Market segmentation

There are five types of products that are, or might be, directed at the graph space. Firstly, there are relational databases that support graph processing. A good example is Pivotal Greenplum, which while still strictly relational, supports a variety of parallelised graph algorithms. However, there are many graph algorithms that do not benefit from parallelisation. Secondly, there are relational vendors that have adopted open source graph languages. Examples here include SAP HANA with support for OpenCypher and, in its latest release, IBM Db2 with support for Gremlin. The same graph algorithms that cannot be parallelised – typically those involving iterative self-joins – will not perform well when using these relational databases as they are not true graph products but only limited subsets thereof. For this reason, these offerings are not discussed in this Market Update.

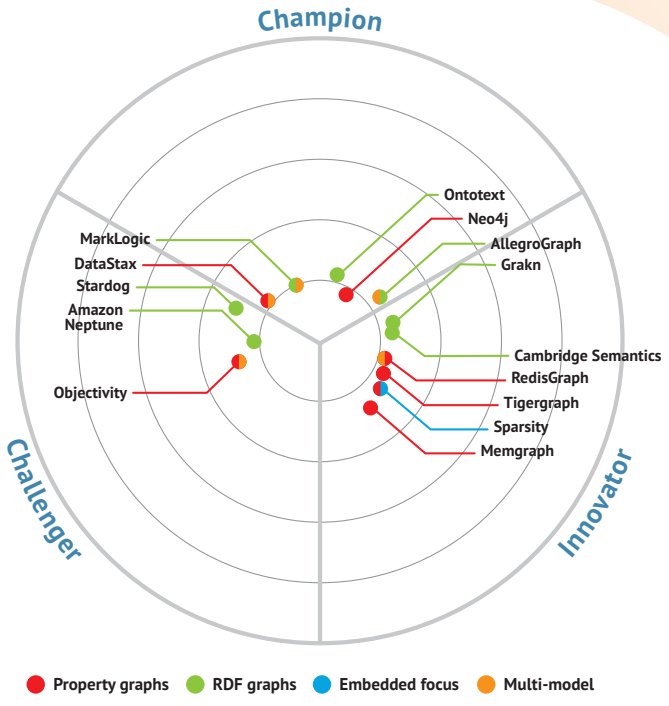
The third class of product that offers graph processing are multi-model databases, as typified by MarkLogic, Redis, DataStax, and so on. These are all discussed here. What is not considered is Microsoft Cosmos DB. Despite Microsoft's claims to the contrary we do not consider this to constitute a true multi-model offering. Specifically, Cosmos DB differs from other products in this category because it requires a different API for each model, whereas other vendors support the use of a single API across all supported models. We believe that the use of a single API is fundamental to the definition of a multi-model database as well as to those RDF graph vendors, such as Franz and Ontotext, that support a document model as well as graphs..

Finally, there are property graphs and RDF graphs. With the advent of RDF* and SPARQL*, which are proposed standards that allow RDF graphs to add labels to relationships, as opposed to reification and other techniques that are either complex or result in node proliferation, there are indications that these spaces are moving closer together, with RDF suppliers adding support for OpenCypher or Gremlin to support graph traversal. This will suit developers, while leaving the underlying model to support the semantics that are often favoured by information architects.

With respect to segmentation, we identified in our last report that there was a distinct difference between vendors focusing on analytics as opposed to those who are more targeted at operational environments. Needless to say, there is significant overlap here. The growth in support for knowledge graphs has led to even further differentiation, with some suppliers focusing on this for its own sake and others that see it primarily as an enabler for data virtualisation (or vice versa: data virtualisation and knowledge graphs are symbiotic). And finally, there are Sparsity Technologies and Grakn, where the former focuses on using the Sparksee graph database as an embedded database, for example, in mobile devices, automobiles and edge devices; and the latter offers a hypergraph.

In the Bullseye diagram we have differentiated – via colour coding – between RDF and property graphs, and between those vendors that are focused particularly on graphs and those that are multi-model databases with graph functionality. The positioning of the latter on the Bullseye diagram relates specifically to their graph capabilities, rather than their overall capabilities. Sparsity (Sparksee is a property graph) has its own colour because it focuses on a different market segment from other vendors. We have colour coded Grakn as an RDF graph because it is commonly used instead of, or to replace, RDF graphs.

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 or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.



Market trends

In 2019 Gartner predicted that the graph database market will be growing at 100% CAGR by 2022. Also last year, Markets and Markets reported and estimate for the overall size of the market to be $2.8bn by 2024. So we are pleased to see that other organisations are endorsing our continued research into the graph database market, this being the 4th Edition of this report.

On the vendor front, there have been two significant changes since our last Market Update. The first is that, predictably, SAP is focusing on HANA as its graph offering rather than OrientDB. The latter has therefore been omitted from this report. The second change is that Objectivity, while still marketing ThingSpan, is focusing more on its underlying object-oriented database with ThingSpan being simply an implementation option for relevant use cases. DataStax is taking a similar approach with DSE. It should be noted that this in no way detracts from their respective offerings: it is simply that they see more use cases where graph is part of the answer but not all of it.

More generally, there is the wholesale adoption of RDF* and SPARQL* by RDF vendors such as Ontotext, Cambridge Semantics and Stardog, though not by Amazon, even though it does support Gremlin as a graph traversal language. The advantage of this approach, whereby you can have both SPARQL and either Gremlin or openCypher running against the same database, is that you don't have to choose which underlying storage engine to use. Of course, Neo4j supports SPARQL also (and Gremlin) but if you haven't got a semantic model underneath that is going to be of limited value. You also won't get the inferencing capabilities that an RDF graph provides.

Another significant trend is in making graph application and query development simpler. For example, multiple vendors are now supporting GraphQL as an API so that you don't need to know SPARQL (for RDF graphs) to build your queries. In a similar fashion, TigerGraph has introduced a no-code graphical development environment which hides the complexities of GSQL from business analysts and other users that want self-service query capabilities. We expect this trend towards ease of use and self-service to grow and expand.

There are some other trends that are starting to emerge. There is obviously the shift to cloud-based provisioning and increased support for managed graph databases as a service. There is also increasing support for Zeppelin and Jupyter notebooks. Migration tools, both from rival graph vendors and from relational sources, are becoming more common. And geospatial support is starting to be implemented by several vendors.

Then there is the question of languages. Both openCypher and Gremlin have significant support while the development of GQL as an ANSI standard language for property graphs continues and is supported be a variety of vendors. However, we suspect that these efforts, though worthy, may be made (largely) irrelevant by companies introducing support for graphical user interfaces and GraphQL that takes away the pain of learning new languages, as discussed above.

Finally, as always, there is an ongoing focus on performance and scalability. The latter seems to have been a major area of development by a number of companies, particularly Neo4j and Franz (AllegroGraph), since our last report, while Cambridge Semantics, with its analytics and *"graph OLAP"* capabilities is even positioning AnzoGraph as a graph data warehouse.

We spent a significant part of our last Market Update discussing benchmarks, something to be treated with a large pinch of salt. We do not intend to repeat this so interested readers are referred to the 3rd Edition (this is the 4th) of this report. We should add, as an example of how witless some marketing people are, that in our research for this Market Update we had one company extolling to us how good their product was at one-hop queries!

Knowledge Graphs

Knowledge graphs deserve special mention, as they have becoming an increasing area of focus. Unfortunately, there is no agreed definition of a knowledge graph. What they essentially allow you to do is to visualise and explore networks of related things. But this is precisely what graphs are, so some authorities qualify this by saying *"things of interest"*. The problem with this suggestion is that often the whole point is that you don't know what is of interest until you start your exploration. In any case, even if you can sensibly filter out entities and relationships that are not of interest, what you end up with is a (sub-)graph. Perhaps it would be better to describe a knowledge graph as an interactive *"view"* of your broader graph, in the same sense that you have views into your relational database.

Whatever they are actually, there is no doubt that knowledge graphs are increasingly popular and a number of graph database providers (particularly those

with RDF graphs) are targeting their construction. This has resulted in various additional terminologies. For example, so-called identity graphs, which support functions such as recommendations; and entity-event knowledge graphs, which are structured to emphasise temporally contextualised events and the entities they relate to (for example, this person took this medication at this time).

More generally, knowledge graphs are not just being used for their own sake, but also to support the creation and reuse of training data for machine learning purposes, and for data virtualisation. In the case of the former, there are two major reasons to leverage knowledge graphs. Firstly, the actual creation of training data is facilitated when you understand the relationships that exist between the data elements you are exploring. Secondly, one of the drawbacks of the usual data science process is that the data, and its relationships, are collected in order to support the task at hand and then, once the development process is complete, get thrown away for lack of anywhere to support its storage. Graph databases enable this and so support reuse.

As far as data virtualisation is concerned, it is worth commenting that in the general-purpose data warehousing market support for data virtualisation, or at least some form of query federation, is now more or less table stakes. And then there are independent offerings in the space from Denodo and TIBCO (Composite) as well as (Starburst) Presto. So this is a crowded market. That said, graph databases have an intrinsic benefit when it comes to data virtualisation in that they can map the relationships that exist between data in different sources.

Metrics

We used eight different scoring dimensions. In alphabetic order these are:

Analytics – the extent to which the product supports analytic capabilities, especially complex analytics. For RDF databases the support for inferencing (both forward and backward chaining) is relevant. The provision of pre-built graph algorithms will be an advantage as well as support for third party graph-based analytic libraries. Also relevant is the ability to assign probabilities to relationships. Note that all products have some degree of analytic capability.

Ease of Use – should be self-explanatory: includes administrative tools, graphical visualisation capabilities and so forth. It will be useful if the product supports both schema and schema-free environments. Availability of the product as a managed service will also be a factor here as will facilities for supporting the creation of knowledge graphs.

Features – measures additional capabilities such as whether a property graph includes labels or whether an RDF database has been extended to support properties and RDF*. Also includes facilities such as specialised importing capabilities from relational or other environments. The ability to track how a graph has changed over time will also be useful in some instances. More general features include high availability, security and so forth.

Integration – how well the product extends beyond graphs per se. For example, support for text (JSON, XML) processing, integration with search engines, and semantics. Also, the ability to leverage (geo-) spatial data. Support for data virtualisation is relevant in this category as is integration with Jupyter and Zeppelin notebooks. The ability to integrate with third party visualisation tools is also relevant or, in some cases, vendors provide their own tools. Integration with traditional BI tools such as Tableau is a bonus.

Language – what is the extent of language support? SPARQL and OWL in the case of RDF, and openCypher, Gremlin or other in the case of property graphs. Also including extensions to support RDF* and SPARQL*. Support for other language bindings is important as well as is GraphQL capability. Furher, there is an overlap with ease of use with respect to the provision of IDEs that hide the complexities of the underlying language.

Operations – the extent to which the product supports operational capabilities, including ACID compliance and immediate consistency. It is worth commenting that almost all products have some sort of operational capabilities (just as they do analytics) but that does not necessarily mean that they are optimised for that purpose.

Performance – this covers not just run-time performance for both operations and analytics but also ingestion rates. While having a *"native"* graph database has theoretical advantages in performance terms, everything depends on the implementation. The capabilities of the database optimiser (where appropriate) are relevant here.

Scalability – not just scale up/out but also scale down/in. Some products may be fine at the top end but would not be cost effective for small scale projects, especially if embedded. We should further comment that there is a difference between scaling up the number of user queries (read) that you can support simultaneously, scaling for high availability purposes, scaling ingestion (write) and the scale of the graph (number of nodes and edges) that you can support.

Conclusion

Some parts of the market are converging: RDF graphs adding Gremlin or Cypher; everybody supporting knowledge graphs and many advocating data virtualisation; and vendors that previously offered limited scalability introducing new architectures that support massive distributed environments, thereby allowing them to compete more effectively with companies that have historically tended to focus on use cases with high-end scale and performance requirements. As far as this last point is concerned, this suggests that at least some suppliers have reached a point at which their products could be described as mature.

On the other hand, some smaller vendors are focusing on particular market segments that are not well addressed by the major players. In this last category we would put Sparsity Technologies (embedded graph databases in, for example, smart cars), Memgraph (which is focusing on extremely complex environments, often where multiple graph algorithms have to be used in conjunction, for instance in managing chemical plants or gas distribution networks), and possibly Grakn as a platform for building cognitive applications.

There is no doubt in our minds that graph databases are becoming more mainstream and that there are a broader range of use cases for which graph databases are being used. We expect this to continue. While it is encouraging to see vendors such as IBM add graph support in Db2, it only goes to validate the market. And while there are a number of graph algorithms that can be parallelised there are many for which relational databases cannot easily achieve adequate performance, largely thanks to the iterative (self-joining) nature of many graph queries. We therefore think that the support of graphs by the likes of Oracle, IBM and SAP is only nibbling at the problem around the edges and that true graph databases are much to be preferred in all but limited instances.

About the authors

PHILIP HOWARD
Research Director:
Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

DANIEL HOWARD
Senior Analyst:
Information Management and DevOps

Daniel started in the IT industry relatively recently, in only 2014. Following the completion of his Masters in Mathematics at the University of Bath, he started working as a developer and tester at IPL (now part of Civica Group). His work there included all manner of software and web development and testing, usually in an Agile environment and usually to a high standard, including a stint working at an 'innovation lab' at Nationwide.

In the summer of 2016, Daniel's father, Philip Howard, approached him with a piece of work that he thought would be enriched by the development and testing experience that Daniel could bring to the table. Shortly afterward, Daniel left IPL to work for Bloor Research as a researcher and the rest (so far, at least) is history.

Daniel primarily (although by no means exclusively) works alongside his father, providing technical expertise, insight and the 'on-the-ground' perspective of a (former) developer, in the form of both verbal explanation and written articles. His area of research is principally DevOps, where his previous experience can be put to the most use, but he is increasingly branching into related areas.

Outside of work, Daniel enjoys latin and ballroom dancing, skiing, cooking and playing the guitar.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**