How Semantic Technology Drives Agile Business

To achieve sustainable competitive advantage and facilitate operational agility, organizations must speed the time to business value of newly acquired data assets from months to weeks or days. Semantic technology provides the architectural foundation for getting there.

Executive Summary

Business decisions are becoming increasingly dependent on timely and effective analysis of an ever-greater volume of data generated by a growing number of sources. Today, data is available from public sources, social media, data aggregators, business partners and internal systems, opening up more choices for sourcing and managing information.

Moreover, mobile technology now provides immediate access to data whenever and wherever it is needed. Users, customers and business partners await answers to impending strategic challenges, so organizations must radically reduce the time required to collect, understand and analyze the data needed to provide those insights.

In spite of significant investments in IT systems and other technologies, many organizations have still not fully integrated existing data assets into their enterprise information architectures. For example, a consumer products company might manufacture high-precision products from batches of raw materials sourced from various suppliers. To perform root-cause analysis for defects in batch formulation and production, the manufacturer must sift through data in its data warehouse and transactional systems, including inventory reports, material analysis reports, batch formulation records and instrument-produced test results — a manual process that can take days or even weeks.

Systems of record may include external data, such as stock and commodity prices, currency exchange rates or weather patterns. Because this data is continuously updated, it is impractical to store it in the data warehouse if it is only queried intermittently for the most current values. However, many organizations lack external visibility — as well as the ability to rapidly analyze data sourced from third parties — leaving them highly vulnerable in today’s hyper-competitive global economy.

The traditionally lengthy development cycles involved with integrating new data into the data warehouse diminish the ability for many organizations to respond to immediate analytics needs. Semantic technology (also known as the Semantic Web) provides rapid data onboarding and integration, connecting external data to the data warehouse, extending the leverage of
existing data assets, and embedding institutional knowledge and expertise into the data — all of which create opportunities to deliver insights faster. The use of semantic technology in data warehousing and business intelligence (DW/BI) can help the entire organization sense and adapt to ever-changing challenges and opportunities, in real time, to achieve and maintain competitive advantage. With the proper use of semantic technology, an organization’s ability to integrate and access data can, for the first time, match the velocity of today’s business environment.

The Speed of Business Outpaces DW/BI
The R&D organization of a pharmaceuticals company needs to quickly understand the progress of drug trials it is conducting through multiple contract research organizations (CROs). An industrial parts distributor must continually and instantly update, organize, standardize and present product images and specifications from hundreds or thousands of suppliers. A retailer must frequently monitor a mix of blogs, social media and Web sites to assess customer perceptions and the popularity of various promotions. All these examples portray the need for modern enterprises to increase the agility, flexibility and speed with which they can sift and rapidly analyze the ever-growing volume, variety and velocity of data. A big pharma enterprise, for example, must uncover which university or contractor will find a promising new treatment. A retailer analyzing Web traffic must predict which mix of Web logs, social media “likes” and user interface elements will reveal the secret to higher conversions. And in a global supply chain, anything from natural disasters to political upheaval to currency changes can suddenly produce the critical data entities that steer decision-making.

Yet it is precisely this ability to analyze new and different data entities and models that is missing in conventional DW/BI tools and practices. Data warehouses emerged in the late 1980s and early ‘90s when organizations needed more in-depth analysis of critical information, while shielding their transactional data from the performance impact of analytics queries. Unlike the well-planned queries performed during business transactions (such as inventory or price checks), analytics queries by their nature are less predictable. Business users need the freedom to ask questions involving a combination or range of data to solve a problem or find a new opportunity. This raised the risk that long-running queries would slow the transaction processing database and interfere with revenue-producing operations.

The answer was data warehouses — relational databases populated by data drawn from operational systems but designed and tuned to manage the ad hoc, unpredictable queries required for business analytics. These warehouses successfully addressed numerous historical analytics challenges and provided deeper understanding of customers, markets and business trends. Such challenges included the continuously updated data in transaction processing systems and the varying lengths of time that systems of record stored data. Both of these challenges made it more difficult to ensure “apples to apples” comparisons of data over time to measure costs, sales, inventory turns and other vital metrics.

Data warehouses also provided users with added control over querying and reporting, and they managed updates in a controlled, auditable fashion. Business units could dedicate servers and storage to data integration and reporting, and adjust capacity to meet their performance and cost needs. Similarly, users who conducted in-depth analyses could decide whether and when data was offloaded and archived, without having to negotiate with those responsible for keeping day-to-day transactions running smoothly.

The first generations of data warehouses worked relatively well for a variety of reasons. Often, the data was generated within and managed by the business, and was well-understood. When new data sources were needed, the process was often manageable and performed on a prioritized and scheduled basis, typically providing months in which to analyze the data and organize the data model. In addition, most of the data in the warehouse originated from relational databases, allowing the use of well-defined relational design practices.

Many data warehouses were implemented before the ubiquitous use of mobile devices, the flood of data from Facebook posts and Twitter tweets and the need to include e-mails and Weblog data to understand everything from demand trends...
to problems with Web applications or customer support. The success of first-generation data warehouses – and the business analytics and reporting they enabled – has, ironically, exposed their limitations (see Figure 1).

Organizations that have used data from their own logistics and point of sale systems to predict demand and fine-tune production must now do the same with data from businesses they have acquired or with which they have merged, without waiting months or years for cumbersome integration. They need to tap new types of data, such as browsing patterns on tablets or comments on social media, to track brand perceptions and customer service issues. This data is often unstructured, will need to be analyzed in unpredictable ways, and contains no description of how the various data elements relate to each other.

**DW/BI Limitations**

The limitations of current relational DW/BI tools and methodologies include:

- **The need to define data descriptions and detailed models early in the analysis process.** This means delays, costs and extra effort each time the definitions and models must be changed to meet new analysis requirements.
- **The velocity of new requirements** that can cause extra work and cost. Often, new requirements demand new data warehouse structures, introducing redesign and rework of data loading and analytics processing, thus creating delays in DW/BI deliverables and ultimately piling on costs.
- **The need for all data to be brought into the data warehouse** to make it available for analysis.
- **Their inability to quickly and easily cope with new data quality issues**, such as those caused by changes in the systems or processes that generate the data (i.e., a new CRM or order entry system). Even more traditional data quality issues, such as missing, blank or improperly formatted data elements, require

### Figure 1: Data Warehousing: Traditional vs. Semantic

<table>
<thead>
<tr>
<th>Business Need</th>
<th>Traditional Data Warehousing</th>
<th>Semantically-Enabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onboard New Data</td>
<td>• Thorough analysis, design and build.</td>
<td>• Rapid, light integration, with initial delivery in one to five business days.</td>
</tr>
<tr>
<td>Connect External Data to the Data Warehouse</td>
<td>• External data is collected and loaded into the data warehouse. The data warehouse is refreshed on a scheduled frequency.</td>
<td>• External data can be sourced from databases, spreadsheets, Web pages, news feeds and more; external data is mapped to ontology elements; internal and external data is queried through common methods, with real-time values delivered at query time.</td>
</tr>
<tr>
<td>Integrate Data Between Business Units or Business Partners</td>
<td>• Governance activities establish common vocabulary, data definitions, data model and interface specifications.</td>
<td>• Systems of record publish existing data specifications or ontology; each organization defines data in a manner that is best suited for its business.</td>
</tr>
<tr>
<td></td>
<td>• Shared data is copied to an integrated database.</td>
<td>• Federation and virtualization features provide choices as to which data to copy and which data to retain in the system of record.</td>
</tr>
<tr>
<td></td>
<td>• Organization-specific definitions may require duplicating certain data.</td>
<td>• All models can be supported through a single copy of the data, maintained in the system of record.</td>
</tr>
<tr>
<td>Capture and Embed Expertise</td>
<td>• Expertise is often captured in reporting and analytics; change management challenges exist when updates are required.</td>
<td>• Expertise captured in the data definitions; single, shared definition minimizes change management efforts.</td>
</tr>
<tr>
<td></td>
<td>• Reference values, such as average order size, are recalculated each time a report is run.</td>
<td>• Slowly changing reference values are periodically refreshed and referenced by reporting, improving overall performance.</td>
</tr>
</tbody>
</table>
expensive, complex and time-consuming tests before new data is loaded into the warehouse.

The extended enterprise data ecosystem has evolved from an in-house focus on systems of record to include large stores of external data whose volumes outstrip most organizations’ data warehouse capabilities (i.e., maintaining and managing the raw data). Rather than bringing every piece of data in-house to make it available for analysis – as was the case in previous DW/BI implementations – organizations need the ability to select which data will be stored in-house and which will be accessed from external sources.

Such an “externally-aware” data warehouse can provide a single point of access to the data ecosystem, querying internal and external data to answer users’ questions.

Another limitation of conventional data warehouses is that they are often organized to support a known domain of questions. The introduction of new data or new categories of questions often results in the most important insights. However, with conventional warehouses, the data onboarding process often requires lengthy knowledge acquisition, redesign and redevelopment activities. To speed this process, organizations must rely on techniques and technologies that leverage existing skills, while easily adapting to new perspectives as their knowledge and expertise grows.

As effective as they have been, current DW/BI technologies are too rigid to support today’s fast-moving, iterative and “learn as you go” business models.

As effective as they have been, current DW/BI technologies are too rigid to support today’s fast-moving, iterative and “learn as you go” business models. In an age when airlines regularly change code-sharing partners and integrate data about their customers’ travel patterns, manufacturers instantly shift global supply chains, and previously unknown mobile apps can skew buying patterns overnight, data warehousing and business intelligence must embrace and enable rapid change.

Rather than requiring a perfect upfront understanding of data models before moving forward, users need to instantly infer a “good enough” understanding of a new data set, construct a model that reflects that understanding to get started, and easily refine it as their knowledge improves and the needs of the market change. This is precisely where semantic technology can help.

### Semantic Technology Delivers Smart Data Integration and Analytics

Semantic technology enables faster, more agile analysis of more varied types of data because it integrates and uses the information required to understand how individual elements relate to each other and to the real world. It allows users to easily build new vocabularies, write new rules for handling new data types and define the relationships among data entities. In other words, a rich description of our understanding of the data is embedded with the data, to guide the access and analysis. Figure 2 depicts many of the standards, components and concepts that make up semantic technology.

#### Defining Semantic Technology

![Figure 2 - Defining Semantic Technology](image-url)
The heart of this new technology is the use of self-organizing databases that describe data entities in the form of “triples.” Each triple comprises:

- The entities to be tracked.
- A description of those entities (an attribute, such as a name or e-mail address) or their relationship to each other.
- The value of the attribute or the identifier of the other entity to which it is linked. These entities may be within the same database, in separate databases inside the organization, or in external databases on the Internet.

Data queries leverage both the data and the descriptions, making them “smarter” than traditional queries. As users add new attributes, descriptions and links to the database, they add context, based on their expertise, which makes it easier for others to understand the data.

Semantic technology delivers four new capabilities that are essential to driving business agility:

- **Extensible ontologies.** An “ontology” is a description of the set of data within a domain. An example includes the demographics, medical interventions, diagnoses, therapies, specimens, lab results and other categories of data that make up an electronic medical record (EMR), and the relationships between these data entities in the EMR. An “extensible ontology” is, thus, a description that users can easily add to or change as they learn more about the data with which they are working, and how elements of that data relate to each other. Using an extensible ontology, for example, medical researchers can identify the information they need from multiple databases to track the efficacy of a drug. Similarly, corporate data analysts can more easily identify the common data that needs to be shared among organizations involved in a merger.

- **Linked data.** Linked data connects information contained in different databases, allowing queries to find, share and combine information so insights can be identified across the Web, in much the same way that users can. This allows organizations to rapidly use databases that were developed independently, without the need to become an expert in them before realizing value from the linked data. Linked data is exposed, shared and connected internally and on the Semantic Web through universal resource identifiers (URIs) and resource definition framework (RDF) graph databases that embed the links across multiple databases.

Using linked data, for example, an auto manufacturer can easily allow users to conduct complex, natural-language searches for product information that span not only its own site but also those of other organizations, such as “find all new and used vehicles, both at your dealer and at used car sellers, between $19,000 and $23,000 and with more than 225 horsepower, a stick shift and four doors, and a user rating of 4.0 or greater.” Healthcare providers can more quickly and easily distribute the latest industry research in machine-readable form via a standards-based ontology that all users can share, without needing to agree on common, standardized decision-support platforms or data formats.

- **Embedded expert knowledge that allows users to integrate their expertise within the data.** The coexistence within the database of both data and descriptions, including rule-based logic, allows users to determine data values for certain attributes based on the current state of the data, even without specialized domain expertise. In healthcare, for example, analysts who are not familiar with all of the rules involved with diagnosing a disease state, such as high blood pressure, can query a dataset to determine whether that disease is growing in a particular population. If, for example, the database tracked the diagnoses and medications used by patients, a subject matter expert can define the logic as, “The patient has one or more of these diagnoses, but not one of those diagnoses, and is not taking this medication that causes false positives.” Therefore, indications exist for the selected disease state, and the logic can be associated with an attribute that the nonexpert analyst can query for a true or false value.

- **Provenance provides data traceability so that users know the origin of the data and how it was processed to achieve the presented results.** This increases both trust in the data and satisfaction with the results.

**Semantic Technology Benefits**

Semantic technology boosts business agility by making it easier to add or change the attributes of the organization’s data and descriptions, and by sharing and combining data from multiple sources to identify insights across the Web.
and definitions of the data that an organization stores and analyzes, as well as the relationships among the data entities.

In the relational database world, it can be difficult, expensive and cumbersome to add new attributes or descriptions to a data entity. It can potentially require a new data structure, revising the movement of data from one database to another, and the rewriting of any associated queries. In a semantic database, it is possible to automatically add new descriptions to existing entities as data is loaded into the graph database and to add rules applying to the new attributes and existing attributes.

Assume, for example, it suddenly became important to analyze the attributes of an existing list of customers based on whether they own or rent their homes. With semantic technology, a user can easily add “home ownership status” to an existing database of addresses by creating a new “triple” consisting of the customer’s identifier, the term “home ownership” and a value of “Rents” or “Owns Home” for each customer. Alternatively, it would be easy to create a rule that sorted all existing addresses and inferred that if the customer’s address contained the words “Apt.” or “apartment,” then the person associated with that address is likely to be a renter. While less precise than researching and loading the actual “Rents” and “Owns Home” data values for all customers, a few minutes spent defining the rule would extend the expertise embedded in the data, making it more valuable for end users.

Semantic technology reduces the cost and schedule for loading newly onboarded data into a database. One way in which it achieves this is through the rapid creation and deployment of rules to cope with data quality issues, such as “hide this field if the value is missing or doesn’t comply with current data quality rules.” This allows the organization to rapidly deliver value from analysis of the new data without testing every incoming data element during data loading, as is often required using data quality management techniques with traditional ETL (extract, transform and load) technology.

Semantic technology can also reduce the difficult, slow and expensive process of integrating data from multiple databases, making it possible to perform queries against a wider range of data. The existence of common entities among databases, with links between them, allows users to create queries without having to know exactly where the data resides or how it is organized. This is useful in scenarios that include merging data, facilitating corporate restructuring, consolidating data from multiple internal business units for analysis and deriving insights from multiple, independent systems that use different database models and data formats.

New Practices for Agile Data Warehousing

While semantic technology can deliver rapid and dramatic benefits, it also requires new ways of thinking about data, as well as new skills and tools. Some best practices include:

- **Focus modeling efforts on an initial set of high-value entities that are required to support a set of analytics.** Next, identify which data categories and points of integration (common attributes) to include in the initial data entities. For example, if one business group needs to analyze a group of healthcare customers from a clinical standpoint, and another from a financial perspective, the common connection point might be a patient identifier.

- **Engage an ontologist (semantic data modeler) to ensure use of the proper data entities; define the links between data and associated entities; and create inference rules to embed intelligence captured from subject matter experts.** The ontologist can create meaningful attribute synonyms so that users in different groups can continue to use familiar terms without cumbersome negotiations between the groups to create new, common terms.

- **Identify mappings from existing systems to the semantic store, so the relational data can be accessed through semantic queries.** Connecting a semantic access interface to each relational database can often be accomplished in just a few hours. Mapping from the semantic database to relational databases can allow users to start querying against relational data without waiting for data loading processes to be developed. It also allows for ongoing “agile” improvements to the data model as users gain experience with querying the data.

- **Apply data quality rules to entity and attribute definitions, rather than as a set**
of filters at the data element level. This makes it possible to achieve finer-grained data quality that aligns with different requirements at various levels of the organization, as depicted in Figure 3. Data that does not comply with one function can be hidden while remaining available for other functions that are less restrictive, rather than being totally removed from the database. As data quality rules change, updates can be applied rapidly without extensive data management reprocessing. This eliminates the difficulty of retroactively applying new data quality rules after the data is loaded, which is often experienced when data quality management is performed as data is loaded into the data warehouse. It also avoids the use of data quality rules that are insufficient for some uses and overkill for others.

- **Streamline data movement using semantic technology to reduce data latency and cost.** This practice replaces the series of movements performed by traditional ETL, one of the most expensive processes performed by the data warehouse. Figure 4 (next page) contrasts data load processing using traditional ETL methods with semantic data loading processing. Traditionally, data is first extracted from the source system and deposited in the landing zone, a part of the data warehouse that speeds the extract process and minimizes the schedule for pulling new data from the source system. Data quality management is performed as the data is moved from the landing zone into the staging area. Data that passes data quality tests is integrated with data already in the integrated store. Finally, the data is organized to support domain-specific analytics in the analytics layer.

As data quality rules change, updates can be applied rapidly without extensive data management reprocessing.

In some cases, this analytics layer is a simple remapping of the integrated store, leveraging the virtualization capabilities of the business intelligence software and does not require data movement. In other cases, the analytics layer is a separate set of data tables in the data warehouse and thus requires additional movement of the data.

Data marts contain subsets of the data in the data warehouse, usually to support a particular business unit. The data is reorganized into a conceptual model defined by the business unit,

---

**Fine-Grained Data Quality Management**

![Diagram of Traditional Data Warehouse vs. Semantic Data Warehouse](image)

*Figure 3*
with data elements often renamed to use terminology familiar to its users. In this case, the new data loaded into the data warehouse must be moved to the data mart(s).

Semantic technology moves the data only once, from the source system to the semantically-enabled data warehouse, leveraging the ontology definition to organize the data to meet the diverse needs of data warehouse users. The ontology can also manage the definition of synonyms to match the terminology that is familiar to end users. Multistage data warehouses and separate data marts are often not required. This approach reduces the cost and time required for populating the data warehouse, while organizing the data to support end users’ analytics needs.

**Agile Data Warehousing and Business Intelligence**

The first wave of data warehouses showed the potential of standalone databases that were optimized for analysis and insight, not transactions. But they were held back by the limitations of the relational technology on which they were based, such as rigid data models that required lengthy “big bang” deployments before the first users could begin generating business value from their data queries.

Today’s agile businesses need to create, deploy and obtain answers from data warehouses far more quickly and with greater flexibility. Semantic technology’s use of self-organizing database technology enables developers and business users to more quickly and easily add new data, descriptions of data and the relationships among that data.

This capability allows users to begin asking questions earlier and obtaining answers more quickly, providing feedback to refine the ontology model (and its database) based on its success in answering users’ questions. Organizations are now able to ask and answer new questions, and uncover fresh insights that drive lasting business benefits.
Footnotes


About the Author

Thomas Kelly is a Director in Cognizant’s Enterprise Information Management (EIM) Practice and heads its Semantic Technology Center of Excellence, a technology specialty of Cognizant Business Consulting. He has 20-plus years of technology consulting experience in leading data warehousing, business intelligence and big data projects, focused primarily on the life sciences and healthcare industries. Tom can be reached at Thomas.Kelly@cognizant.com.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 162,700 employees as of March 31, 2013, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.