

# The shifting continuum

The increase in semi- and unstructured data means changes for your data warehouse.



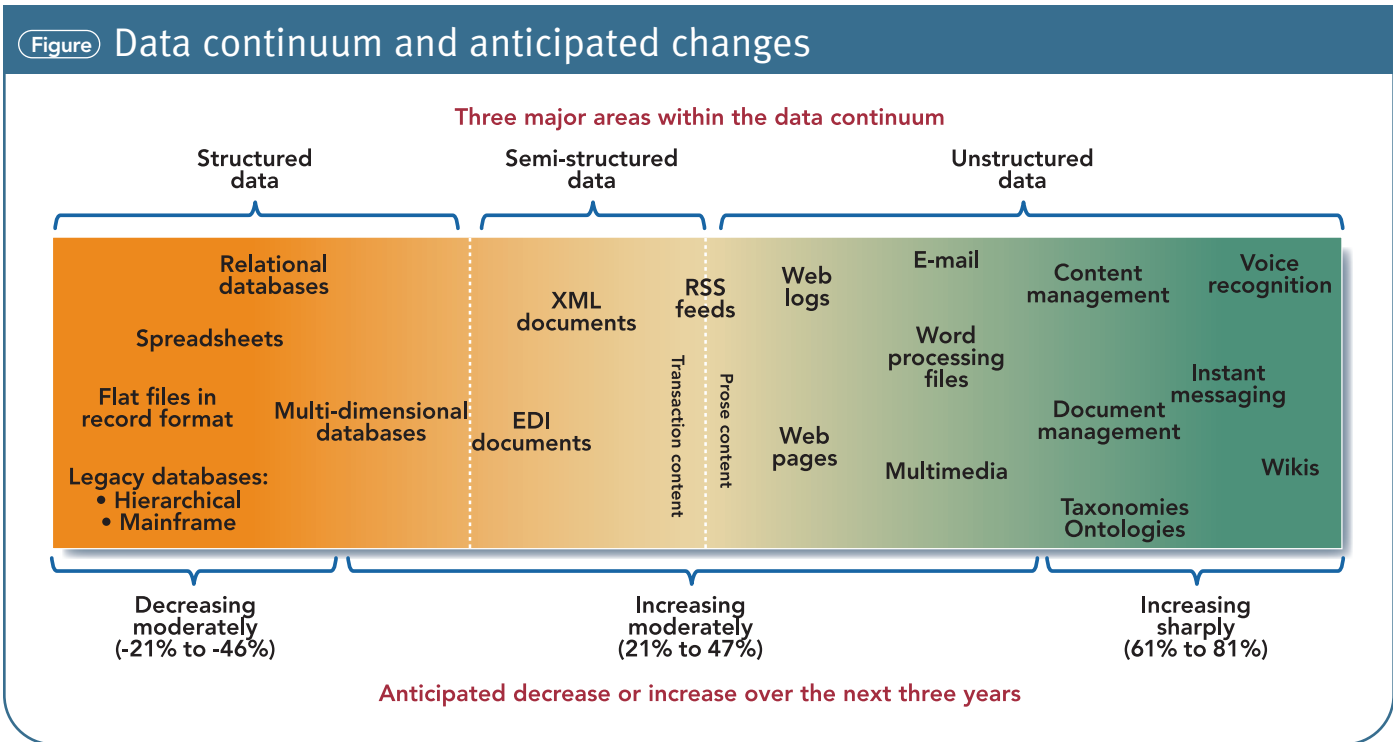
Structured data has long dominated the information content of data warehouses. The source data that comes from operational applications and databases is almost exclusively structured.

The data warehouse itself is highly structured, with its tables, star schema and other structured data models. Most reporting and analysis tools demand structured data accessible via structured query language (SQL). Data integration routines transform source data into the data structures required by the data warehouse and reporting tools.

But having such a narrow focus on structured data excludes the mass of valuable information in unstructured and semi-structured data sources, which

in turn hinders decision making and organizational performance supported by the data warehouse. In other words, conventional wisdom says that a data warehouse should be “a *single* version of the truth,” so that all decision makers work from the same information. However, a data warehouse is not the *whole* truth without representation from semi-structured and unstructured data.

A change is coming that will correct the current imbalance of data sources. Before we drill into the change and what



The various sources of data for a data warehouse plotted and sorted from most structured (on the left) to most unstructured (on the right).

it means to your data warehouse, let's step back and define our terms.

## The data continuum

The data continuum is simply the spectrum of available data sources. It categorizes and quantifies sources for a data warehouse. The data continuum breaks into three broad areas, or data types. Each is associated with specific data sources:

- > **Structured data.** At one extreme of the data continuum, structured data is commonly found in the mostly numeric data of database management systems (DBMSs).
- > **Unstructured data.** The other extreme includes documents of mostly natural-language text such as word-processing files, e-mail and text fields from databases or applications.
- > **Semi-structured data.** The area between the two extremes includes semi-structured data in spreadsheets, flat files in record format, RSS feeds and XML documents.

Note that some media are hybrids, in that they manage multiple data types. For example, most database records consist primarily of fields of numeric data, yet some fields contain text. Also, RSS feeds can transport prose (unstructured data) and transactions (semi-structured).

## Gaining equilibrium

The balance of the data continuum is currently tipped sharply toward structured data. Recent research from TDWI reveals that three-quarters of the information content of the average data warehouse today comes from structured data sources, while one-quarter is drawn from semi- and unstructured sources. However, the balance is expected to shift significantly over the next three years toward a fifty-fifty split between

structured sources and semi- and unstructured sources.

To illustrate this change, the figure on page 14 plots various sources of data for a data warehouse and sorts them from most structured (on the left) to most unstructured (on the right). Note that the percentage increase for data sources plots from left to right, too. Databases and other traditional sources of structured data (on the left) will decrease moderately as a percentage of overall data warehouse content, while sources of unstructured data that are new to data warehousing—such as content management systems and voice recognition—will increase sharply over the next three

years. In the middle, semi-structured data sources (especially XML and RSS) and some unstructured data sources (including Web pages and e-mail) will increase moderately.

In a nutshell, the vast majority of data warehouse feeds today come from structured sources. These won't go away but instead will be joined by more semi- and unstructured data sources, as the balance of the data continuum shifts. **T**

 For more information on this topic, see the TDWI report "BI Search and Text Analytics," available in its entirety at [www.tdwi.org/research/reportseries](http://www.tdwi.org/research/reportseries).

## Prepare your data warehouse

Anyone who works with or depends on a data warehouse or business intelligence (BI) solution should be aware of upcoming changes and prepare for them.

- > **Prepare for a deluge of unstructured data.** Many data warehouses struggle to scale up to burgeoning volumes of even structured data. Increasing the number of semi-structured and unstructured data sources will bring new challenges to scalability.
- > **Unstructured and semi-structured data must be transformed into structured data.** After all, data warehouses and reporting tools require structure. You may need new tools for text mining, text analytics or data integration that can impose structure on semi-structured and unstructured data.
- > **Data models will need adjustments.** A few data models in data warehouses will require

tweaking to accommodate the structured data coming from semi-structured and unstructured data sources. Similar adjustments are required when loading raw unstructured data into a data warehouse.

- > **Training—and learning—are in order.** Data warehousing professionals currently have little or no experience with unstructured or semi-structured data sources. Therefore, additional training is needed, and because of the minimal experience, the learning curve will be long and flat.
- > **Represent more unstructured and semi-structured data in the data warehouse.** You need to close the gap between structured and unstructured data; otherwise, your data warehouse will remain a single version of the truth, but not the whole truth.

—P.R.