance based on causal relationships between nodes.
- Key challenges for Connection Analytics is making the approach accessible to the existing base of enterprise SQL developers and business analysts without requiring knowledge of specialized programming languages, and the ability to partition a problem to make it solvable through scale-out, massively parallel compute engines.
- Teradata is the first player to introduce an approach that makes Connection Analytics accessible to SQL developers and business analysts, with an architecture that effectively leverages an MPP analytic platform.
- Teradata's Aster Graph engine enables connection analytics through SQL and pre-build algorithms around social network analysis, fraud detection and machine learning delivering powerful insights to address use cases like customer retention, marketing attribution & product affinity

BIG DATA EXPANDS ANALYTIC HORIZONS

Beyond SQL

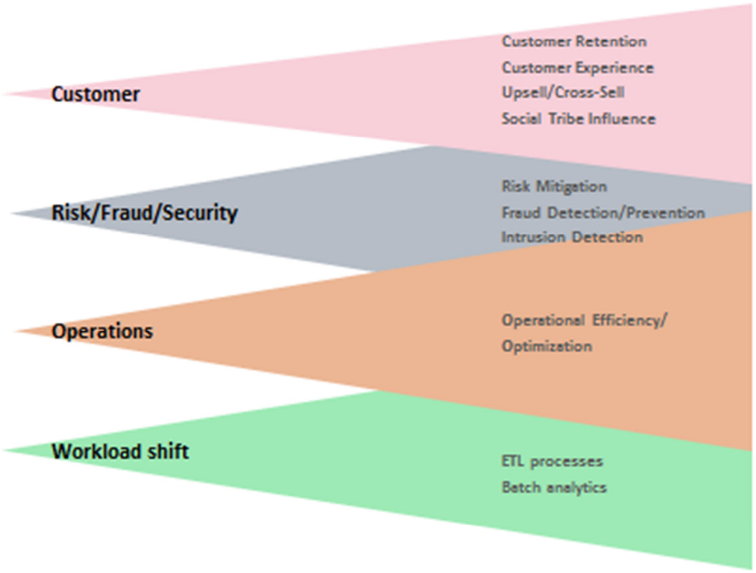
SQL, the de facto standard query language for enterprise transaction databases and data warehouses, has been a huge success story because it provides a syntax allowing non-programmers to query databases. SQL relational data warehouses have in turn provided the platforms on which enterprises continue to rely for storing and managing structured data aggregated from operational systems, and generating the reports and analytics that have helped organizations meet their strategic and operational goals.

However, explosion of data from nontraditional sources outside enterprise transaction systems promises to complement the insights that organizations have realized from their data warehouses. Data sources such as social media text, messaging, log files (such as clickstream data), and machine data from sensors in the physical world present the opportunity to pick up where transaction systems leave off regarding underlying sentiment driving customer interactions; external events or trends impacting institutional financial or security risk; or adding more detail regarding the environment in which supply chains, transport or utility networks operate.

Big Data analytics emerged as a result of a perfect storm. Phenomena akin to Moore's Law for processors is now making increase storage capacity, connectivity, and networking bandwidth accessible on economic, off-the-shelf commodity hardware. New data platforms and compute frameworks are capable of harnessing the capacity and power of massive scale-out cluster architectures that evolved out of Internet data centers. No longer need enterprises make tradeoffs between rich and reach; they can submit highly complex queries against enormous, highly varied data sets without concern that new data with varying structure will break their queries. These innovations are also bringing the capability for asking questions that, in the SQL environment, would otherwise require tens or hundreds of table joins.

While Big Data analytics originated with Internet companies seeking to solve problems unique to their business such as ad optimization or search indexing, enterprises are now seeking to harness the same capabilities for analyzing data at Internet scale to yield new insights on some very familiar problems, as shown in Figure 1.

Figure 1. Typical Big Data analytics use cases



Source: Ovum

Introducing Connection Analytics

Big Data analytics encompasses a large umbrella of analytic approaches. Although initially centered on MapReduce, a powerful framework for writing analytic problems to run on massively parallel, commodity clusters, numerous approaches are now emerging that perform different tasks, such as search and real-time streaming analysis. These analytics will complement, not replace traditional SQL query processing.

Among the most promising of the new Big Data analytic approaches is "Connection Analytics," a method for exploring data that analyzes changing relationships between people and/or "things." Connection analytics can be performed using multiple techniques like graph, machine learning, data mining and others.

Graph analytics is a prime example of Connection Analytics

While there are many ways to analyze connections, graph analytics provides one of the most powerful and intuitive approaches for gaining insight where there are many-to-many relationships. Such forms of relationships are often the most accurate way of portraying the real world. For instance, most people have relationships with other people in multiple social groups. Cars, busses, and trains in a city interact at neighborhood and regional level. Likewise, drugs that interact with other drugs have different impacts with different human organs. As such, this report will focus specifically on how graph computing can be used to better understand how relationships between connections impact business outcomes.

While traditional analytics can yield insight on why things happen, graph analytics provides a more granular view of the chain of events and players that influence an outcome. For marketing, connection analytics supplements focus group research. This approach can provide a highly intuitive approach to answering perennial questions that C-level managers in sales, marketing, product development, or operations have long been asking, such as:

- Who are the thought leaders that are driving market acceptance or rejections to new products, services, prices or promotions?
- What is the pattern of events that occur when new products go viral?
- What are the cascading effects of political events on markets and levels of financial risk among certain classes of customers?
- What are the causal relationships that underlie drug interactions for specific patients or classes of patients?

Deriving insight from relationships

By focusing on many-to-many relationships, Graph Analytics probes the factors that collectively influence an outcome. For instance:

- People – A person has relationships, not just with a single person, but with different groups and things. For instance, a person may be an opinion leader within one or more circles of friends that follow a certain music group but may be a follower when it comes to different domains, such as buying insurance or retail banking services. Or the picture may be mixed within the same domain, such as music or entertainment – a leader in some circles and a follower in others.

- Products – Products succeed on the perceptions of buyers within their target markets. In turn, the perception of a product will be affected by entry of new products to market. For instance, Nokia was a brand leader in feature phones, but lost that leadership when Apple's introduction of the iPhone created a new smartphone product category, and with it, perception of leadership. Likewise, the introduction of new services, such as gaming or social networks, has had close dependencies to the mobile platforms on which they are based and other phenomena, such as endorsements by celebrities.
- Things – Events that occur in the world are rarely ever truly isolated. For instance, heavy traffic on a highway may cause an accident to occur, which in turn could bottleneck traffic coming from adjacent neighborhoods or towns. Likewise, the rise or fall or a particular stock may in turn lead to different patterns of buying and selling across a category of stocks – depending on the nature of the event and whether that stock is considered a bellwether for a specific sector of the economy.

Graph Analytics use cases

Influencer analysis

This answers questions such as who are the thought leaders, and which people should marketers assign more priority to cultivating? This is an age-old question that has traditionally been addressed using focus groups. Graph analytics provides a more efficient means for identifying the relationships of people to different groups, and the nature of those relationships; it is well-suited to representing a complex picture where individuals are members of multiple circles, and in many cases, may play different roles or have different levels of influence from one to the next.

Product affinity analysis

Products are rarely bought in isolation. A generation ago, a Teradata retailing customer gained fame for "Beer and Diapers" market basket analysis insights that drew attention to the possibility that actual consumer buying patterns in some cases may be counter-intuitive. Connection analytics provides a more comprehensive picture that takes a different approach compared to classic market basket analysis; rather than look at aggregated data to draw inferences on product affinities, connection analytics probes individual interactions and pieces together relationships or product affinities from the ground up.

Network analysis

This can apply to people, processes, or things. It can be used for dissecting the activities of person-to-person networks over time (such as when awareness of an event goes "viral"); or examine the cascading effects of disruptive events, as detected from sensory data, to supply chain networks, smart grids, smart urban infrastructure; telecommunications networks, and so on.

Ripple effects

This gauges the downstream impacts of an event based on the degree of interconnectedness. This can be used to predict the impact of occurrences such as:

- Housing foreclosures, where the impact of a single foreclosure can be graphed against housing prices on the rest of the block, or in aggregate where the incidence of foreclosures is analyzed at a neighborhood level.
- A merchant suddenly drops the price of a product in a brick-and-mortar store or online venue like eBay; what is the likely impact of the price change on the same and related products?
- Water supply contamination, where the site of a leakage, the direction and strength of water flow, and connections with other distribution pipes in a network can predict the downstream impact of a toxic leak.

Fraud and money laundering

This examines the impact of questionable financial transactions is examined across networks, institutions, and customers. While identities can be spoofed, the behavior of entities (people, botnets, or institutions) provides the real picture on suspect patterns of transactions. The capacity of Big Data analytics platforms allows the sample set to be expanded out to identify behaviors that would otherwise not be apparent with smaller time slices.

HOW THE TERADATA ASTER DISCOVERY PLATFORM SUPPORTS CONNECTION ANALYTICS

Bridging the skills gap

To embrace Big Data analytics, enterprises faced a dilemma; while their IT organization had well-established SQL skills, Big Data analytics required knowledge of new platforms and processing frameworks such as Hadoop and MapReduce. The rate by which new analytics approaches are emerging compounds the skills challenge. Addressing the gap, Ovum believes that analytic data platforms are converging by adding support for multiple data storage engines and analytic engines. Teradata has staked an aggressive position to makes many of these new approaches accessible to the core base of SQL developers.

Aster Discovery platform bridges SQL with Big Data analytics

The Aster Discovery Platform is a NewSQL analytic database designed for in-database analytics through a massively parallel, shared-nothing architecture that supports end-to-end parallelism for all operations. Teradata positions the Aster platform as a "discovery" platform within its Unified Data Architecture (UDA); the Aster Discovery platform complements the Teradata data warehousing platform for core, repeatable analytics, and Hadoop as the data lake for Big Data. Ovum's latest Decision Matrix for analytic databases cited the Teradata Aster Discovery platform for its support of multi-structured data exploration, and for its strategy to unify the different data-processing engines within its product family.

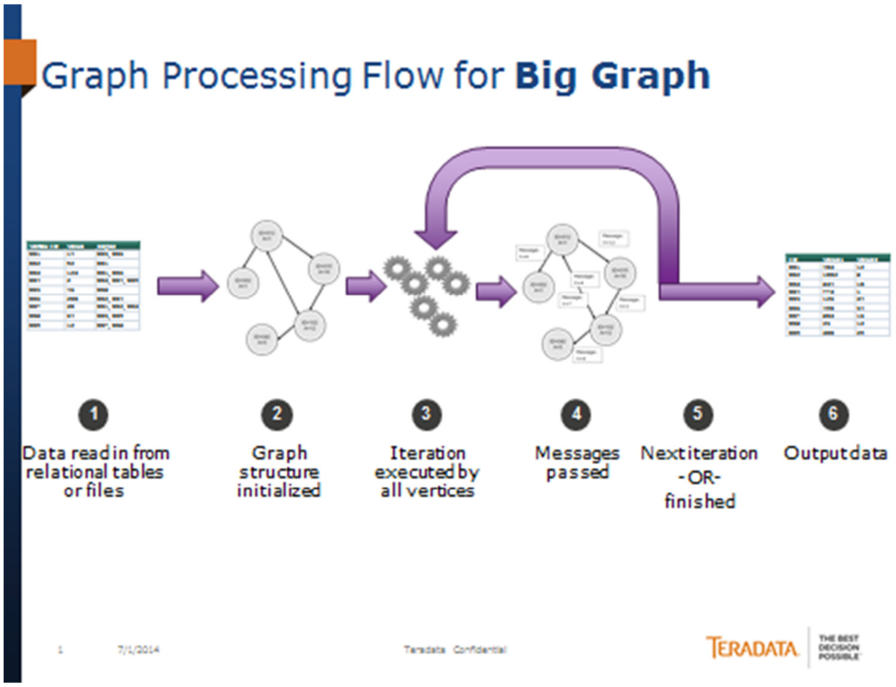
SQL GR -- A SQL entryway to Connection Analytics

Teradata offers several analytic engines, data stores and functions that make analytic approaches, such as MapReduce processing, statistical analysis, graph analysis, text processing, path analytics, and time series queries, accessible through SQL. It does so by allowing the user to specify SQL commands (e.g., "Group By" for selecting data), then bridging to Java to partition the data sets and redistributed them across the cluster for processing and returning the results. For instance, Teradata's patented SQL-MapReduce allows any analytics code running in-database or any MapReduce function to be incorporated into an analytic application through a SQL statement; in turn, Teradata's nPath analytic function is a sequential and trending pattern analysis built on SQL MapReduce that discovers relationships between rows of data.

The latest addition to this portfolio is Teradata's patent-pending SQL based Graph engine called SQL-GR ™ , which makes graph processing and Connected Analytics through a similar SQL-oriented approach (see figure 2). As the first analytics provider to make graph analytics accessible to SQL developers, Teradata's strategy is to simplify the process so analysts can focus on solving the problem rather than building algorithms. As part of SQL-GR, Teradata provides six pre-packaged graph algorithms to analyze specific types of relationships, including:

- Page Rank -- Made famous by Google, this counts the number and quality of links to a page to estimate the importance of a website.
- Eigen Centrality -- This assesses the strength of a person or thing's(a "vertex") connection to other highly connected people or things in a network. For instance, this technique can be used for analyzing which people are members of the most social tribes; these actors often play influential roles in helping events, perceptions, or other phenomena go viral.
- Betweenness -- This finds the shortest paths between different people or things in a network, which can be helpful when discovering how different communities or groups of things are connected.
- Closeness -- This determines the closeness of a relationship between two specific people or things, which can be used to estimate how quickly information will travel between them.
- Local clustering coefficient -- This metric quantifies how close a person or thing's neighbors are to a complete group.
- All pairs shortest path -- A measure of the shortest distance between every pair (of people or things) in a graph.
- Loopy belief propagation -- This quantifies the degree of uncertainty in a relationship between any two neighboring people or things; this approach is used when information on relationships or interactions is spotty, and is used to help fill in the blanks.

Figure 2. SQL-GR's integration with SQL



Graph Processing Flow for **Big Graph**

1. Data read in from relational tables or files
2. Graph structure initialized
3. Iteration executed by all vertices
4. Messages passed
5. Next iteration -OR- finished
6. Output data

1 7/1/2014 Teradata Confidential

TERADATA THE BEST DECISION POSSIBLE

Source: Teradata

SQL-GR is not the only graph computing engine; there are a number of frameworks and technologies (many of them open source) that support graph computing. Many current alternatives are full-blown graph databases. While graph databases can persist relationships, the downside is that data must be stored in specialized "triple stores" that require that objects (people or things) and predicates (attributes) that are accessible only with specialized languages (e.g., SPARQL). SQL-GR differentiates from these approaches, not only through its ability to directly apply graph analytics on relational of file based data and its accessibility to SQL developers, but also its capabilities to massively scale graph compute problems to take advantage of Teradata Aster's MPP engine via a Bulk Synchronous Processing (BSP) approach. That can significantly reduce processing lead time and overhead, as SQL-GR can be directed to address just the relevant portion of a scenario, rather than recomputing the graph of the entire data set. Furthermore, by continuing to rely on a relational database, the Teradata approach does not require organizations to learn how to deploy and administer unfamiliar data stores such as triple stores, or unfamiliar access languages such as RDF.

Because Teradata's approach is SQL-friendly, users can more readily integrate or embed Connection Analytics to business applications. For instance, the output of a Connected Analytics application run with SQL-GR graph analysis could enrich customer records in a CRM system. Or multiple analytic techniques like social network analysis with sentiment analysis can be combined together to identify influential unhappy customers.

Connecting Connected Analytics

Connected Analytics add a new weapon for discovery because it can add context to data. For instance, a Connected Analytics run can enrich clinical discovery data by factoring potential drug interactions; likewise, it can enrich risk management applications by examining the impacts of external events on the creditworthiness of a customer or the wisdom of completing a financial transaction. Connected Analytics can also work in conjunction with other analytics techniques to generate insights. For instance, a graph computing run can identify the relative levels of influence of specific customers and groups; then that can be combined with sentiment analysis from a text analytics process to gauge the sentiment of key influencers, which enables to enterprise to take actions to raise the level of satisfaction of its most influential customers.

Teradata Aster's SNAP framework was developed with such scenarios in mind. SNAP allows the user to orchestrate the running of multiple analytic engines such as SQL, MapReduce, or Graph, to solve a problem. It allows analysts with SQL and BI/Visualization skills to solve analytical challenges by combining the right data with the right analytical tools.

RECOMMENDATIONS FOR ENTERPRISES

Until now, the technical barriers to graph computing made it difficult for enterprises to take advantage of Connection Analytics; the languages were too specialized, and the compute overhead significant. Furthermore, the structure of graph models that underlie Connection Analytics sounds extremely cryptic and abstract: nodes and vertices.

However, translated to ordinary English, graph structured are very intuitive, consisting of entities and the nature of relationships that connect one entity to another. Entities can be persons, places, or things. With the structure of graph compute constructs demystified, the next step of making Connection Analytics accessible to SQL developers and analysts clears many of the remaining hurdles aside. The challenges are no longer primarily technology, but the ability to understand how and where to employ connection Analytics.

There are clearly use cases in areas that are already familiar to enterprises, whether that is in areas of customer engagement, risk mitigation, fraud prevention, security, clinical research, and network operations. The hard part is not necessarily the understanding of how to structure the model, because entities and relationships are often baked into the design of conventional querying and analytics. As with any new form of analytics, the challenge -- and the solution -- is taking an iterative approach to building the models, and then determining whether the answers make logical sense.

APPENDIX

Author

Tony Baer, Principal Analyst, Ovum IT Information Management

tony.baer@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum’s consulting team may be able to help you. For more information about Ovum’s consulting capabilities, please contact us directly at consulting@ovum.com.

Disclaimer

The facts of this report are believed to be correct at the time of publication but cannot be guaranteed. Please note that the findings, conclusions, and recommendations that Ovum delivers will be based on information gathered in good faith from both primary and secondary sources, whose accuracy we are not always in a position to guarantee. As such Ovum can accept no liability whatever for actions taken based on any information that may subsequently prove to be incorrect.