Bloor

MarketUpdate

Graph and RDF databases 2015

Market basics

Graph databases represent a significant growth area. Indeed, research suggests that it is the fastest growing segment of the database market. There are arguably three reasons for this growing interest and each pertains to a particular sub-sector of the graph market. The first is that graph databases are designed to handle many-to-many relationships, which relational databases are not, so graph products have particular advantages in operational and transactional environments where this is the case. The second is with respect to semantics, which is of growing importance in its own right, and which graphs handle with ease. The third is that many graph analytic algorithms require a significant degree of iteration, which is not available with MapReduce and, where many-to-many relationships are involved, are not easily parallelised in conventional warehousing environments.

Not surprisingly, as an emerging market there are vendors coming at this space from a variety of different directions: with purpose-built native graph stores or with graph implementations on top of other types of database or file store, and with products that are aimed at different functional requirements.

As a result the graph database market is not homogeneous. Previous attempts to classify the market have distinguished between "graph databases" on the one hand and "graph compute engines" on the other, where the former tend to be more operationally focused and the latter are targeted at data warehousing and analytics. However, we believe that this is too simplistic and think that it is appropriate to distinguish between RDF (resource description framework) databases, which are often targeted at semantic applications or environments that involve semantics, and graph databases, which are less semantically oriented. Further, when the original two-tier distinction was proposed (by Neo4j) there were no graph databases per se in the data warehousing space. That is no longer true, so we need a new description. We have opted for RDF databases, operational graph databases and analytic graph databases. There is, of course, overlap between these categories, as explained in the following brief descriptions:

- **RDF databases.** Often semantically focused. Often, but not always, based on non-graph underpinnings (including relational databases). For use in operational environments but have inferencing capabilities. Require indexes even in transactional environments. Often ACID compliance.
- **Operational graph databases.** Tend to be native graph stores or built on top of a NoSQL platform. Focused at transactions (ACID) and operational analytics. No absolute requirement for indexes though these will typically be offered in order to improve query performance.
- **Analytic graph databases.** Some vendors focus on solving "known knowns" problems (the majority) where both entities and relationships are known, while others are more focused on known unknowns and even unknown unknowns. Multiple approaches characterise this area with different architectures including both native and non-native stores, different approaches to parallelisation, and the use of advanced algebra.

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is over 2.5 and Challenger if it is less than 2.5. The exact position in each segment is calculated based on their combined innovation and overall score.

A further distinguishing factor is in the languages supported by the different vendors. Most vendors support SPARQL (SPARQL Protocol and RDF Query Language), which is a W3C standard declarative language, but users often prefer to employ other options.

There are three other declarative languages available from different vendors, including one that is an extended form of SQL. Most graph products also support traditional languages such as Java while there are also specialised graph traversal languages such as Gremlin. A number of vendors with triple stores targeted at semantic processing support OWL (Web Ontology Language – so named because Owl in Winnie the Pooh misspelled his name as WOL).

For a detailed discussion of the types and architectures and uses of graph products see the Bloor Research Spotlight paper: ***"All about graphs: a primer"***.

Market trends

The biggest trend is simply towards graph products in general, as more and more products appear on the market. In particular, more of the major vendors are getting involved in this market. IBM, for example, has triple store support in DB2, has a graph database known as System G in R&D, and is embedding third party graph databases (at least two) into forthcoming products. Informatica and others are also embedding graph databases. Microsoft too has been researching graph technology. Similarly, MarkLogic is touting its triple store capabilities. Teradata has been offering graph analytics based on its Aster Data platform for some time, Oracle has been similarly active both for semantics and analytics, and SAP has introduced a graph engine into HANA. A corollary to this burgeoning interest in graphs is that there is not enough space in the market for all the vendors that are in it. Over time we can expect many of the current incumbents to either be acquired or to disappear altogether.

A secondary trend, which may be more of a marketing concept than a technical reality (depending on the vendor), is for RDF databases to make themselves over as operational graph databases. That is, the suppliers of these products are trying to expand beyond the confines of semantic processing to offer more general purpose operational graph processing, for example, by adding support for property graphs. Precisely what sort of operational graph applications these will be suitable for will depend on the product. For example, some products only support eventual consistency, which will not be suitable for certain types of transactional applications.

The market is split three ways in terms of SPARQL support: true believers, vendors (usually major ones) that support it because they think should but don't really care, and those that positively think that there are better alternatives. In practice, both KEL (Knowledge Engineering Language from Lexis Nexis) and Cypher (Neo4j) are more advanced in terms of functionality and performance, though that does not mean that these are without flaws either. Moreover, while both of these are open source they only work with their respective databases and until and if either of these is made to work more widely, then SPARQL will dominate the field for declarative languages. At present there is still scope for competition to arise because SPARQL is quite limited. You cannot, for example, add a time stamp to a relationship because this requires predicate attributes, which are not supported (though you can define relevant additional triples, which means that the graph expands). Recommendations have been made to the W3C that this be incorporated in the next version of this standard. If these sorts of additions can be added sooner rather than later then SPARQL will come to dominate this space.

Finally, it is worth commenting on the use of graph algorithms in analytics. Some vendors, such as Teradata, Franz and Oracle (through Parallel Graph Analytics: PGX) offer pre-built algorithms for analytics such as page ranking, finding shortest path, predicting future edges and so on. However, most suppliers rely on TinkerPop, which is a developer group working on an open source stack for graphs. Among its offerings are Gremlin, the Blueprints API and Furnace. The last of these is a graph algorithms package though it is only suitable for property graphs (that is, graphs that allow properties to be associated with the vertices and edges of the graph). There are around a hundred known graph algorithms but only a relative handful are available in a pre-built fashion at present. We expect this number to grow significantly.

Vendors

We have not attempted to analyse every graph or RDF product on the market, not least because there are so many of them. Those that are covered here have been included based on our own judgement and based on recommendations from Bloor Research subscribers. Because it is important to understand the positioning of the various offerings and their focus, the following provides a brief outline of those vendors/products that are included here.

Aladyn

Aladyn is a German company that markets Aronto, which has been under development for more than

ten years. Aronto is not a graph database per se but a self-service development environment (that is, there is no code and it is suitable for use by business domain experts – you build applications by linking ontological concepts using graph-based visualisations) that runs on top of the company's own graph database. It is not ACID compliant but existing deployments include applications such as asset management, configuration management, fleet management, service ticketing and so on.

Algebraix

SPARQL Server from Algebraix is not generally available yet and is currently in beta. It is interesting because data is represented algebraically within the database and goes to a level of mathematics that underlies set theory, including category theory (set theory for mappings). In theory, this level of algebra effectively provides a multi-model capability so that you can instantiate any database you like. As its name suggests SPARQL Server is specifically a graph product and it is targeted primarily at analytics, with SPARQL as the relevant query language.

Complexible

Complexible Inc. (which used to be known as Clark & Parsia after the founders of the company) are the developers of Stardog. Stardog is (currently) an RDF database with strong support for SPARQL and OWL (it supports all of OWL 2) and the company has embedded the Lucene search engine into Stardog. The database is ACID compliant and supports two-phase commit. A focus is on (model-driven) integration and analytics. The database uses query time reasoning that does not require the materialisation of inferences. It has a built-in optimiser for SPARQL. A major feature is that it provides graph versioning so that you can track changes to a graph, both for auditing and analysis purposes. In the company's forthcoming 3.1 release Stardog will be adding property graph and graph traversal capabilities along with support for TinkerPop and Gremlin, as well as graphing algorithms.

Cray

Urika-GD from Cray is what used to be known as YarcData Urika. It is an in-memory graph database that is delivered as an appliance (that is, it is completely pre-built and pre-installed). In fact, many graph databases make use of memory but not at the scale of Urika-GD, which can grow into hundreds of terabytes of memory (512TB). Urika-GD is targeted at the most intractable graph analytic problems and is specifically targeted at discovery analytics. While it can certainly handle known-known analyses it is aimed at environments where there are a lot of unknowns in large datasets. Examples include uncovering potential terrorist plots and medical research.

Datastax

DataStax has just acquired Aurelius, the developer of the Titan graph database. DataStax has committed to taking the product to version 1.0 and then hopes that the community (Titan is open source) will advance the product thereafter. DataSTax will, meanwhile, be developing its own graph capabilities. Titan is available to run on various NoSQL engines, including HBase, BerkeleyDB and Cassandra and, in principle, any version of BigTable. It is primarily targeted at operational environments but ACID compliance is dependent on the underlying database. Titan has also been integrated with Hadoop (generating MapReduce) to support more complex, batch-based analytics. You can run a single cluster with part of the cluster being used for operational purposes and part for analytics. Unlike some other products in this space, Titan employs a database schema, which has advantages when partitioning the data and for processing graph algorithms. Like other graph databases (but not triple stores) it support "index-free adjacency" – which means that each node points to adjacent nodes – but also supports the definition of additional indexes for performance and functionality reasons. The preferred programming language is Gremlin.

Franz

Franz Inc. has been in business for around 30 years and is the world's leading supplier of Lisp compilers. It started to develop AllegroGraph more than a decade ago. This is actually a quad store which you can implement as either an RDF database or to support property graphs, according to requirements. The product is cloud enabled. Its approach is to automatically index everything and it uses column-based index compression to reduce disk requirements. While it supports RDF and SPARQL the extensive indexing means that the company tends to focus more on analytics than transaction processing despite its ACID compliance. Text indexing is included as well as Solr and Lucene integration. The product includes reasoning: both forward chaining and backward chaining based on PROLOG. Unusually, AllegroGraph comes with its own browser-based visualisation and discovery engine, Gruff, which includes a graph query builder. In its latest release (5.0) the product includes "nDimensional" support which means that you can query against any combination of time, location, temperature, pressure and so on. Graph algorithms and social network analytics are provided out of the box. Security is implemented at the triple level.

IBM

While IBM has a number of development projects involving embedded graph databases (including at least two of the other vendors in this Market Update) the only currently available product in this area is DB2, which can act as a triple store. Unlike other relational implementations DB2 stores triples as encoded vectors. SPARQL is supported. However, the triple store support is otherwise limited and the company has no plans to develop this further. It has limited capabilities at present: IBM has promised to implement an inferencing engine but this is not yet available.

It is perhaps worth commenting on IBM's G2 algorithm Sensemaker, which is about discovering complex non-obvious relationships in entity-relationship environments. However, this is an algorithm (for example, it runs with InfoSphere Streams) and does not offer storage so it has not been scored as a part of this Market Update. G2 is only in limited release mode at present. It should also be noted that IBM Watson (also not scored) has overlapping capabilities with some graph products. Finally, IBM has a graph database (System G) in R&D. There are no plans to release this (which we believe to be a mistake) but some of the fruits of this research may appear in other products or solutions in due course.

Lexis Nexis

Lexis Nexis provides HPCC (High Performance Computing Platform), which is a multi-model database that can be made to look like more or less like anything you want. In its most common instantiation it is a direct competitor to Hadoop although in many ways it is a superior offering and it is generally deployed in analytic environments (including as a conventional data warehouse). It is an open source product. As a graph implementation HPCC uses KEL (Knowledge Engineering Language) to access data. This is a powerful and very terse declarative language and there is a relevant database optimiser. Over recent months Lexis Nexis has implemented a number of significant features designed to support graph processing and it also leverages advanced mathematics (notably linear algebra and matrices) to enhance query performance. SPARQL support is planned for a future release.

MarkLogic

MarkLogic is both the company and the product. As a product it is an XML database that has had significant success within target industries. In the latest release (version 8) the product includes both native JSON and RDF triple storage. Inferencing is support via backward chaining. You can combine SPARQL queries with full-text search, structured search, geospatial, and so on, in a server-side programme written in either XQuery or JavaScript. You can also embed triples into XML or JSON documents and then annotate them appropriately. MarkLogic has in-built search capabilities and you can combine this functionality with general, geospatial and RDF indexing in a single query. There is also bi-temporal support (two time stamps: for example, one when something was true and the other when you knew about it). OWL-Horst is supported. The product is ACID compliant and the company has a history of provision to major organisations so you can expect features such as high availability, resilience and so forth. The product is available in the cloud (AWS) as well as on-premise. It is arguable that the product should be shown as green/blue in the accompanying Bullseye diagram because it has features comparable to an operational graph database because of its non-graph capabilities. However, from a purely graph perspective that is not the case.

Neo4j

Neo4j is the market leader in this space in terms of deployments and name recognition and it is the oldest established (it was founded in 2000 in Sweden: it is now based in the US) graph database vendor. It is an operational graph database with native graph storage that provides ACID compliance. While SPARQL is supported the vast majority of Neo4j's customers use the company's own declarative language Cypher. In its forthcoming 2.2 release the company will extend its optimiser from a rules-based optimiser to include cost capabilities (in other words, it will be collecting statistics). In addition to its own direct customers it has a significant partner and OEM base. For example, Pitney Bowes uses Neo4j as the basis for its MDM (master data management) offering.

Objectivity

Objectivity provides InfiniteGraph. This is a graph layer implemented on top of the Objectivity distributed object database and it can best be described, within the context of this paper, as an operational graph database. It is cloud enabled. The company has been around for over twenty years as has its database although InfiniteGraph is a more recent addition. This longevity suggests an enterprise ready product, which may not be true for some other offerings. As a graph database (and this also applies to Objectivity itself) the company is focused on very large graphs (billions of nodes) that are often refreshed in real-time and where analysis and traversal needs to be run across the whole graph rather than sub-graphs. As a result most of the company's deployments are in government, law enforcement, the military, and also in fraud detection. InfiniteGraph supports Gremlin though not (yet) SPARQL.

Ontotext

Ontotext was one of the first vendors into this space, having been originally founded in 2000 (in Bulgaria) to investigate semantic technologies. It is the probably the most widely used pure-play RDF database. Its GraphDB (previously known as OWLIM) is an RDF database and therefore Ontotext can justifiably claim to be the oldest established vendor in this space (as opposed to Neo4j's graph database). GraphDB integrates with various search technologies and, unlike most other vendors in this space the company has developed specific solutions for various industry sectors, including publishing and media, recruitment, life sciences and healthcare, museums and archives and, more generally, for compliance and document management. GraphDB's inference engine employs forward chaining and the company has a patented method for retracting materialised inferences. One pre-defined graph algorithm is available (for PageRank) and others are planned.

OpenLink Software

OpenLink Software provides Virtuoso, which the company refers to as a "Universal Server". This is a multi-model database that supports the storage of relational data (accessed via SQL), RDF data (triple or quad store accessed via SPARQL and variants thereof) and content (XML, JSON and so forth). Virtuoso also includes a Web Server and other elements that mean that it is more than just a database. It is ACID complaint and open source.

Orient Technologies

OrientDB from Orient Technologies is an open source graph database built on top of a document store. In fact it is a hybrid graph-document database. This has some obvious advantages in document-oriented environments, which makes it a potential competitor to RDF databases (OrientDB is ACID compliant) as well as graph databases. It supports Gremlin and Blueprints but, more importantly, it uses an extended form of SQL for query processing. The product is schema-free and uses sharding for distributing data across a cluster.

Oracle

Oracle first announced its triple store and semantic capability (it supports OWL) with Oracle Database 11g. Since that time it has extended its capabilities into analytics with its Parallel Graph Analytics capabilities though these will be limited to "known-known" queries. The company offers a number of pre-packaged graph algorithms. The product uses forward chaining in its inference engine but has no retraction capability. Needless to say, with Oracle you are going to get an enterprise-class product. Security is implemented at the triple level.

SAP

SAP HANA has a graph engine in its latest version, which is in "controlled early adoption". It is an ACID-transaction compliant graph database built on SAP HANA's in-memory columnar storage architecture. SAP HANA graph uses a property graph data model as the central data structure providing directed, attributed (vertices and edges) multi-relational graphs. At present, the graph engine uses a declarative language called Graph Exploration & Manipulation (GEM) for data query and manipulation.

A number of graph algorithms are provided out of the box, which can be configured by means of a graph API. It is notable that Simple Logistics, the second module within S/4HANA, leverages the graph engine in HANA in order to support bill of materials. We can expect future S/4HANA modules to take a similar approach.

Teradata

Teradata uses its Aster Data platform to store the vertices of a graph in a table and the edges in a second table, which can then be queried using SQL and conventional joins. This, of course, limits analytics to known-known questions but the vast majority of graph analytics fall into exactly this category. Being able to use SQL has the significant advantage of being able to use traditional BI tools. It is worth pointing out that it is relatively easy to partition the data appropriately when you are dealing with graphs whose relationships are known. It is when that is not the case that partitioning starts to be a problem. Teradata implements a BSP (bulk synchronous parallel) architecture to support graph processing. This is especially useful in graph problems because it improves performance for iterative processes that are common with graph algorithms (a number of which are supplied by Teradata). It is worth noting that Pregel (from Google) and Hama (an Apache project) are both BSP-based.

Others

Other vendors/products in the RDF space include (but are not limited to): 4Store (the developer – which has also created 5Store – has been acquired by Experian and it is being used internally rather than being developed on behalf of end users), BrightStarDB, CubicWeb, Dydra, Mulgara, Redland, RedStore, SparqlDB and Strabon.

In the graph database arena there is Amisa Server, ArrangoDB, FlockDB (an open source product originally developed by Twitter), Giraph (an Apache project), GlobalsDB (interesting because it is the core engine of InterSystems' Caché database, which has been deployed as an underpinning for a variety of NoSQL projects), Graphbase, HyperGraphDB, InfoGrid, SparkSee

(formerly DEX) and Sqrrl. Also notable are Google's Pregel, Cayley (a development by an individual Google employee, which has been endorsed by the company) and the Apache Hama project.

Finally, we should mention SPARQLverse from SPARQL City. This is an analytic product that was launched in 2014 and which Actian (which is an investor in SPARQL City) has been re-selling. We had initially included SPARQLverse in this Market Update. However, we have decided not do so at this time as the company is currently evaluating alternative development and distribution models.

See also http://en.wikipedia.org/wiki/Graph_database.

Comments

As a general principle we prefer native implementations to ones based on other database platforms. Moreover, in our view, NoSQL implementations are preferable to relational ones. However, both of these statements are dependent on the application. If you are exploring known-knowns in a purely analytic environment then storing edges and vertices in relational tables should provide perfectly acceptable performance. In other environments, for example when supporting transactional and operational processing, we would expect relational products in particular to perform poorly compared to native and even NoSQL-based implementations.

With respect to the various products shown on the following Bullseye Chart, we have sometimes used product names and sometimes vendor names. In general we have used the name with which we believe readers will be most familiar. It should be noted that SPARQL Server from Algebraix is currently in beta and the final release version may differ from that which is currently available. Titan, similarly, has yet to reach 1.0 status. The graph engine in SAP HANA is also new and, as yet, relatively unproven, although it shows promise. As will be seen the various product/vendors are colour coded so that we are comparing apples with apples. Even so, the following additional comments are relevant:

- **RDF databases –** both Oracle and Stardog are placed in this category despite the fact that Oracle provides analytics and Stardog will shortly be introducing graph database capabilities. They have been scored bearing these facts in mind.
- **Operational graph databases –** InfiniteGraph frequently does not compete with either Titan or Neo4j (or any other graph provider for that matter). This is because of its focus on very large, whole graph applications, especially where these graphs change rapidly.
- **Analytic graph databases –** both Cray and Teradata are outliers in this category: Cray because it focuses on environments where there are a lot of unknowns and Teradata for the exactly opposite reason, because it focuses on known-known analytics. The other products in this category are more general-purpose.

It should be clear from the above that some of these apples are eating apples and some of them are cooking apples and some may even be crab apples! We could have used seven different colours in our Bullseye Chart but that would have been over complicated. Just bear in bear these distinctions when viewing the following chart.

Conclusion

As has been noted there are lots of open source and development projects within the graph space. We have focused on those that we believe to be enterprise-ready. That is to say, we expect features such as high availability, resilience, security, scalability and performance as well as features that are specific to the graph and RDF markets.

With the exception of IBM, which has not yet got its act together with respect to graphs, all of the products included in this Market Update have significant strengths. The difficulty for potential users is identifying the particular types of use case for which each product is most suitable. This is one of the reasons why this is a rather longer Market Update than is typical: because we have wanted to give some indication as to the focus areas of the different vendors. As always, ultimately users should conduct proofs of concept both with respect to functionality and performance.

Bloor

2nd Floor
145–157 St John Street
LONDON EC1V 4PY
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.BloorResearch.com**
email: **info@BloorResearch.com**