SYBASE
An SAP Company

ZEMENTIS
Adaptive Decision Technology

IBM Business Partner

WHITE PAPER

PMML: Accelerating the Time to Value for Predictive Analytics in the Big Data Era

Big Data is a big dud without tools to model it and analyze it. Predictive analytics would not be as effective without those underlying intricate mathematical models combing through vast data sets to extract insight and value. Creating predictive models takes experience and skill. And moving the models into a database for execution takes more talent and time.

Too much time, in fact, which is why the business intelligence industry developed the Predictive Model Markup Language (PMML) standard to cut the time it takes to make the models useful to organizations. With PMML the time to value inherent in these predictive models is dramatically shortened.

Sybase IQ is the most comprehensive, highest performing purpose-built engine for business intelligence that can run these complex predictive analytics models. As the world's leader in columnar database technology deployments, Sybase IQ delivers unmatched capabilities for enterprises that are embracing the advantages of predictive analytics for business — everything from real-time fraud detection and instant online retail promotions to extremely high-volume, low-latency transaction applications and in-depth strategic reports for the corporate boardroom. These and many more business applications depend on Sybase IQ's unique capabilities to execute predictive analytics models in the blink of an eye, literally, in many cases. (See page 4: Shared Everything: A Modeler's Advantage.)

But Sybase IQ's impressive performance for predictive analytics applications is maximized when the models have been imported into its database. Up until now, there has always been a lengthy bottleneck between the application developer's completed analytics model and its intended users working environment. Frankly, it has taken far too much time to move a model from its creator's development environment over to its ultimate execution environment. This delay is costly, forcing some business decisions to remain in limbo while technical experts work to recreate the model in an enterprise data warehouse.

With its adoption of PMML, Sybase is removing one of the key barriers that slows the implementation of game-changing analytics applications. Users now have one of the fastest, automated tools available for importing models into Sybase IQ.

At their heart, these applications exploit analytical models that are built with rigorous mathematical and algorithmic precision. Predictive analytics is too complex an exercise to conduct using basic SQL queries against a data warehouse. Developing effective data mining and statistical models for predictive analytics requires talented and experienced individuals, who use advanced modeling tools to build solutions.

Once completed by the developer, a model must be changed to run in a production database that is different from the environment in which the model was built. That difference between where a predictive analytics model is developed and where it's run is the problem. Traditionally, after a model is built the next step would be to recode it from scratch for the production database while transforming the data once again to work in the new environment. Although each IT organization's capabilities and processes differ, this step could take weeks, months, even a year or longer to complete.

During that time the compelling need for the predictive model may have diminished or changed substantially. As a result, some models languish in irrelevance because they could not be deployed in a timely manner. Worse, business opportunities are often lost forever while waiting for deployment of the model.

The Predictive Model Markup Language is an industry standard that assures that compliant modeling applications and databases can export and import complex analytic models quickly. A process that was once measured by a calendar can now be tracked in many cases with a wristwatch.

PMML: A MODEL STANDARD

Based on XML, PMML is the work of the Data Mining Group, an industry-led consortium that includes companies like IBM/SPSS, MicroStrategy, SAS, and many others. The PMML standard has received contributions from technology heavyweights such as Microsoft, SAP, BusinessObjects, and Tibco. The standard is well-established, field-proven, and robust, having been in the marketplace for 10 years and, as of December 2011, is now in version 4.[1]

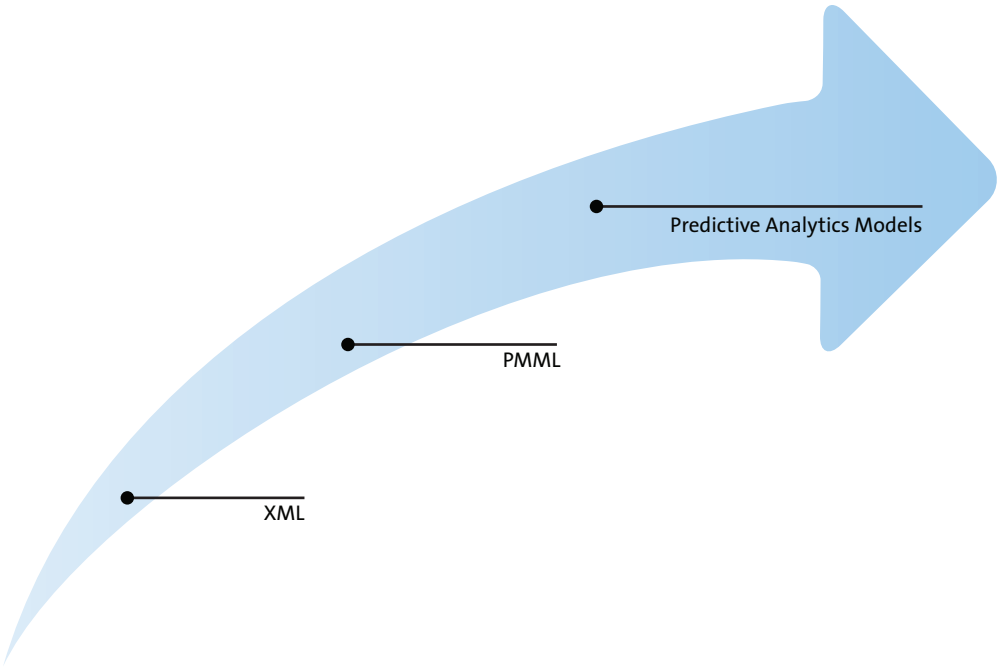

Figure 1. Based on the industry standard eXtensible Markup Language (XML), PMML will propel the deployment of predictive analytics models in a variety of markets.

As the de facto standard markup language used to represent statistical and data mining models, PMML reduces the time-consuming, iterative development process that all-too-often engulfs analytics projects as they move from model creation to execution in the production database. By automating the recoding of the model to match the data warehouse execution environment, model implementation time is dramatically reduced. Another, more subtle reason for the faster deployment time is that because of the widespread adoption of PMML within the industry, analysts can continue to use the tools they are familiar with to create models as fast as possible.

The PMML standard is robust and comprehensive.[2] It includes a **header** that describes the PMML document itself such as the version of the standard used. It has a **data dictionary** that defines the fields used in the predictive model and a **mining schema** that lists the fields specific to a certain model. The standard covers a model's **taxonomy**, **statistics**, **targets**, **output**, and **functions.** There are also a variety of **data transformations**, such as normalization, aggregation, and value mapping. And, of course, there is the **model** itself, which will include attributes such as its name, function, and algorithm among other elements.

The PMML standard is ready for the world of Big Data. It offers several arithmetic and logical operators that can be combined to represent complex pre-processing steps to assure the data in the development phase is representative of the real world where the model will execute. In our era of massive data sets this is a significant benefit. The data used in the model creation must be complete enough to be able to handle the large data sets extant in a production data warehouse to be of true business value.

Exporting a PMML-compliant model from the plethora of tools on the market is simple. Once the model has proven itself in the development environment, the analyst uses built-in steps to output PMML. Although the steps vary from tool to tool, it's often as simple as pulling down a menu and choosing "Save As..." to invoke the PMML export function.

Figure 2. Developers can use their tools of choice and export their models into a PMML standard file, then use the Sybase Universal PMML Plug-In to run the model in Sybase IQ.

Sybase implements its support for PMML through the Universal PMML Plug-In for Sybase IQ. It automatically recognizes compliant models that support all versions of the PMML standard from 2.0 to the current release. The software was developed by Zementis Inc., a leader in the PMML community.[3] The depth and breadth of the Plug-In is considerable. (See box: Depth and Breadth in Models.) Analysts can create many different kinds of models with a wide variety of development tools and effortlessly export their models to a Sybase IQ production database and almost immediately begin seeing results.

The speed at which models can now be deployed means that as business conditions change models can be updated and quickly put back into operation. For example, after a corporate merger or acquisition among, say, retailers, the analytic model used by management to estimate overall sales can be adjusted quickly to account for sales forecasts that affect stores within the new company that now have overlapping geographies.

Depth and Breadth in Models

With the Universal PMML Plug-in for Sybase IQ, analysts can build a wide range of predictive analytics models for high performance scoring, including:

- Decision Trees for classification and regression
- Neural Network Models: Back-Propagation, Radial-Basis Function, and Neural-Gas
- Support Vector Machines for regression, binary and multi-class classification
- Linear and Logistic Regression (binary and multinomial)
- Naïve Bayes Classifiers
- General and Generalized Linear Models
- Cox Regression Models
- Rule Set Models (flat decision trees)
- Clustering Models: Distribution-Based, Center-Based, and 2-Step Clustering
- Scorecards (point allocation for categorical, continuous and complex attributes)
- Association Rules
- Multiple Models (model composition, ensembles, and segmentation)

It also implements the definition of a data dictionary, missing and invalid values handling, and data pre-processing.

PMML lets you easily share predictive analytic models between different applications. That is, you can train a model in one system that might be used to project net usable product from an assembly line, then express it in PMML, adjust it and test it in development, and then very quickly move it to another system where you can use it, for example, to predict the yield of a product in a different manufacturing process.

PMML is also an extremely flexible standard, designed to meet the needs of modern business intelligence professionals employing advanced modeling techniques. For example, because each predictive model has its own advantage for a given problem, sometimes one model won't be comprehensive enough for the solution. So PMML lets you build applications with multiple models, including model ensembles. Each model can be exported via PMML and executed in the production environment as intended within the application.

One more benefit of using PMML is staff turnover. Sometimes the expertise in creating a particular predictive analytics model is locked inside the mind of its developer. Without being able to express the model in a common, well-understand standard, the knowledge of how a given model functions disappears once the expert who created it leaves an organization. If the model needs to be changed to meet new business conditions, it can be all-but-impossible to alter it. With PMML, the knowledge to create the model is saved for business continuity and posterity.

SHARED EVERYTHING: A MODELER'S ADVANTAGE

The architecture of Sybase IQ gives it a distinct advantage over other approaches when executing analytical models. By leveraging the unique PlexQ™ shared-everything, massively parallel processing (MPP) columnar architecture, analytical models never encounter resource boundaries. The compute, memory, and storage capabilities of the system can scale to match performance requirements of any model.

As such, model developers do not need to consider finite resource environments when building PMML-compliant predictive analytics applications. Unlike shared nothing architectures, PlexQ utilizes a shared everything approach that dynamically manages and balances query workloads across all the compute nodes in a PlexQ grid. PlexQ's automatic workload re-balancer aggressively works to avoid contention among users for system resources, thereby providing predictable high performance and resource efficiency for a broad spectrum of concurrent workloads.

At the heart of PlexQ is its Distributed Query Processing capability. DQP can improve the performance of a query by breaking it up into pieces and distributing those pieces for concurrent execution across multiple Sybase IQ servers in a grid. For example, with PlexQ's DQP design, you can scale out storage requirements independently while concurrently scaling out the volume of users and their demands on the system, all while retaining performance metrics established in organizational service-level agreements (SLAs).

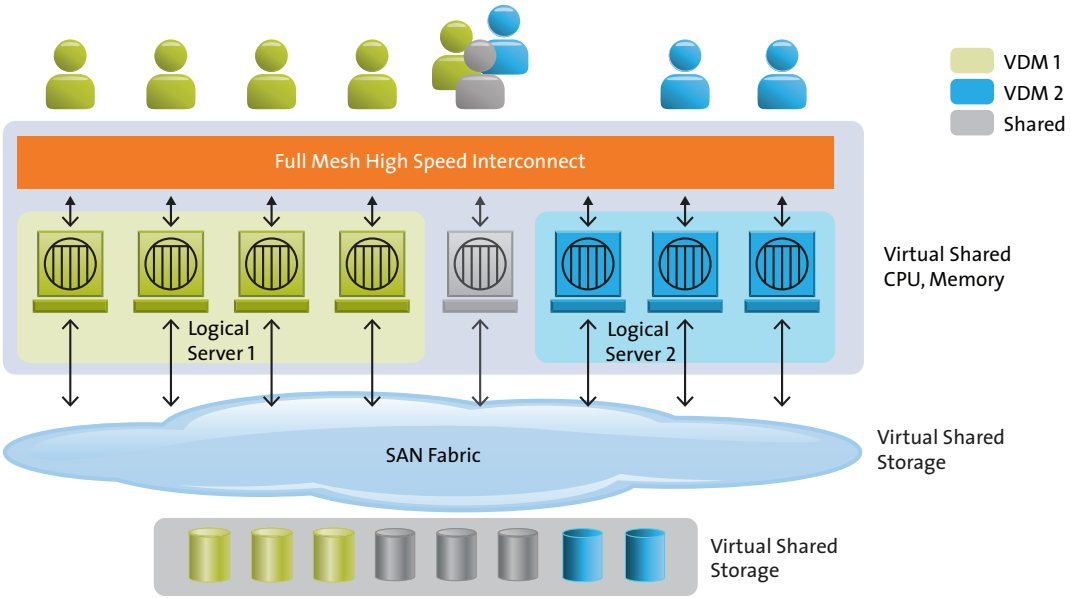

Figure 3. The shared-everything MPP architecture of the Sybase IQ PlexQ design can scale out and up for any PMML predictive analytics model.

USE CASE: HEALTHCARE

PMML can speed the deployment of analytics models across all industrie s— from telecommunications and manufacturing to financial services and retailers. These models deliver insight to virtually every department up and down each enterprise's organization chart. Looking at just one market, healthcare, the benefits of the PMML standard makes more than good business sense by speeding up how fast analytics solutions get into management's decision process, it literally can accelerate life-saving choices.

As noted, Sybase's support for the PMML standard embraces numerous models used in predictive analytics applications. Artificial neural network (ANN) models is just one example with a proven track record in the healthcare field.[4] Medical researchers and practitioners use ANNs to fuse data from multiple cardiovascular sensors to get an overall view of a patient's system. ANNs also are excellent models for quickly analyzing medical images to determine the existence of tumors and other anomalies.[5]

A shared-everything architecture is essential to enterprises where a strategic goal is to maintain a central repository of data — a single version of the truth. Without the scaling limitations of other architectural designs, PlexQ technology permits analysts in every department inside an organization to build complex, resource-intensive analytical models that can be used frequently by as many users as necessary on massive shared data sets. These huge data sets are common in numerous markets such as healthcare.

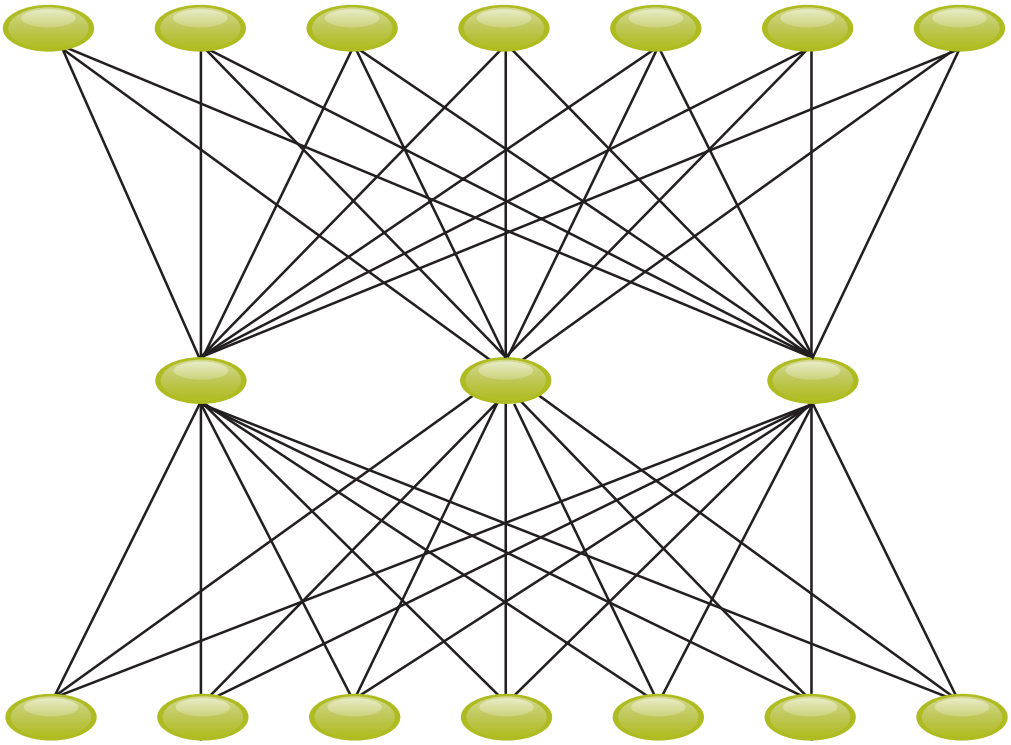

Figure 4: A simple artificial neural network with an input layer on the bottom, a hidden layer in the middle, and an output layer on top.

Real-time predictive analytics is also possible. Doctors and nurses in intensive care units are already applying predictive analytics models to track vital signs on premature babies from numerous sensors. The analytics software can predict potential problems for the newborns before a healthcare worker would notice anything from glancing at various discrete thresholds in the data.[6]

Analytic models are equally valuable for healthcare administrators. For example, the Patient Protection and Affordable Care Act of 2010 requires Medicare and Medicaid services to implement a new readmissions policy that favors more outpatient treatment for certain medical conditions and will reduce reimbursement payments to hospitals for some readmissions of patients.[7] According to a report in the *New England Journal of Medicine*, unplanned patient readmissions cost Medicare more than $17 billion in a single year. This new policy begins October 2012. Ideally, hospitals will be able to predict which patient groups fall under the new readmissions policy and how they can adjust treatment to assure reductions in readmissions. Developing and quickly deploying analytics applications that, for example, can identify which portion of their patients will respond best to outpatient services would be extremely useful to hospitals seeking to meet the government's aggressive new readmission standards.

Fraud detection is another area where analytics can help improve the economics of healthcare. The Federal Bureau of Investigation says healthcare fraud costs the U.S. economy $60 billion annually. The National Health Care Anti-Fraud Association (NHCAA), which estimates health care insurance fraud to be around 3% of industry expenditures annually, says, "The majority of health care fraud is committed by a very small minority of dishonest health care providers."[8] That means the possibility of identifying those few criminal transactions among so many legitimate ones is beyond human detection. Advanced, predictive analytics applications are the ideal way to catch fraudulent claims before they are settled.

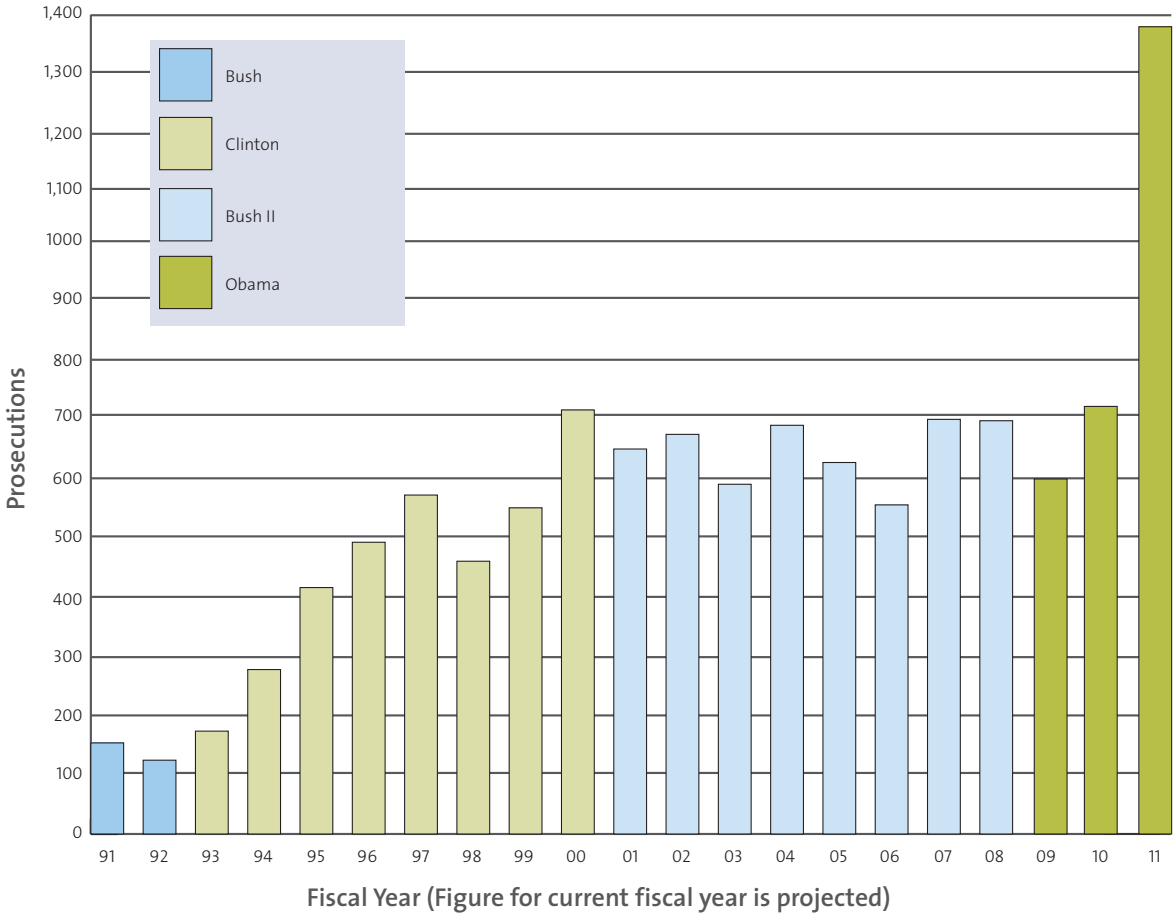

Figure 5: Federal prosecutions for healthcare fraud over the past 20 years in the U.S.

As apparent in Figure 5, the U.S. government has begun to aggressively prosecute the perpetrators of healthcare fraud and get restitution for agencies such as Medicare. And, according to the NHCAA, for every $2 million invested by private insurers in fighting fraud, organizations get $17 million in return.[9]

Even esoteric applications in the field will benefit from predictive analytics. The *New York Times* reports that the U.S. government will begin analyzing Twitter feeds in Latin America as an experiment to identify and track diseases that could develop into a pandemic.[10] By reviewing the posts of millions of people in the region using the microblogging service the researchers' hypothesis is that they can build an analytics model that will predict if and when a major contagion will develop. With this knowledge, health experts hope to be able to stop virulent outbreaks at the local level before they can become global catastrophes.

In all these situations, speed to deployment is critical. Whether administrators detecting fraudulent claims or pushing ahead to meet pressing government readmission timelines, health professionals trying to save lives in the ICU, or disease trackers monitoring deadly infections across the globe, time to success is everything. Without PMML, the time necessary to recode the model to work in the new environment would be unacceptable. After all, it's not just money at stake in healthcare. In some instances, time is of the essence because another premature baby may die unnecessarily or a pandemic may even be getting underway.

[1] Data Mining Group (2011). PMML version 4.1, http://www.dmg.org/pmml-v4-1.html.

[2] Predictive Model Markup Language. Wikipedia: The Free Encyclopedia. Wikimedia Foundation.

[3] A. Guazzelli, W. Lin, T. Jena (2010). PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics. CreateSpace (available on Amazon.com – http://www.amazon.com/dp/1452858268).

[4] Christos Stergiou and Dimitrios Sinanos, *Neural Networks*, http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Conclusion.

[5] Harjit Singh, editor, *Artificial Neural Networks in Medicine and Biology,* Springer-Verlag, 2000.

[6] Alex Guazzelli, *Predictive Analytics in Healthcare*, Nov. 2011. http://www.ibm.com/developerworks/industry/library/ind-PMML3/.

[7] Robert Kocher, "Hospital Readmissions and the Affordable Care Act," *Journal of the American Medical Association*, October 2011. http://jama.ama-assn.org/content/306/16/1794.extract.

[8] NHCAA, *The Problem of Health Care Fraud*, http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_centr&wpscode=TheProblemOfHCFraud.

[9] Coalition Against Insurance Fraud, *Go Figure: Fraud Data*, http://www.insurancefraud.org/healthinsurance.htm.

[10] John Markoff, "Government Aims to Build Data Eye in the Sky," *New York Times,* October 10, 2011. http://www.nytimes.com/2011/10/11/science/11predict.html?pagewanted=all.

For more information,
visit **www.sybase.com**

Sybase, Inc.
Worldwide Headquarters
One Sybase Drive
Dublin, CA 94568-7902
U.S.A
1 800 8 sybase

www.sybase.com

SYBASE®
An SAP Company