TDWI HOT TOPIC

THE CHANGING ROLE OF THE DATA SCIENTIST

// BY STEPHEN SWOYER

Because data scientific expertise is so rare, vendors will focus on automating as much of the analytic process as is possible. A market for data scientific services will likely flourish, too.

There's a small pool of professionals who possess the skill set that makes for a good data scientist. With many of them physically located in Silicon Valley and charging high premiums for their services, vendors are focusing on how to automate as much of the data science processes and skills as possible.

The revamped Aster Discovery Platform that Teradata Corp. announced at February's TDWI World Conference in Las Vegas is just one harbinger of things to come. The company bills its new Aster Discovery Platform v5.1 as a "data science solution in a box." Given the complexity commonly associated with predictive analytics, this might sound like the company is promising too much. David Stodder, director of TDWI Research for business intelligence (BI), argues that Teradata's solution could be an effective and innovative approach to an extremely difficult problem.

For one thing, Aster Discovery 5.1 automates the most time-consuming aspect of predictive analysis: data acquisition and preparation. Today, these processes consume the bulk of a data scientist's or data analyst's time, Stodder notes. So, too, does the development, selection, and testing of algorithms for statistical and behavioral analytics, which Teradata's solution automates. Finally, Teradata provides visualization capabilities to represent the results of analytics. All these analytical processes (data acquisition, preparation, analytics, and visualization) can be fully implemented simply by leveraging the existing SQL skills within the organization with 70+ out-of-the box SQL-MapReduce functions.

"Teradata is doing a really good job of aligning that concept [i.e., of a discovery platform] with the traditional data warehouse. The two [use cases] have very different requirements; Teradata's been intelligent about how they're going about automating time-consuming or repeatable tasks in the discovery [use case]," Stodder explains.

Intelligent automation is key, he suggests. In the future, vendors will focus on automating as much of the analytic process as is possible. In addition to data access and preparation, this means automating (or making suggestions about) the selection of analytic algorithms to solve specific problems; managing the process of developing predictive models; and—especially with respect to still-maturing technologies such as Hadoop—automating the scheduling of big data analytic workloads.

All of this is a necessity, Stodder claims, given the dearth of data scientific expertise. You want your costly data-scientific talent working on predictive models, *not* coding complex data transformations or MapReduce jobs. "If you can automate that [preparatory heavy lifting], data scientists can focus on [selecting] algorithms and [optimizing] models," he comments.

"The root of the problem is that [companies] can't find enough people [with data science expertise]. In some ways, there's been a bit of a step back from the mania about data science, partly because there just aren't enough people out there."

This begs a question: for decades, large enterprises have maintained statistical analysis and data mining practices. Today, data mining or predictive analytic products are used throughout *Fortune* 500 companies.

Doesn't this mean that enterprises *already have* considerable data scientific expertise in-house? Yes and no, Stodder says.

"The SAS statistical analyst is a very close match to what people are looking for in a data scientist. What's different, for one thing, is the *technological* difference. Many organizations [today] are looking at Hadoop and MapReduce and the kind of big data they can put in those sources and how you analyze that data."

He points to another key difference: the data scientist, as distinct from the statistical analyst, typically is more of a hybrid—i.e., a kind of *business technologist*.

"There's this emerging, different idea of [what] data science [is]. I've found that companies would like people who have a little more than just statistical analysis skills. They [want people who] have a better understanding of the business and [who] can present their ideas to business executives around the organization," Stodder points out, adding that this is as much a function of cultural as of technological change.

"Companies would like their data science people to connect to business leaders throughout the organization. Many data scientists are very tightly involved in marketing, [e.g.] for marketing analytics and improving the performance of marketing systems. In some cases, they report directly into the CMO. That's one example of how [data science is] different."

It's also an indication of why data science is so hard: there's a small pool of people who possess the kind of hybrid skill set—e.g., statistical and mathematical expertise, domain-specific business knowledge (or knowledge that cuts across business domains), and the ability to communicate it all—that makes for a good data scientist.

"It could be that [companies will have to] step back a bit from the revolutionary thought of revamping their entire organizations with data science," Stodder speculates. "The task is to help companies figure out what they can *practically* do with data science. We usually solve problems like this with technology or services, and I think we're going to see a flowering of data science services out there, but the vendors can and will do their part, too." ●

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.