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 rela-tionship 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 val-uable 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 posi-tioned 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

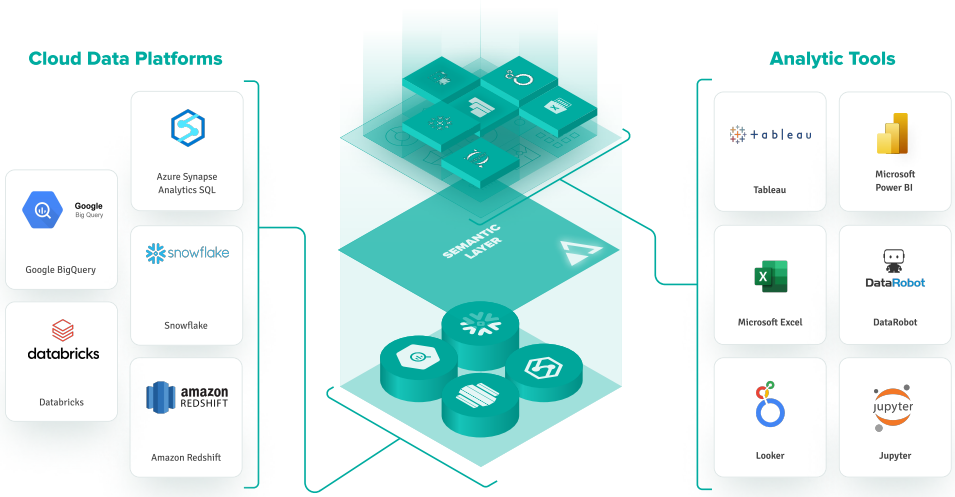

Figure 1: AtScale's semantic layer

Back in BIG Supplies, Deborah would have avoided her misinterpretation of cus-tomer 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 infor-mation 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 dis-cover 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 prop-er 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.