TDWI HOT TOPIC

BIG DATA AND THE DATA WAREHOUSE: A QUESTION OF ADJUSTMENT

// BY STEPHEN SWOYER

Retooling for the era of big data isn't a question of starting from scratch. It doesn't require radically reinventing the data warehouse (DW). Philip Russom, director of data management for TDWI Research, says that what's needed are "adjustments" to how the DW is structured and managed.

Despite the big data phenomenon, the scalability issues that historically constrained IT systems have to some extent been neutralized. Vendor platforms are more adaptable and scalable; user practices are likewise more diverse, incorporating techniques and methods that are appropriate to the era of big data. As a result, organizations have the unprecedented ability to mine and analyze big data for valuable business insights.

"To get more business value from big data, companies are ... experimenting with analytics of all kinds. [They're experimenting] with everything except OLAP because a company of any size has a sizable investment in OLAP already. The growth is in the other forms, which are sometimes called discovery or exploratory analytics," Russom comments.

In fact, organizations need to move beyond OLAP—in spite of the fact that OLAP is and will continue to be a critical analytic technology. This is the first such adjustment, according to Russom.

"An important adjustment to big data for your data warehouse [environment] would be to acquire analytic tools that are new to you, [such as tools] that do data mining and statistical analysis, so that you can do discovery and exploratory analytics," he explains, pointing to the growing importance of natural language processing (NLP), which is used in analyzing text data. Bleeding-edge advanced analytics technologies use sophisticated algorithms and (especially if they're being used against unstructured data) tend to de-emphasize SQL. Nevertheless, a SQL-driven approach to advanced analytics—what Russom calls "extreme SQL"—makes sense for certain kinds of big data.

"SQL is ancient; it's been with us forever; yet it's so pliable that it actually becomes an effective, iterative methodology," Russom notes. "SQL obviously only works with structured and, in fact, relational data—but for many companies, their big data *is* structured, it *is* relational, or it *can be* easily processed to become that."

If the scope or function of analytics is changing, so, too, is that of the enterprise data warehouse (EDW). A single EDW can't always provide all the services that users expect, but instead can be augmented with special-purpose systems—in other words, a collection of diverse data storage and analytics systems along with the EDW at its center.

The difference—the *adjustment*—is heterogeneity, says Russom. "The acronym 'EDW' means a centralized data warehouse that has enterprise-scope data: it has data representing a lot of processes from many different business units and departments. This is the single version of the truth for most decision-making data. You can still have that, but as the core of a distributed architecture.

"It's not one DW," Russom continues. "It's actually lots of different platforms possibly from many different vendors. We have more data-processing workloads today than ever before. We have newer workloads around real-time analytics and unstructured data that the data warehouse was not designed for. That's okay, because you can have secondary platforms within the extended environment that are well suited to those workloads."

These secondary or "edge" platforms include (among others) NoSQL databases, which are used for storing semi-structured or unstructured information; graph databases, which are used for specialty analytics; and RDBMS appliances, which are ideal for extreme SQL offload scenarios.

"A lot of these edge systems are for extreme SQL [offload processing]," Russom explains. "The typical SQL statement is perhaps a dozen lines or so long per routine, but applied to analytics—where you're actually doing a lot of the data transformations in the SQL statement [itself]—you can

actually wind up with [SQL statements] hundreds or thousands of lines [long]."

There's also Hadoop, which has emerged as both a useful platform for data processing *and* as a kitchen-sink solution for landing or ingesting data.

In other words, big data isn't just shaking up analytics and recasting the data warehouse; it's significantly transforming data integration, too.

This requires another adjustment, Russom indicates: "With big data, the data staging areas have to have a greater storage capacity. They also have to have some optimizations for file-based data. That's why a lot of organizations are thinking about using Hadoop as their next-generation data staging area."

There's a final big-data-related adjustment that organizations are only beginning to come to grips with.

"Another thing we have to think about is economics. Big data costs money. If you add up all the costs around the data warehouse, it is expensive; it's typically your highest dollar-per-terabyte [cost] versus other platforms. At the other end, dollars per terabyte with Hadoop is typically very low," Russom concludes, noting that the cost of specialty appliances is "about in between" those of the DW and Hadoop.

"We need to preserve ... the data warehouse for workloads that will only run in an optimal way *on* the data warehouse. If you can take a workload to another platform that will run it cheaper and better, then do it. As companies are rethinking their data [management] strategies to accommodate big data, I wish more of them would think about the economics of it." ●

Stephen Swoyer is a contributing editor for TDWI.

ABOUT OUR SPONSOR

TERADATA. THE BEST DECISION POSSIBLE™

teradata.com

Teradata is the world's largest company focused on integrated data warehousing, big data analytics, and business applications. Our powerful solutions portfolio and database are the foundation on which we've built our leadership position in business intelligence and are designed to address any business or technology need for companies of all sizes.

Only Teradata gives you the ability to integrate your organization's data, optimize your business processes, and accelerate new insights like never before. The power unleashed from your data brings confidence to your organization and inspires leaders to think boldly and act decisively for the best decisions possible. Learn more at teradata.com.

ABOUT TDWI

tdwi
THE DATA WAREHOUSING INSTITUTE

tdwi.org

TDWI, a division of 1105 Media, Inc., is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, on-site courses, certification, solution provider partnerships, an awards program for best practices, live Webinars, resourceful publications, an in-depth research program, and a comprehensive website, tdwi.org.