Knowledge Graphs:
A Key Technology Enabling Efficient Information Exchange at Scale

Government agencies have long recognized the need to share and consume information to provide effective services. However, information sharing has revealed a new challenge: efficiently understanding data from many different sources. Organizations must process disparate data efficiently and effectively to provide accurate and timely information to their customers and partners. However, a recent Dresner Advisory survey reported that over 90% government sector workers ranked *data integration* as “critical” or “very important” for their success – the highest ranking of all industries in the survey.(1)

Unfortunately, a separate survey by the Government Business Council reported that only 6% of respondents are highly satisfied with the data access their organization provides them to support the decision-making process. 57% agreed that the top reason for poor data access is inadequate systems and technology.(2) A new, modern approach to data discovery and data integration is required.

This paper briefly explains why traditional data management solutions such as data warehousing and ETL, and even newer technologies such data lakes, are insufficient to keep up with today’s fast-growing informational needs.

This brief will inform you what you should look for when evaluating data discovery and integration technologies, while avoiding solutions that require you to “rip and replace” existing systems or implement “black box” solutions that result in costly vendor lock-in.

You will also discover how a semantic and graph-powered enterprise data platform is uniquely capable of integrating, blending and presenting your existing disparate data sources – structured and unstructured, internal and external – in ways other solutions cannot match.For many government and defense entities, knowledge workers remain one or more degrees of aration from the many sources of data they need to make key, time-sensitive decisions.

© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.

In a commonly occurring scenario, a knowledge worker may have to submit a request for information (RFI) to a centralized reporting and analysis center. The data analysts receiving the RFI must canvas the enterprise to determine what relevant data sources are available. Then, analysts catalog applicable sources. Next, they submit their queries to the sources, and retrieve the results — typically copying the results to spreadsheets. Analysts compile query results from each source and then develop an aggregation of the results into a large spreadsheet. They send the aggregation, along with data from each source, to the knowledge worker who only then can begin to interpret the compiled data.

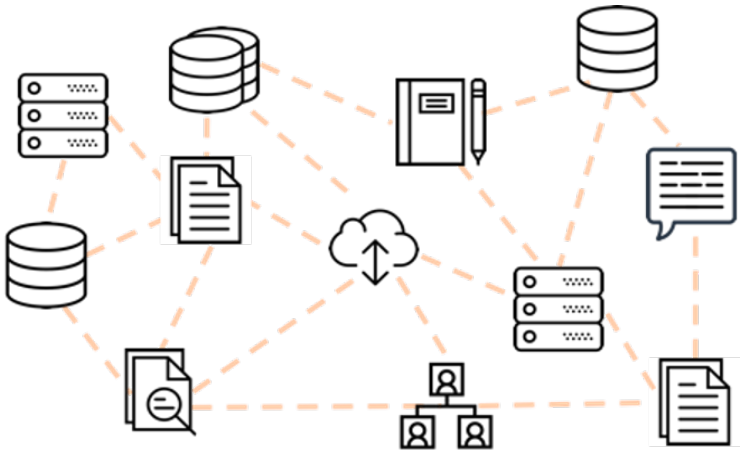
Analysts must search multiple sources, aggregate query results and then synthesize the data in a manually intensive, transactional workflow.

This process is manually intensive, error prone and time consuming. In practice, the time from RFI submission to response is on the order of weeks.

For years, implementing data warehouses has been touted as the solution to this problem. A data warehouse is commonly defined as central repositories of integrated, processed data from one or more disparate sources, containing current and historical data that is used for creating analytical reports for operational users throughout the enterprise.

The data warehouse approach experienced some success, but also had three main issues:

© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.



SIGNIFICANT UPFRONT COSTS.

Because the data warehouse provides an enterprise view into structured data to answer a specific set of use cases based on operational need, the data must be integrated from multiple systems into a single repository, which leads to significant upfront costs:

- The process to copy, cleanse, and enhance data from systems to the data warehouse, referred to as extract-transform-load (ETL) or some variant thereof, is expensive to develop
- The "boil the ocean" enterprise approach requires significant collaboration from subject matter experts (SMEs) across various business functions.

INCREASING TOTAL COST OF OWNERSHIP (TCO) OVER TIME.

ETL processes are typically defined in a vendor-proprietary format, with increasing maintenance costs due to vendor lock-in. Additionally, the static data warehouse schema is based on the current organizational mission rules and data models. When these rules or models change, or the operational use cases change, or when the data changes, the warehouse requires extensive changes, including remodeling and reloading.

LACK OF DATA DISCOVERY

The use cases that a data warehouse can address are limited to the model and business rules already integrated into the model. By default, only data that conforms to the model is integrated; therefore, the warehouse only addresses use cases supported by the model. If something is unknown to the model, analysts cannot use the data warehouse for discovery.

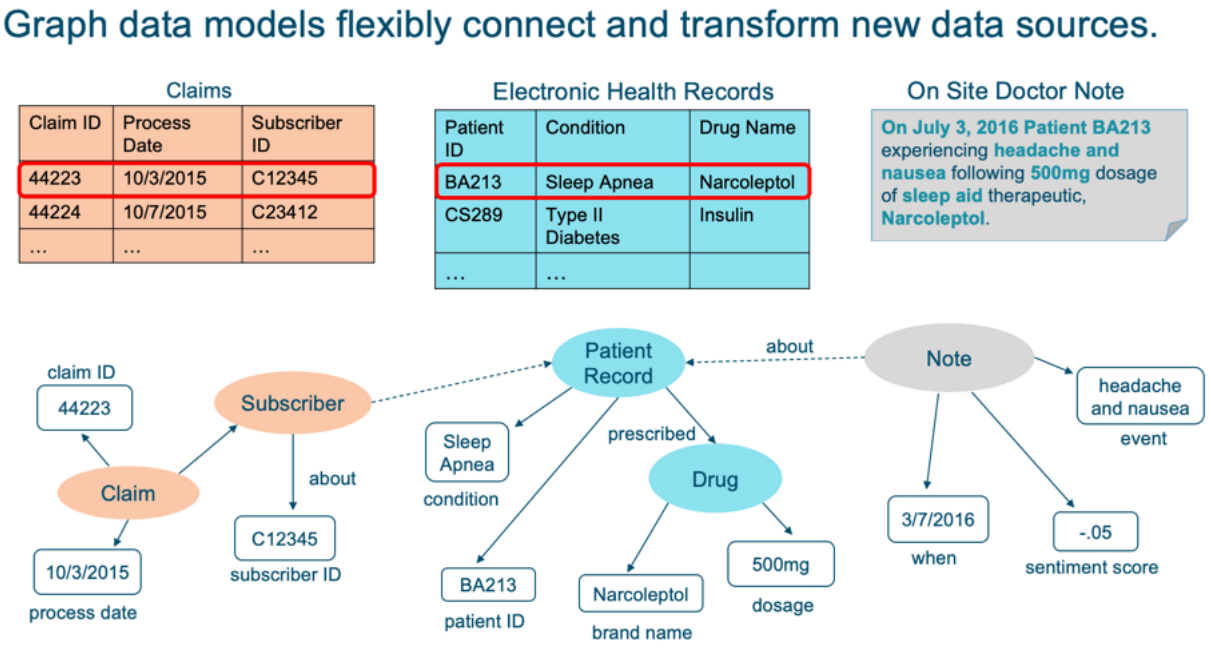
© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.

The solution to reduce the decision cycle is a new, modern data discovery and integration platform that enables an enterprise *data fabric*.

The data fabric normalizes the different syntaxes and structures of enterprise data sources into a graph model. Then, the data fabric harmonizes the concepts and their important relationships into a semantic layer. This process creates a cross-enterprise view of integrated information. In other words, the "system" aggregates and synthesizes information into analytics-ready data assets, accessible on demand by knowledge workers.

A **semantic layer** is a human-readable representation of data that uniquely identifies and connects data with common business terms that helps end users access data autonomously, securely and confidently.

Using such a semantics approach to represent any data asset and its relationships to other data assets enables the creation of a connected knowledge graph capable of spanning organizational context (see example below).

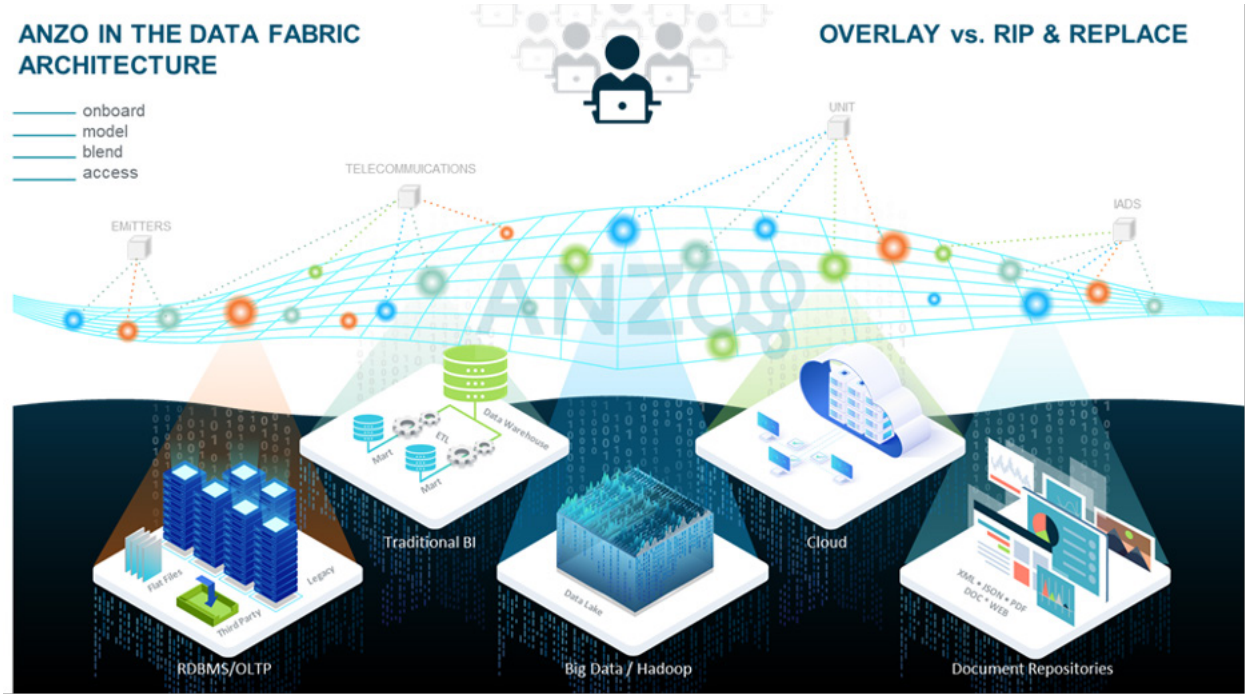


With a data fabric, instead of knowledge workers submitting RFIs, relying on human analysts to discover, access, aggregate and summarize data, the knowledge worker is empowered to ask questions on demand. The decision cycle shrinks from weeks to "now."

© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.

Meanwhile analysts shift from transactional data preparation tasks to higher order cognitive tasks to ensure the data fabric is complete, current and relevant.

At enterprise scope and scale, the data fabric is not “rip and replace;” instead, it is an overlay to existing data management investments:



Anzo is a scalable knowledge graph platform that simplifies and accelerates the integration, modeling, and blending of data to support rich insights and advanced analytics.

Map And Explore Enterprise Data. Anzo maps enterprise data to document its location, content, and contextual business meaning; exposes connections between datasets; and enables rapid visual data exploration and discovery.

Build Blended Analytic-Ready Datasets. With Anzo data scientists and other data consumers in the business build blended analytic-ready data sets by iteratively cleansing, transforming, aligning and linking data from multiple previously disconnected enterprise data platforms.

Enterprise-Ready Data Management. Anzo includes a robust set of enterprise-scale governance capabilities, making it easier and faster to ensure that all data is completely protected and secured from data ingest through delivery.

© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.

Anzo is built on a high-performance graph database engine that uses an in-memory MPP processing model to execute queries against datasets extremely quickly, enabling agile data integration, transformation, and visualization at enterprise scale. Anzo leverages graph standards including W3C's RDF, OWL, SKOS, and SHACL to build knowledge graphs of metadata and data which can be powerfully explored, transformed and analyzed, while also ensuring open data interoperability and easy integration with other systems. Anzo is an open overlay platform that allows users to build knowledge graphs against the underlying data resources without displacing or disrupting any existing processes or platforms. Anzo integrates with enterprise metadata, governance, security controls and policies, and includes a RESTful API for lights-out integration into other processes.

Cambridge Semantics Inc.

Cambridge Semantics Inc. is a modern data management and enterprise analytics software company. Our solutions transform siloed data into enterprise-scale knowledge graphs, revealing previously hidden insights, fueling pervasive analytics, and making previously unanswerable questions answerable.

Cambridge Semantics solution Anzo® is a scalable knowledge graph platform for modern data integration and analytics. Anzo dramatically simplifies and accelerates the integration, modeling, and blending of siloed data into insight-rich knowledge graphs at enterprise scale. Anzo is built on AnzoGraph®, the fastest & most scalable knowledge graph engine supporting data integration, graph algorithms, data warehouse-style analytics, feature engineering for Machine Learning, and more. The company delivers solutions that enable IT departments and business users across Life Sciences, Financial Services, Government, Manufacturing, and other industries to accelerate data delivery and provide meaningful insights across the organization at hyper-speed and scale.

Learn more at www.cambridgesemantics.com or contact us at info@cambridgesemantics.com

© Cambridge Semantics Inc. All rights reserved. Unauthorized reproduction is prohibited.