ebook

DECMEBER 2016

DELIVERING ADVANCED ANALYTICS IN THE ERA OF BIG DATA

Sponsored by:

STATISTICA® TERADATA. WhereScape®

tdwi
Transforming Data With Intelligence™

INTEGRATION, DATA PROCESSING, AND ANALYSIS AT SCALE

Advanced analytics is an umbrella term for a new class of technologies that apply the techniques of statistical and mathematical analysis, along with other cutting-edge domains, to common business problems. Advanced analytics makes use of new and existing data sources—as well as new technologies for managing, processing, and analyzing data at scale—to accelerate, and, in some cases, to automate, the development of analytics models and the identification of analytics insights.

Advanced analytics technologies can be embedded in a self-service user experience (UX), permitting the new breed of data scientists, statisticians, and citizen data scientists to work more productively. Conversely, some kinds of advanced analytics (machine learning, deep learning) are designed to run without interactive human intervention.

The shift to a broad adoption of advanced analytics is made possible by a convergence of three trends:

- A self-service–enabled—and statistically savvy—community of users.
- Massive improvements in storage and data processing efficiency, coupled with the availability of vendor-developed and open source machine learning and statistical analysis technologies.
- Fundamentally new economies of scale. As a function of self-service, technological advancement, open source innovation, and ongoing commoditization, advanced analytics projects are more doable, practical, and affordable than ever before.

"I saw data and analytics projects years ago that took federal grants to execute. It used to be that most

companies couldn't afford to experiment with, let alone adopt, predictive analytics, machine learning, or other even more sophisticated forms of analytics," says Shawn Rogers, director of global marketing and channels with Statistica. "Now, commoditized hardware is in play, open source software is in play, and self-service is in play. All of this contributes to this new economic advantage that allows everybody to get involved with advanced analytics."

Getting Started with Advanced Analytics

Designing and implementing an advanced analytics program can seem like a daunting task. After all, advanced analytics makes use of a slew of esoteric technologies, including machine learning (ML), data mining, rule-driven automation, function-specific artificial intelligence (AI), and deep learning. The good news is that a growing number of software platforms substantively abstract the complexity of these technologies, exposing a self-service, wizard-driven user experience.

There's another wrinkle here, too. In all likelihood, a viable advanced analytics program will also include some familiar technology components—namely, your existing business intelligence (BI) and/or data warehousing investments. Contrary to what you might have heard, advanced analytics is not meant to be a rip-and-replace "upgrade" for BI. Rather, the relational data that is grist for most BI-analytics use cases will also be used in advanced analytics applications.

Imagine a situation in which an analysis of data streaming from a cluster of remote Internet-of-Things (IoT) sensors points to a problem with a specific upstream component. The obvious solution is to replace the faulty component, but what is the window for doing so? Is the necessary part in inventory? If so, where? What are the logistics of getting it out to the remote site and getting a skilled technician out there at the same time? What is the predicted timeframe before the device fails? What is the cost of downtime? To what extent should the company expedite this replacement?

This information exists in either the company's operational systems or their data warehouse. This is an oft-overlooked point. We fixate on the potential of advanced analytics to disclose as-yet-unglimpsed patterns or trends—e.g., hidden trends or phenomena or game-changing business opportunities—but its most powerful application is its potential to transform day-to-day business operations. That can only happen when advanced analytics insights are combined—*contextualized*—with BI insights.

How, then, do you get started with advanced analytics?

First, be skeptical of prepackaged, all-in-one "solutions." Now as ever, vendors are keen to pitch you all-in-one or "turnkey" infrastructure or platform packages for advanced analytics. These offerings may be infeasible for several reasons, starting with the fact that—when it comes to advanced analytics—there is no such thing as "turnkey." More important, you *already have* a data management and BI-analytics infrastructure—one with which your new advanced analytics program must coexist. An all-in-one infrastructure "solution" may not get on well with your existing infrastructure investments.

Second, think of advanced analytics as a domain with four well-understood problem areas:

Data integration and preparation. Before advanced analytics practitioners can ask questions, test hypotheses, or discover patterns and insights, they must first have data to act on. In the background, the cutting-edge algorithms, functions, and visualization technologies used by data scientists, analysts, and other consumers are enabled by unseen, but no less cutting-edge, data integration, data preparation, and data provisioning technologies. The challenge is to accelerate the integration of data—automating the means by which it is transformed and provisioned—to permit data to be processed and analyzed at scale.

Data and analytics processing at scale. In the same way, the data processing platforms that underpin advanced analytics research and discovery must be optimized for churning through structured and unstructured data as quickly and efficiently as possible. Speed—the ability to complete an analysis as rapidly as possible—is everything. The more quickly data can be processed and analyzed, the more quickly advanced analytics practitioners can iterate. The more quickly they can iterate, the more quickly they can

reject faulty hypotheses or, conversely, identify meaningful insights and patterns.

A self-service user experience that abstracts complexity. The challenge is to expose highly esoteric advanced analytics technologies in the context of a guided, self-service user experience. This front-end UX must be smart, flexible, and deft enough to support a diverse range of human analysts, from less-statistically-savvy business analysts to data scientists and statisticians. Self-service amenities should be tailored for different users with different skills and expectations; business analysts will need more guidance than data scientists.

A (synthetic) data architecture that minimizes data movement. It is inefficient (and in some cases impossible) to move data around at big data scale. Processing must instead be pushed out to the platforms on which data lives or (for real-time applications) the entry points at which it is ingested into the enterprise. This is true of processing in conjunction with both data integration and analytics. If data lives in a data warehouse, data lake, document store, or operational data store, processing should be pushed out to those platforms. Advanced analytics software from IBM, SAS Institute, Statistica, as well as open source R code, can run in the context of many database engines (or in data processing environments such as Hadoop or Spark), thereby accelerating analytics development. Advanced analytics is time critical. The faster you can iterate, the faster you can fail. Failure is prelude to success.

Third, take stock. Reconcile what you need to do with what you've already done. This means taking a good, hard look at your existing BI and data warehouse infrastructure. It is absolutely critical that you drive inefficiency out of your existing data management investments.

Advanced analytics is a relatively immature domain, with comparatively few self-service or management automation amenities, at least under the covers. You are going to need more people—and more smart, creative people, at that—to design, build, and maintain your advanced analytics investments. As you've probably discovered, smart, creative people tend to be priced at a premium.

Fourth, use self-service as a means to empower users and bridge gaps. IT no longer has to be an *Übermensch*. To invoke a time-honored metaphor, a self-service user experience gives the end user a means to fish for herself. IT's priorities must shift accordingly. IT should focus on doing what it can do to promote the self-service user experience and support self-service users.

For some kinds of self-service user roles, IT should focus on rapidly provisioning data that's *good enough*—not perfectly cleansed and consistent. For others, IT should make it a priority to give users access to data rather than developing and deploying customized views. IT should think of self-service as a proving ground, too; if a self-serving user creates something—a metric, analytic, data flow, etc.—that's useful and valuable, consider making it available to all users. "The challenge is to make data available in a way that supports self-service. You cannot rely on waterfall approaches to do this. Nor can you move lots of data around, from system to system, at big data volumes," says Chris Twogood, vice president of product and solutions marketing with Teradata.

"The biggest challenge for data scientists and self-service users is that they want to integrate their data with production data. They need access to data. If they can join [their data] together with production data, they can more easily prove out the different hypotheses they're trying validate against. The second [challenge] is to take what [self-service users have] done ... if it's something that can benefit the business as a whole, to make it repeatable and put it into broader production."

Look to Leverage Automation When and Where Possible

Teradata's Twogood gets at another key point—the advanced analytics machine is fed by data integration processes. Starve analytics processing platforms of the data flows they need and you'll bring an advanced analytics practice to a grinding halt. The rub is that human beings (individually and in organizations) are generating more data than ever before. A disproportionate share of this data is being generated by machines, however. By 2020, says International Data Corp.

(IDC), the global amount of data will increase by a factor of 10. Much of this increase will be fueled by an explosion in machine-generated data, IDC predicts. The upshot is that we no longer live and work in a world in which data can be ingested, processed, and analyzed at predictable intervals.

In our new world, data continuously pulses and streams, arriving in trickles, bursts, and deluges. Nor is relational data the only (or even the primary) game in town. Data comes in different shapes and sizes, from semistructured JSON files, event messages, and log files to multistructured multimedia and binary files.

This deluge of data is too much for data architects, data engineers, ETL developers, business analysts, statisticians, and data scientists to process and integrate on their own. The larger point is that it is neither cost-effective nor practicable to employ human beings to perform many data engineering tasks. This is especially true for tasks such as building and maintaining the data flows that feed advanced analytics practitioners. To the extent possible, many aspects of this work must be automated. The same is true for data and analytics modeling, analytics model training, and other tasks.

Automation in this context isn't an excuse to replace or to eliminate people. It is, instead, *mandated* by an exponential increase in the size, heterogeneity, and volatility of data.

You're automating so that you can free up your human talent to focus on the kinds of valuable, difference-making activities that require human creativity, ingenuity, and imagination.

Automation is an overlooked tool in traditional BI and data warehouse environments. After all, most DI and RDBMS vendors ship platform-specific automation features—e.g., pre-fab source connectivity and transformation wizards; data model design, generation, and conversion tools; SQL, script, and even procedural code generators; and scheduling facilities—with their respective databases.

Several vendors likewise market platform-independent data warehouse automation (DWA) tools that purport to automate many aspects of data warehouse design, deployment, and maintenance. DWA technologies can accelerate scoping and data model design; generate platform-optimized SQL code and stored procedures; make use of platform-specific features, such as loader technologies; and automatically generate metadata and documentation. Custom-fitting and design of some kind will always be required. In connection with data warehouse systems, however, much of the day-to-day heavy lifting of data integration can be accelerated, and in many cases automated.

The same is true, albeit to a lesser degree, of data and analytics processing. In the first case, the priority is to minimize data movement. This means pushing data transformations down and out to the platforms on which the source data resides: the equivalent of TEL, as distinct from ETL.

In the second case, the priority is to facilitate access. This means leveraging data fabric-like capabilities—e.g., three-part names in DB2; database links in Oracle; in the Teradata world, QueryGrid is a much more sophisticated data fabric technology—to make data more easily available to self-service consumers. Using these technologies (individually or together), self-serving users can transparently query against tables in remote database systems. Teradata's QueryGrid likewise permits users to query against data in NoSQL platforms such as Hadoop and MongoDB.

In the third case, the priority is to exploit in-database (or, in the case of Hadoop or Spark, in-data-processing-platform) analytics functions and algorithms. Not only does this help to minimize data movement, it can accelerate both data integration and analytics processing. Advanced analytics functions and algorithms from commercial vendors such as Fuzzy Logix, IBM, SAS, and Statistica—along with open source software libraries, including R and many kinds of ML algorithms—can run in the context of the major database engines (and can be scheduled to run in Hadoop and Spark).

This is critical for both data integration and analytics. In the future, data integration technologies will incorporate statistical functions and libraries both for the sake of convenience and out of necessity. For example, it's much easier to use statistical techniques to correlate IoT events

across both the space and time dimensions during the data integration process itself. In the same way, IoT events can only be correlated using statistical techniques such as curve-fitting and regression analysis.

There's another wrinkle here, too. Just as an ability to "fail fast" is key to rapid iteration and progress in advanced analytics, it's no less critical to data engineering and data warehouse development, too.

"It's the same philosophy as in data science. It isn't how fast you can build it, it's *how many times* you can build it," says Michael Whitehead, CEO of data warehouse automation specialist WhereScape.

"With automation, you're actually giving someone a license to fail because you're making it much easier for them to experiment," he argues. "You have an environment that makes it possible for you to fail quickly so that you can improve just as quickly. That's the real value of automation, whether it's with machine learning in data science or agile data warehouse design."

Deliver a Self-Service Experience That Empowers Users and Bridges Gaps

An advanced analytics practice can take root and thrive if the following conditions are met:

- Users have ready access to data as well as the ability to blend, explore, and manipulate data
- The user experience can be tailored to the strengths and weaknesses of each self-serving user

This is easier said than done. Advanced analytics technologies such as ML, deep learning, neural networks, function-specific AI, and rule-driven decision automation—to cite just a few—are incredibly esoteric. Comparatively few people understand how to apply these concepts; vanishingly few are capable of actually understanding the math (or theoretical assumptions) behind them.

Advanced analytics platforms from Statistica and other vendors help abstract the complexity of these technologies behind a guided self-service user experience. On the front-end, they expose question-based wizards and other, similar ease-of-use features to help simplify the selection of algorithms and functions. They also use data visualization technologies—in combination with built-in data profiling and analysis algorithms—to accelerate the process of exploration and discovery. Finally, they automate the design and training of analytics and predictive models, as well as the preparation of data flows that mix streaming data from (for example) IoT sensors with data from operational systems or the data warehouse, or with data from geographic information systems, and so on.

Because too few potential users actually understand the concepts and mathematics behind advanced analytics technologies, such "smart" front-end tools are essential, argues Statistica's Rogers.

In Statistica's case, the underlying platform helps accelerate the development of analytics models and the identification of analytics insights. It also makes it possible to push analytics processing out to Teradata, Hadoop, and other data processing platforms. In the future, Rogers argues, a larger portion of analytics processing will shift from the data center to the "edge"—i.e., to the points at which data is ingested into the enterprise, or to the remote locations where sensors and signalers live. "This is becoming a necessity. If you can run the analytics where the data lives, that's an advantage. Just doing analytics at the core, in the data center, is no longer good enough. It is essential to be able to push the analytics out to where the data is living," Rogers says.

Smart tools and enabling platform technologies are one piece of the overall puzzle. Rogers urges organizations to consider making structural changes to better support their advanced analytics initiatives. "The most successful customers are actually standing up centers of excellence for analytics," he continues. "What they are trying not to do is to be solely reliant on the skill set of a single or a couple of data scientists, so they're trying to put together a smart supportive team, people with deep domain

expertise, sometimes along with people from the application development side, too."

Conclusion

The challenge of advanced analytics is two-fold. First, it poses significant problems with respect to data access, preparation, and provisioning, particularly at big data scale, when data flows to the enterprise from many different vectors. Second, there's the challenge of *democratizing* advanced analytics—of repackaging the tools and techniques of an extremely arcane field in a UX that makes them intelligible, insofar as possible, to as many users as possible.

The upshot is that advanced analytics—even more than BI and data warehousing—takes an ecosystem. No single data management, data processing, or statistical analysis platform is ideal for all advanced analytics applications. With respect to data management and data processing, data processing is no longer confined to batch intervals and no longer centralized in a single location.

With respect to statistical and advanced analytics front-end tools, the open source software (OSS) community is a locus of innovation. New libraries, functions, algorithms, and other assets usually appear first in the OSS world. Any sufficiently sophisticated advanced analytics practice is going to be a hybrid of sorts. "Doing advanced analytics in the era of big data requires an analytics ecosystem and that ecosystem needs to include multiple different analytics engines and an assortment of [front-end] tools. No single analytics engine can suffice for all workloads," argues Teradata's Twogood. "Each of these [processing engines] has different uses. The core value around having an analytics ecosystem is that it makes data access and [analytics] processing seamless to the user. They're not worried about where the data is or what they need to do to access it."

WhereScape's Whitehead makes a similar point. "No one is going to give you the complete solution. If [a vendor] tell[s] you they can do that, run, don't walk, away from them," he says.

"You don't buy IBM or SAS or Teradata or WhereScape and expect it to be *the* answer," Whitehead concludes. "This stuff is hard. Unless you're Google or Facebook, you don't have the luxury of building it yourself. You don't have the talent to build it yourself. More important, you don't have the luxury of starting from scratch. You begin where you're at—like basically everybody else."

STATISTICA®

www.statsoft.com

Statistica provides a comprehensive and robust platform for data and text mining, and predictive analytics that can deliver solutions ranging from fraud detection in finance to process optimization in manufacturing.

The Statistica solution delivers

- Leading-edge predictive analytics: sophisticated algorithms to build models that provide the highest accuracy and best ROI
- Enhanced text analytics: an advanced text mining tool leverages unstructured and textual data within the model building process
- An enterprisewide solution: the secure multiuser, role-based Statistica Enterprise platform is a truly collaborative environment for building, testing, and deploying the best possible models for fraud detection
- Reflexive models for real-time needs: Live Score processes new claims as they happen and updates fraud models with turn-around times made possible only by Statistica's integrated solutions
- Integrated workflow: Statistica Decisioning Platform provides a streamlined workflow that uses business rules and industry regulations in conjunction with advanced analytics to build powerful predictive models

TERADATA®

www.teradata.com

Teradata empowers companies to achieve high-impact business outcomes through data and analytics. With our focus on business solutions and with our industry-leading technology and architecture expertise, we are able to unleash the potential of great companies. This is the result of our unique approach, which is centered on three core capabilities—business, architecture, and technology.

Our Business Analytics solutions leverage data and analytics to drive business outcomes in customer experience, finance transformation, risk mitigation, supply chain intelligence, product innovation, and asset optimization. Our Ecosystem Architecture Consulting provides trusted advisors with deep industry knowledge and expertise in open source and analytics ecosystem architecture strategy and design. Lastly, our Hybrid Cloud solutions deliver Teradata's core technology in a hybrid cloud architecture implemented on a combination of private, public, on-premises, or managed cloud environments, all orchestrated to work together.

Teradata is more than just a data warehousing company—we are a data and analytics leader. We're focused on mapping the right path for our customers to yield high-impact business results through analytics at scale on an agile data foundation. To learn more, visit www.teradata.com.

WhereScape®

www.wherescape.com

With more than 700 customers, WhereScape is the pioneer and leading provider of data-driven automation software for profiling, prototyping, developing, loading, extending, and managing analytics data hubs, data warehouses, and business intelligence, advanced analytics, and big data environments.

WhereScape RED is an integrated development environment that eliminates complex hand coding, automates development, creates a simplified infrastructure, and helps deliver faster time to value and dramatically lower total cost of ownership. WhereScape's highly scalable ELT architecture leverages the processing power of the target system, further increasing the value of your server investment. Working through WhereScape RED's user interface, users simply drag and drop to develop objects, build tables, generate code to populate the tables, and create HTML documentation simultaneously from within WhereScape RED.

Organizations such as Canadian National Railway, Cornell University, Delta Community Credit Union, Tesco, and United Rentals use WhereScape automation software to accelerate the development, deployment, and delivery of fully documented analytics data hubs, data warehouses, data lakes, data vaults, and big data solutions. To learn more, visit our web site, email us at info@wherescape.com, or give us a call at +1 503-466-3979.

tdwi

Transforming Data With Intelligence™

tdwi.org

TDWI is your source for in-depth education and research on all things data. For 20 years, TDWI has been helping data professionals get smarter so the companies they work for can innovate and grow faster. TDWI provides individuals and teams with comprehensive business and technical education and research that allow them to acquire the knowledge and skills they need, when and where they need them.

TDWI advances the art and science of realizing business value from data by providing an objective forum where industry experts, solution providers, and practitioners can explore and enhance data competencies, practices, and technologies.

TDWI offers six major conferences, topical seminars, onsite education, a worldwide membership program, business intelligence certification, live webinars, resource-filled publications, industry news, an in-depth research program, and a comprehensive website at tdwi.org.