The power of enterprise integration

Part 3: Combining data to create a full spectrum of information

By Terri Rylander

Building an analytical data mart in support of an individual business unit used to be enough, but not anymore. Today's organizations have had their appetites whetted and now have an insatiable hunger for information that crosses business units. End users are now asking for analytical information to be combined with operational data, as well as unstructured data like e-mails, sales contracts and call center notes. End users are even asking for internal data to be combined with external data from the public domain, suppliers and partners. This integration of data of all types from all sources is known as enterprise information integration (EII).

Data integration used to be synonymous with the term ETL (extract, transform and load). This often meant developers would write SQL scripts to pull data from source systems and load it into tables in reporting systems, doing some cleanup along the way to make the data more workable for end users. As end users have become more sophisticated, their demands have increased. They are no longer satisfied with insight into one data source, such as a billing system or enterprise resource planning (ERP) system. End users are demanding access to the whole picture.

Managing the growing number of data sources and ETL scripts puts ongoing pressure on scarce IT resources. Adding visibility to new data sources is time-consuming, is labor-intensive and adds to the ongoing support burden. EII helps IT organizations move toward a service-oriented architecture environment, enabling them to more quickly roll out new applications. EII also facilitates enterprise application integration by providing a means to integrate the data portion of the applications.

Moving beyond traditional ETL

EII has evolved beyond traditional ETL in that it not only offers integration by moving data from the sources to a central database but also offers integration through data federation or virtual data sources. EII integrates not only relational data and flat files but also extensible markup language (XML) sources and even unstructured data.

Data federation is defined as joining data from multiple sources without moving any data from its original source. Federated queries can pull and combine data from any source without the user needing to create special queries, because the source data and SQL access are homogenized through a middle tier. This middle tier relies on XML to unify the semantics used to define the data, which removes the need for proprietary SQL. Fortunately, a number of EII products that can help with this are on the market.

Though EII products can assist IT professionals in moving forward with data integration, the success of EII depends largely on having supporting disciplines in place. The three disciplines that play key supporting roles are metadata management, master data management (MDM) and customer data integration (CDI), while the foundation for success relies on quality data. EII and its supporting disciplines are enabled by XML and sound data management practices.

XML: The great enabler

Probably the single greatest advance in enterprise integration is the use of XML. It allows data to be tagged with information that describes format, content type and other attributes. Tying this all together is something called document type definition. Attributes are defined in XML coding so that the data will adhere and translate to the proper format. With XML sitting between two disparate data sources, data can be defined and essentially mapped to provide that layer of translation.

Using XML, data can be tagged either upon creation or at a later time, providing greater flexibility. These tags can also be extended or modified to suit changing needs. XQuery, an emerging query language based on XML, is similar in syntax to SQL. It allows for direct queries on collections of XML data. Consider XML the universal translator—the cornerstone of all enterprise integration.

Principles of common definition
EII is founded on the principle of having common definitions of data, and several disciplines support this principle. Though the concepts are not new, metadata management, MDM and CDI are getting a fair amount of press, and for good reason. Each in its own way plays a role in successful data and information integration, supporting common definitions of data.

Metadata management provides detailed data definitions and attributes for each data element. Think of it as a card catalog for data. Having good metadata reduces IT development time and reduces the time spent reconciling data by the end user. EII is moved along more quickly with metadata in place. However, metadata management is not easy to implement. It takes time and effort to set up good corporate metadata, and to date no silver-bullet applications exist that simplify the process. Because of this, many companies have found metadata management to be elusive.

MDM is a close sibling to metadata management. TDWI defines MDM as "the practice of defining and maintaining consistent definitions of business entities (e.g., product and customer) and data across multiple IT systems and possibly beyond the enterprise to partnering businesses."

EII depends on having MDM practices in place. Providing and managing consistent definitions of common data elements such as customer, product, financial metrics, employee and geographic location allow data to be combined and shared across systems and business units. Having this common set of reference data helps support internal and external audits as well as reduces the data-quality issues that stem from irreconcilable data.

CDI and MDM are close relatives, and some might argue they overlap. CDI provides a common view and definition of "customer" across multiple channels, business lines and systems. Identifying a single customer, regardless of the many types of interactions he or she might have, has always been difficult. CDI represents processes and automation to integrate the various views of a customer into a single, authoritative customer record.

Importance of data quality
Data quality plays a huge role in EII, because data-quality issues present in just a single system and business unit are now exposed to the enterprise. It is imperative that appropriate attention and resources be given to correct any issues before the point of integration, so as not to replicate the problems.

The issue of data quality existed long before EII and is still one of the top challenges faced by the enterprise. Dealing with legacy systems, redundant systems from mergers and acquisitions, and ever-increasing demands for new systems have left precious little time to address data quality. Now, EII places an even bigger demand on the need for data quality and IT resource time, as those systems are considered for integration.

Enterprise data management is not optional
With data sources and applications crossing the boundaries not only of business units but also between corporations and their suppliers, practicing enterprise data management is no longer optional. At the very least, new government regulations such as Sarbanes-Oxley, Basel II and HIPAA drive the need to manage and monitor data to ensure compliance and meet audit regulations.

At the heart of enterprise data management is the discipline of data governance. Data governance takes on many forms but is primarily responsible for establishing the data strategy, data quality and data ownership, including data definition. Data governance helps answer such questions as what to integrate, when to integrate, how to handle integrated data-quality issues and who owns the definitions of master and customer data.

When it makes sense to integrate
Deciding when and which data to integrate depends on many factors, including the complexity of any transformations, the volume of data to be integrated and the number of data sources involved. EII proves valuable when integrating both structured and unstructured data. By leaving data in place and leveraging XML, these two types of data can be quickly married at a relatively low cost.

EII, however, is not the answer to all data-integration needs. While federated, virtual views and queries might seem logical, the query performance is greatly reduced by the number of data sources named in the query. Frequent queries involving multiple data sources are best done in a single data warehouse. Also, integrating highly transactional (insert, update and delete) data with data-integrity constraints across data sources could cause the processing to fail.

EII: The bottom line

Though not a simple undertaking, EII can provide significant returns. These returns are maximized when used as part of an overall enterprise data management strategy that includes metadata management, MDM and CDI. As EII products mature in the marketplace, it will be easier to satisfy the demand for data that crosses internal and even external corporate boundaries. An integrated information environment will give rise to new business opportunities. Integrated information is the ship's guiding light in the sea of global competition. **T**

This is the third in a series of four articles on enterprise integration. The first article discusses the strategic advantages of integrating the enterprise and an overview of the various integration components. The second article takes the mystery out of enterprise application integration and how technology has evolved. The fourth and final article will take a look at the new information worker and how advances in technology help employees leverage their human capital.

Terri Rylander is former Director of Business Intelligence at a Fortune 500 company. She is currently a business consultant and freelance writer.

Teradata Magazine-December 2007

http://www.teradata.com/tdmo/v07n04/Features/CombiningData.aspx