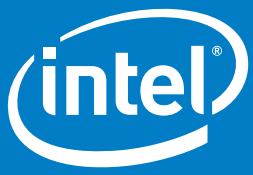

Accelerate Insight and Impact Using Big Data Graph Analytics

The Power of Relationships in Data Analytics

The ability to capture and mine massive data stores—both structured and unstructured—is changing business models and accelerating the pace of discovery for research and industry. Current analytic technologies only begin to tap the insight latent in Big Data. New sources of data (including some previously discarded, such as log files) are now affordably captured, offering a wealth of new opportunities for data-driven impact. Examples range from deeply understanding and servicing an individual customer's preferences, to spotting subtle patterns that indicate serious fraud and security risks, to uncovering critical operational trends and preempting anomalies that improve service quality and efficiency, to uncovering previously unseen relationships between data and trends in scientific analyses.

Distilling the important signals buried in diverse data becomes easier and more efficient when we can understand and explore the relationships connecting individual data points. This context is diverse, depending on the real-world system being modeled. Examples include physical and temporal proximity, social and familial relations, preferences and attitudes, communication and physical paths, and connections, hyperlinks, and references. High degrees of connectedness are especially prevalent in many of the growing new areas of data constituting Big Data sources, such as Web content, system logs, IMs, emails, tweets, blogs, call records, and images. The quantity of highly connected data continues to explode, with newly digitized information and connected devices such as data from sensors, wearable devices, digitized health records, video cameras, and increasingly affordable personalized genomic sequencing.

In the world of linked data, analyzing how data objects relate to each other and what the patterns of connections reveal is as important, or more so, than simply classifying and summarizing just the individual data objects. And the number of connections are frequently more numerous than the number of data objects they connect. The power to mine and predict behavior by uncovering the hidden relationships in inherently networked information is increasing interest in and use of **graph analytics**, the science of applying algorithms to tackle the unique challenges of analyzing connected data.

Graph Analytics

"Graphs" are data models that structure data as entities (called "vertices") and their connections ("edges"). Because this data is structured as a network of connected objects, it can be visualized like a network or a tree and, in fact, whenever you have a system of information that can be naturally visualized or expressed in network form you are seeing a system that lends itself to being modeled as a graph. Modeling data as a graph is a very intuitive way of structuring data, because it aligns well with how the human mind works.

Graph analytics is the science of applying algorithms to answer questions about how the objects in these graphs are connected. Graph analytics tools allow us to visualize and mine connected data in ways that are significantly faster, more intuitive, or in some cases not practical, using traditional database and data exploration tools. For instance, in social network services, graph analytics algorithms underlie the capability to quickly identify the shortest path relating people of similar interests; generate recommendations about new personalities to follow because of common interests; or auto-complete a search request that insightfully predicts intent before the query is completely entered.

Beyond search and social networking, graph analytics on Big Data is now poised to impact a broad range of vertical industries and endeavors as use of connected data grows and the tools evolve to be more accessible and affordable across industries. Use cases are already deployed or being explored for targeted and personalized recommendations for commerce, increased network security through better predictions about unknown websites or downloads, subtle pattern detections indicating advanced persistent threats, fraud prevention based on uncovering clusters of related incidents, and many applications to uncover insight across a wide range of scientific and medical data.

Moving from Broad Correlations to More Precise Predictions

Let's take a closer look at recommender systems, a use case that showcases the value of graph analytics. Tailored recommendations generate a much higher response rate and therefore are much more valuable to vendors. (Examples include the high-quality recommendations from Netflix or Amazon.)

It can be difficult to tailor recommendations when the vendor knows little about the customer, such as their purchase history. That's where graph analytics comes into play: filling in the "missing" pieces in a customer's history by predicting interests using inferences from relationships contextually specific to that user and transaction type. For example, similar products used by people influential to that consumer may be a good predictor of interests, with the chosen relationship contexts defining similarity and "influence." Different product preferences may be indicated by social, family, or work relationships and demographics, life situation, or geographic proximity. Any number of relevant relationships may provide valuable factors to making a superior recommendation, and thus achieving a competitive advantage. This is in stark contrast to generalized recommendations, such as using the mean of a population of consumers to infer preferences. In contrast, graph analytics targets granularity that can weigh specific relationships around an individual.

"Graph analytics on Big Data is now poised to impact a broad range of vertical industries and endeavors."

It isn't hard to envision how technology perfected for tailoring product purchase recommendations could be extended to tailoring personalized offers. Or, a personalized course of medical treatment could be recommended based on healthcare outcome data that is captured and mined for patterns of insight.

Storing and accessing huge data sets across multiple sources magnifies familiar datacenter challenges, such as scalability, data security and privacy, network performance, and storage costs. As the scope of analytics expands to billions and trillions of data points, usability and scalability requirements are pushed to new levels. Working with graph data brings new complexities, such as speedily traversing and searching across objects that are arbitrarily connected, often across several dimensions. This challenges the underlying data placement and indexing optimizations of traditional database solutions, and highlights the opportunity to use the growing category of graph database and analytic engines that are optimized to work with graph data. Such solutions enable powerful new analytic and transactional models and introduce new query and analytic interfaces, rather than use paradigms familiar today, such as relational database and statistical analytics.

Another challenge is visualization. A powerful use for graph models is intuitively seeing the connections and interactively traversing from object to object along the relevant edges—as seamlessly as clicking on hyperlinks when Web surfing. Traditional databases and visualization tools simply cannot support this level of interactivity. These graph visualization tools meld together the ease of interaction with seamless, behind-the-scenes use of graph analytics to narrow the visual scope of the information. For example, showing only the most relevant data by utilizing graph analytics algorithms to discover and display only strongly connected vertices, while filtering out less relevant and more distant connections (preventing massive visual clutter).

These tools—graph databases, graph analytics engines, and graph visualization—all start with the premise of a relevant set of graph-structured data on which to operate. Much of the time data scientists and analysts spend before ever performing analytics or writing reports goes to manipulating data. Some data scientists estimate the majority of their time goes into the laborious task of merging data sources, cleaning outliers, extracting model parameters from raw data, or manipulating raw data to create the desired parameters (called feature engineering). To these "data wrangling" challenges, graph analytics adds a new one known as "graph construction," which is structuring vast amounts of raw data into a well-structured form of vertices, edges, and their properties, and correctly stitching together the related objects with the desired connections.

Lastly, when using graphs that capitalize on the connections in Big Data, timely throughput requires performing all these data preparations, graph constructions, and analytics using cluster computing, bringing another layer of deployment, performance optimization, and management challenges.

Enabling Solutions for Accessing Big Data Insight

To address all of these challenges, Intel, along with the wider ecosystem, is developing effective, efficient graph data tools and solutions that unlock Big Data advantages in organizations worldwide. Intel works closely with the open source community to enhance existing software tools and applications and contribute innovative new capabilities back to the open source community. Intel® Datacenter software products and ongoing collaboration with the Big Data ecosystem at each layer of the software stack is resulting in maximizing performance and efficiency on standardized Intel® architecture. Additionally, Intel® platform technologies, such as processors, 10GbE converged Ethernet adapters, Solid-State Drives (SSDs), and fabrics are being optimized for Big Data applications.

Intel® Graph Builder for Apache Hadoop* Software v2

Intel Graph Builder open source software libraries simplify the creation of graph data models, enabling data scientists to focus on solving the business problem at hand instead of formatting data. To simplify data wrangling and graph construction, Graph Builder provides prebuilt, scriptable automation routines that simplify and speed the process of extracting and formatting Big Data into graph form at scale using Hadoop cluster computing. Data scientists or analysts using graph tools, whether graph databases, graph analytics platforms, or graph visualization tools, can use Graph Builder libraries to more simply create their Hadoop application to clean, transform, and connect unstructured, semistructured, and structured data into graph form.

Intel® Graph Builder for Apache Hadoop* Software v2 provides:

- Popular and convenient scripting (using Apache Pig*) to clean and transform data at scale, including string and big table manipulation, null checks, and math operators
- Automation routines to transform data into rich graph representations that model real-world problems with support for property graphs, multiple edge labeled graphs, directed or undirected graphs
- Flexibility to provide output to a wide range of graph analytics and visualization tools using the widely supported Resource Description Format (RDF)
- Parallel execution for fast throughput and iteration up to trillion-edge networks

Ecosystem activities include Intel Graph Builder support for bulk loading into the scalable Titan* graph database. Titan complements the Intel® Distribution for Apache Hadoop* software and Graph Builder by providing an open source, scalable graph database for storing and querying graphs containing hundreds of billions of vertices and edges distributed across a multimachine cluster.

"Intel, along with the wider ecosystem, is developing effective, efficient graph data tools and solutions that unlock Big Data advantages."

Intel Distribution for Apache Hadoop Software

Intel is committed to developing a platform on which the entire ecosystem can build next-generation analytics solutions, including graph analytics. Intel believes using Hadoop across the analytics toolbox simplifies the infrastructure and reduces costs (versus using multiple siloes of independent solutions). The Intel Distribution for Apache Hadoop software is the only distribution built from silicon up to enable the widest range of data analysis on Apache Hadoop. It is the first with hardware-enhanced performance and security capabilities, and the only open source platform for Big Data with support from a Fortune 100 company. The code has been optimized for the latest hardware platform technologies, including crypto acceleration, SSD storage, and 10GbE networking, enabling deployments that support data confidentiality at minimal encryption overhead. Management capabilities have also been added to make Hadoop easier to deploy and operate.

Quick Glossary

Graph Analytics: The application of algorithms on graph data to solve business or computing problems, as well as a description of the category of tools that provide these capabilities.

Graph Builder: Intel® open source software libraries that run on Hadoop* for constructing large-scale graphs.

Graph: A way of representing data as networks of relationships (technically as vertices and edges).

Graph Construction: The process of choosing how to connect data points based on modeling how data is related in different contexts, and restructuring the raw data into structured lists of vertices, edges, and their properties.

Machine Learning: A form of computing intelligence that learns from patterns in data, and can apply those learnings for the purpose of making predictions.

Intel® Datacenter Software

Intel® Distribution for Apache Hadoop* Software
Distributed processing and data management for enterprise applications analyzing massive, diverse data.

Intel® Expressway Product Family
Software appliance delivering cloud service brokerage capabilities with secure, scalable APIs.

Intel® Enterprise Edition for Lustre* Software
Tap into Lustre's power and scalability, with simplified installation, configuration, and monitoring from Intel.

Intel® Cache Acceleration Software
High performance media caching.

Intel® Datacenter Manager (Intel® DCM)
Provides accurate, real-time power management and monitoring for datacenter servers.

Intel® Virtual Gateway
Tools for virtual server management.

Learn more about Intel® Datacenter Software tools at: intel.com/datacentersoftware

Explore Intel® Graph Builder for Apache Hadoop* Software v2 and other resources for graph analytics at: intel.com/graph