The Bloor Group

The Data Fabric Journey: Cutting to the Chase with Anzo

How Cambridge Semantics' Anzo Deals with Dark Data

Eric Kavanagh &
Robin Bloor Ph. D.

WHITE PAPER

The legendary Johnny Cash once sang of assembling a first-class automobile one piece at a time. He fantasized about working at General Motors, incrementally sneaking away the nuts and bolts to create a masterpiece of machinery. In true country-rock form, he immortalized the concept of building the perfect beast, all while cleverly circumventing rules of the road.

Oh, the best-laid plans! Somewhere along the way, life threw a few wrenches into the works. The transmission dated its motor by two decades, creating an impedance mismatch of epic proportions. The bolt holes disappeared at some point, requiring a loose coupling of critical components. Even the headlight architecture changed, altering visibility in fundamental ways.

Complicating matters further, to quote the Cash man himself: "The back end looked kinda funny too," apparently referencing the mainframe that still holds an image of the original fishtail design. Of course, in the fairytale land of song, everything worked out just fine: the wife came out, saw the car, and said, "Honey, take me for a spin."

Fabric of Success

If we use our imagination, we can see a fascinating analogy here to the modern information management industry, especially for large organizations with a long tail of legacy systems. Just like Cash's piecemeal Cadillac, many companies now manage an array of information systems that span entire eras of development in the world of data.

To wit: mainframes still run many of the Fortune 2000. Quality COBOL coders are practically worth their weight in bitcoin. The remnants of Service-Oriented Architectures still dot the landscape. Netezza (and other appliances) still roam the random forests of data centers everywhere. How many shades of virtualization? Oh, and those zombie instances of Hadoop!

... the new solution for gleaning valuable insights from data really needs to come into focus soon... very soon.

The stark reality on the ground (and in the cloud(s)) for most companies these days? The complexity of their information architectures demands a new approach, just as Internet giants are disrupting (read: threatening) practically every industry. In other words, the new solution for gleaning valuable insights from data really needs to come into focus soon... very soon.

The Warehouse Era

In the past, dating back to the 1990s or so, many companies went the warehouse route. Data was pulled out of operational systems like Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) via the tried-and-true process of Extract-Transform-Load (ETL). Once in a warehouse, the data could be sliced and diced to glean insights.

That model worked well for many years, and was dominated by the likes of Teradata, a company that still holds enough market share to make any Hadoop vendor (well, either one, really) green with envy. IBM and Oracle and even SAP still run data warehouses on-prem for a wide range of clients, and those systems fuel targeted insights right this very moment.

The data warehouse model has now even extended into the cloud with upstart solutions like Snowflake, or Amazon's RedShift, among other options. Clever moves like separating compute and storage have made cloud warehousing more affordable; while an array of advances in automation and documentation have enabled rapid-fire building of warehouses. But the warehouse model still focuses heavily on structured data, the kind that fits neatly within a relational array of two-dimensional tables.

That kind of analytics can still generate significant value, but only through a certain type of lens. In today's wildly heterogeneous world of big data and beyond, a better solution is needed to enable the next generation of business analysts.

In today's wildly heterogeneous world of big data and beyond, a better solution is needed to enable the next generation of business analysts.

Data Lakes and Middle Earth

With the onset of so-called Big Data, savvy information warriors realized their weaponry was seriously obsolete. There was just no way to ETL those massive sets of data into a warehouse. And besides, JSON doesn't really jibe with the almighty Structured Query Language (SQL).

Oh, sure, the data lake brochure said "schema on read," but what it failed to mention was the labyrinthian file system. Turns out, finding what you wanted, without knowledge of the lake's topography, took a little more effort than cobbling together your garden-variety SQL statement. And then, after several years of much-ballyhooed NoSQL, the worm turned, and suddenly, all the cool kids were talking about SQL on Hadoop. The problem was, largely, that SQL on Hadoop was, and still is, only a shadow of a real SQL engine; and as a result, we were back to the old world of workarounds and band-aids.

All in all, the data lake movement has made clear that the consolidation of data, in and of itself, will not solve critical business needs. The lacking component has been the appropriate array of lenses through which to view and better understand that data. While the data visualization vendors went a long way to giving the right kind of visibility into enterprise information, there's something to be said for the middle ground: that space between where data lives (databases, streams, files), and the point at which it's viewed (data viz).

All in all, the data lake movement has made clear that the consolidation of data, in and of itself, will not solve critical business needs

Tolkien may well have had this world in mind (metaphorically, at least), when he devised Middle Earth. It's a vast, unwieldy place that arguably occupies a full 80% of the overall data pro's landscape. There are many languages to be understood, concepts to be mastered.

Key point here? The Middle Earth of data won't be conquered with a single tool, not even the Ring of Power! A whole collection of technologies must be marshaled strategically. The good news? That can now be done thanks to Anzo, from Cambridge Semantics.

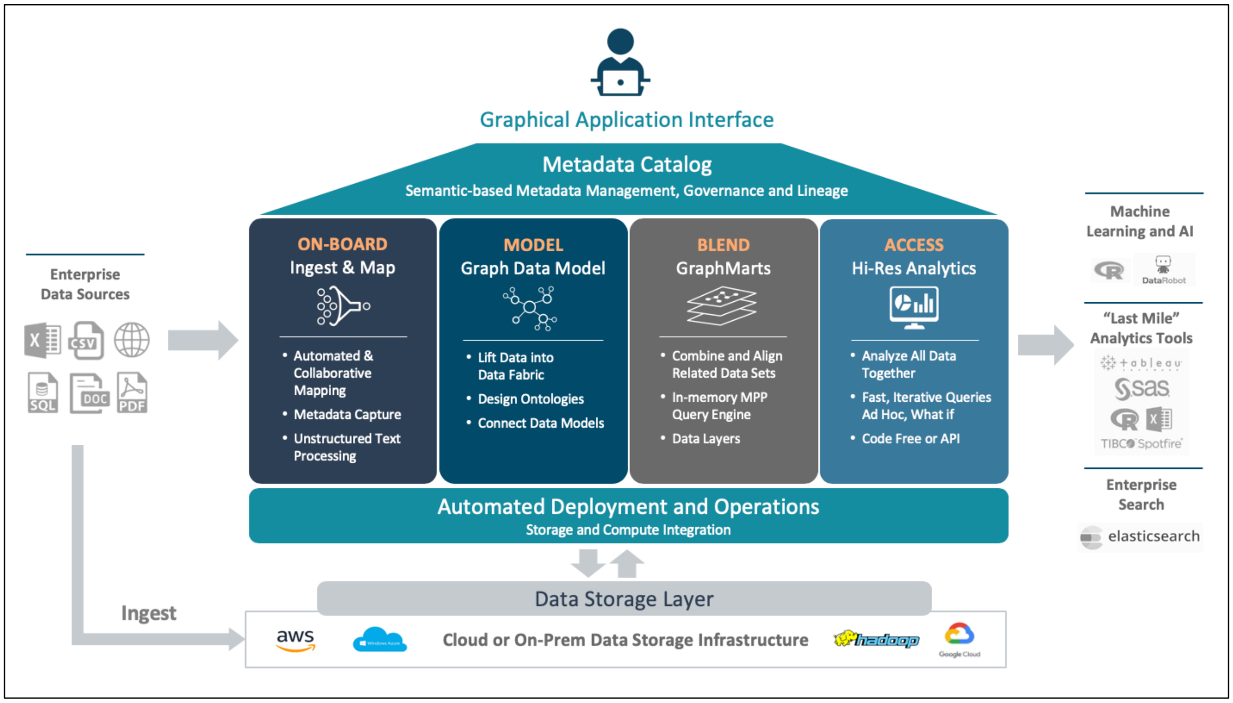
Graphical Application Interface
Metadata Catalog
Semantic-based Metadata Management, Governance and Lineage
Enterprise Data Sources
X
CSV
SQL
DOC
PDF
ON-BOARD
Ingest & Map
• Automated & Collaborative Mapping
• Metadata Capture
• Unstructured Text Processing
MODEL
Graph Data Model
• Lift Data into Data Fabric
• Design Ontologies
• Connect Data Models
BLEND
GraphMarts
• Combine and Align Related Data Sets
• In-memory MPP Query Engine
• Data Layers
ACCESS
Hi-Res Analytics
• Analyze All Data Together
• Fast, Iterative Queries Ad Hoc, What if
• Code Free or API
Machine Learning and AI
DataRobot
"Last Mile" Analytics Tools
tableau
sas
TIBCO Spotfire
Enterprise Search
elasticsearch
Automated Deployment and Operations
Storage and Compute Integration
Ingest
Data Storage Layer
aws
Windows Azure
Cloud or On-Prem Data Storage Infrastructure
hadoop
Google Cloud

Figure 1. Anzo Data Discovery and Integration Platform

Anzo: The Big Picture

The purpose of Anzo, pure and simple, is to make it easy to blend any mutli-sourced collection of enterprise data into a ready-to-roll data products for ML, AI and even BI to have their wicked way with. As such, Anzo is a powerful dyed-in-the-wool semantic engine crafted for data discovery and data integration.

The beauty of semantic technology is that it, and it alone, can store and represent any kind of data, from an itty-bitty data item to blockbuster Marvel movie. It does so using the RDF standard (Resource Description Framework), brilliantly conceived by the World Wide Web Consortium and implemented from the get-go by Cambridge Semantics to provide a fully functional semantic layer over data.

Welcome Aboard!

Figure 1 above provides a clear and comprehensive overview of Anzo. As indicated at the bottom of the diagram, Anzo is a platform that sits on top of the actual data, which may be ingested from anywhere and reside anywhere: in the cloud, in on-prem databases or Hadoop data lakes—and it may arrive in streams or in batches.

> ***The beauty of semantic technology is that it, and it alone, can store and represent any kind of data***

On the first part of the Anzo journey, analysts will onboard data. This process can incorporate and untangle just about any kind of data imaginable. Yes, to warehouse data; yes to .csv files; yes to data marts and Excel spreadsheets; yes, even, to PDFs, word documents; yes to any dark unstructured data you can shake a fist at.

Anzo captures both metadata and sample data from every source, and maps it on a master graph. This is the cohesive substrate of the data fabric that Anzo will construct, depicted in *Figure 1* by the Model activity. It is (or can be) a whole universe of data, modeled and mapped from a multitude of diverse sources. It embodies the nodes and edges which will form the nuts and bolts of the finished data products which may be constructed over time and become fodder for ML and AI algorithms.

When you put the On Board and Model components together you get the capability to build real-time data pipelines and, you can do this dynamically.

Blend and Access: Creating Finished Products

The finished products that Anzo creates are data marts—Cambridge Semantics calls them graph marts, which is perhaps a more accurate term, given that they are defined by slices of the graph data model that Anzo builds. Users can add data layers to these graph marts to carry out structural enhancement or maintenance (data cleansing, transformation, relationship linking and semantic model adjustment) and to define access control.

The Blend component embodies AnzoGraph, which enables the maintenance and querying of highly scalable in-memory Knowledge Graphs. This engine performs concurrent complex ad-hoc or OLAP interactive or batch queries across the whole data resource, achieving bewildering performance at very large scale.

Anzo provides a highly versatile semantic layer across large and diverse collections of data, providing users with on-demand access to explore and analyze the data at in-memory speeds.

The Access component enables users, independent of IT, to explore and analyze all the data that Anzo has mapped—using there own favorite tools if they like through OData and SQL JDBC/ODBC access, as well as Anzo's Hi-Res graph aware dashboard builder. The capability, made for business users, analysts and even citizen data scientists, allows users to pose questions of complex data structures and receive real-time answers. It is a lightning fast intuitive interactive capability. Bringing it all together Anzo provides a highly versatile semantic layer across large and diverse collections of data, providing users with on-demand access to explore and analyze the data at in-memory speeds.

Graph Technology and the Data Fabric

The term data fabric refers to the whole of the corporate data universe, no matter where it resides and how it is manipulated. The fundamental technology question Cambridge Semantics resolves with Anzo is how to map this extensive resource so that it can be used in a fully integrated manner. The solution it brings is founded on graph technology.

You are probably familiar with the usefulness of graph technology when it comes to mapping relationships between populations of people or the relationships between nodes of large networks. You may be less familiar with the existence of knowledge-graphs where queries that filter and traverse the edges and nodes of graphically stored data can furnish and even discover knowledge.

Cambridge Semantics employs graph technology in an even more sophisticated way to enable data integration and harmonization. Key to this is the ability to apply complex JOIN/FILTER style aggregate queries to very large scale OLAP data. It's a highly efficient capability that addresses data integration problems that were previously impossible, because of the processing time and resources required. In short, this capability empowers organizations to carry out data integrations in situations where it was not previously viable to bring data together, never mind maintaining or extending the data resource it created. Additionally the knowledge graphs that Anzo creates will offer far richer and more detailed data models than was previously practical.

Cambridge Semantics employs graph technology in an even more sophisticated way to enable data integration and harmonization.

The goal of the Anzo platform is not to boil the data ocean and accurately chart its every trough and trench. It is to provide a mapping capability for the corporate data resource that can create rich data "products" that that can reap dividends, particularly to analytics users.

Sean Martin, founder and CTO of Cambridge Semantics, comments: "We see the corporate data fabric as embodying an architecture that has distinct layers, which combine together. We think of Anzo as contributing to rather than defining the entire architecture. In some areas it can create data products that deliver real and unique value."

"We provide an overlay on the existing data resource that allows you to describe all the structured and non-structured data sources, so that they become elements of a massive graph. No matter what the form of any data store: relational database, a corpus of documents, a Hadoop file system or data lake, there are mappings that can lift the data and describe it as a graph."

"You can build a data product that focuses on a given problem by picking the data you want from wherever it is: data from documents, or databases or flat files or data streams. Some of this with some of that and that, and so on. You add in the graph-based pipelines, for transforming, cleaning, linking, exposing the right kind of views. The product you create this way allows you to focus precisely on a problem you wish to address."

"You can then put this in the cloud, so that it is pay-as-you-go and unconstrained by resources. A single user can form their own dedicated instances with multiple subgraphs that are instantiated on the fly, available while needed and which evaporate when you have finished with them. This is an ephemeral capability; you create what you need when you need it.

You do not need to, and should not put the entire knowledge graph into the cloud."

The Bottom Line

Large organizations do not need to be told that their extensive data universe is fragmented, poorly integrated and drenched in complexity. Neither do they require schooling in the value that could be derived from that data fabric if they could assemble, explore and analyze organized subsets of that data. What Cambridge Semantics brings to the party with Anzo, is the ability to do just that, in a scalable and practical way.

Anzo maps the whole data resource, permits the unconstrained creation of "data products" and provides data scientists and analysts with a unique data service. It is a remarkable capability; one that large organizations would do well to investigate.

About The Bloor Group

The Bloor Group is a consulting, research and technology analysis firm that focuses on open research and the use of modern media to gather knowledge and disseminate it to IT users. Visit both www.TheBloorGroup.com and www.InsideAnalysis.com for more information.

Austin, TX 78720 | 512-524-3689