Bloor

Market Report

Market Report Paper by Bloor
Author **Philip Howard**
Publish date **November 2016**

Graph and RDF databases 2016

There have been some significant changes with respect to open source projects, most notably that Tinkerpop is now an Apache project, Sesame is an Eclipse project (RDF4J) and Titan is now an open source project that is no longer supported directly by a vendor.

Author **Philip Howard**

Graph and RDF databases 2016

Market segmentation

It is important to appreciate that not all graph databases are created equal. Briefly, databases can be categorised in the following ways:

- **Property graph versus RDF** (resource description framework) databases (sometimes known as triple stores). Some products offer both property graphs and RDF while some RDF databases have been extended with property-like features.
- **Native versus non-native databases.** Some products have been specifically built with native graph engines and some have been built on top of other database engines (Hadoop, Cassandra, relational, object oriented, XML and so on). A priori you wouldn't expect non-native implementations to perform as well as native engines, but there are a lot of other factors involved. While it makes a good competitive story for those vendors who offer native engines, in practice it is a spurious argument: the real issue is performance, not how you get that performance.
- **Multi-model versus single model databases.** Multi-model databases are those that have been designed to support different model types. For example, a common possibility is a three-way option of document store, key value store or RDF/graph store. One vendor offers an alternative between relational, RDF or property graph. The advantage of multi-model databases is precisely that they are not limited to a single type of data. Pure-play vendors will argue (not necessarily correctly) that this will mean that your environment will not be as optimised for graphs as it would be otherwise. Most multi-model databases are based on native engines (albeit optimised across datatypes) but some are not.
- **Analytic databases versus operational databases.** Needless to say, all operational databases support some degree of analytics but some are more performance oriented than others. In addition, not all products support immediate consistency so these should be regarded as operational but not transactional (which also has relevance for real-time analytic environments such as fraud detection). Conversely, some products only target analytics. Thus, transactional databases include operational databases which include analytic databases.

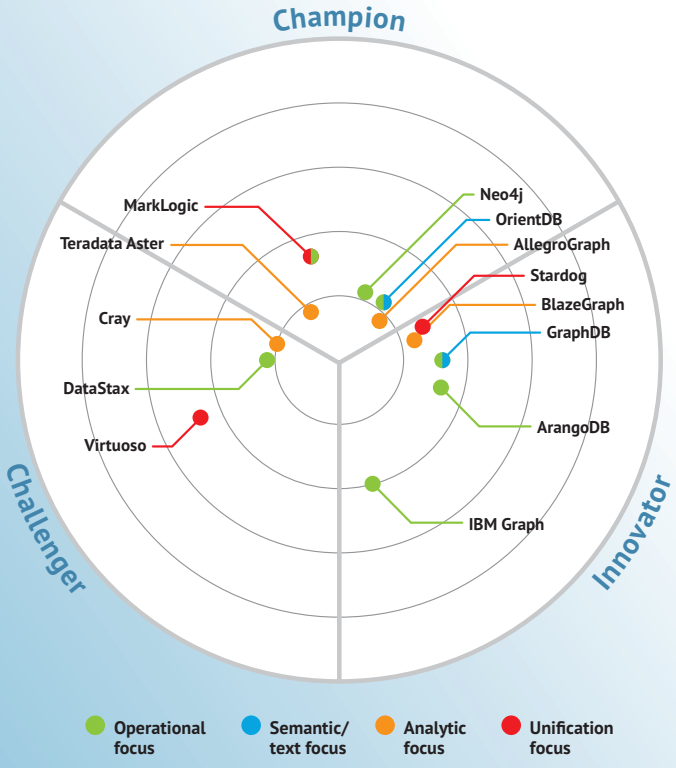


Figure 1:
The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is 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. It is important to note that colour coded products have been scored relative to other products with the same colour coding. Comparisons across colour codes are not necessarily valid.

In addition to these categories, there is an emerging market segment whereby some graph database vendors are concentrating upon providing what is sometimes known as a "unification platform". That is, a database that sits as a middleware layer over and above existing relational and non-relational data stores that can act as a virtualised query platform that spans those various sources. As a true database, some of the underlying data may be moved into the graph database using conventional ETL processes but otherwise such information is retained in situ. The big advantage that graph databases have in this environment (as opposed to conventional data federation tools) is that the relationships between the different data sources and the data they contain, are directly expressed and understood through a graph of the environment.

There are a lot of choices and combinations here, and it should be clear that it is not simply a question of our not comparing apples with pears but with not comparing apples with any other fruit in the fruit bowl. This is relevant because the use cases supported by the different approaches are different. The following table highlights the distinctions between products. Products asterisked in the second column are RDF databases that optionally support properties. AllegroGraph and BlazeGraph have two entries in the last column because they offer extended capabilities for supporting analytics, while Stardog and Virtuoso (and, to a lesser extent, MarkLogic) are focusing on unification (data virtualisation) even though they support conventional transactional environments. Note that Cray Urika is not built on top of Hadoop but both the graph engine and Hadoop are implemented within Urika-GX.

Database	Property or RDF?	Engine type	Use cases
AllegroGraph	RDF*	Native	Transactional/ Analytics
ArangoDB	Property	Multi-model	Transactional
BlazeGraph	Either	Native	Transactional/ Analytics
Cray	RDF	Native + Hadoop/Spark	Analytics
DataStax	Property	Multi-model on Cassandra	Transactional
GraphDB	RDF*	Native	Transactional
IBM Graph	Property	Titan on Cassandra	Transactional
MarkLogic	RDF	Multi-model	Transactional/ Unification
OrientDB	Property	Multi-model (native)	Transactional
Neo4j	Property	Native	Transactional
Stardog	RDF*	Native	Transactional/ Unification
Teradata Aster	Neither	Multi-model on Hadoop with BSP	Analytics
Virtuoso	RDF-based Property	Multi-model (relational)	Transactional/ Unification

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"***. The 2nd edition of this paper has been published alongside this Market Update.

Market trends

We reported in our last update that the biggest trend was simply towards graph products in general. This remains true. However, since last year the most obvious move has been towards consolidation. This is typified by vendors going out of business, being acquired or focusing on niche markets, all of which are now apparent. Sparql City has gone out of business, Aurelius (we reported this last year) has been acquired by DataStax and three other vendors with graph products: Algebraix, Lexis Nexis, and Objectivity are all de-focusing on graph and are emphasizing Hadoop and/ or Spark instead. In the case of Objectivity, InfiniteGraph is no longer being developed. Similarly, development of FlockDB has been abandoned, while Sqrrl has moved from what was ostensibly a general-purpose graph product to one that is specifically focused on cybersecurity. This is not especially surprising as Sqrrl was an NSA spin-off in the first place.

It is also noticeable that there is a significant trend towards multi-model approaches. For example, to use a database that can be either or both of a document store and a graph database. This means that you can sell your product against a wider range of use cases than graph alone. It's a good way to survive if you can't get enough traction for graph-only applications. Of course, there are use cases that are particularly suited to these sorts of hybrid environments and some products were designed to address these from the outset. On the other hand, some products have been retro-fitted as multi-modular, which you might take to be a warning sign.

There have also been some significant changes with respect to open source projects, most notably that Tinkerpop is now an Apache project, Sesame is an Eclipse project (RDF4J) and Titan is now an open source project that is no longer supported directly by a vendor. In this last case, DataStax has taken the IP from Titan and embedded it into its own product but the company is not contributing to the Titan open source community. Moreover, neither is anyone else, so Titan look as if it is dying. IBM has embedded Titan into IBM Graph, and both InfoTrellis and Global IDs have embedded the database into their respective products, but unless one of these takes on the role of investing in the underlying technology – and there is no sign of this –

then we expect Titan to die fairly quickly.

The other major trend we have noticed is the extent to which other software vendors are embedding graph databases in their own products. We have already mentioned several of these that have embedded Titan but the other popular database used for this purpose is OrientDB (which is an RDF database where Titan is a property graph database), which has been embedded in their products by Informatica, Diaku and Dell Statistica, to name just three examples. It is also worth mentioning Reltio that has engineered its own graph database on top of Cassandra: it remains to be seen if it ports to the new DataStax graph product. And finally, Pitney Bowes embeds Neo4j in its master data management (MDM) product. Note that both InfoTrellis and Reltio are direct competitors of Pitney Bowes for MDM so we expect more vendors in this space to adopt a similar approach. Further, all the other suppliers we have mentioned are active in the wider data governance arena and we expect to see graph capabilities becoming more widespread within this milieu. We are also aware of user organisations leveraging graph technology to build their own data governance capabilities, for example, to support metadata management.

Finally, it is worth discussing the 800lb gorillas. Oracle offers Oracle Spatial and Graph and has implemented a property graph approach in addition to its historic RDF capabilities. However, the truth is that we are not impressed with Oracle's graph capabilities: it seems to us very much like an add-on rather than something that is intrinsic to the Oracle environment. SAP also has graph capabilities built into SAP HANA but, again, we are by no means enthusiasts. Indeed, we understand that some of SAP's development teams feel much the same way. Microsoft has a product in an experimental stage and as far as IBM is concerned DB2 supports RDF triples with even less depth than Oracle does, it too has an experimental product (System G) and it offers a managed service, called IBM Graph that runs on IBM Bluemix and which is based on Titan. This has specific features to make it easy to use for developers but is otherwise quite limited. To be frank, it is only included here because it is new, whereas both Oracle and SAP were discussed in our previous Market Update. In practice, the specialist vendors in the graph and RDF spaces are significantly ahead of the major database vendors in terms of their capabilities at present, and it is why we have focused on these providers. This may change in the future but only if, and when, the big boys start to take this market seriously.

Differentiating use cases

Simplistically, RDF products are typically employed where either semantics or text processing is a key requirement. This often involves documents, search and similar capabilities. Perhaps the simplest way to describe the use of property graphs is that they are essentially addressing relational problems that are too complex for relational databases to handle efficiently. In other words, they are primarily targeted at structured data, perhaps combined with unstructured data, but the former is the focus rather than the latter. Often, these are use cases where understanding the relationships between data elements are more important than the data elements themselves. In fact, another way to put this would be to say that with RDF databases you typically query the graph whereas with property graphs you traverse the graph though this is, of course, dependent on the product in question. Nevertheless, while some vendors might (will) disagree, it is typically the case that RDF databases compete amongst themselves whereas property graphs compete with non-graph products.

From an analytics perspective, all graph products have some query capability and some have significant abilities, including the ability to support PageRank and other graph algorithms. However, it is typically the case that when it comes to analytics, vendors target operational applications such as recommendations (next best offer) or fraud detection, or hybrid operational and analytic environments. Relatively few vendors specialise specifically in analytics, and of those that do, most of them are very different architecturally, and often do not compete with one another.

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. As will be seen the various product/vendors are colour coded so that we are comparing apples with apples.

“The other major trend we have noticed is the extent to which other software vendors are embedding graph databases in their own products.”

ArangoDB

ArangoDB is a German company that was founded in 2014 as a spin-off from a consulting organisation that was itself constituted in 2004. In practice, the original company developed its first database capabilities from its inception and this gradually evolved into what is now the ArangoDB product. It is a multi-model database that supports JSON, key-value and property graph capabilities with one database core and one declarative query language. The database is ACID compliant and supports immediate consistency. Support is provided for Apache TinkerPop but the company also offers its own AQL declarative query language, which works across all the ArangoDB supported data models and even lets one combine different data models in one query. Also notable are the Foxx framework and the product's support for Apache Mesos. The former is a Google V8 engine based JS framework enabling the creation of JavaScript-based micro-services. The latter is an open-source cluster management solution and the company is also a partner of Mesosphere, which provides container orchestration based on Mesos and which helps to support very large datasets. ArangoDB is one of the very few databases capable of sharding a graph to a cluster and handling graph-traversals in a performant manner (via ArangoDB SmartGraphs). The company focuses on operational and hybrid operational/analytic environments such as recommendations (next best offer), network and wire management, fraud detection and so forth, as well as more esoteric applications such as exploring genome-based data. There are a number of clients listed on the company website, of which the most well-known is Liaison Technologies. Another notable user is the Karlsruhe Institute of Technology.

Strengths

- Clustering support is significant for both scale-out purposes and high availability.
- ArangoDB scales both horizontally and vertically.
- The use of AQL across different datatypes. This is significantly faster than using Gremlin.
- The partnership with Mesosphere is a substantial differentiator.

ArangoDB

ArangoDB
Zülpicher Platz, Cologne, Germany

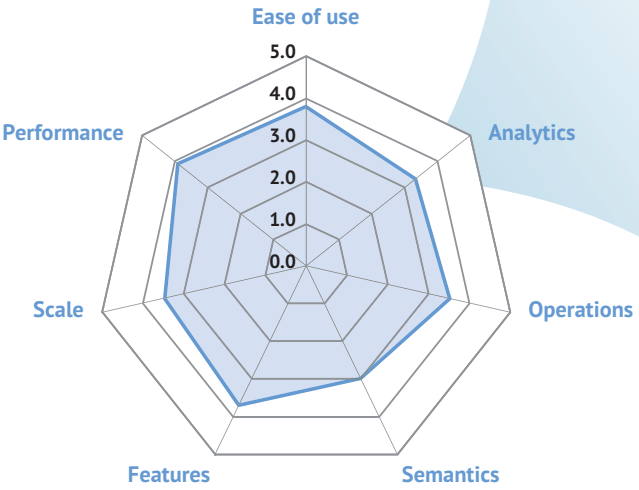
www.arangodb.com

Threats

- Relatively small company and not particularly well-known.
- The Enterprise Edition (with added security, encryption at rest, and auditing) has only just been launched.

Summary

ArangoDB is less well-known than some of its competition but, especially with the introduction of the Enterprise Edition, it shows significant promise.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Blazegraph

SYSTAP LLC is the developer of the Blazegraph RDF database, which has been under continuous development since 2006. Unlike the majority of vendors in this market, which tend to target operational and hybrid operational/query environments, SYSTAP is squarely focused on graph analytics and query, especially large scale, complex graph analytic environments, particularly where relationships are not known in advance. In addition to Blazegraph, the company markets Blazegraph GPU, which is an add-on to Blazegraph enabling graph analytics to be accelerated using NVIDIA graph processing units. The company also has a product called DASL (pronounced "dazzle") that supports the development of analytic and statistical algorithms that will specifically run using GPUs. The database itself is an (extended) RDF graph database that has property graph features. It supports SPARQL 1.1, Apache Tinkerpop 3 (both Gremlin and the Blueprints API), OWL (web ontology language), the Sesame (now an Eclipse project called RDF4J) API, a graph mining API and the Lucene search engine. In particular, Blazegraph supports the development of domain specific languages whose syntax is converted into SPARQL queries at run-time. Further, SPARQL queries are translated into suitable code for the GPUs by the software so there will be little or no change required when upgrading to a GPU-based environment. The product is available for both cloud-based and on-premises deployments and either as an open source version or with an enterprise license.

Strengths

- Deployment with GPUs provides two orders of magnitude price/performance benefit compared to solutions from other vendors, for relevant analytic requirements. Moreover, there are relatively few vendors in this space that focus specifically on analytics.
- Wikimedia selected Blazegraph (without GPU acceleration) and published its comparison spreadsheet that determined the selection of Blazegraph over other competitive products (see *https://docs.google.com/spreadsheets/d/1MXikljoSUVP77w7JKf9EXN40OB-ZkMqT8Y5b2NYVKbU/edit#gid=0*). While this is slightly out-of-date (May 2015) it is still a useful reference point for Blazegraph.
- The product has strong high availability characteristics.

Blazegraph
1875 Connecticut Ave NW,
Washington, DC 20009, USA

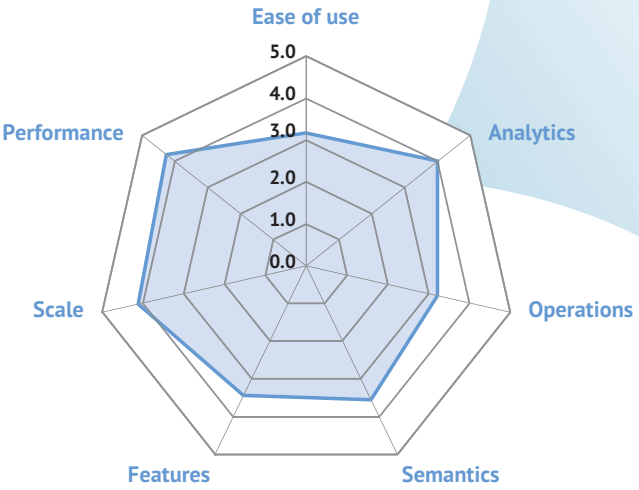
www.blazegraph.com

Threats

- GPU acceleration is not a panacea. It is best suited to graph algorithms and analytics that can easily be partitioned. This is because of the relatively limited amount of memory in each GPU.

Summary

For appropriate analytic applications BlazeGraph will achieve much improved price/performance – thanks to its use of GPUs – compared to its competition.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Product Sheet

Cray

Cray first entered the graph market when its subsidiary YarcData introduced a product called Urika. This was an in-memory RDF database delivered as an appliance. Subsequently, YarcData was taken in-house, Cray launched Urika-GD. Around the same time, the company introduced a Hadoop appliance called Urika-XA and what the company has now done (May 2016), is to combine these two products into a single offering called Urika-GX, which supports Hadoop, Spark and graph processing. This is not an appliance per se because, although the software is pre-installed and ready to run, you are not limited to what you might subsequently install. A particular advantage of including Hadoop within Urika-GX is that you can leverage Hadoop to transform and load data into the graph database much more efficiently than would otherwise be the case. The product is specifically targeted at the most intractable analytic problems and use cases include cybersecurity, the Internet of Things, machine learning, research into new drugs and new uses of existing drugs, and to uncover risk and compliance issues within financial services environments. The product uses SPARQL for query purposes and there are many built-in functions, including geo-spatial functions. Both post-graph and pre-graph analysis can be performed using R.

Strengths

- Undoubtedly the performance leader in this space.
- The availability of Hadoop for converting existing data formats into RDF is a significant plus, as well as for its ability to support non-graph analytics.
- Extensive pre-built analytic functions and support for R, as well as both machine and deep learning capabilities.

Threats

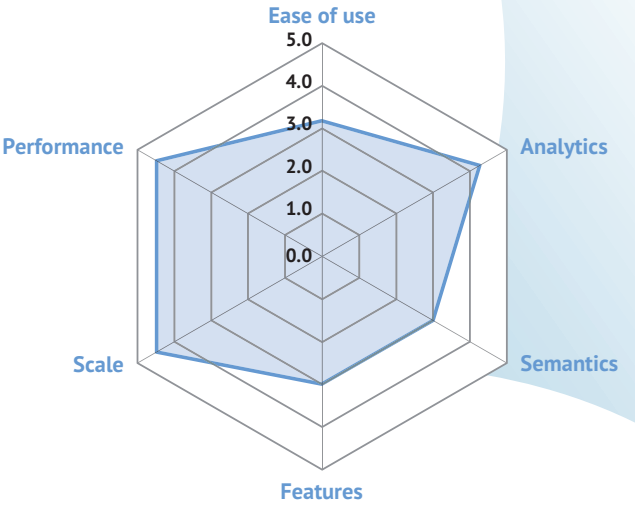
- Potential users may not necessarily associate Cray with commercial capabilities within this space.
- Competes with cognitive computing in some of its use cases.
- Reputation (not necessarily deserved) for being expensive.

Cray
901 Fifth Avenue, Suite 1000
Seattle WA 98164, USA

www.cray.com

Summary

Cray Urika-GX has the highest ranking of any product we have examined across three of our categories.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

DataStax

DataStax acquired Aurelius, the developer of the Titan graph database, in 2015. DataStax has optimised it specifically to run on the Cassandra database engine within DataStax Enterprise (version 5.0, launched June 2016). Options depend on the edition, with graph transactions available in the Standard Edition and analytics and search (based on Solr) added as options in the Maximum Edition. It is a property graph solution that supports either or both of CQL (Cassandra Query Language) and Gremlin, as well as conventional languages. The company is also a leading contributor to the Apache Tinkerpop project of which Gremlin forms a part. There are, in fact, two processing engines: one for transactional and operational purposes and one, based on Spark, used for analytic processing. You choose which of these to use through DataStax Studio, which provides a development environment that is similar to the visual development tools that users will be familiar with from relational environments. We are particularly impressed by DataStax Studio. Initial users of the graph capabilities of DataStax Enterprise are especially around supporting 360° views of customers, products and so forth. Recommendations (next best offer), and fraud detection are other targeted use cases.

Strengths

- DataStax has a significant existing user base and the company is therefore likely to rapidly acquire a significant graph user base.
- As a multi-model vendor DataStax supports a variety of different types of data that can be used together or separately for both operational and analytic purposes.
- Graph processing inherits the existing robustness, scale out capabilities, transaction support and performance of Cassandra.

Threats

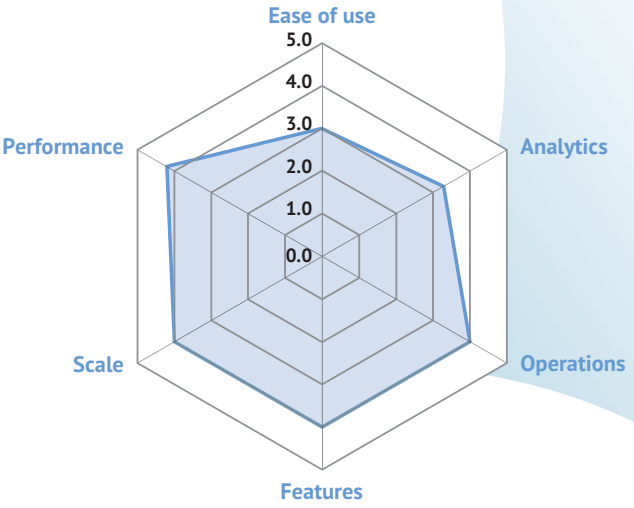
- Graph processing is functionality essentially built on top of Cassandra. While we regard it is a somewhat spurious argument this means that DataStax is open to the criticism that it does not provide a native graph engine.
- Will potential users see DataStax Enterprise as a database for graph only applications or will they see it as merely an add-on to existing and future Cassandra implementations?

DataStax
3975 Freedom Circle, 4th Floor,
Santa Clara CA 95054, USA

www.datastax.com

Summary

DataStax is the leading vendor of the popular Cassandra database and the addition of graph database capabilities should further increase this popularity.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Product Sheet

Franz

Franz Inc. originated with the initial Artificial Intelligence boom and still provides its Lisp compiler to numerous Fortune 500 companies. The company started to develop AllegroGraph more than a decade ago at the request of U.S. DoD and IC customers. It is a quad store which you can employ 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. AllegroGraph supports transaction processing with ACID compliance and immediate consistency. However, many customers also use it for analytic applications. Text indexing is included as well as SOLR and Lucene integration. The product includes reasoning: both forward and backward chaining and it also includes full PROLOG support for logic reasoning. Unusually, AllegroGraph comes with its own browser-based visualisation and discovery engine, Gruff, which includes a visual graph query builder. The product includes "nDimensional" support which means that you can query against any combination of time, location, temperature, pressure and so on. Another major feature is that you can associate a probability with a relationship within a graph. In other words, you can estimate how likely a relationship is to be true. This is exactly the sort of functionality that cognitive computing provides. Graph algorithms and social network analytics are provided out of the box. Security is implemented at the individual triple level.

The other major differentiator for Franz is its focus on vertical market sectors where it has built (along with clients) specific ontologies, for example, for Healthcare. In collaboration with Montefiore Medical Center, Franz developed the first Semantic Data Lake for Healthcare (SDL). The SDL platform integrates complex information for daily healthcare management, clinical, population, community, environmental, behavioural and wellness research data. The SDL is an example of how AllegroGraph can be used as the analytics platform for a wide range of different applications instead of deploying multiple data marts thereby simplifying the computing environment. Because of these capabilities Franz increasingly refers to its database as a "semantic data lake": an apt description.

Franz
2201 Broadway, Suite 715, Oakland CA 94612
www.franz.com

Strengths

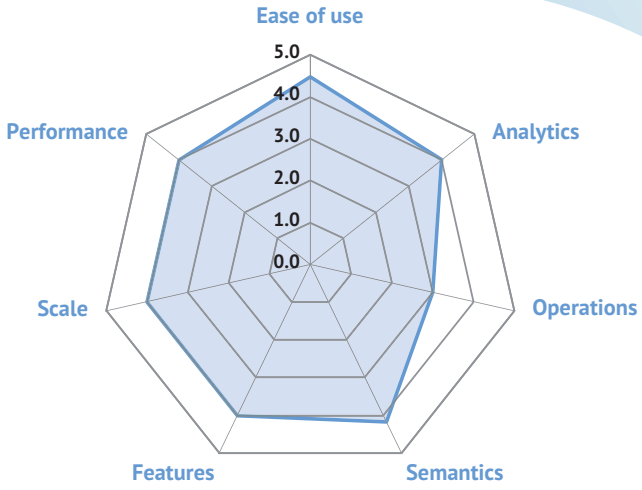
- Gruff is a major differentiator. It provides by far the easiest way of developing graph queries that we have seen from any vendor. The company also partners with graph visualisation vendors such as Linkurious.
- The analytic support provided is extensive and Franz is one of relatively few vendors that is serious about complex analytics. The nDimensional support is also a differentiator.
- We particularly like the semantic data lake concept as well as the ability to associate probabilities with relationships.

Threats

- Vertical market focus is a good thing if you are in a relevant space, but it can be off-putting if not. In practice, it doesn't take long to implement appropriate ontologies so this is a perceived rather than a real threat.
- It is impressive that Franz can compete in the cognitive computing space and we especially like the fact that it is not a black box (in other words, you can see what is happening). Nevertheless, this does bring the company into competition with heavyweight vendors that already occupy this space.

Summary

AllegroGraph is the highest ranked product in its class and, thanks to Gruff, we rate it as the easiest product to use.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

IBM

IBM has three potential graph-based options. Firstly, you can store RDF triples in DB2. However, IBM has not implemented anything very sophisticated within DB2 to leverage this storage capability. Secondly, it has an experimental graph database called System G. This looks seriously impressive (we would like to see it released as soon as possible) but it has been under development and/or in experimental mode for a long time and it is not clear when or if it will be made generally available. Thirdly, there is IBM Graph, which has recently been released (summer 2016). This is a cloud-based (IBM Bluemix) managed property graph database that is available as a service. There is no on-premises option. It is based on Titan (though some of the datatypes support by Titan – for example geo-shapes – are not supported by IBM), running on Cassandra, Apache Tinkerpop (Gremlin) and Electric Search. At present the focus is on operational applications with an intention to add advanced analytics capabilities. The target market for IBM Graph is developers building applications running on Bluemix and the emphasis for IBM has been on creating interfaces to IBM Graph that make graphs easy to use and develop for those who are otherwise unfamiliar with graphs.

Strengths

- The very fact that this is IBM entering a market populated by far smaller vendors has to be a plus.
- There are few other managed service graph products in this market.
- The emphasis on making graphs easy (easier) for developers is laudable.

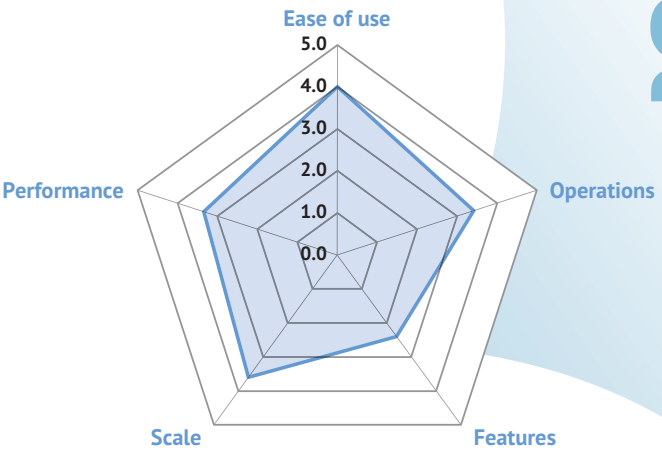
Threats

- IBM does not appear to have a coherent strategy with respect to graph products.
- There is no on-premises option other than an implausible DIY implementation.
- Titan is under threat because no-one, including IBM, is contributing significantly to the code base. Titan is therefore likely to have to be replaced. Provided that this is by something with a native graph engine then we would have to say that that is a good thing.
- While IBM is backing Tinkerpop – something we applaud – it is the only vendor in this entire space not to offer any sort of declarative language. It is true that some aspects of Gremlin have declarative properties but as a whole you would not describe it as declarative. Performance is therefore likely to suffer.

IBM
Armonk, North Castle, New York, USA
www.ibm.com

Summary

IBM Graph is a good choice for developers wishing to use graph-based technology on the Bluemix platform. However, we await the release of System G with bated breath.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

MarkLogic

MarkLogic is the name of both the company and the product. As a product it was originally an XML database though it would now be more accurate to describe it as a multi-model database. JSON support, for example, is now part of the core engine. From a graph perspective RDF triples are embedded into either XML or JSON documents and then accessed via a triple index, which you can query either via SPARQL (version 1.1 is supported) or by other means. SPARQL queries may also be called directly from server-side code written in XQuery or JavaScript. Inferencing is supported via backward chaining. There are 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 – this feature also supports versioning). The product is ACID compliant, with immediate consistency, and there are enterprise grade features such as high availability, resilience and so forth. In the most recent release the company has added significant security capabilities including improved encryption and key management, enhanced role-based security, redaction and even data masking capabilities. OWL is also supported. The product is available in the cloud (AWS) as well as on-premises. Historically, the company was traditionally successful within media and publishing but is now seeing its main growth in financial services and healthcare, typically for operational and hybrid operational/analytic applications. These often involve combining data from different sources within a unification (virtualisation) layer.

Strengths

- MarkLogic is the leading vendor – at least by size of company and number of accounts – in the RDF database space.
- The company has a history of providing mission critical software to major enterprises.
- The security and privacy controls being introduced in the latest release (version 9) are ahead of its competitors, as is the bi-temporal capability provided.

MarkLogic
999 Skyway Road, Suite 200,
San Carlos CA 94070, USA

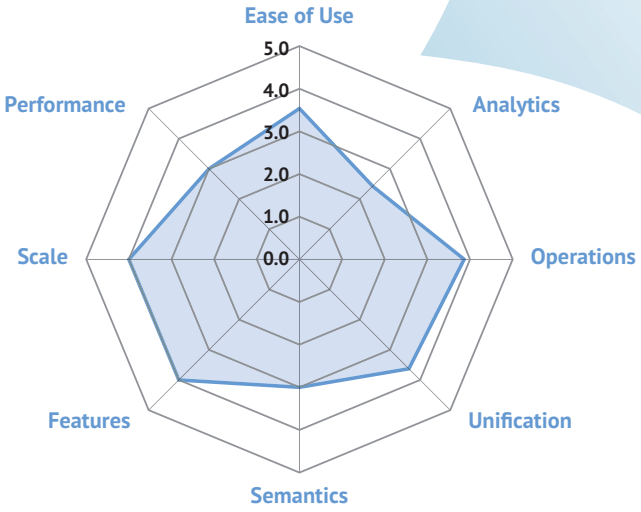
www.marklogic.com

Threats

- While multi-model approaches have a number of benefits, like all multi-model vendors MarkLogic is subject to the claim that implementations on native engines can offer superior performance.
- Unification is a growing market but is less well-known as a space in which graph databases operate. So a significant degree of evangelism is required.

Summary

Unusually for a vendor in this market – where suppliers tend to focus on a single area – MarkLogic is a leading vendor into two areas: for both operational environments and as a unification platform.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Neo4j

Neo Technology was founded in 2007 in Sweden and moved to the United States in 2011. Its product, Neo4j, is a labelled, property graph database with a native engine that is targeted at operational and hybrid operational/analytic use cases. It is ACID compliant and supports immediate consistency. Unusually for a property graph SPARQL is supported. So too is Gremlin (part of the Apache Tinkerpop project). However, most users employ Cypher or OpenCypher (the open source version), which is the declarative language developed by Neo4j. It is notable that both Databricks and Oracle have publicly endorsed OpenCypher.
As with any declarative language this is best implemented along with a database optimiser and the company has devoted considerable resources to this, extending beyond an original rules-based optimiser so that it is now primarily cost-based.
In the latest release (3.0) the optimiser is used to optimise writes as well as reads. In addition, the optimiser supports Cypher queries running against Spark environments as well as Neo4j. Historically, the company has prioritised performance over scale but the latest developments introduced by the company (and those planned for the next release, due later in 2016) mean that this is no longer the case. Finally, it is worth stating that the company has a significant partner base in addition to its direct customers and some of these have extended the product's capabilities. For example, Structr.org has extended Neo4j to act as a JSON document database while GrapheneDB is a fully managed cloud-based version of Neo4j running on both AWS and Heroku.

Strengths

- Neo Technology is the leading vendor of property graphs and Neo4j is the most well-known database in the entire graph space.
- Performance is a major focus: Neo4j not only has a native graph engine but the company has also invested significant resources into its database optimiser.
- The longevity of the company and its product mean that Neo4j is likely to be more stable, more robust and more enterprise ready than some of its more recently introduced competitors.

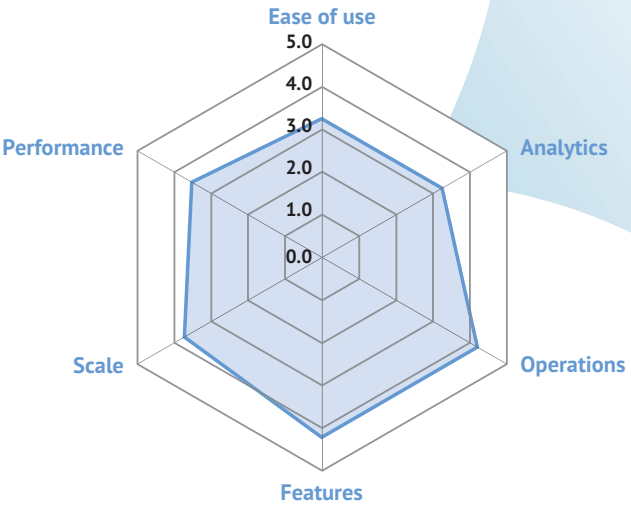
Neo4j
111 E 5th Avenue San Mateo, CA 94401, USA
www.neo4j.com

Threats

The only potential threat to Neo4j's leadership position is if one or more of the 800lb gorillas (IBM, Microsoft, Oracle, SAP) get their graph act together. At present there are no signs of that happening anytime soon.

Summary

Neo4j is the clear leader in the property graph space when it comes to operational and hybrid operational/analytic environments.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Ontotext

Ontotext was one of the first vendors into this space, having been originally founded in 2000 (in Bulgaria) to investigate semantic technologies. Head office remains in Sofia but the company also offices in London and New York. Its GraphDB product (previously known as OWLIM) is an RDF database with dynamic indexing that integrates with various search technologies, as well as text mining. 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. GraphDB's inference engine employs forward chaining and the company has a patented method for retracting materialised inferences. As far as features go, GraphDB includes a number of capabilities that extend beyond the database, notably ontology visualisation, connectors to a variety of third party environments (mostly search engines such as Solr, Lucene and ElasticSearch), partnerships with companies like TopQuadrant and Semantic Web Company that build semantic models that are implemented on top of GraphDB. There are also the sort of database administration tools and similar features that you would normally expect. Also notable is support for geo-spatial constraints. There is a developer edition (GraphDB Free Edition) that is free but limited by the number of concurrent users, and there is also a SaaS version available on AWS. Target environments include reference and master data management, metadata-based content management, knowledge management, information and relationship discovery, and content management solutions that involve text analytics on top of big knowledge graphs.

Strengths

- The company and product have proven longevity.
- Taking the brand name "GraphDB" is a substantial advantage.
- Offers a much broader set of capabilities – both in terms of ancillary tools and industry solutions – than typical competitive solutions.
- The company offers a one-stop shop for both the database and text mining. The latter capability is especially strong because of the way that it works in conjunction with big knowledge graphs.

Ontotext
Polygraphia Office Center fl.4, 47A
Tsarigradsko Shosse, Sofia 1124, Bulgaria

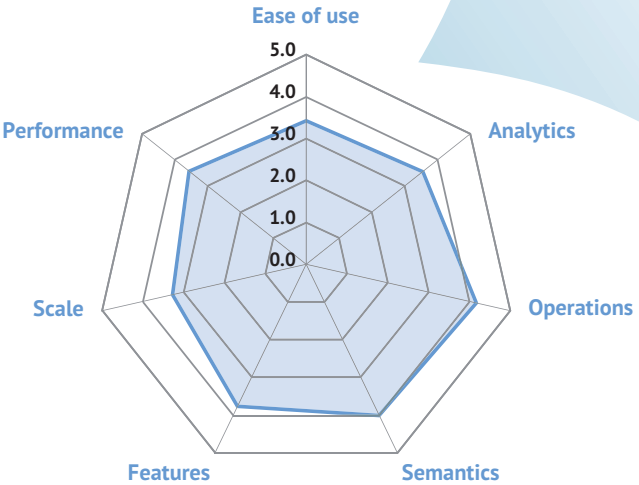
www.ontotext.com

Threats

- There is a very definite focus on text, content and related areas. This may be perceived (falsely) to mean that GraphDB is not suitable for more general operational and hybrid operational/ analytic environments.
- The company initially engaged Americans to help them break into the US market. This was not as successful as the company would have liked and it has now sent Bulgarians across the water. It remains to be seen if this will work.

Summary

Ontotext is the product of choice for text mining and associated activities.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Bloor

Product Sheet

OrientDB

OrientDB from OrientDB Ltd is a multi-model database with extended property graph database capabilities. It is an open source offering although there is an Enterprise Edition available with additional features such as monitoring, auditing, incremental backups, multi-data centres, and so on. Cloud-based options are also available for AWS and Microsoft Azure. The company claims that OrientDB was the first multi-model graph database to be launched (in 2009). Originally, it was designed as a hybrid document/graph database, but since then the core (native) engine has been extended to include objects, spatial and key-value elements. The product is ACID compliant and supports strong consistency though eventual consistency is an option. It supports Apache Tinkerpop and, especially, Gremlin. However, more importantly, it uses an extended form of SQL for query processing that leverages MapReduce under the covers. The product can be used in full schema, schema-free or hybrid schema environments and uses sharding for distributing data across a cluster. In addition to the database itself, there is a native ETL engine that can be used to import and export JSON documents and, moreover, Teleporter (the product name) provides transformation capabilities for mapping from a relational to a graph model. There is also a Studio product that allows viewing and editing of the environment, and there is a JDBC connector to support integration with various (partner) visualisation tools. The product is targeted at both operational and hybrid operational/analytic environments and it is also worth commenting that the company has considerable success with third party technology companies (most notably Dell and Informatica) embedding its database into their products.

Strengths

- Other vendors have developed their own declarative languages so supporting (extended) SQL is a significant benefit.
- We especially like the capabilities offered by Teleporter. Converting traditional data formats into graphs can be onerous and anything that will ease this process has to be a good thing.
- The multi-model nature of OrientDB gives the company a larger addressable market than would otherwise be the case.

OrientDB
Unit 702, Salisbury House, London Wall,
London EC2M 5QQ, England

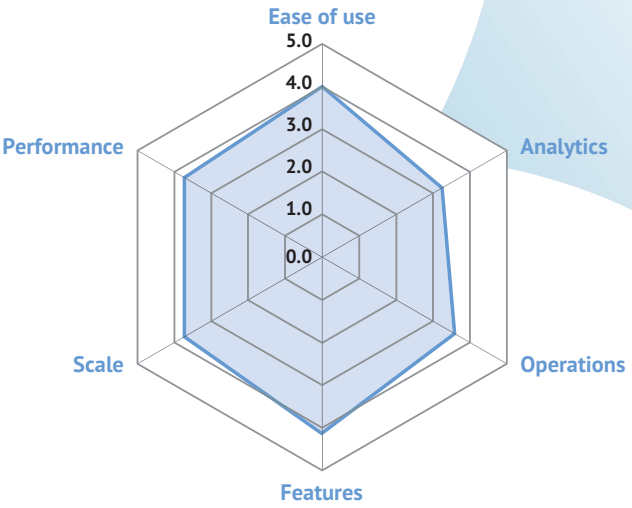
www.orientdb.com

Threats

- The company has historically devoted limited resources to sales and marketing, relying on downloads and word of mouth. We are pleased to hear that the company plans to expand its efforts in these areas but it remains to be seen if the company is successful in that effort.

Summary

Across all scores OrientDB there is no higher ranking graph (or RDF) database targeted at the operational and hybrid operational/analytic markets.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Stardog

Stardog is developed by Stardog Union (previously Complexible, and before that Clark & Parsia). Historically, the company was self-funded but it recently raised its first round of venture capital. Technically, Stardog is an RDF database with strong support for SPARQL and OWL (it supports all of OWL 2) but extended with property graph and graph traversal capabilities as well as support for Tinkerpop (and Gremlin). The Lucene search engine has been embedded into Stardog. The database is ACID compliant and supports immediate consistency. The database uses query time reasoning that does not require the materialisation of inferences. It has a built-in optimiser for SPARQL and there is graph versioning so that you can track changes to a graph, both for auditing and analysis purposes. The emphasis in the product has always been on (model-driven) integration and analytics and although it has historically tied itself to the graph database bandwagon – and offers conventional transactional/analytic graph database capability – its real focus is on unification and query processing (what is sometimes known as data federation or virtualisation) across multiple data sources. Graph databases – given relevant capabilities – are especially suited to this environment because of their ability to understand relationships across data sources. In the next release (version 4.2) the company intends to implement significant support for unstructured data, including text mining capabilities.

Strengths

- Graph databases have significant benefits compared to other technologies which support unification architectures. We would go so far as to suggest that, given appropriate features, they represent a next generation version of data federation/virtualisation.
- Stardog is explained and positioned as a unification platform in a much more succinct and clear fashion than competitors in the graph market.
- The company already has a substantial user base amongst prestigious customers although these are typically departmental deployments spanning up to around a dozen data sources.
- The product name is the most memorable in this space.

Stardog Union
1400 Crystal Drive, Suite 660
Arlington VA 22202, USA

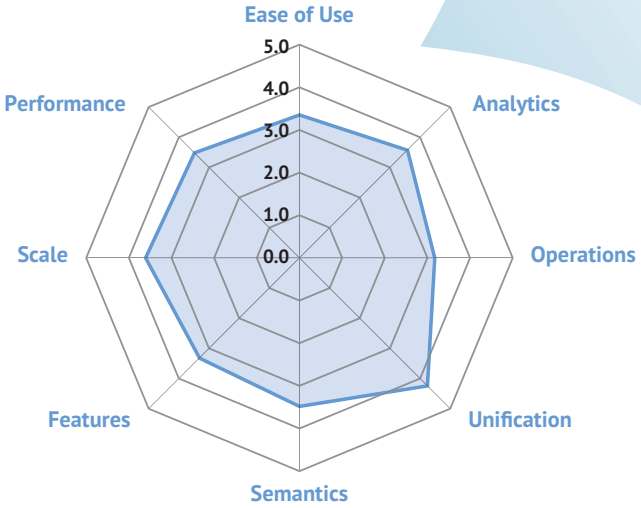
www.stardog.com

Threats

- Stardog needs to scale up its platform if it is to engage at an enterprise as opposed to a departmental level. In general, the number of sources is in double digits though we understand that the company does have users accessing as many as a hundred sources. Nevertheless, there is further work to do. We are pleased to hear that the company is working on this.
- Stardog faces significant competition from incumbent non-graph providers of data federation and data virtualisation.

Summary

While it has already capability, Stardog is the highest ranked product for unification, on which the company focuses.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Teradata

Teradata uses its Aster Analytics platform to store the vertices of a graph in a table and the edges in a second table, which can then be queried using what Teradata calls SQL-GR. Under the covers, this SQL based engine makes calls to a number of pre-built graph functions. Alongside this, analytics against both relational data and text is supported, thus providing what Teradata describes as "multi-genre analytics". Until the most recent release (version 7, announced in August 2016) Aster has relied on a proprietary storage engine. However, with version 7 Teradata has decoupled the product from its persistence layer. The first instantiation of this is on Hadoop. However, while the persistence is in HDFS, distributed processing is implemented (as it was previously) using a BSP (bulk synchronous parallel) architecture riding on the SQL foundation. 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). Typical use cases include recommendations (next best offer), churn predictions, social graph analytics for (say) criminal networks, and so on.

Strengths

- Being able to use SQL has the significant advantage of being able to use traditional BI tools.
- Teradata is a big beast in the analytics world and has a well-deserved reputation that will encourage potential users.
- Multi-genre analytics offers significant advantages when processing hybrid queries.
- Aster graph analysis does not require new skills to build a semantic data model nor extensive data preparation to load data into that model. It uses existing RDBMS or HCatalog tables.

Threats

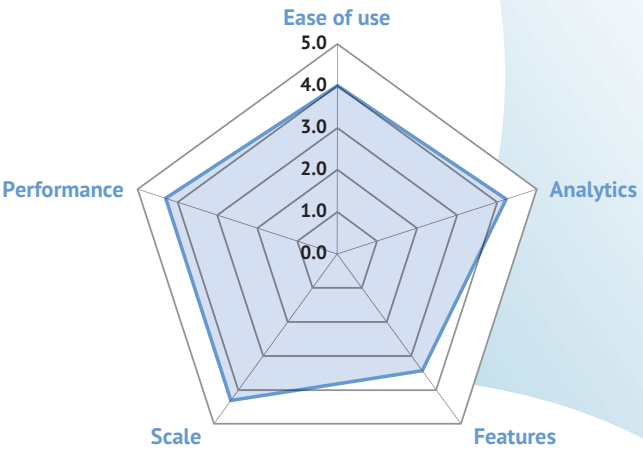
- For all its capabilities Aster is not a graph database. It is designed to enable large-scale graph analysis, not to enable semantic queries or SPARQL.
- Aster treats graph data as data about relationships. This is valid. But graph data can also include semantics. As a result, Aster can handle graph analytics but is inappropriate for semantic analysis or inferencing.

Teradata
10000 Innovation Drive,
Miamisburg, OH 45342, USA

www.teradata.com

Summary

While Teradata Aster is not a graph database per se it is the highest ranked vendor for graph (but not semantic) analytics, especially where multi-genre analysis is required.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Product Sheet

Virtuoso

Virtuoso is provided by OpenLink Software. The product started life as data virtualisation software (hence the name) supporting federated queries across heterogeneous environments. Over time, and to better support query performance, the company added persistence capabilities that eventually evolved into what the company refers to as a "Universal Server", which addresses the market for unification software. In our view it would be more accurate to describe the product as a multi-model "relational" database where, by "relational" we mean anything that handles relationships, including RDF- based property graphs (triple and quad stores accessed via SPARQL) and also relational tables (accessed via SQL). The product directly supports mapping from relational to RDF structures and it includes a SPARQL to SQL gateway. The environment also supports documents (both XML and JSON). Virtuoso also includes a Web Server (supporting both SOAP and REST) and other elements that mean that it is more than just a database. It is ACID compliant and supports immediate consistency. There are both open source and enterprise editions where the latter includes security, custom inferencing rules, data virtualisation, clustering and HADR. There is a significant emphasis on the semantic richness of the product, both in terms of the use cases that are supported and with respect to technical advantages (for example, semantic richness means that you can often avoid expensive joins, which is very important in federated query environments).

Strengths

- The product is very strong when it comes to semantics.
- We are especially impressed with the security provided and, indeed, with the potential use of Virtuoso specifically for security, where you can capture the characteristics of users and use that as the basis for defining access. This is much more granular than is typically the case.
- The multi-model nature of Virtuoso has advantages when it comes to querying across different types of data.

OPENLINK SOFTWARE
Making Technology Work For You

Virtuoso
OpenLink Software, Inc., 10 Burlington Mall Road, Suite 265, Burlington MA 01803, USA

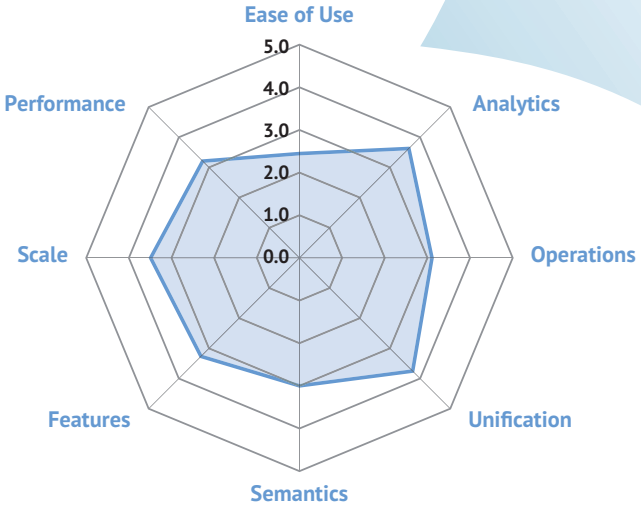
www.openlinksw.com/

Threats

- Virtuoso is complex and the company does not do a good job of explaining exactly what it is. This is particularly damaging because the product has few direct competitors. As a result, it is difficult to appreciate where it has advantages over other products that do not have its breadth of capability.
- More generally, there seems to be little investment in either sales or marketing.

Summary

Virtuoso may be a much better product than its scores suggest. However, the company does not do a good job of explaining its technology and that is reflected in these scores, especially ease of use.



It is important to note that these scores are in comparison to a theoretical ideal product rather than in comparison to any competitive product.

Conclusion

As has been noted there are lots of open source and development projects within the graph space. We have focused on products that we believe to be enterprise-ready. 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.

The products included in this Market Update have significant strengths, though some more than others. The difficulty for potential users is in identifying the type or range of use case for which each product is most suitable. As always, ultimately users should conduct proofs of concept both with respect to functionality and performance.

About the author

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 also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and 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.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channels.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

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

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