Graph Analytics and Big Data

Author, Ajit Singh

A Data Science Foundation White Paper

May 2019

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

Ajit Singh

Assistant Professor

Patna Womens College

Bihar, India

Communication Author : Ajit Singh (www.ajitvoice.in)

ABSTRACT

Graph analytics, which is an analytics alternative that uses an abstraction called a graph model. The simplicity of this model allows for rapidly absorbing and connecting large volumes of data from many sources in ways that finesse limitations of the source structures (or lack thereof, of course). Graph analytics is an alternative to the traditional data warehouse model as a framework for absorbing both structured and unstructured data from various sources to enable analysts to probe the data in an undirected manner.

Big data analytics systems should enable a platform that can support different analytics techniques that can be adapted in ways that help solve a variety of challenging problems. This suggests that these systems are high performance, elastic distributed data environments that enable the use of creative algorithms to exploit variant modes of data management in ways that differ from the traditional batch-oriented approach of traditional approaches to data warehousing.

Keywords: Graph Analytics, Big Data, Graph Analytics Algorithm, Graph Algorithm Features

1. INTRODUCTION GRAPH ANALYTICS

It is worth delving somewhat into the graph model and the methods used for managing and manipulating graphs:

What constitutes graph analytics?

- Types of problems that are suited to graph analytics.
- Types of questions that are addressed using graph analytics.
- Types of graphs that are commonly encountered.
- The degree of prevalence within big data analytics problems.

This motivates an understanding of its utility and flexibility for discovery-style analysis in relation to specific types of business problems, how common those types of problems are, and why they are nicely abstracted to the graph model. In addition, we will discuss the challenges of attempting to execute graph analytics on conventional hardware and consider aspects of specialty platforms that can help in achieving the right level of scalability and performance.

THE SIMPLICITY OF THE GRAPH MODEL

Graph analytics is based on a model of representing inpidual entities and numerous kinds of relationships that connect those entities. More precisely, it employs the graph abstraction for representing connectivity, consisting of a collection of vertices (which are also referred to as nodes or points) that represent the modeled entities, connected by edges (which are also referred to as links, connections, or relationships) that capture the way that two entities are related.

The flexibility of the model is based on its simplicity. A simple unla-beled undirected graph, in which the edges between vertices neither reflect the nature of the relationship nor indicate their direction, has limited utility.

Among other enhancements, these can enrich the meaning of the nodes and edges represented in the graph model:

- Vertices can be labeled to indicate the types of entities that are related.
- Edges can be labeled with the nature of the relationship.
- Edges can be directed to indicate the “flow” of the relationship.
- Weights can be added to the relationships represented by the edges.
- Additional properties can be attributed to both edges and vertices.
- Multiple edges can reflect multiple relationships between pairs of vertices.

REPRESENTATION AS TRIPLES

In essence, these enhancements help in building a semantic graph—a directed graph that can be represented using a triple format consisting of a subject (the source point of the relationship), an object (the target), and a predicate (that models the type of the relationship).

A collection of these triples is called a semantic database, and this kind of database can capture additional properties of each triple relationship as attributes of the triple. Almost any type of entity and relationship can be represented in a graph model, which means two key things: the process of adding new entities and relationships is not impeded when new datasets are to be included, with new types of enti-ties and connections to be incorporated, and the model is particularly suited to types of discovery analytics that seek out new patterns embedded within the graph that are of critical business interest.

GRAPHS AND NETWORK ORGANIZATION

The concept of the social network has been around for many years, and in recent times has been materialized as communities in which inpidual entities create online personas and connect and interact with others within the community. Yet the idea of the social network extends beyond specific online implementations, and encompasses a wide variety of example environments that directly map to the graph model. That being said, one of the benefits of the graph model is the ability to detect patterns or organization that are inherent within the represented network, such as:

Embedded micronetworks: Looking for small collections of entities that form embedded “microcommunities.” Some examples include determining the originating sources for a hot new purchasing trend, identifying a terrorist cell based on patterns of communication across a broad swath of call detail records, or sets of inpiduals within a particular tiny geographic region with similar political views.

Communication models: Modeling communication across a community triggered by a specific event, such as monitoring the “buzz” across a social media channel associated with the rumored release of a new prod-uct, evaluating best methods for communicating news releases, or correlation between travel delays and increased mobile telephony activity.

Collaborative communities: Isolating groups of inpiduals that share similar interests, such as groups of health care professionals working in the same area of specialty, purchasers with similar product tastes, or inpiduals with a precise set of employment skills.

Influence modeling: Looking for entities holding influential positions within a network for

intermittent periods of time, such as computer nodes that have been hijacked and put to work as proxies for distributed denial of service attacks or for emerging cybersecurity threats, or inpiduals that are recognized as authorities within a particular area.

Distance modeling: Analyzing the distances between sets of entities, such as looking for strong correlations between occurrences of sets of statistically improbable phrases among large sets of search engines queries, or the amount of effort necessary to propagate a message among a set of different communities.

Each of these example applications is a discovery analysis that looks for patterns that are not known in advance. As a result, these are less suited to pattern searches form relational database systems, such as a data warehouse or data mart, and are better suited to a more dynamic representation like the graph model.

2. CHOOSING GRAPH ANALYTICS

Deciding the appropriateness of an analytics application to a graph analytics solution instead of the other big data alternatives can be based on these characteristics and factors of business problems:

Connectivity: The solution to the business problem requires the analysis of relationships and connectivity between a variety of different types of entities.

Undirected discovery: Solving the business problem involves iterative undirected analysis to seek out as-of-yet unidentified patterns.

Absence of structure: Multiple datasets to be subjected to the analysis are provided without any inherent imposed structure.

Flexible semantics: The business problem exhibits dependence on contextual semantics that can be attributed to the connections and corresponding relationships.

Extensibility: Because additional data can add to the knowledge embedded within the graph, there is a need for the ability to quickly add in new data sources or streaming data as needed

for further interactive analysis.

Knowledge is embedded in the network: Solving the business problem involves the ability to exploit critical features of the embedded relationships that can be inferred from the provided data.

Ad hoc nature of the analysis: There is a need to run ad hoc queries to follow lines of reasoning.

Predictable interactive performance: The ad hoc nature of the analysis creates a need for high performance because discovery in big data is a collaborative man/machine undertaking, and predictability is critical when the results are used for operational decision making.

3. GRAPH ANALYTICS ALGORITHMS AND SOLUTION APPROACHES

The graph model is inherently suited to enable a broad range of analyses that are generally unavailable to users of a standard data warehouse framework. As suggested by these examples, instead of just providing reports or enabling online analytical processing (OLAP) systems, graph analytics applications employ algorithms that traverse or analyze graphs to detect and potentially identify interesting patterns that are sentinels for business opportunities for increasing revenue, identifying security risks, detecting fraud, waste, or abuse, financial trading signals, or even looking for optimal inpidualized health care treatments. Some of the types of analytics algorithmic approaches include:

Community and network analysis, in which the graph structures are traversed in search of groups of entities connected in particularly “close” ways. One example is a collection of entities that are completely connected (i.e., each member of the set is connected to all other members of the set).

- Path analysis, which analyze the shapes and distances of the different paths that connect entities within the graph.

- Clustering, which examines the properties of the vertices and edges to identify characteristics of entities that can be used to group them together.
- Pattern detection and pattern analysis, or methods for identifying anomalous or unexpected patterns requiring further investigation.
- Probabilistic graphical models such as Bayesian networks or Markov networks for

various application such as medical diagnosis, protein structure prediction, speech recognition, or assessment of default risk for credit applications.

- Graph metrics that are applied to measurements associated with the network itself, including the degree of the vertices (i.e., the number of edges in and out of the vertex), or centrality and distance (includ-ing the degree to which particular vertices are "centrally located" in the graph, or how close vertices are to each other based on the length of the paths between them).

These graph analytic algorithms can yield interesting patterns that might go undetected in a data warehouse model, and these patterns them-selves can become the templates or models for new searches. In other words, the graph analytics approach can satisfy both the discovery and the use of patterns typically used for analysis and reporting.

4. DEDICATED APPLIANCES FOR GRAPH ANALYTICS

There are different emerging methods of incorporating graph analytics into the enterprise. One class is purely a software approach, providing an ability to create, populate, and query graphs. This approach enables the necessary functionality and may provide the ease-of-implementation and deployment. Most, if not all, software implementations will use industry standards, such as RDF and SPARQL, and may even leverage complementary tools for inferencing and deduction. However, the performance of a software-only implementation is limited by its use of the available hardware, and even using commodity servers cannot necessarily enable it to natively exploit performance and optimization.

Another class is the use of a dedicated appliance for graph analytics. From an algorithmic perspective, this approach is equally capable as one that solely relies on software. However, from a performance perspective, there is no doubt that a dedicated platform will take advantage of high-performance I/O, high-bandwidth net-working, in-memory computation, and native multithreading to provide the optimal performance for creating, growing, and analyzing graphs that are built from multiple high-volume data streams. Software approaches may be satisfactory for smaller graph analytics problems, but as data volumes and network complexity grow, the most effective means for return on investment may necessitate the transition to a dedicated graph analytics appliance.

5. CONCLUSION

The concept of the social network has been around for many years, and in recent times has been materialized as communities in which inpidual entities create online personas and connect and interact with others within the community. Yet the idea of the social network extends beyond specific online implementations, and encompasses a wide variety of example environments that directly map to the graph model. That being said, one of the benefits of the graph model is the ability to detect patterns or organization that are inherent within the represented network.

The graph model allows you to tightly couple the meaning of entity relationships as part of the representation of the relationship. This effectively embeds the semantics of relationships among different entities within the structure, providing an ability to both invoke traditional-style queries (to answer typical “search” queries modeled after known patterns) and enable more sophisticated undirected analyses. These undirected “discovery” analyses, include inferencing, identification of interesting patterns, and application of deduction, all using an iterative approach that analysts can use to discover actionable knowledge that was previously unknown. This allows the analysts to rapidly seek out emerging patterns and enable real-time knowledge-driven decision making in the context of how these newly discovered patterns impact the corporate business value drivers.

5. REFERENCES

1. http://www.w3.org/RDF
2. http://www.w3.org/TR/rdf-sparql-query
3. Big Data Analytics, By Ajit Singh, Amazon KDP LLC
4. http://www.tutorialpoinns.com
5. http://www.researchgate.com

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation