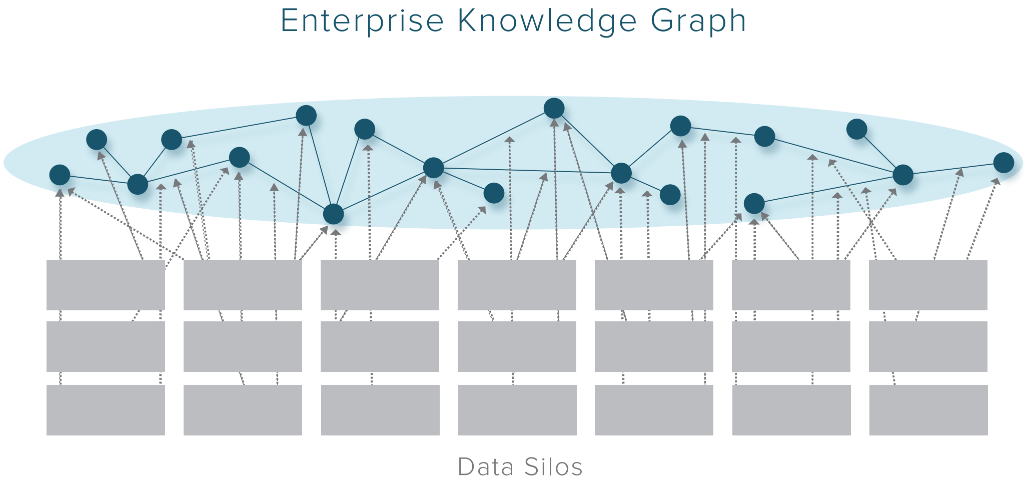
STARDOG UNION

INDUSTRY PERSPECTIVES: **INNOVATION**

# What is a Knowledge Graph?

## Unconnected Data is a Liability



Stardog is the world's leading Knowledge Graph platform for the enterprise. But what is a Knowledge Graph, and why should you want one?

### The Problem With Data

Enterprise data is both the disease and its cure. Data will save us, and data will kill us all. At the same time. Enterprise data is the world's most strategic asset going forward, while on the ground it's painful—diverse, heterogeneous, and distributed.

My favorite metaphor (to overcome) is the ***silo***. Which is where farmers store grain—keeping it safe from blight, pests, and the weather—so that no one starves when winter comes. Real silos have high utility in farming. And in the enterprise, data silos allow local control and governance in a way that may be valuable. Legal and regulatory considerations may require that some silos remain as silos.

But a silo is a disconnected thing that prevents larger structures from being composed easily. Silos impede everything: app dev, data science and analytics, reporting, compliance, and the AI-singing robots of our fever dreams.

Data silos mean unconnected data—and unconnected data ***sucks***.

## Connectedness is the Solution

A knowledge graph is the only realistic way to manage enterprise data in full generality, at scale, in a world where connectedness is everything. This shouldn't really surprise us given that graphs are all about ***connections*** and ***connectedness***.

The CIO of an enormous American bank told me recently that his organization annually spends 1/3rd of its IT budget on the Enterprise Data Silo Problem. That's just unacceptable.

Among Stardog's current production customers there are many motivating examples where crucial decisions depend on connecting otherwise unconnected data:

- Global manufacturer calculates market penetration (for all global markets) for each of its thousands of SKUs, turning 20+ PLM silos into a single PLM
- Global consumer org populates map application from dozens of 3rd party data sources
- Global ecommerce site turns many disconnected, partial product catalogs into a single product catalog
- Global bank turns trillions of counterparty risk instruments into a single Financial Knowledge Graph
- Pioneer of space exploration unifies disconnected supply chain and systems engineering data into a single Knowledge Graph for lunar mission planning

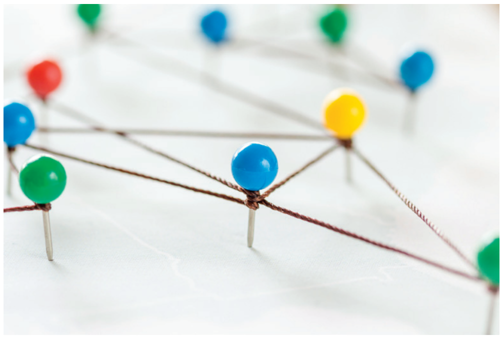
Each of these is radically different but, from our point of view, also identical. At the same time.

How is that possible? Two large-scale historical technology trends are crucial here: first, ***virtualization of everything***; second, ***graph is the data model for the next 20+ years***. The value proposition of Knowledge Graph lives at the intersection of these trends. Let's talk about both of them.

# Graph is the data model for the next 20+ years.

## First Trend: Virtualization

What do I mean by virtualization? Consider, by way of analogy, Storage Area Networks (SAN). A SAN is a lie, a useful fiction that we agree to tell ourselves. A SAN says there is one infinite-sized hard  drive that everyone in the enterprise can read from and write to whenever they want without regard for the laws of physics. That's a useful lie; which is to say, it's an ***abstraction***.



From one perspective, a SAN is fake news. Sad! Physics always matters. And yet with the proper implementation, technical know-how, and capital investment, we can just act as if there is one big hard drive and...**it just works**.

Can we generalize this analogy? Yes. The right abstraction often has exactly this kind of “useful fiction” feel—SAN virtualizes storage; Cloud virtualizes compute; DCOS virtualizes infrastructure.

So what virtualizes data? What virtualizes the silos?

How would you build a thing that made it appear as if the data silos didn’t exist but the data–and databases, data sources, data-providing services, and so on–in them still did? Virtualization here doesn’t just mean virtual or federated query. It means the same kind of useful fiction about data as we have with compute, storage, etc. Virtualization means the appropriate abstraction for the task and in the context.

In the context of treating enterprise data, ***all an enterprise’s data***, as an actionable asset and as a whole, the only workable abstraction is the Knowledge Graph.

## Second Trend: The Rise of Graph

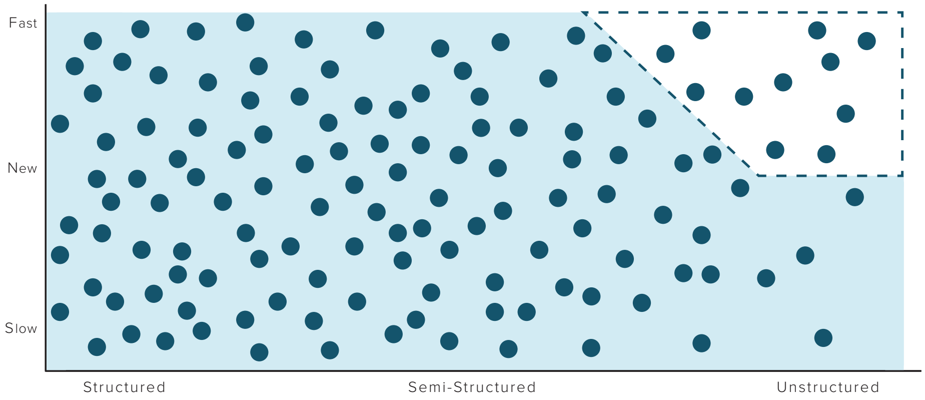
So Graph is the other trend that matters here. At the largest scale, Google, Facebook, and LinkedIn are fundamentally ***knowledge graph enterprises***. Which means at scale there’s one and only one sane way to manage—think: extract, query, analyze, monetize, re-use—enterprise data ***in full generality. The natural data model of the Enterprise is a Knowledge Graph.***

Why is “full generality” important? Because two spectrums define the enterprise data landscape:

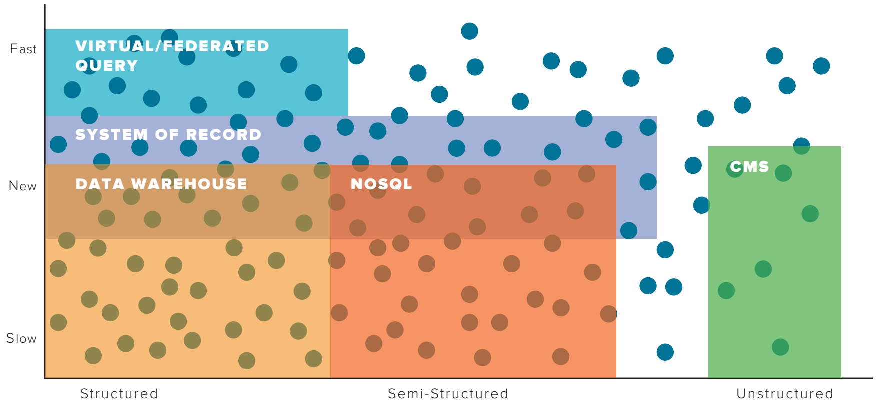
- **Data Type**: structured, semistructured, unstructured
- **Data Velocity**: fast, new, slow

## The natural data model of the Enterprise is a Knowledge Graph.

If you think of these as X and Y axes, you can plot all enterprise data as a combination of data type and data velocity. We can generalize this idea, too, to think about different kinds of enterprise data management systems. Which kind of data is some particular kind of system best equipped to handle? For which part of the enterprise data landscape is it optimized?



The value proposition of a Knowledge Graph for the enterprise is that ***all data, data sources, and databases of every type can be represented and operationalized by the Knowledge Graph***. It doesn’t count if you only handle *some* of the data. And it doesn’t help if you can only do basic, low-level, or primitive things ***with*** all the data.



When you consider the actual enterprise data landscape, it's obvious why a data warehouse or a federated query system or CMS or NoSQL store alone isn't enough. Everything in the Knowledge Graph or what's the point?

Increasingly, and for as far out as anyone can see, the world's largest enterprises need a software platform that will solve the problem of unconnected data once and for all. That platform is a Knowledge Graph.

## Putting Knowledge in Graph

Graph databases are awesome. We think they are so awesome that we built one, from scratch. But that wasn't enough to make a Knowledge Graph platform.

Graph database vendors, including Neo4j, which is the leading graph database, are busy turning relational silos into graph silos. We applaud that effort. Graph silos are often better than relational silos.

But at the end of the day, a silo is still a silo, whether it's got tables, key-value pairs, or nodes and edges inside of it. **Unconnected data sucks**.

## Plain Graph Databases

Plain graph data stores like JanusGraph, CosmosDB,and DSEGraph, which are all lovely systems, are really more like ***data structure*** servers than databases. They support graph as a data structure. They tightly couple traversal code and the graph itself, with no real abstraction between. Their primary operation is traversing that data structure, that is, a graph ***traversal*** API. Traversal is a universal, but low-level interaction pattern. Imagine giving someone directions across the country without using street or highway names or any cardinal directions.

But a data model is more than a data structure, and a database should support a data model independently of its implementation. Actual graph databases—for example, Neo4j, MarkLogic, or GraphDB—support evaluation of an actual query language, an interaction pattern that treats the graph as a data model rather than merely as a data structure.

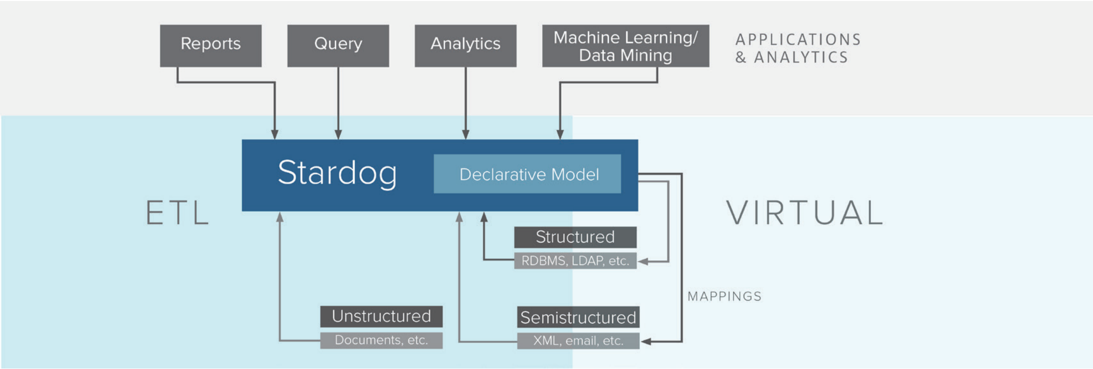
In contrast, a Knowledge Graph platform enriches and amplifies ***graph as a data structure*** and ***graph as a data model*** into something greater than the sum of its parts. And it does this by adding a ***Knowledge*** Toolkit to a ***Graph*** Database.

## Operationalized Knowledge

What distinguishes a Knowledge Graph platform from a plain old graph database? A Knowledge Graph is a Knowledge Toolkit deeply integrated with a Graph Database. By deeply integrating a Knowledge Toolkit with a Graph Database, a Knowledge Graph platform supports a much wider and deeper range of services than a plain graph database. So there is more to a Knowledge Graph platform than just a graph database.

So what's in a real Knowledge Toolkit? It starts with an expressive graph data model with open and closed world semantics; structured, semistructured, and unstructured data unification by way of virtual graphs; logical reasoning, including inference, explanation, and model checking; machine learning, including statistical inference and probabilistic reasoning; semantic search; geospatial semantics and query; and knowledge graph construction services.

Bonus points if **a Knowledge Graph platform provides these services in a declarative way that's integrated within an expressive, fast query language**.



## Norvig's Other Law

Peter Norvig said, famously, that more data beats smarter algorithms. We think that's true. And it's the other way a Knowledge Graph beats a plain graph database. In other words, **a Knowledge Graph about X knows everything about X that's worth knowing**.

A Knowledge Graph about X knows everything about X that's worth knowing.

A Knowledge Graph platform supports traversals and queries of the graph data structure and data model, respectively, too. But it adds a layer of machine understandability by supporting a richer semantics for the graph. A Knowledge Graph knows the difference between graph as data structure—there's an edge between node `A` and node `B`—and graph as something more—for example, a symmetric or reflexive or transitive property between a `Person` and an `Organization`.

But the power of going beyond data structures or models to graph as ***knowledge representation*** is further enhanced by having access to all, or even most, of the relevant data. A Knowledge Graph, unlike a plain graph, adds more to the data by turning it into knowledge, and it does that to and with ***all*** the data, which, if Norvig's Other Law is correct, creates another layer of enterprise value.

A Knowledge Graph doesn't (necessarily) destroy silos. Some should be consolidated; some should be left in place. There is no universal answer to that question. It always depends on the silos and the situation. Solutions built on the wrong abstraction inevitably dictate a fixed answer to that question. That's the wrong kind of disruption.

A Knowledge Graph, built on the right abstraction, supports the widest possible means to manage data silos. And it can do that because it can query data silos, or even just parts of data silos, in place or pull their data via ETL in any arbitrary combination that best suits business needs.

A Knowledge Graph makes it safe to proceed as if the silos don't even exist and, thus, lets the enterprise act as if there's just unified, connected, actionable knowledge.

## The Knowledge Graph Enterprise

A Knowledge Graph platform should look and act like the other large-scale virtualization systems (SAN, Cloud, DCOS, etc) as much as is possible. It should look and act as if there is one graph in which all enterprise data finds the right, operationalized place and is consumed there as such.

At least three great technology companies are built in part on a knowledge graph: Google, Facebook, and LinkedIn (and Microsoft Graph, too).

What about the other two great technology companies? Amazon appears to be working on the Amazon Product Graph. Apple's recent public embrace of AI, not to mention some Apple job listings I've read lately, suggests something is happening there, too.

What does that add up to? It looks like the five most important software companies in the world—not to mention 3 trillion of market cap—embracing Knowledge Graph ***full-on***.

Software companies are eating the world, but the fuel that software runs on is ***data***. If it's true that every company is turning into a software company, then it's also true that every company needs a Knowledge Graph to manage ***all its data as a connected whole***.

The Big Five have got that sussed already. For every other big enterprise on the planet, where the data is wild and the knowledge is scarce, there's Stardog: the Knowledge Graph platform for the enterprise.

**About Stardog Union**

Stardog is an Enterprise Knowledge Graph platform that allows customers to query massive, disparate, heterogeneous data regardless of structure with simplicity of implementation. Stardog's enterprise customers include Fortune 500 companies in finance, healthcare, life sciences, energy, media, and government.

The company is headquartered in Arlington, Virginia. Stardog is a privately held, venture-backed company.

For more information, please visit **Stardog.com**