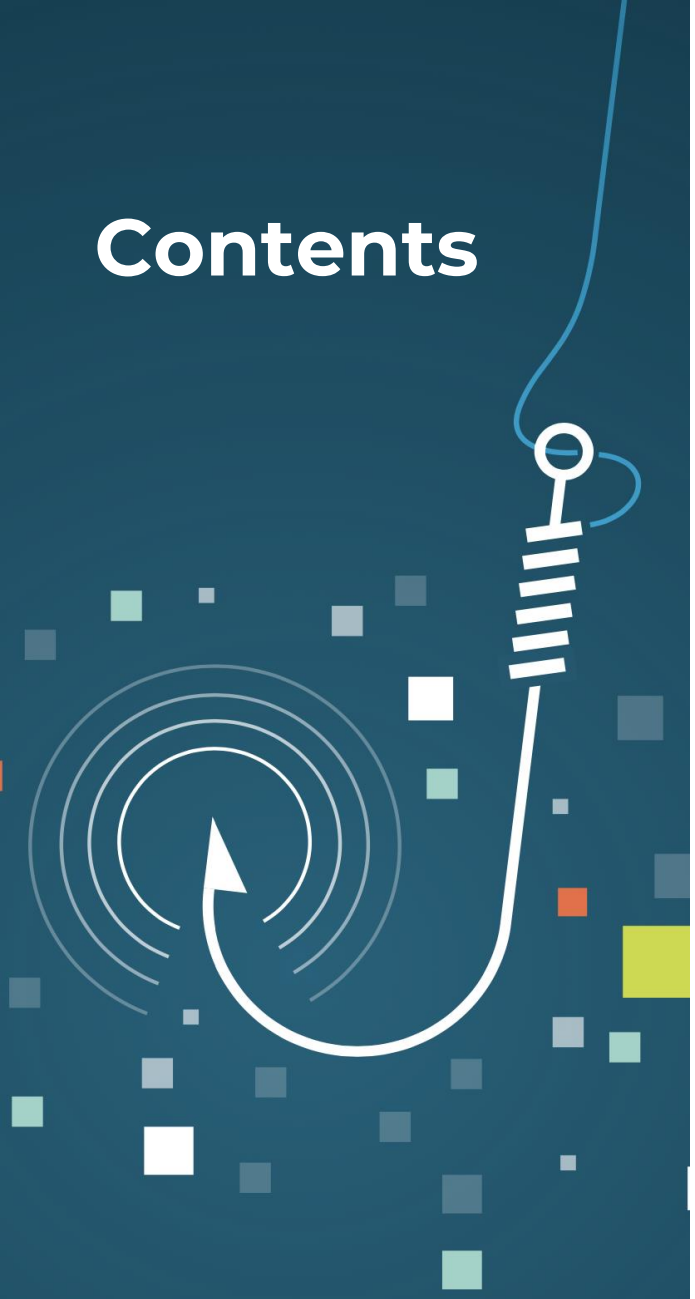
Maximize the ROI of Your Databricks Lakehouse with Stardog

A STARDOG E-BOOK

Contents



Introduction

With the ability to store a full variety of data types and handle ever-larger data volumes generated at high velocity, the Databricks lakehouse has become *the* modern enterprise data architecture.

Data-driven organizations have chosen the lakehouse to streamline data management, derive new insights, and accelerate machine learning. A semantic layer that harmonizes the lakehouse with data that must live outside of it further improves ROI, helping to answer complex questions, connect relevant data elements, and provide greater context to data.

A semantic data layer provides the ideal platform to accelerate returns from lakehouse investments. By connecting enterprise data and overlaying business semantics, Stardog's Enterprise Knowledge Graph platform facilitates more agile data operations, reduces the cost of data integration, and helps generate powerful insights into complex business challenges.

Throughout this e-book, we'll discuss what you can accomplish by combining Databricks with Stardog, the leading Enterprise Knowledge Graph platform:

- Semantically enrich your consolidated data for more robust analysis
- Enable federated querying and analytics on your lakehouse
- Accelerate your machine-learning projects
- Connect data outside your lakehouse for a full information landscape
- Make sense of real-world complexity

Chapter 1

Semantically enrich your consolidated data for more robust analysis

Innovative companies like Databricks and their Data Lakehouse approach are helping many organizations co-locate data from across the organization with cost-effective approaches for storage as well as opportunities to operate on all that data at the computational layer to leverage the benefits of AI. These landing zones have effectively reduced the need for organizations to maintain expensive brittle ETL pipelines against traditional structured and costly data warehouses on-prem.

While access to all the data has become a reality, it is far from being democratized and available to the very users that need rapid insights to keep pace with the change in business dynamics and consumer preferences, in other words, enabling non-technical users to self-serve and collaborate.

Hooking up a BI tool like Tableau directly to Databricks might seem to accomplish that last mile, but there's a lot to be desired in terms of reducing latency, promoting collaboration and re-use with an easy-to-understand vocabulary, providing context by connecting data across domains, enabling self-service through data exploration, and enriching analytics by inferring new insights.

*A **Semantic Layer** is a data layer that operates between data storage and analytics, represents a logically-enriched view of information described as a set of interrelated business concepts, and results from the implementation of a knowledge graph.*

The idea of a semantic layer is not new.

It has been around for over 30 years, often promoted by BI vendors helping companies build purpose-built dashboards.

However, broad adoption has been impeded, given the embedded nature of that layer as part of a proprietary BI system, often too rigid and complex and suffering from the same limitations as a physical relational database system which models data to optimize for its structured query language rather than how data is related in the real world—many-to-many.

Boehringer Ingelheim Uses Stardog to Transform Its Data Lake

The world's largest privately held pharmaceutical company implemented the Stardog Enterprise Knowledge Graph platform as a semantic layer over its data lake, making the information much easier to navigate, query and analyze.

Enhanced discovery
Bioinformaticians use Stardog to navigate up to 90% of the company's R&D information, with all relevant data logically connected. It has allowed users to search for specific diseases, studies, or genes, and easily explore how individual elements relate to one another.

Faster time to insights
The new linked data dictionary enabled users to fetch data directly from the data lake and put it into R, ready for analysis.

Simplifying complex data analysis
By increasing the connectedness of its data, Boehringer has been able to answer complex questions more quickly and effectively.

Learn more

An enterprise knowledge graph powers a "missing middle" semantic data layer that enables your organization to explore and exploit connections across your data universe with business context. You gain a more complete and accurate understanding of any given scenario. Specifically, end users can:

1. **Ask questions based on business concepts and the inter-relationships between them.** Those concepts, in turn, map to the underlying metadata (tables, views, attributes) that can help facilitate rapid pipeline development for sharing data across applications through the creation of metadata-informed data pipelines.
2. **Run fast and flexible federated queries between Databricks and data in other sources**—structured, semi-structured, or unstructured—in support of ad hoc data analysis. Linking and querying data in and outside of Databricks enables just-in-time cross-domain analytics for richer, faster insights without creating data sprawl challenges for the organization.
3. **Save time and money spent on data wrangling and data movement activities**, easily sharing your findings through visualization that promotes data story-telling and enable self-service analytics directly inside your existing investments like Tableau or Jupyter notebooks through the re-usable semantic layer.

Chapter 2

Enable federated querying and analytics on your lakehouse

Just as the DevOps movement has driven greater automation of the software development lifecycle and increased the speed with which developers can get code into production, it also promises to increase the speed with which data can be provisioned to support both production applications (through regularly scheduled pipelines) and ad hoc analysis.

Many organizations recognize the high cost and latency associated with large teams of data engineers continuously wrangling data to ready it for analysis. Being able to automate some of those processes sounds ideal. Unfortunately, much of the know-how to bring data together is not systematized in most organizations.

An enterprise knowledge graph provides the means to turn DataOps from pipe dream into reality. By logically connecting enterprise data, an Enterprise Knowledge Graph maps out key data elements in a way that makes sense to the business.

In fact, analytics applications that point to a semantic data layer powered by an enterprise knowledge graph can provide better and faster services to end users than those which reach directly into the data lake by drastically reducing latency for queries, often from tens of seconds to hundreds or dozens of milliseconds.

Stardog returns queries in milliseconds *that currently take tens of seconds when they're directly against a data lake that lacks a semantic, caching layer.*

Chapter 3

Accelerate your machine-learning projects

Machining learning (ML) development strategy generally focuses on ensuring data is accessible, reusable, interpretable, and high quality, which is often a challenge even with infrastructures that include a data lakehouse.

Most organizations have incorporated statistical learning through data science projects. But rule-based AI is growing, and this approach includes everything from making intelligent inferences about schemas to expedite data integration to assembling techniques for text analytics or Natural Language Processing (NLP).

The best data science projects come from combining more than one source of data, and that data science nightmare has become less terrifying with the advent of solutions such as Databricks. When it comes to combining data sources and datasets, ontologies and context help get machine-learning projects through the last mile. This means platforms like Stardog assist tremendously.

According to the 2020 AI in Organizations Survey, 23% of organizations deployed graph techniques in their artificial intelligence (AI) projects. Large technology companies, public-sector organizations, financial solutions providers, and healthcare led the way, using graph technologies to enhance data search, information retrieval, and recommendations.*

*As cited in Gartner's "How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications: May 27, 2020.

Knowledge graphs make it easier to feed better and richer data into ML algorithms. They do this by helping you leverage industry-standard models and ontologies, model your domain knowledge, and connect disparate data sources across the enterprise. You can maximize the use and reuse of your internal content by laying the foundation for AI and semantic applications—ultimately, showing meaningful relationships between your data.

How Knowledge Graphs and Machine Learning Work Together

The inherent traits of knowledge graphs posit them as a top tool of modern AI and ML strategy. Let's examine a few ways in which they help.

To learn about how Stardog leverages AI to solve common challenges, and how the LLM-powered **Stardog Voicebox** can make knowledge graph adoption simple, check out our blog.

Learn more

Enable Highly Productive Data Workers

A significant portion of data scientists' time is spent cleansing data to produce the desired model or expected results. With a knowledge graph, they can train models directly on unified data.

Stardog's Inference Engine allows you to resolve conflicting data definitions without changing or copying the underlying data. Capture your business and domain rules in the data model; the engine intelligently applies these rules at query time.

Work Seamlessly with Your Existing Tools

Knowledge graphs that contain virtualization (not just graph databases acting as a knowledge graph) work well to maintain data accuracy and the security of existing tools. A knowledge graph that is a true data layer does not require you to change anything you're doing today.

Stardog's tools comprise things like Python support, R libraries, etc. Additionally, the output of your models can also be put back into the knowledge graph.

Get Started in AI and Machine Learning Quickly

If you need to predict something, classify something, or see if things are similar, an enterprise knowledge graph platform can help you get started quickly.

Stardog ships with built-in predictive analytics and similarity search, supporting rapid model development and iteration for data analysis. You can extract patterns from your data and make intelligent predictions based on those patterns.

Infer New Facts

Machine learning complements logical reasoning. Inference expresses all the implied and predicated relationships and connections between your data sources, creating a richer, more accurate view of your data. Better data means better learning. And better means providing context, not just volume.

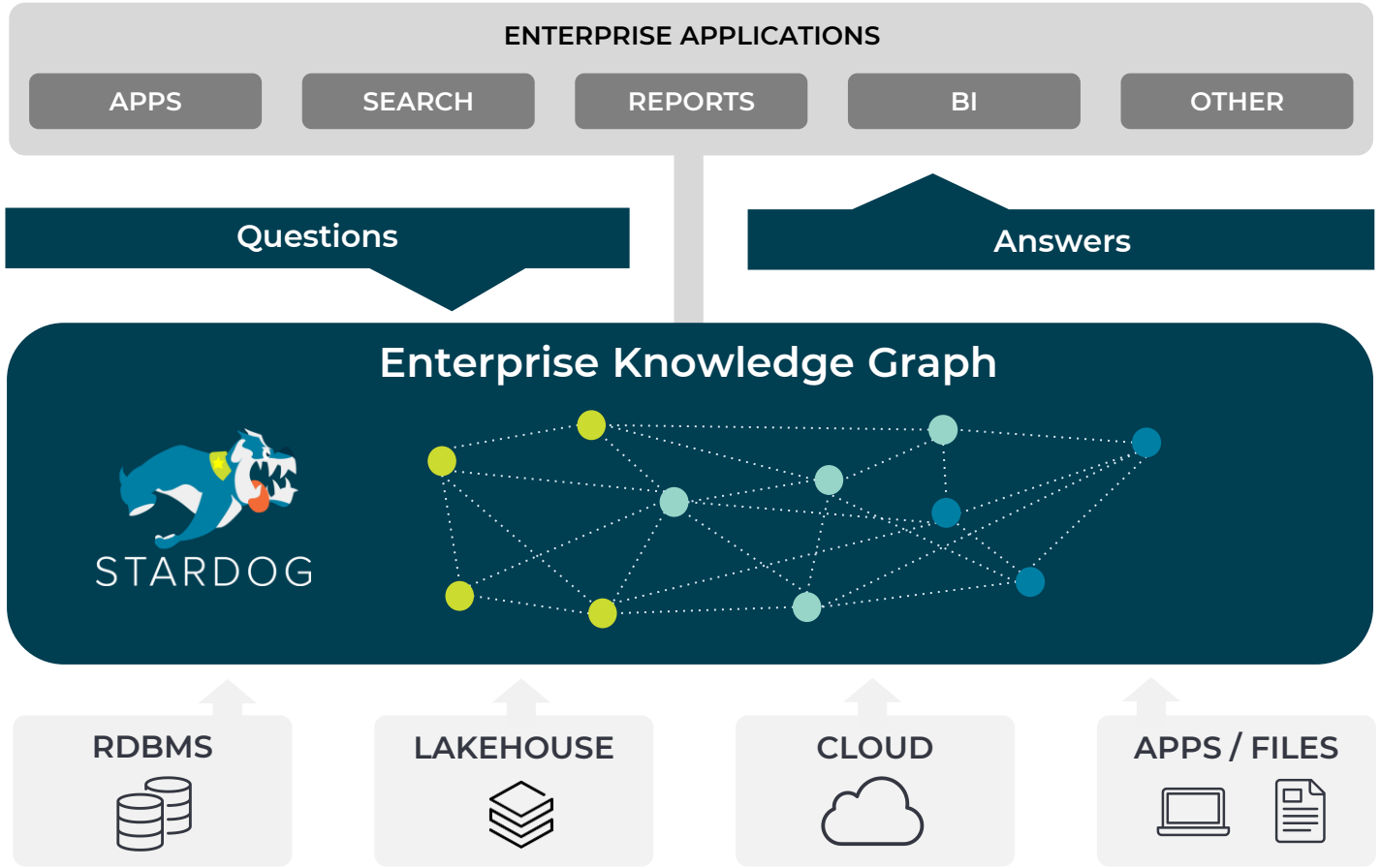
Adding the best-in-class inference engine in Stardog to your tool suite, whether you use its built-in ML libraries exclusively or in concert with third-party libraries, allows you to capture facts and infer new facts.

Chapter 4

Connect data outside your lakehouse for a full information landscape

While a lakehouse provides the ideal view of the data lake beside it, it doesn't usually capture the full vista of an organization's data landscape. The reality is that some data may remain in silos outside the data lake, often for valid reasons, such as multi-cloud apps and regulatory or sovereignty concerns.

Adding a semantic layer to your data lake means enjoying lakehouse scalability and lower cost-per-GB storage, while speeding up unification to reduce or eliminate the need for complex ETL pipelines and manual data mart stewardship. This combination generates a direct ROI over data warehouses lacking the semantic layer. The business drives value by using the fabric with an ecosystem of tools, dashboards, and systems, including data science notebooks and Tableau.



Implementation: Big-Picture Thinking, One Step at a Time

The problem with many enterprise technology implementations is that, by their very nature, they are large and complex. Without a clear focus on delivering business benefits, it is easy for projects to become white elephants. To prevent that from happening (or to rescue existing white elephant projects), it is important to remember that there is never a need to eat an elephant in one bite.

An enterprise knowledge graph offers a highly flexible and extensible data model. But that does not mean that everything needs to be modelled at once. Each organization must prioritize implementations according to their unique business requirements. In other words, they must look at which business questions they need to answer most urgently.

Each organization will have its own unique set of priorities. However, it is vital that each organization focuses on delivering business goals rather than simply technical capabilities. Tackling specific data domains in their entirety while focusing on achieving business objectives—for example, attaining a more complete understanding of customer data to help improve service levels and drive retention or identify cross- and up-sell opportunities—enables organizations to direct resources in a more productive manner.

> “Our primary objective is to provide data at a higher quality and relieve the heavy lifting up front so our data scientists can actually work with the data.”
>
> **— Head of IT Research, Top Global Pharma**

Chapter 5

Make sense of real-world complexity

The world we live in is complex. Many specialized systems have emerged to help individuals organize their day-to-day responsibilities. However, the complexity faced by organizations is compounded many times over.

It's difficult to follow our instincts and explore data to find the root cause of issues

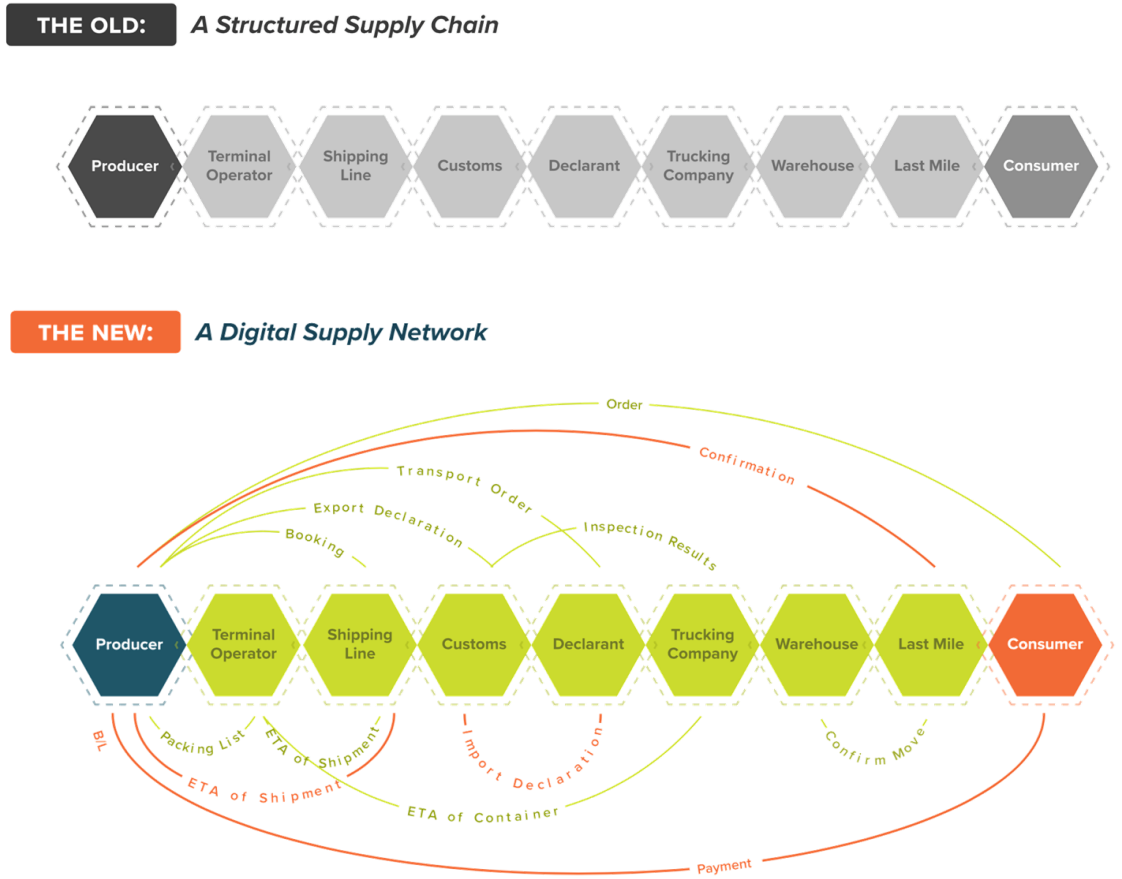
Data sources define compounds in different ways, though they're the same entity

How do we detect fraud across the entire network of actors, assets, controls, and systems?

While data lakes are very effective at consolidating all kinds of enterprise data, they do not necessarily help to make sense of that data. In order to do so, organizations need a way to model the kind of complexity that we have just illustrated.

Enterprise knowledge graphs capture business complexity in an intuitive manner. That is because they are not encumbered by the physical location or structure of data (i.e., where and how it is stored). Rather, they focus on modelling data in a way that makes sense to the business by capturing the relationship between different data elements. This business-centric approach makes it much easier to tackle the kind of complex use cases previously described.

In the case of supply chain analysis, organizations can make complex associations of data across the span of the operation. Manufacturers, for example, can instantly see the connections in the chain to enable scenario planning based on complete data, rapid identification of needed operational changes, and organizational resiliency to drive long-term customer value and higher profit margins.



> Data and analytics leaders rely on graphs to quickly answer complex business questions which require contextual awareness and an understanding of the nature of connections and strengths across multiple entities. Gartner predicts that by 2025, graph technologies will be used in 80% of data and analytics innovations, up from 10% in 2021, facilitating rapid decision making across the organization."
>
> **— Gartner, 2021**

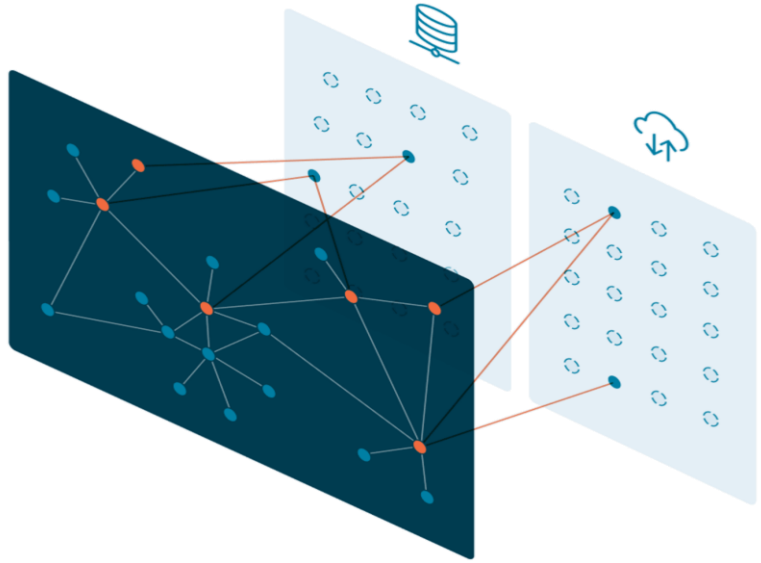
Conclusion

Maximize ROI from data lake investments using an enterprise knowledge graph

A Databricks lakehouse is very effective at consolidating data that has naturally amassed in silos. However, the biggest ROI for an organization comes when that data—along with data that must live outside the lakehouse—can be leveraged for demonstrated business value.

The combination of Databricks and Stardog helps data leaders:

- Unify data through business meaning
- Connect to all relevant data
- Enable data exploration and discovery



> Formally transitioning from a relational model to that of linked data was a huge strategic benefit to the bank. We are now able to design and link domain models across organizations and silos."
>
> **— Executive Director, Top 5 US Bank**

Accelerate Your Lakehouse Investment With Stardog

The Stardog Enterprise Knowledge Graph platform connects data based on business meaning to get better insight faster. Organizations like Boehringer Ingelheim, Schneider Electric, and NASA rely on Stardog's flexible, reusable semantic data layer to accelerate insights from data lakes, data warehouses, or any enterprise data source. Learn more at stardog.com.

Contact us:
stardog.com/company/contact/

Learn more:
stardog.com

Follow us:
@StardogHQ

STARDOG

STARDOG