REPORT REPRINT

Stardog rocks on with improved performance for data unification

AUGUST 15 2019

By Matt Aslett

Enterprise knowledge graph provider Stardog has announced the launch of version 7 of its eponymous underlying graph database, adding a new storage engine based on RocksDB for improved balance between read and write query performance.

Introduction

Stardog has announced a major new update to the eponymous graph database that underpins its platform for data unification and the creation of enterprise knowledge graphs. Stardog 7 marks the completion of a two-year development effort to create a new low-level storage engine for Stardog. Based on the RocksDB persistent key value store, the new storage engine is designed to improve the balance between read and write query performance, to the extent that the company is claiming write performance improvements of 10-20x. The latest version also adds virtual transparency and schema multi-tenancy to improve the unification of data from multiple sources.

451 TAKE

We see significant potential for both the knowledge graph concept and Stardog's approach to it, given the increasing need for enterprises to overcome the limitations of data silos. One of the challenges faced by the company (and other proponents of knowledge graphs) is that, the larger the scale of the project, the greater the benefits – meaning that small-scale pilot projects may be underwhelming. Fortunately for Stardog, it has a growing list of reference customers that have deployed the company's technology at scale. These will likely be increasingly important as the company seeks to persuade a growing number of customers that it is possible to unify data in disparate systems and locations without having to first move it to a single platform.

Context

We noted in 2018 that the enterprise knowledge graph (EKG) had emerged as a concept to enable enterprises to overcome the challenges posed by storing data in multiple databases in multiple execution venues, managed by multiple departments, business units and geographic entities. One of the companies at the forefront of the emergence of the EKG as a market segment is Stardog. The company was founded in 2006, and can trace its origins back to the University of Maryland research lab, where its cofounders, Kendall Clark (CTO), Evren Sirin (chief scientist) and Michael Grove (VP engineering), cut their teeth on semantic web standards.

While the company was initially focused on semantic web projects for US government agencies (including NASA), and then the initial development of its eponymous graph database, its value proposition lies in the combination of that graph database functionality with a knowledge toolkit, as well as support for virtual graph queries that enable the unification of data from multiple silos. It is this unification of data from multiple data sources and the virtualization of graph queries that enable enterprises to create the knowledge graph – a virtual repository of knowledge that spans what would otherwise be individual data silos.

Stardog concedes that pitching the benefits of such a holistic approach to data management can be challenging compared with more traditional options to data management (e.g., ETL-based data integration), which often start with solving a niche data integration challenge and can grow into a complex web of integration pipelines that can be difficult to manage.

Greater adoption of the EKG is helping. Stardog now has more than 50 customers and can point to the likes of NASA, Bosch, Ericsson and Morgan Stanley as reference customers. Although the company had early success in financial services, manufacturing and government, it is now seeing traction in other areas, including high tech.

Stardog has more than 50 employees, compared with more than 30 in early 2018. The company has raised $11.3m in funding from Core Capital, Boulder Ventures, Grotech Ventures and Tenfore Holdings.

Technology

While Stardog is known for its eponymous graph database technology, it is important to note that the key to the company's ability to enable EKG deployments is its abstraction of the business logic layer from the data storage layer. Stardog does not use graph as a data model for storage, but to connect and map the relationships between data in multiple underlying data models.

Graph queries can then be run against the unified data, with Stardog's virtual graph engine translating SPARQL and GraphQL queries into the appropriate queries (including SQL) for execution by the data storage environment, which could be a database (such as relational, key value, document or graph databases) or unstructured content (PDFs, emails, text files, images, etc.).

The virtual graph engine is one of the key features provided by Stardog that makes the creation of knowledge graphs possible, along with its BITES ingestion system, which incorporates natural language processing functionality such as entity extraction and linking, as well as relationship inferencing, to enable the querying of unstructured data sources. Reasoning and inference are of significant importance in the development of the semantic data model, reducing the time taken to define the queries and map the higher-level semantic concepts to underlying databases and tables.

Those capabilities are part of the company's Knowledge Toolkit, which complements the Stardog database with functionality including logical reasoning for inference, explanation and model checking; statistical inference and probabilistic reasoning; and full-text search, geospatial semantics and query, and knowledge graph construction services. The platform also includes Stardog Studio, which was introduced in early 2018 as a development environment that provides a SPARQL query notebook and tools for managing Stardog databases, users and roles.

As noted above, the delivery of version 7 of the Stardog database is the fulfilment of a two-year development project to create a new underlying storage engine (called Mastiff) based on the RocksDB embeddable persistent key value store. The primary advantage of moving to Mastiff is to improve performance, particularly concurrent write performance across a horizontally scalable cluster.

Another new feature in Stardog 7 is Virtual Transparency, which essentially removes the previous requirement to code the location of the data being queried into a virtual graph query. Removing this requirement means that queries no longer need to be rewritten as data locations change. There is also support for schema multi-tenancy, which removes the previous requirement for multiple departments and applications within an organization to agree on and adhere to a single schema in order to share and reuse data connected via Stardog.

Competition

Stardog is by no means alone in pitching the concept of the enterprise knowledge graph. There are a number of providers that have evolved from a focus on the semantic web to address knowledge graphs, including Cambridge Semantics, Ontotext, Siren and Franz. Maana is another potential competitor, albeit with a focus in the oil & gas sector.

There are also several vendors focused on the challenge of overcoming data silos through the unification of data, although not necessarily pitching their offerings using the knowledge graph terminology. Examples include Io-Tahoe, with its data discovery platform; Gemini Data, which combines graph visualization with zero-copy data virtualization; Tamr with its machine-learning-driven, probabilistic-model-based approach to data unification; and Apollo with its GraphQL-based application development platform.

Perhaps greater competition comes from more traditional (ETL-based) approaches to data integration, as offered by the incumbent data management vendors. Stardog sees its greatest competition coming from data management heavyweight Informatica, as well as Reltio, which delivers a combination of graph-based master data management and industry-specific applications.

SWOT Analysis

STRENGTHS

Stardog boasts undoubted expertise in relation to graph technologies, as well as a differentiated offering that provides a potential solution to the age-old problem of data silos.

WEAKNESSES

Pitching the overall advantage of the knowledge graph approach requires some evangelizing, particularly since the true advantages may not be evident from a small-scale pilot. A growing roster of reference customers is helping.

OPPORTUNITIES

While it is not new, the need to overcome the challenges of data silos is, if anything, growing in significance given the growth of nonrelational databases and cloud computing.

THREATS

The company faces a number of competitors, while the data management market is dominated by traditional incumbents that will be quick to respond as adoption of the knowledge graph approach expands.