

Rapid analytics

Data mining technology drives predictive intelligence into the business process. *by Arlene Zaima and Robert Cooley*

Advanced analytics, or data mining, provides a wealth of intelligence to help businesses make better decisions. The cost limitation of data mining, however, often requires selectivity when determining what data to analyze.

Companies are collecting more and more data on their millions of accounts. A wireless provider gathers more than 450 different variables for each of its 47 million customers. Just trying to understand and analyze this volume of data can put a tremendous strain on a company's analytic infrastructure, not to mention its analyst.

Also, businesses are less satisfied with having to wait long periods for analytic results. Clearly, yesterday's tools and processes can't keep up with today's demands. Businesses need a solution that simplifies, automates and scales the data mining process to provide analytics for all of their users.

Teradata, the leading data warehouse solution provider, and KXEN, a visionary in predictive analytics, have teamed up to deliver an analytic environment that enables rapid analytics. Rapid analytics is the practice of maximizing an existing data infrastructure and data mining technology to drive predictive intelligence into the business process.

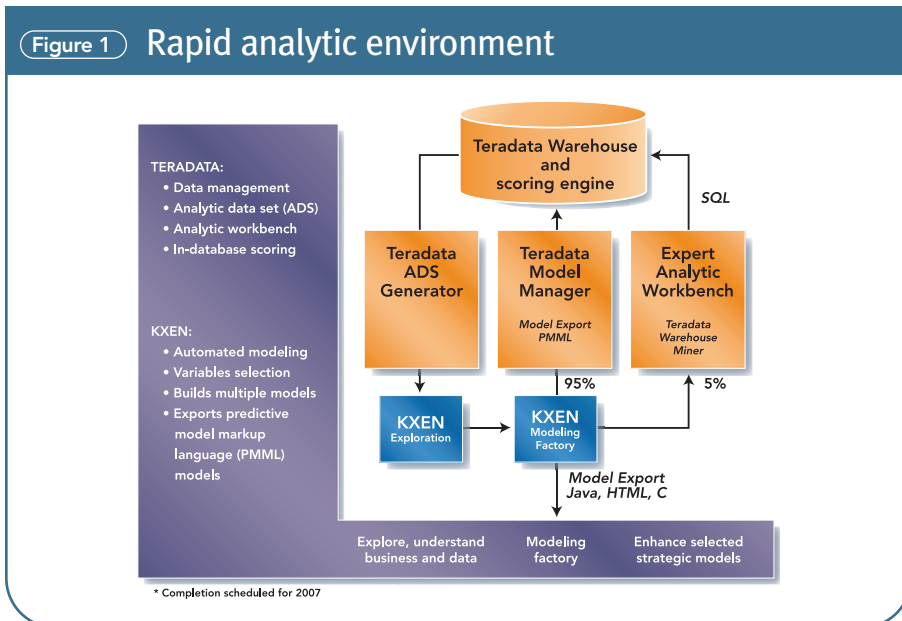
The ultimate goal is to improve the profitability of analytic projects by reducing cost and accelerating the data mining process. (See figure 1.)

Analytic environment

For effective analytics, an organization's data must be integrated, analyzed and properly prepared to meet data mining criteria. The data mining process requires an analytic data set (ADS), which can be referred to as the customer analytic record. It contains all of the data elements that will be fed into the data mining tool for analysis. Typically this is a flat, denormalized data structure where each record represents the subject modeled, with columns listing the variables.

In many analytic environments, numerous data marts are created for data mining. This is generally due to the proprietary nature of the data mining tool, siloed data strategies, antiquated processes and politics. Multiple silos may work for businesses where there is limited use of analytics, but maintaining numerous data marts has become a growing cost and source of pain.

The iterative process of building the ADS requires complex transformations, aggregations and the creation of new analytic data. These tasks cause additional concerns, such as the cost of sourcing, moving, loading and integrating the data and ensuring referential integrity. Building the ADS is not simple, and using individual data marts only adds to the difficulties. For example, when using



The combination of the Teradata Warehouse, Teradata Warehouse Miner and KXEN provides an optimal environment for analytics.

separate data stores, you must constantly move data for each iteration of analysis.

Many Teradata customers who have integrated predictive analytics into their business process are, instead, building the ADS directly into their Teradata Warehouse to accelerate analytic development. These organizations can then leverage the parallel architecture of the Teradata Database and lift data limitations.

Build once, use often, refresh anytime

When analysts create their own analytic data for their customer analytic model, they replicate tasks, creating redundancy in the process and inconsistency throughout the models. But many analytic models, such as customer segmentation, customer attrition or propensity to buy, that predict or describe customer behavior share many predictive variables. Consolidating many isolated ADSs into one that is shared across the modeling community can significantly improve the data mining velocity.

Creating the ADS accounts for at least 70% of the analytic modeling process; once this is developed, the time to build the models is reduced by half. This is the backbone for rapid analytics.

The rapid analytic environment begins with a consolidated, reusable ADS called the enterprise ADS. The enterprise ADS includes all analytic variables for models describing the same subject, such as customers. With the ADS created directly in the data warehouse with Teradata Warehouse Miner, the cost of refreshing the ADS is reduced by eliminating

data movement, leveraging the power of the Teradata Database and automating the ADS refreshes with batch SQL programs. Now analytic modelers can create relevant and accurate models with current data instead of dealing with the consolidation, integration, transformation, aggregations and the rest of the tasks required to build an ADS. Consequently, the enterprise ADS can be summed up by the following: Build once, use often and refresh on your own terms.

Automation with KXEN



Once an enterprise ADS is created, the next step is to build a set of predictive models to improve one or more business metrics—customer value, for example. The act of

building predictive models with data mining technology is referred to as “model training.”

With many traditional data mining products, a number of tasks must be performed on an ADS before the model training phase. These tasks usually require deep statistical knowledge and involve analysis of potentially thousands of variables with a different set of techniques. The modeler selects which variables to use and recodes them appropriately before the actual model training. The list of tasks and techniques is virtually endless and can take a statistician a significant amount of time, depending on the size of the ADS.

The automation of these tasks is one of the key features of the KXEN technology. Once an ADS has been created, reliable high-quality models can be trained without any intermediate steps, even when there are thousands of variables.

The real business benefit of training any one model in an automated fashion comes from using the time savings to train many more models. Obvious candidates are campaigns that were left unmodeled because of a lack of resources. But even if every cam-

paign goes out the door with some sort of targeting, there is often a lot of unrealized value by not taking advantage of regional, product or segment differences.

Conventional wisdom says a handcrafted model trained by an expert will out-perform a model trained with automated methods. The difference is we care about three high-level metrics: quality, reliability and speed. Quality relates to business value, and reliability deals with consistency of the results over time. One handcrafted model has the potential to perform at 100% quality compared with, say, 98% quality for the automated model. However, in many situations there simply isn't enough time

Rapid analytics

Combined Teradata and KXEN capabilities allow businesses to build models within days, not weeks, by:

- > Optimizing data preparation via analytic data sets (Teradata)
- > Rapidly building models (KXEN)
- > Automating model deployment (Teradata)


to train a handcrafted model. Ten models at 98% quality of a hand-crafted model will provide far more business value than a single model at 100%.

In the comparison figure 2 on page 68, the top diagram identifies the standard tasks to build an analytic model. The bottom diagram shows the optimizations by Teradata and KXEN to rapidly build multiple models.

Once the model is created, the next challenge is to get it into production and integrated into the business process. Two things must be considered: the linkage between the model and the data, and automating the refresh process of the production data.

Models trained in a proprietary environment within a separate data mart and using variables that do not exist in the data ware-

Vodafone D2 found that, if it looked at churn by product instead of overall by customer, it needed 480 models instead of 36.



house are expensive to put into production. Even if the scoring code can be translated into an appropriate language, it can take significant effort to re-create the variables from scratch.

KXEN simplifies the process without creating new variables during modeling. Rather, through predictive model markup language (PMML), KXEN produces scoring code that includes all of the processing required to go from an ADS to a set of model scores.

Many traditional data mining products produce a scoring code that excludes steps such as missing-value handling or attribute re-coding, leaving a major gap in the processing chain even if no additional variables were created.

Automating the refresh step

An enterprise ADS that contains hundreds to thousands of variables to serve many analytic models—the build-once, use-often concept—will ensure data consistency. But let’s say you have a propensity-to-buy model that needs only 15 of those variables. To avoid having to refresh the entire enterprise

One company builds and executes its **predictive analytics** in less than a day. For most companies, this procedure typically takes weeks and even months to develop.

ADS when you want to run that model, you would have to re-create a production ADS that includes only those particular variables.

This process is automated with Teradata Warehouse Miner and Teradata Model Manager. By creating your ADS with Teradata Warehouse Miner, the metadata defining the enterprise ADS is stored in Teradata. If you create a model using a partner tool such as KXEN (which supports PMML) you can export it as a PMML model. Teradata Warehouse Miner will automatically read the data definition of the PMML and traverse through the enterprise ADS producing a SQL script that refreshes only the variables included in the model.

In short, you can build your ADS once and reuse it for multiple models, signifi-

cantly reducing the data preparation phase for subsequent models. KXEN’s innovative technology has automated the bulk of the model development process by identifying which variables are best for your model and quickly building many models with minimal analyst intervention. Teradata Warehouse Miner and Teradata Model Manager automate deployment and provide a way for business users to access the model, thereby reducing the execution gap.

Integration into the business process

Several ways exist to integrate analytic models and results into the business process. An application can automatically score and use the results. But the cost of building an application can far exceed the cost of building the models.

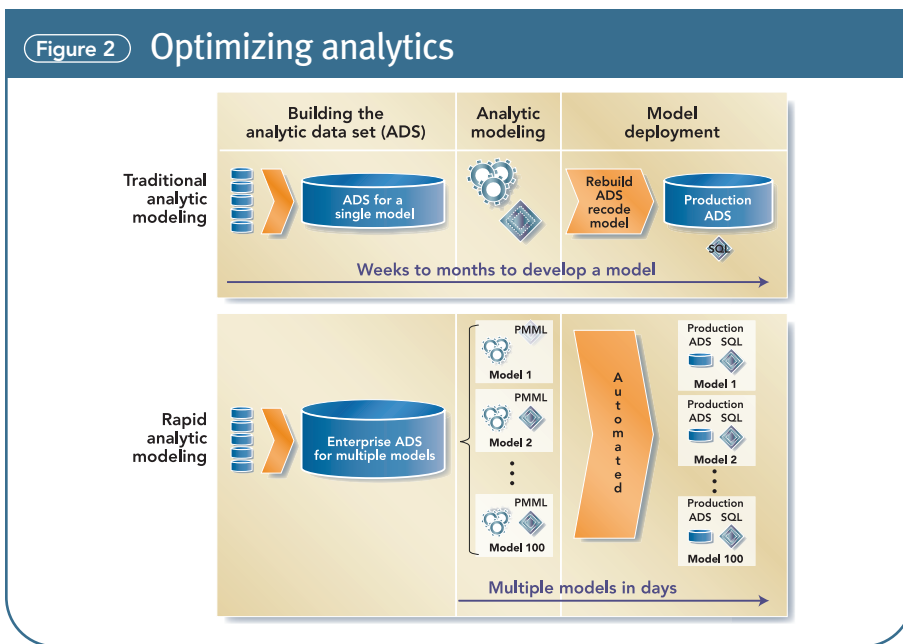
Models can also run manually; however, business users have to rely on IT or their analytic team to run the models, or they must run them themselves—an unlikely event.

Alternatively, Teradata Model Manager provides an easy-to-use interface that automatically refreshes the ADS with current data from the organization’s data warehouse. With the updated ADS, models are scored with a click of the button. In addition, Teradata Model Manager provides scheduling, tracking and a number of management features.

By reducing the cost of analytics, automating key tasks and increasing the data mining velocity, businesses can now leverage the predictive intelligence for all analytic applications and campaigns, thus enabling better, faster and smarter business analytics. **T**

Arlene Zaima is the Teradata Strategic Intelligence Program manager with a focus on predictive analytics.

Dr. Robert Cooley, vice president of North American technical operations for KXEN, is a 10-year veteran in data mining technologies.



Top: In traditional analytic mining, the analyst works autonomously in a serial process to create one model. Bottom: The rapid analytic modeling process delivers an automated environment that allows the same number of analysts to create hundreds of models in the time it took to create one with the traditional mining process.