WHITEPAPER

THE IMPORTANCE OF A SEMANTIC LAYER FOR AI & BI

A Perspective From Legendary Best-selling Author Dr. Barry Delvin

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing

ATSCALE

THE IMPORTANCE OF A SEMANTIC LAYER FOR AI & BI

In modern AI and BI applications, a semantic layer is vital in allowing businesspeople to find and use relevant information, as well as supporting IT in high quality and cost effective data delivery.

As requests from the VP of Marketing go, this one was pretty straightforward. Deborah Dee, head of data science at BIG Supplies, could immediately envisage the algorithms needed to calculate the likely customer churn percentage when shipping charges increased next month in line with the surging cost of fuel. Back at her desk, she gathered up the customer data required from the new company CRM system and began the process of combining it with the detailed historical sales from both the Web and in-store order and sales management apps. With the recently installed cloud-based analytic environment, she knew she'd have the answer before lunchtime. With a self-satisfied smile, she hit "enter" and sat back to await the result.

That smile was short-lived. 42%. The VP would be apoplectic. That would kill the business. It couldn't be correct! Deborah reran the process from scratch, checking the scripts, reviewing the input variables, everything. Nothing helped. The answer was 42, still.

It was Bill Prior, the old-timer who built the company's first data warehouse, who noticed she was tearing her hair out. His eyes lit up as she explained her problem. "Customers is the problem," he said ungrammatically.

Just because a data item has the same common name in different places doesn't confirm it has the same meaning.

"In the Web app, 'customer' is someone who bought something. In-store, 'customer' also includes someone who bought and returned an item. But the big issue is with the CRM system: it keeps both prospects and ex-customers in the same table as current customers, distinguished only by a flag. So, if you took your customer list raw from the CRM systems, you have a significant overcount."

Deborah smiled ruefully and returned to redo her data preparation routines. If only we could bottle Bill's semantic knowledge, she pondered... today was his last day.

Semantics is... as semantics does

Forrest Gump said the same about "stupid." But semantics is a lot cleverer, at least in principle. Wiktionary offers perhaps the simplest definition: the study of the relationship between words and their meanings.

In a business setting, the meanings of many common words and phrases depend very much on the department using them. Often, it's the most common of words, such as customer—as seen above—or profit that have the greatest variety of meanings. The result: when a business analyst or data scientist draws data from multiple sources that have been built by different departments, they quickly encounter these differences in meaning. The result is the old meme of "garbage out."

In this case, however, the problem is not "garbage in." In fact, the source data may be perfect within its own context—the department that originally created it and first uses it. Data always has an implicit creation context that defines its original meaning. Data also requires an explicit usage context—and there may be more than one—that allows someone who doesn't know its original meaning to use it correctly and with confidence. Context, meaning, and semantics are intimately related, and it's vital to take this into account in modern analytical work.

Lesson one: Data is the commonly used word, but information is what business really needs. The difference is context, and semantics plays an increasingly central role in delivering value.

Make way for the semantic layer

Metadata has been talked about since the earliest days of data warehousing. Sadly, much of it has been nothing more than talk. Metadata was supposed to provide the context for data: data about data. In practice, early metadata was mostly created in ETL (extract, transform, and load) tools and was largely technical in nature. More recently, the focus has turned to business metadata, stored in data catalogs, mainly as a result of the prevalence of context problems polluting data lakes.

I suggested in 2013 that what we really need is **context-setting information**[1] (CSI)—metadata on steroids—as a way of refocusing attention on the true breadth and importance of the context through which basic data becomes valued and valuable information. CSI is pervasive throughout the information environment, but to make it truly useful and usable for business and manageable for IT, it must be positioned centrally in the data delivery architecture.

Knowing the context of data creation and use is key to its valid use, especially when using data together from different sources or in different applications.

[1]Barry Devlin, Business unIntelligence, 2013, Technics Publications, New Jersey, http://bit.ly/BunI-TP2

Enter the semantic layer—simplistically a business representation of data that provides business users easy, understandable access to that data

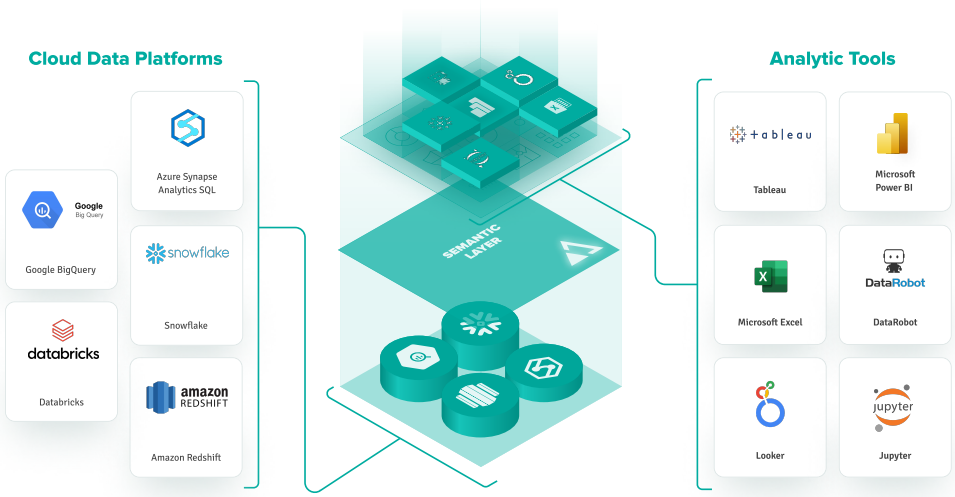

Figure 1: AtScale's semantic layer

Back in BIG Supplies, Deborah would have avoided her misinterpretation of customer if a semantic layer had been interposed between her analytics environment and the various data sources. When the Marketing VP saw the correct answer, he ordered that the number should be included in the weekly management information pack. Talking to the BI (business intelligence) reporting team, Deborah was now wise enough to ask what they meant by customer and was unsurprised to discover another couple of definitions. The semantic layer she desired must support many-to-many relationships of meaning—catering for multiple and varying usage contexts—between source and target systems. A semantic layer did exist for users of the business intelligence too—but only there—and it was now clear that a proper semantic layer must support everybody from spreadsheet users to data scientists with their analytic and AI (artificial intelligence) tools.

Lesson two: A semantic layer must work for all potential users of data and for whatever purpose the data is used. It is a common, shared resource across the whole business.

Deborah took a deep breath and headed toward the office of the Chief Infor-mation Officer (CIO). It was only at that level in the organization could such a common resource be promoted and implemented. A semantic layer depends on and impacts all the data (and information) sources and targets of the business. It also must clearly benefit all the producers and users of data if it is to be successfully implemented.

This is certainly not just semantics

BIG Supplies' CIO, Ted Bruin, was often called "the Bear"—behind his back, of course—because he got very grumpy if infrastructure didn't clearly deliver business value as well as saving IT some money. Unusually, Deborah's pitch to him was about business value. A semantic layer allows the business to avoid serious mistakes in decision making and significantly speeds up time to value in BI and AI by ensuring everybody speaks the same language. Faster and better decisions have clear business value, although it may be difficult to quantify precisely.

A semantic layer is a key component in successful self-service initiatives for both BI and AI, as well as bridging the gap between them. Understanding the meaning of data in context is a great starting point. However, BI users often struggle to build valid and performant queries on their own.

As BI expands to AI and analytics, a semantic layer is key to bridging data across these differing environments to deliver consistent results.

A semantic layer can present data structures in a form more familiar to business users—such as OLAP cubes—and both check and confirm SQL validity, as well as optimizing the query for improved performance. Recent benchmarks by AtScale of their platform show reductions of some 76% in SQL query complexity when using a semantic layer vs. accessing various cloud data warehouses directly.

Moving from BI to analytics and AI involves an often-significant increase in complexity of data structures and processing needs. A semantic layer offers the opportunity to simplify these new environments for businesspeople. Both business ana-lysts and data scientists want to be more self-sufficient in data use. They want more easily to find and understand the available data, to obtain it from the best approved and governed sources, and to create the best queries and analyses. They also know that increasingly they will have to work together to operationalize analytic discoveries and apply AI to existing BI reports. The semantic layer is key to such self-sufficiency.

Ted was pleased. He saw immediately the prospect of reducing the IT bottleneck in delivering data-centric solutions. But—as ever—he wanted more. "What about the underlying IT costs?" he grumbled.

Lesson three: A semantic layer provides not just meaning and context for data, but also support for the creation of valid and performant queries and analyses of that data.

Semantic layer to the rescue

As seen in Figure 1, a semantic layer resides in a pivotal position in the data delivery architecture. This is because of its role as an intermediary in data meaning between source data stores and target user tools. As we've just seen, these semantics extend also to the users' processes of querying and analyzing the data. This knowledge of data access semantics can also benefit IT, particularly if more performant data structures for specific users' queries can be constructed and managed in the underlying data stores.

Consider the OLAP or "slice-and-dice" approach to analysis of sales data that de-pends on data structured as cubes or in star schemas. When fact tables, such as customer or sales transaction, in such schemas become very large, many common queries may result in expensive table scans.

A semantic layer offers the ideal location to automate the creation and management of derived data structures to enhance query performance.

Traditionally, this problem has been addressed through the creation and ongoing management of summary tables structured according to the IT department's best guess as to which queries are most likely and most expensive. Aggregate tables such as sales-by-branch or customers-by-region are defined for these queries. They are then calculated and loaded once per day, for example, and used repeatedly in queries, reducing the processing costs for such queries significantly.

This approach, although widespread, is not without its issues. How do IT know which queries will be most common? If the frequency changes over time, how will IT know, and update the aggregates being built accordingly? These problems relate not to the aggregation approach itself but to the management of these aggregates initially and over time. This is where a semantic layer that is actively involved in routing queries to the data stores can help. The semantic layer monitors user queries and determines which are most common and if that changes over time.

It can then cause appropriate aggregate tables in the data warehouse to be defined, loaded, used when appropriate, and retired when no longer useful. Automating such processes is becoming ever more important as data-driven decision making becomes ever more widespread.

Lesson four: A semantic layer allows active and automated management of data structure optimization through its pivotal position between users and the underlying data stores.

The result is that IT gains significant performance benefits, without incurring the costs of manual design and ongoing maintenance. Recent benchmarks by AtScale have demonstrated 3-11x improvements in query performance, 11-31x improvements in user concurrency, and 1.7-16x reduction in compute costs over the four most common cloud data warehousing platforms when their semantic layer mediates some common TPC-DS queries accessing a 10TB database.

Key characteristics of a semantic layer

The four lessons listed above cover some of the most important benefits of a semantic layer. However, there's more. The following characteristics provide a broader view of what a semantic layer, fully and properly implemented, can offer:

1. **Data understanding for self-service decision making:** Businesspeople who want to be self-sufficient in finding the right data and using it correctly need the function of a semantic layer to automatically ensure that they "get it right first time" irrespective of their experience of the original source data meaning and lineage. In addition, where IT does need to be involved, a semantic layer reduces time and effort for them in supporting the business as well.

2. **Consistency in data usage across the organization and irrespective of BI or AI tooling:** As organizations implement AI, BI, and analytics applications in parallel, moving seamlessly between them and accessing the same data from many places, a semantic layer provides the required consistency of data meaning and usage across all approaches.

3. **Reduced "time to insight" for businesspeople:** Time is of the essence in many digital business situations and a semantic layer reduces delays in finding data and in running the required processes to prepare and query it. Even in non-real-time processes, such as month-end reporting, a semantic layer can accelerate closing the books.

4. **Support for businesspeople in creating valid and meaningful analyses:** Beyond simply describing data, a semantic layer contains appropriate context around that data to support and, in some cases, automate the generation of valid SQL and other queries, reducing the danger of producing incorrect results.

5. **Automation of the management of data structure optimization by IT:** A semantic layer can reduce IT costs by optimizing data structures for maximum performance, while reducing setup and maintenance costs by automatically deriving and managing these same data structures as business needs change.

6. **Monitoring of data lineage and usage in support of security and governance:** As an intermediate layer in the data delivery stack, a semantic layer can track all access to data at a detailed level—by whom, from where, and when—as input to enterprise governance and security tools and processes.

A semantic layer is at the core of the delivery, automated management, and use of meaningful and valid information to the business.

In brief, a semantic layer is at the core of the delivery, automated management, and use of meaningful and valid information to the business.

Conclusions

Back at BIG Supplies, Ted Bruin has now commissioned a semantic layer. The business is humming along with largely self-service data analyses, confident they know what data they have and what it consistently means irrespective of its source. They can also engage IT when necessary, who have access to the same semantics and can work productively with the business. IT have less work in optimizing database performance and maintaining performance-enhancing data structures. The data warehouse is also humming along. Ted is in a happier place and is seldom referred to as "the Bear" today.

As data volumes grow, and as information is put to an increasing number and variety of uses, defining and maintaining a record of its diverse contexts and meanings is vital. Data catalogs have already become common sight. The semantic layer is the next frontier in placing information with active context and meaning at the finger-tips of businesspeople and automating the delivery of well-managed and performant data by IT.

Dr. Barry Devlin

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.

Barry is founder and principal of 9sight Consulting. He specializes in the human, organiza-tional and technological implications of deep business insight solutions combining all as-pects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor on Twitter (@BarryDevlin), LinkedIn, and more, Barry is based in Cornwall, UK, and operates worldwide.