ATSCALE

WHITEPAPER

# THE SEMANTIC LAYER: WHY IT'S CRITICAL TO ANALYTICS SUCCESS AND ALWAYS HAS BEEN

Andrew Brust, Founder & CEO of
Blue Badge Insights

**Andrew Brust** is Founder/CEO of Blue Badge Insights, providing strategy and advisory services to data, analytics, BI and AI companies, as well as their partners and customers. Andrew covers the data and analytics world for VentureBeat and The New Stack, and is a lead analyst for GigaOm in that same space. He also co-chairs the Visual Studio Live! series of developer conferences, is a Microsoft Regional Director and Data Platform MVP, an entrepreneur and consulting veteran.

# The Semantic Layer: Why It’s Critical to Analytics Success and Always Has Been

While the term “semantic layer” has been trending in popularity within data and analytics circles, it is hardly a new term and certainly not a new concept. The semantic layer may seem like a new solution particularly necessary in the modern world of cloud-scale analytics, it’s really a solution to a persistent challenge: giving individual business domains control over data without leading to chaos.

Before we dive into what semantic models are, it’s important to understand the need for them and what led to their development. In this whitepaper, we’ll discuss the evolution of data analytics architectures. where the future is headed, and the impact this has on an organization’s analytics success.

More specifically, we’ll cover:

- The early days of analytics with OLAP systems
- Why there was a shift to data lakes in the Big Data era
- The revival of data warehouses in the cloud
- The need for semantic modeling
- The rise of metrics stores
- The importance of a semantic layer
- Drawbacks of a pure data mesh operating model
- Delivering actionable insights at scale

# The Semantic Layer: Why It’s Critical to Analytics Success and Always Has Been

In the early days of analytics and business intelligence, most data was modeled in a rigid and structured manner. There were specialized backend databases for analytics that were called Online Analytical Processing (OLAP) systems. This differed from the operational databases, which were called Online Transaction Processing (OLTP) databases.

These engines would optimize analytics queries by understanding the data model and the kind of queries to expect, then organize both the low-level data and pre-calculated data aggregates into “cubes,” enabling the calculation work to happen in advance, before users went to query it. This enabled the systems to service those queries faster than most alternatives at the time, such as conventional data warehouses, OLTP databases, and, of course, spreadsheets.

While this OLAP approach worked well for small domains of analysis and small volumes of data, it was expensive, complex, and required rarefied skill sets. This limited organizations to answering only the "known unknowns": discovering answers to predefined questions. There was still a high barrier to asking new questions and bringing in additional data to answer those questions.

## Then Came the Data Lake

To overcome the limits of the traditional OLAP approach, many organizations shifted to large-scale monolithic data lakes, based on "big data" technologies starting with Hadoop. This approach solved the data volume limits of the past, enabling companies to be much more inclusive with bringing in new datasets and avoid throwing data away. Data storage was also simplified because everything was disaggregated – meaning the data was largely stored in its raw format rather than reorganized and pre-aggregated.

The ability to store and access large quantities of raw data enabled enterprises to perform exploratory analytics – answering questions based on "unknown unknowns" – and do so at scale, because there was so much data available in its original form. Using "schema on read" and applying structure to the data only at query time was much more flexible than pre-modeled data. But it still had its drawbacks.

The primary challenge with the monolithic data lake approach was lack of support for domain-specific analysis, which was the focus of traditional OLAP in the past. While data lakes enabled exploratory analytics and answering unknown unknowns, they made still-important production analytics and answering known unknowns harder and slower. While the "schema on read" approach eliminated pre-built models, it essentially required models to be created for each analysis at query time, which slowed time-to-insight and created new inefficiencies. That meant domain-specific analysis was limited to a small subset of specialists rather than increasing data analytics adoption across the organization.

In short, though this approach had more potential because it eliminated data volume constraints, it ultimately had usability and productivity limitations for business users. This led to a revival of more structured data architectures, as we'll see in the next section.

## What's Old is New Again

Since unstructured data lakes didn't fully meet the needs of production data analytics for business users, many organizations looked to data architectures of the past. While data warehouses pre-dated data lakes, and even business intelligence, they've made a comeback in the cloud with solutions like Amazon Redshift, Snowflake and Google BigQuery. Enterprises recognized that data warehouses had useful constructs and query optimization capabilities that were critical for domain-specific analytics.

Another trend is the data lakehouse model, which brings data warehouse query optimization, parallel processing, and other performance optimization techniques to data lakes. This hybrid technology aims to bring back the benefits of the traditional data warehouse in today's Big Data era. Worth noting, however, is that it's still missing one crucial piece of the puzzle: data modeling.

The lack of predefined models has a number of drawbacks:

- **Significant duplication of effort** — Data consumers need to define metrics and dimensions during each analysis in their analysis and BI tools.
- **Greater difficulty doing routine analysis** — The "known unknowns," or domain-specific analysis, require clear business models.
- **Drill-down analysis is much harder** — Hierarchies are constructed at query time rather than beforehand.
- **There's often a mismatch between BI tools and backend data platforms** — BI tools build their own data models locally, creating an extra layer.

## The Need for Semantic Modeling

The missing piece since traditional OLAP has been semantic models – to provide context and meaning around data. Since metrics drive business success, modeling these is the bare minimum necessary for data-driven operations. However, it's also critical to model the dimensions and hierarchies to drill down into this data with the right context.

When you predefine these models – as was done in the OLAP era – business users in specific domains can very easily and quickly perform analysis on the data. In turn, this helps business users get comfortable with using data and helps encourage adoption throughout the organization, which is critical for developing a data-driven culture.

It’s also important to note that having a model for routine “known unknown” analysis doesn’t prevent exploratory “unknown unknowns” analysis. In fact, as we’ll discover later in this whitepaper, the ability to perform both types of analysis from the same data is a powerful tool for bridging the gap between business intelligence and data science teams.

## The Rise of Metrics Stores

In the machine learning world, data scientists use models to generate predictions based on the values of input variables, called features. It’s become popular to create stores to catalog these features and manage the data pipelines for engineering them. Feature stores are tools that curate features, manage their lineage, and ensure features are accessible, discoverable, and reusable by data scientists.

Metrics stores are the business analytics equivalent of feature stores. A metrics store can provide consistent definitions of metrics, even in loosely structured data lakes. In addition, metrics stores can provide metric-by-metric dimensionality and assure analytics consistency across platforms.

However, metric stores still fall short of a full semantic layer. There’s support for defining metrics, but there is no ability to create predefined dimensions and hierarchies. Metric stores also have limits in analytics granularity and are highly dependent on SQL. This hinders the depth and flexibility of business analytics.

## What is a Semantic Layer?

The semantic layer provides a centralized model for AI/ML and data analytics. It can turn a vast array of largely undocumented and undiscoverable datasets into a contextualized and accessible data model.

More specifically, the semantic layer is organized into metrics and dimensions that make sense in a business context. Since dimensions often define hierarchies, it’s also easier to drill down or roll up for deeper insights and root cause analysis. Semantic models are also self-documenting, more discoverable, and readily consumable because they’re based on business concepts and vocabulary, and are designed to support business goals.

Ultimately, the semantic layer ensures the scalability and performance of modern data lakes while also providing the traditional capabilities from the OLAP era necessary for business domain-specific analytics. As we’ll discuss in the next section, the semantic layer can enable a “hub-and-spoke” operating model that’s been popularized with the data mesh.

## The Data Mesh Approach and Hub-and-Spoke Analytics

The data mesh approach has become popular with today's modern data architectures. Since the model prescribes that data be decentralized or distributed throughout the organization, it makes sense to bring data analytics capabilities to the teams with business domain expertise. This requires cross-functional responsibilities to manage their own business data.

However, this decentralized approach often means that centralized IT teams are sidelined. They're responsible for ensuring governance and providing the infrastructure for analytics, but the IT team loses control over performance and consistency when business domains are given too much autonomy.

While the most pure data mesh implementations are may be too decentralized, there is nonetheless real merit to giving business domains autonomy to create and manage data models that are relevant to them. A semantic layer can help enterprises strike the right balance between autonomy at the edges and control in the center for a more effective hub-and-spoke operating model.

### BRIDGING THE GAP BETWEEN AI AND BI

Once you have a semantic layer implemented, this opens the opportunity for generating features from semantic models by business teams. This enables organizations to bridge the gap between AI and BI by coordinating data engineering and ML engineering as well as DataOps and MLOps.

Furthermore, enterprises can extend the data products concept to AI products, with autonomy, federation, coordination, governance, and other principles. This can create a positive feedback loop where operational or descriptive analytics can turn into predictive analytics, which in turn supports prescriptive insights for automatically improving operations. By incorporating insights from predictive analytics, organizations can iteratively improve their performance, enabled by a data-driven approach.

## Delivering Actionable Insights at Scale

The semantic layer enables data to be turned into actionable insights faster, at scale, and with a high level of consistency, availability, and relevance across the organization. Since a domain-specific semantic models can be built atop shared, centralized models, there remains sufficient autonomy for business teams to create data definitions that suit their needs.

At the same time, the organization-wide base semantic layer creates a common denominator for business domains, ensuring consistency and autonomy coexist. Different business teams can also adopt subsets of models from other domains, enabling the reuse of relevant data definitions and metrics. The semantic layer, therefore, creates coordinated autonomy rather than chaos.

This federated model includes the base semantic layer itself as the “hub” and the “spokes” are the individual semantic models, built above it, at the business-domain level. More importantly, the shareability and reusability means there can be bilateral connections between different spokes, and not only between the spokes and the hub. This federated approach is the key to enabling collaboration between business domains without thwarting their independence or innovation in the  data analytics realm.

In short, enterprises that want to deliver actionable insights at scale using a hub-and-spoke operating model need to consider a semantic layer. This is the best way to turn data into a product that’s accessible to business users, effectively democratizing data throughout the organization.

Andrew Brust, Founder & CEO of Blue Badge Insights

**Andrew Brust** is Founder/CEO of Blue Badge Insights, providing strategy and advisory services to data, analytics, BI and AI companies, as well as their partners and customers. Andrew covers the data and analytics world for VentureBeat and The New Stack, and is a lead analyst for GigaOm in that same space. He also co-chairs the Visual Studio Live! series of developer conferences, is a Microsoft Regional Director and Data Platform MVP, an entrepreneur and consulting veteran.