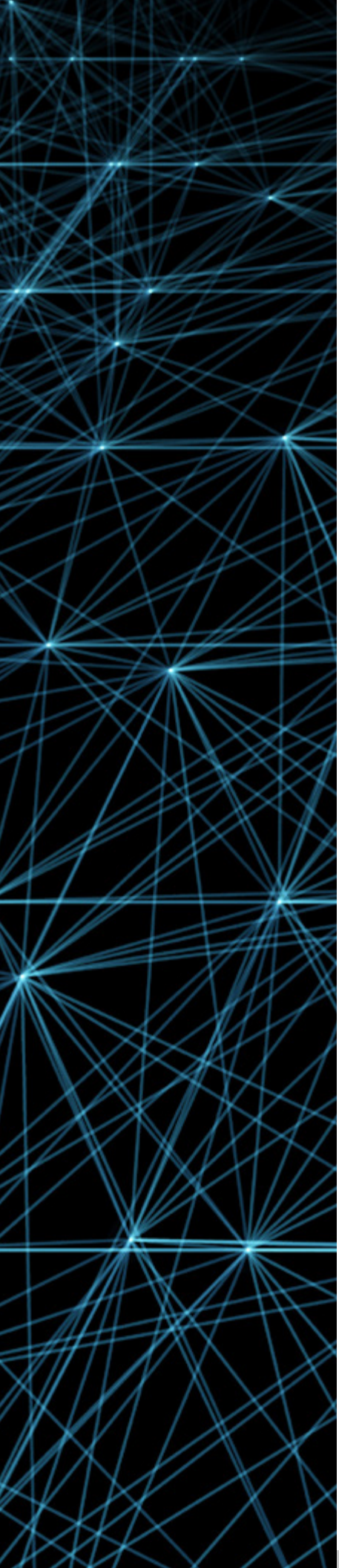
Robust, Agile, and Comprehensive:
The Story of the Data Fabric

WHITE PAPER

early adopter research

CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®



CONTENTS

early adopter research

By using semantic standards, the data integration is even more powerful as are the queries and algorithms that can be applied to the fabric.

INTRODUCTION

The data fabric is a new way to manage and integrate data that promises to unlock the power of data in ways that shatter the limits of previous generations of technology such as data warehouses and data lakes.

The data fabric is a novel construct powered by an inspired combination of technology and standards that casts the widest possible net over all the data in an organization. It is able to model and integrate data at whatever level of granularity desired.

Because it is based on a graph data model, the data fabric is able to absorb, integrate, and maintain the freshness of vast quantities of data in any number of formats. By using semantic standards, the data integration is even more powerful as are the queries and algorithms that can be applied to the fabric.

With such a wide and deep model, it is possible to integrate data and present it for use by analytics, AI, and ML systems in ways that make those techniques, which thrive on abundant, high-quality data, even more powerful. Questions that simply cannot be asked using other approaches are possible, and often easy, to answer. The implicit tax on curiosity that is levied when exploring data is difficult is dramatically lowered, and more people can mine available data and unearth powerful signals.

In addition, when implemented properly, the data fabric takes full advantage of the flexibility of the cloud or runs on-premises infrastructure as determined by the needs of the application.

The data fabric has rightly caught the attention of the analyst community and is celebrated as a major step forward supporting new efforts toward digital transformation as well as super charging existing programs for AI, ML, data science, and business intelligence. Forrester has recognized a new category of products that address this space, referring to these solutions as “big data fabrics.” According to Forrester, “Big data fabric focuses on automating the process of ingestion, curation, and integration of big data sources to enable the analytics and insights that are critical for business success. It minimizes complexity by automating processes, workflows, and pipelines, generating code automatically, and streamlining data to simplify deployment.”

Cambridge Semantics believes that Forrester's vision becomes even more powerful by implementing the data fabric using a graph data model that makes extensive use of semantic standards that allow integration based on what data means in the context of a model that explains the relationships between concepts.

Cambridge Semantics believes that Forrester's vision becomes even more powerful by implementing the data fabric with a graph data model that makes extensive use of semantic standards both to describe data and enable integration based on what data means. Such data is mapped to the language of business, and is less opaque and more understandable by end users.

Data fabrics are addressing some of the most challenging data management, integration, and analytics problems, such as making sense of vast quantities of pharmaceutical clinical trial data, manufacturing data, and data from IoT sensors.

The data fabric essentially transports the desires that led to creating the data warehouse and the data lake into the modern age. It then updates the structure and the implementation of the resulting repository and surrounding set of capabilities to deliver a way of capturing, integrating, modeling, connecting, aggregating, and analyzing data that keep pace with the needs of organizations in the modern world.

By answering the following questions, this paper seeks to provide a sophisticated description of the data fabric so readers can determine whether it fits their requirements:

- How does a data fabric create the value just described?
- How does Cambridge Semantics implement the data fabric?

A DATA FABRIC ARCHITECTURE CASTS THE WIDEST NET OVER ALL YOUR DATA

It is important to acknowledge the progress made by data warehouses and data lakes in advancing data management, data integration, and analytics. That said, the data fabric is breaking new ground, which has led to broad adoption for even the most challenging use cases.

To determine whether a data fabric is right for an organization, it's important to understand why it is more powerful than previous models.

Data in its raw form is highly variable in data quality. Sometimes it is well formed and clean. Other times it is sparse and uneven.

The Modern Problem of Data Management and Integration

Anyone seeking to make use of all of the data in an organization faces challenges that have become more acute in the modern world:

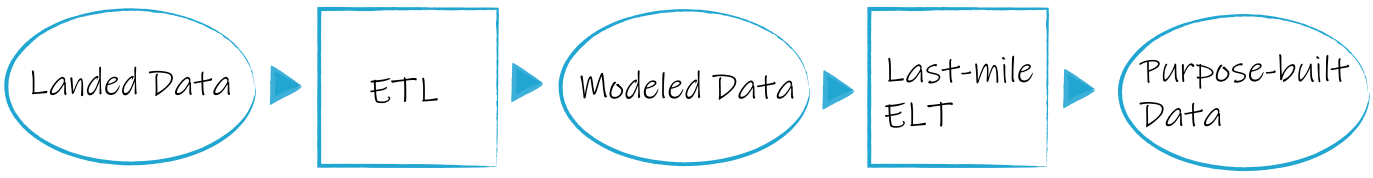
- There is more data than ever from a multitude of both structured and unstructured sources.
- Data in its raw form is highly variable in data quality. Sometimes it is well formed and clean. Other times it is sparse and uneven.
- Data comes in many different (and incompatible) formats.

Access. The first problem is getting access to all this data and bringing it into a format or representation where it can be managed and integrated.

Structure, integration, and transformation. The second problem is bringing structure to the data and integrating and transforming it into new forms. For example, data:

- Is often sparse, with missing values
- Requires consolidation at varying levels, with on-demand drill down to details
- Must support powerful queries so it can be delivered as needed to applications
- Must be able to be refreshed as often as needed (including in near real-time)
- Exists in many different unstructured and flexibly structured forms beyond relational, bringing with them conceptual complexity and literally an explosion in the number of entity types and relationships between them

The following diagram captures broad stages of integration and modeling used in most business intelligence implementations:



The data fabric combines the power of graph technology to capture and represent data in all its rich complexity with semantic standards that allow us to understand what the data means, but also to capture the data and deliver it where needed.

Comprehensive view. The third problem is keeping track of the much wider repository and all the models and transformations created to make the data useful.

Flexible implementation. Finally, it is vital to implement the infrastructure to integrate and transform the data on whatever platform best serves the application, whether that is on-premises, on one cloud, or on many clouds.

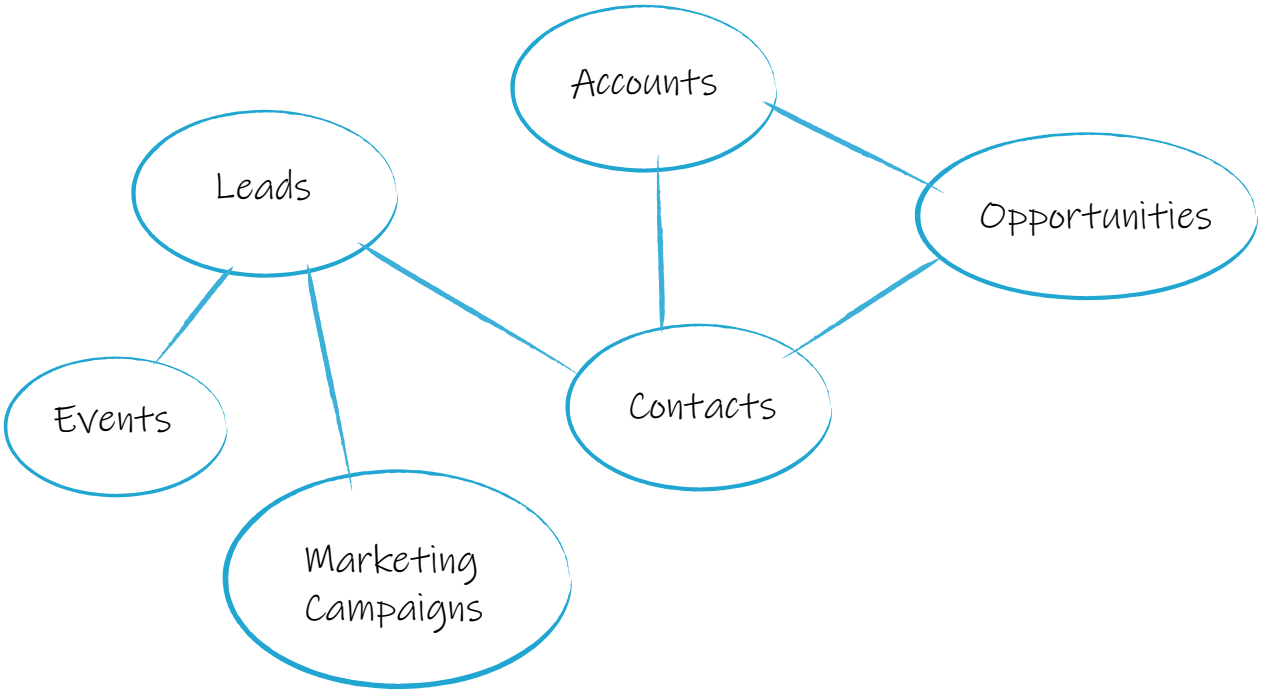
Our explanation will show how Cambridge Semantics' data fabric implementation addresses all of these issues.

How to Build a Powerful Data Fabric

Cambridge Semantics creates a data fabric from a collection of graph data models that are made more powerful using semantic standards. Here's the basic idea.

The data fabric combines the power of graph technology to capture and represent data in all its rich complexity with semantic standards that allow us to understand what the data means, but also to capture the data and deliver it where needed.

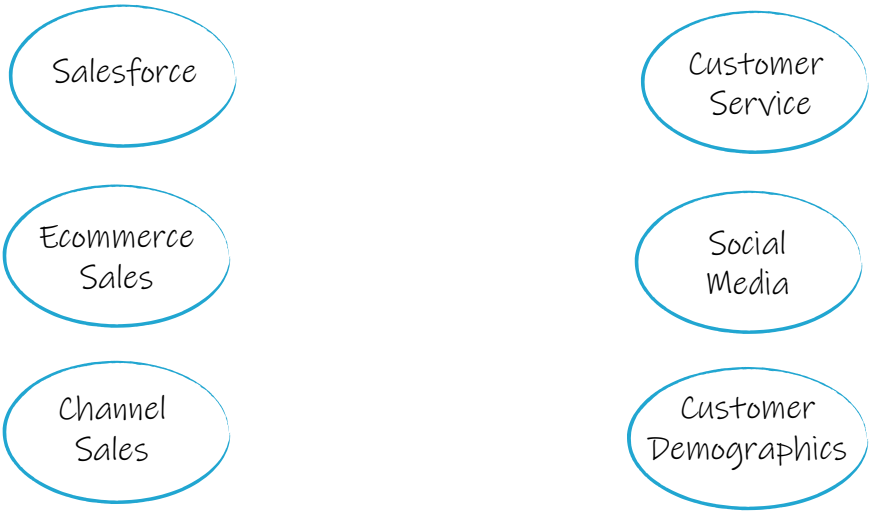
A set of tables in a relational database from an application like Salesforce might look like this:



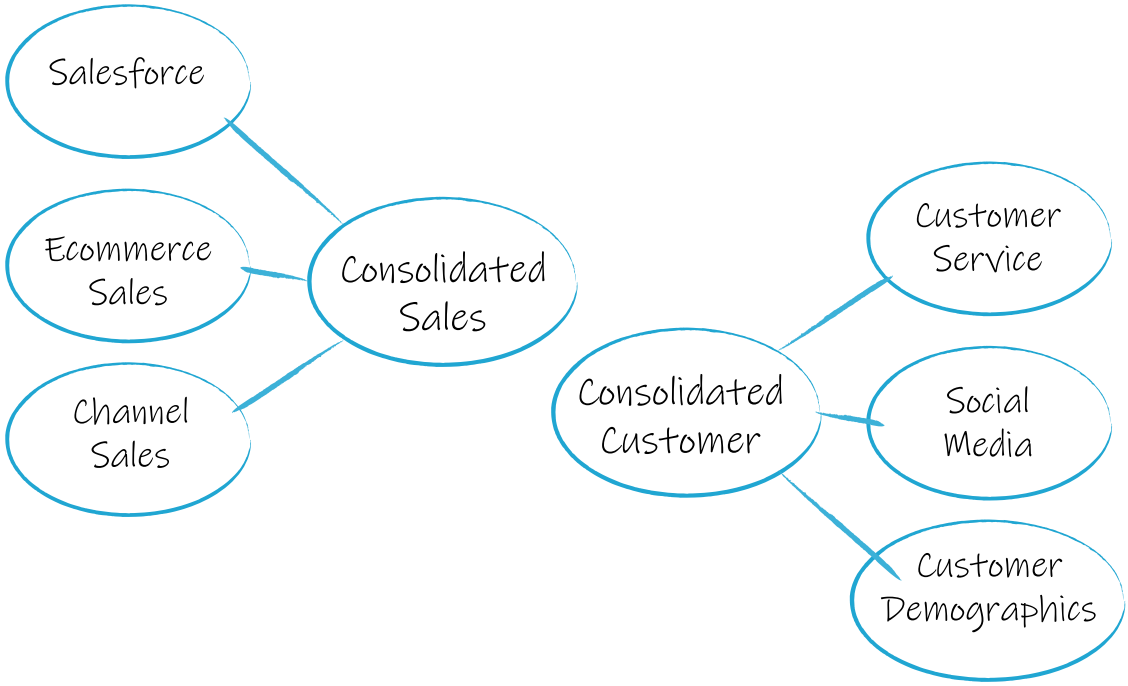
The effort needed to consolidate and integrate data can be trivial or require more substantial effort based on the state of the data and the desired end-result.

That's just one source of data in a company.

All the data related to sales and marketing might involve several sources of data, each of which could be represented as graphs:



These graphs can then be consolidated and made sense of by adding new layers of graphs that bring together information from several different graphs:



The use of semantic standards enables graph ETL and ELT to be defined and maintained without becoming nightmarishly complex, refreshing the data fabric with new and updated data.

This process can be repeated to bring all data into a data fabric so that it can be connected, queried, analyzed, and delivered in the form needed by analytics systems and applications. The effort needed to consolidate and integrate data can be trivial or require more substantial effort based on the state of the data and the desired end-result. Sometimes data doesn't need to be consolidated. Other times, many sources are combined to create a new, complete picture.

The resulting data fabric can then be analyzed at whatever scope is required, from detail level to consolidated information and any combination thereof.

This process works because graphs enhanced with semantics are powerful enough to capture and represent the complexity of today's data. Semantic standards provide a roadmap for distilling data, performing queries, running algorithms, and creating advanced analytics.

The use of semantic standards enables graph ETL and ELT to be defined and maintained without becoming nightmarishly complex, refreshing the data fabric with new and updated data.

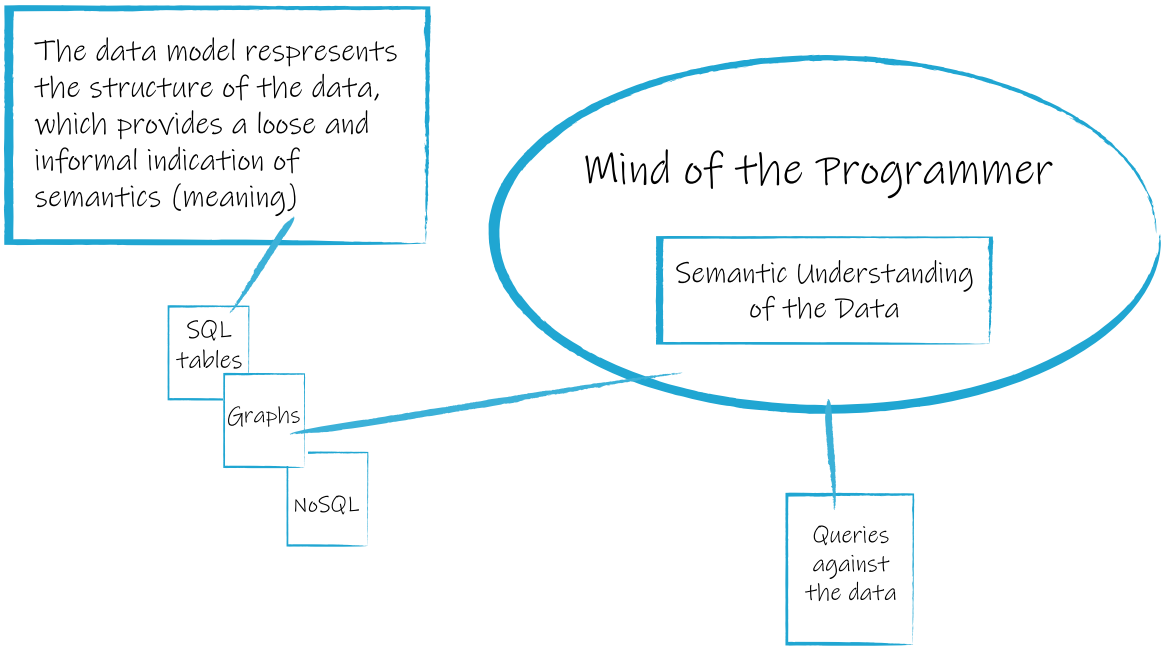
How Semantic Modeling Amplifies the Power of Graphs

A graph data structure is flexible, easy to understand and fast to query. But making data usable doesn't simply mean putting it into a particular structure. To make all the data usable, you need a layer that specifies the meaning of the data and its relationships to other data.

This ability to capture the meaning of the data is referred to as semantics. In the earlier methods of modeling we described, semantics was really in the heads of the programmers or analysts using the data. Anytime a query is written, it was as smart as the programmer or analyst could make it. The data model represents the structure of the data and this represents some part of the semantics, but not in an explicit way. There's nothing wrong with this — using semantics in an informal way has taken us a long way.

Before: The meaning is in the mind of the programmer

Where are the semantics (what the data means)? In the data itself (and the mind of the programmer)



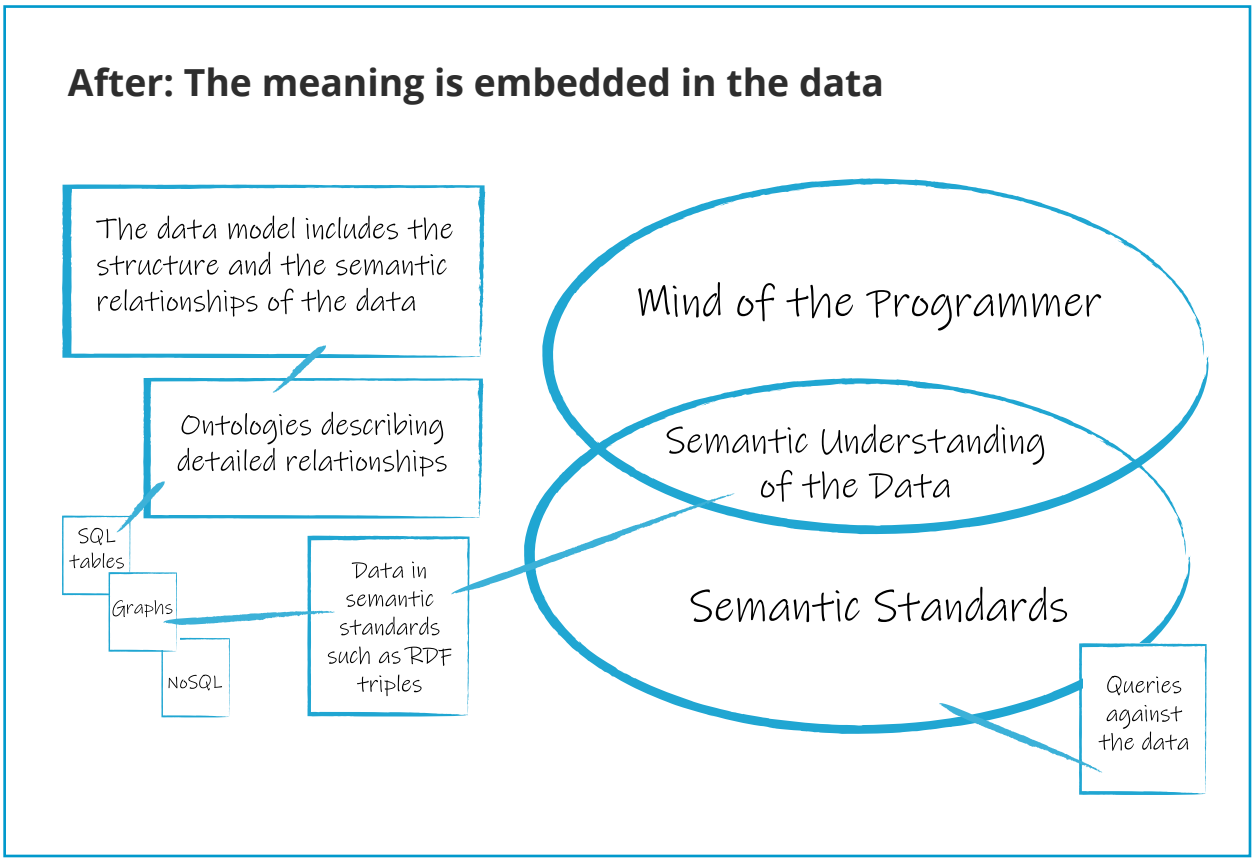
The power of data warehouses comes from the fact that the semantics of the data model create a common language. Yet these semantics are not recorded in the data model but are created by communicating about them in data dictionaries and in common ways of using the data.

If semantics are separated from the data, it restricts the ability of queries, algorithms, and analytics to use the semantics in powerful ways (such as machine learning and AI).

Analysts and programmers have the semantics in their minds and can make use of them, but automated systems cannot.

Anzo's data fabric adds semantics to the power of the graph model. The data model represents the structure of the data relationships using semantic standards such as ontologies and RDF triples. The semantic layer is a layer of metadata that adds depth and meaning to the graph overall so that queries and algorithms can use this information.

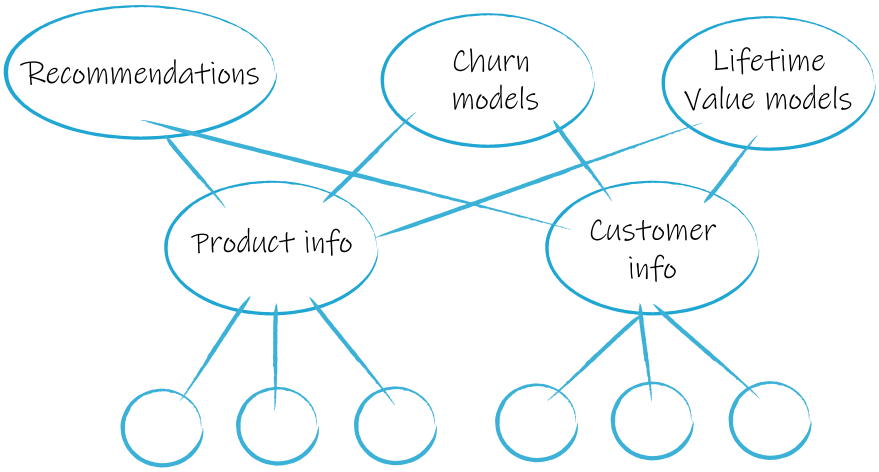
The result is a far more usable and self-explanatory graph.



Semantically powered graphs go beyond graphs that don't use semantics. The semantic information about each node allows many more connections to be made programatically and through inference. In a non-semantic graph, connections are all created explicitly.

But it turns out that eventually, a data fabric has many layers, where you have landed data (data coming in from a variety of sources), modeled data (landed data that is integrated into usable formats) and purpose-built data (designed to support specific analytics in specific applications).

Semantically powered graphs go beyond graphs that don't use semantics. The semantic information about each node allows many more connections to be made programatically and through inference. In a non-semantic graph, connections are all created explicitly.



HOW ANZO CREATES A SCALABLE DATA FABRIC ENVIRONMENT

Efficiency in building and maintaining multi-layered data fabrics from which value can be automatically extracted is vital to success. Reducing the cost of absorbing more data and presenting it in a useful form in effect reduces the curiosity tax[1] that is imposed when unlocking the power of data requires a huge amount of work.

A variety of jobs must to be done to make a data fabric sustainable for enterprise use. This section of the paper explains the needed capabilities and how Anzo handles each of them. While it is excellent to imagine a data fabric that makes all your data usable, that vision won't be realized unless there is a systematic way to:

- Create the multi-layered, semantically rich data fabric just described
- Ingest and add new data sources and evolve the data fabric over time
- Present a guided exploratory experience to end users
- Provide scalability for SQL-style queries, graph queries, algorithms, and analytics
- Deploy the data and compute where it is needed to support various applications

Enterprises don't just need a data fabric; they need a product that can build, maintain, scale, and evolve a data fabric that continues to make all data usable by both people and programs.

1. Dan Woods coined this term. Curiosity tax is high when asking questions of data is time-consuming. If questions are easily answered, curiosity tax is reduced and data is more useful.

Enterprises don't just need a data fabric; they need a product that can build, maintain, scale, and evolve a data fabric that continues to make all data usable by both people and programs.

Here is a summary of how Cambridge Semantics' Anzo implements an enterprise data fabric.

Data Engineering and Graph ETL: In a data fabric, data is mapped to semantic standards. The graph is created based on those standards.

- The data layers in the Anzo architecture provide the flexibility and power to capture complex data, perform transformations, and create a much more dynamic equivalent of data pipelines of whatever size needed for each use case and application.
- When materialized, the landed data may be mapped and transformed to combine many graphs. In this way the data integration takes place. Further combinations may create purpose-built data.
- The SPARQL language plays a key role in implementing and automating the transformations between the layers.
- All of the mappings and transformations are maintained so that when source data changes, the entire data fabric can be refreshed.
- The computing infrastructure for the data fabric that implements a particular application is built just in time on whatever computing platform makes the most sense. Remember, a data fabric is designed to capture all the data in an organization. But no application uses all the data at once. That's why the layers of modeled data and purpose-built data are so important. They are used to select and prepare data for a given application. At runtime, you can choose just the right computing platform for the needed subset of data based on cost, connectivity, and scalability concerns.

Semantic Standards in the Data Fabric: The semantics used in a data fabric are like a data model, but are vastly more powerful and flexible than the metadata and data models used in data warehouses.

- The data fabric puts the data into a semantic structure explained by the ontology, which represents key concepts and how they are related. Semantic standards take the place of a schema in a relational database system.
- Data is represented using a simple construction from an open standard called resource description framework (RDF):
  - Kathleen founded StartUpX
  - VC-firm funded StartUpX

The power of the data fabric to support virtually any use case comes from the ability to use as many layers as needed to transform and integrate the data.

› BigCompany bought StartUpX

This representation is captured as a graph.

- Storing all your data in a form that is self-describing and enriched with context is powerful. Data in this form is easier to find and to combine and recombine with data stored using the same standards. After the data has been mapped to a model, the heavy lifting is done once and for all (unlike traditional ETL where each project is a new effort). It becomes an independent, self-describing dataset that anyone can find and immediately reuse.

Layers in the Data Fabric: The power of the data fabric to support virtually any use case comes from the ability to use as many layers as needed to transform and integrate the data.

- The landed/modeled/purpose-built data construct is a useful summary of how data is divided into layers, but in practice, the number of layers is often more complex.
- There is no practical limit to the hierarchies that can be captured in this way. That's why the use of semantic standards is essential for using the full breadth, depth, and diversity of the data in an enterprise. This is the essence of an enterprise data fabric that encompasses and connects all of your data, making it usable.

Guided Exploration and Discovery: The structural and semantic information about the data makes the data fabric more powerful and more useful. The information about the form and meaning of data can be understood by users, algorithms, and programs. But the user experience doesn't require understanding of semantic standards. Rather the guided process makes use of the standards, reducing (and most of the time eliminating) the need to understand semantic structures, instead providing a series of automatically generated queries that can be tuned and adjusted by the end user.

- SPARQL, the standards-based query language used in Anzo, allows data exploration based on semantic relationships, not just structural ones. This allows for richer data models with deeper layers and subtypes, while still enabling you to move data into tabular form as needed.
- SPARQL is also a powerful platform for automating query functions that can be presented without having to write queries.

Unlike any other data fabric, Anzo allows the graph to be instantiated on-premises, in one cloud, or in multiple clouds depending on the needs of the application.

- The ontology provides a map of what everything means and how everything is connected. SPARQL allows any query, not just queries that have been optimized as in an RDBMS. This capability, coupled with automated query generation, gives users complete freedom to roam the data in any number of directions simultaneously.
- Users need a guided experience through a data fabric, as a way to explore and hone questions and in the context of applications and analytical dashboards for high-value use cases. The metadata available from the semantic standards and the layers is turned into a visual roadmap of the structure of the data that users can explore. Users get relevant answers to queries and suggestions for related information. This guided process can provide a business-relevant version of the voice-driven services we see in the consumer marketplace.

Multi-Dimensional, Just-in-Time Scalability: It doesn't make sense to go all in on an enterprise data fabric if you can't have breadth (many graphs) and depth (layers from landed data to purpose-built). And queries that are extremely complex need to be able to run across this entire landscape to answer any question on the fly.

- A scalable enterprise data fabric platform must provide a lightning-fast response to complex queries. In this way, the data fabric can deliver large amounts of data of the sort associated with OLAP queries, and also scalably execute graph queries, analytics, and algorithms that answer a new domain of questions.
- It must manage a large number of transformations and ontologies so that they can adapt and evolve as needed.
- Users need a platform like Anzo that can operationalize this entire process. It must be possible to create, update, and maintain many graphs in production without ongoing care and feeding by specialists like DBAs. The only way to scale is to allow a vast number of users to access and query the graphs. The Anzo platform and the powerful AnzoGraph virtual data warehouse engine provides this scalable functionality, while most data fabric platforms do not.
- Unlike any other data fabric, Anzo allows the graph to be instantiated on-premises, in one cloud, or in multiple clouds depending on the needs of the application. Data is materialized on a just-in-time basis.

THE ENTERPRISE DATA FABRIC IS INEVITABLE

The challenge we've seen with data warehouses, data discovery tools, and data lakes is that they separate the semantics from the data. Additionally, the power of the queries is limited and it becomes harder to use data in multiple ways.

Raw graph databases put semantics in the transformation programs rather than in a generic form — just like NoSQL databases. They also limit users to graph-like queries and niche use cases. To truly capture the power of the data fabric, you need a platform that provides all the functionality of AnzoGraph to create a data fabric that casts a net over the entire enterprise data model.

AnzoGraph OLAP database and the surrounding components of the Cambridge Semantics portfolio create a data fabric that fully supports the growing needs of the enterprise.

Too often, companies face a difficult choice when adopting open source platforms and open standards. Such decisions require an investment in time and resources to design, maintain, and evolve these architectures.

Anzo offers both fast time-to-value and breadth of functionality while remaining open.

Data is exploding. An enterprise data fabric makes all data usable, supporting the way we think and work. Anzo can answer any question, with data that can be continuously refreshed and augmented to accommodate more information.

Digital transformation depends on using all your data. The enterprise data fabric offers a clear path forward to a data management, transformation, and analysis platform that can keep up with the complexity and scale of the data present in a modern company

The desires that led companies to spend so much time and money on data warehouses and data lakes can now be met in a far more complete and affordable fashion by an enterprise data fabric. To find how how to make this powerful idea work for your company, visit **www.CambridgeSemantics.com**.

This paper was written by Early Adopter Research and sponsored by Cambridge Semantics. Learn more about **Cambridge Semantics** or request a **demo**.

Connect with us

early adopter research