Accelerating Business Analytics

Combining Grid Computing and In-Database Processing to Solve Big Data Problems
EXECUTIVE SUMMARY

In a challenging economic environment, reactive decision-making is not sufficient to sustain competitive advantage. Corporations are collecting more data, dealing with more complex business issues and experiencing heightened global competition. There has never been a greater need to transform data assets into innovation and maximize the productivity of resources to drive sustainable growth. In this Tech Dossier, you will learn how combining Grid Computing and In-Database processing drives proactive, evidence-based business decisions and promotes agile strategies to anticipate and manage change.

Previously, large data sets, complexity of data relationships, and modeling complexity made it impossible to solve the highest-value business analytic computations quickly and efficiently. High Performance Computing, which combines In-Database processing from SAS and Teradata® and Grid Computing from SAS and Platform Computing, solves business problems once thought to be unsolvable. With High Performance Computing, business users see exponential performance gains, an increase in productivity and the ability to streamline their analytics processes. Together, In-Database processing and Grid Computing enable a revolution in business analytics and offer customers immediate competitive differentiation and cost savings.
What drives the need for High Performance Computing?

Most companies today have plenty of data. Creating intelligence and glean ing real insight and value from this data is what continues to elude organizations. Despite years of talk about scorecards and metrics, gut feelings and experience are often still the guides for making important, sometimes critical decisions. This is in spite of current research showing a clear relationship between business performance and the use of business analytics.

Business analytics is the application of analytical techniques to understand, analyze and solve business issues and problems. It provides organizations with a framework for decision making. Business analytics helps organizations solve complex business problems, improve performance and drive sustainable growth through innovation. It anticipates and plans for change while managing and balancing risk.

Using business analytics, one state government avoided $14 million in fraudulent Medicaid insurance claims and discovered an additional $27 million in fraudulent claims that led to the indictments of several suspects. A private insurer saved $11 million in one year alone when using a finely tuned solution. The benefits of business analytics are real and measurable.

Conversely, improper application of business analytics can be costly. A worldwide mortgage bank’s recent campaign lost in excess of one billion dollars ($1B) because they used an outdated analytical model that targeted the wrong mortgage prospects and customers.

Demands on IT to deliver a scalable environment and faster response times continue to increase as more data, users and applications are being utilized to solve complex business problems. Enterprise analytics departments drive the need for:

- Workload management and prioritization.
- Governance of the entire IT environment.
- Optimized performance across all business processes.

High Performance Computing delivers faster business analytics that can have a powerful effect on an organization’s performance. True High Performance Computing melds Grid Computing with In-Database processing, resulting in a powerful solution that enables optimized, timely decision-making.

Clearly the rewards and risks are significant. Not doing business analytics is a missed opportunity few companies can afford.

However, organizations often have questions around their business analytics initiatives:

- Are there technologies available to improve the performance of data and analytics processing?
- Can you gain orders of magnitude improvements by parallelizing your data and workload?
- Will business analytics jobs run faster if they are closer to the data?


High Performance Computing provides solutions to all of these problems. Grid Computing (see sidebar) with SAS and Platform Computing manages and distributes multiple workloads in parallel across a network to reduce the time to process the data. In-Database processing for analytics with SAS and Teradata (see sidebar on page 5) leverages the power of the database for analysis without having to duplicate and migrate the data. Taken separately, each technology can improve business analytics processing tremendously, and the combination of the two has powerful ramifications for enabling superior and more timely decision-making.

**GRID COMPUTING**

With Grid Computing, business analytics are done faster, better and cheaper. Grid Computing distributes computing-intensive SAS applications across hardware resources for improved performance and business continuity. Enterprises can create a managed, shared environment to process large volumes of SAS programs more efficiently.

Grid Computing provides critical capabilities that are necessary for today’s business analytics environments including:

- Workload management and job prioritization.
- High availability.
- Parallelization of business analytics jobs for improved performance.

Workload management allows users to share resources in order to most effectively balance workload and meet service levels across the enterprise. Business analytics jobs benefit by having workflows execute on the most appropriate resource and multi-user workload is balanced within the grid to enable the optimum usage of resources. With ever increasing numbers of analytics users in an organization, it is also critical that resources are available for critical tasks. Grid Computing provides the capability to prioritize jobs, which enables critical jobs to start immediately instead of waiting in a queue. Low priority jobs can be temporarily suspended to enable critical jobs to be immediately processed.

The presence of multiple servers in a grid enables jobs to run on any available server, and, if a server fails, its jobs can be seamlessly transitioned to another server, providing a highly available business analytics environment. High availability also enables IT to perform maintenance on specific servers without interrupting analytics jobs as well as introduce additional computing resources without disruption to the business.

Grid Computing provides standardized workload management to optimally process multiple applications and workloads to maximize overall throughput. In addition, Grid Computing can parse large analytics jobs into smaller tasks that can be run in parallel, on smaller, more cost effective servers with equal or better performance than seen on large symmetric multi-processor (SMP) systems. Parallelization of analytics jobs enables organizations to improve processing speeds by orders of magnitude and deliver exceptional improvements in analyst productivity.

By combining the power of workload management, job prioritization and high availability, Grid Computing accelerates performance and provides enterprises with more control and utilization of their business analytics environment.

Bank of America’s (BoA) Corporate Investments Group (CIG) manages the bank’s available-for-sale real estate portfolio and is responsible for modeling and calculating the probability of default (PD) on the 9.5 million mortgages it services. In addition to calculating PD, CIG also calculates the market value, prepayment speeds, and sensitivity to change in interest rates and then hedges these risks for the $19 billion mortgage-service-rights assets.

At BoA, 9.5 million mortgage portfolios equates to 30 terabytes (TB) of source data that must be processed, scored, prepared and summarized before the correct models are selected. This is normally a time-consuming series of activities in a business where time is critical. By implementing a platform comprised of SAS® Enterprise Risk Management in a SAS Grid Computing environment, Bank of America was able to:

- Cut the calculation of loan-default probabilities from 96 hours to 4 hours.
- Reduce the time to score 400,000 loans from 3 hours down to 10 minutes.
- Decrease processing time by 90 percent.

Using Grid Computing has enabled CIG at Bank of America to minimize losses by yielding timely results around defaults and allowing the bank to stay ahead of the market. Grid Computing ensures business continuity for these mission-critical functions and handles new growth opportunities for the bank’s loan portfolio.
Grid also provides a standardized management system. Administration of the entire compute environment is simplified and routine maintenance tasks like installing software patches or taking machines out of service for routine maintenance can be accomplished without interruption to business users.

Often, business analytics are bogged down due to infrastructure issues: servers are unavailable, there are not enough CPU’s to run the ever-increasing volume of jobs, or other workload and resource constraints get in the way. Grid Computing enables users to automatically optimize their compute environment to provide resources as needed and prioritize workload in order to meet the most critical demands first.

IN-DATABASE PROCESSING
In-Database processing is ideal for two key scenarios. The first scenario is for “big data” enterprise analytics, where the sheer volume of the data involved makes it impractical to repetitively copy it over the network. The second scenario is in complex, organizationally diverse environments, where varying business communities need to share common data sources, driving the need for a centralized enterprise data warehouse. Companies use corporate data governance policies to promote one single version of the truth, minimizing data inconsistency and data redundancy, and aligning data access needs to common business usage.

By definition, In-Database processing is ideal in both of those cases, pushing the analytic processing into the database, which is optimized for high-performance data access, and organized for consistent business utilization. This allows companies to:

- Streamline their analytic processes to gain efficiencies.
- Provide a consistent view of the truth for better outcomes.
- Get faster results through scalable, high-performance data processing.

Analytic users operating in big data environments can significantly benefit from leveraging a centralized enterprise data warehouse for managing their data.

One benefit of having an enterprise data warehouse such as Teradata is a flexible but governed environment that can be used to simplify analytic processes and promote collaboration within an organization. With this approach, multiple analysts can share an “analytic sandbox”, each creating varying datasets that suit their own particular business purpose, but over time, evolve to create a rich, permanent and sharable analytic storage.

The efficiencies of an analytic sandbox greatly reduce the time required to perform repetitive data exploration and data preparation steps. It encourages the adoption of shared data dictionaries for commonly accessed models and enables the analyst to remain focused on model development tasks.

By leveraging In-Database processing, analytics users are leveraging the power of the database platform, designed specifically for highly efficient data access methods, even with enormous million or billion row datasets. SAS enables In-Database processing for a set of core statistical and analytic functions and model scoring capabilities within Teradata, leveraging the Massive Parallel Processing (MPP) architecture for scalability and performance of analytic computations. These capabilities allow analytic computations to be run in parallel across potentially hundreds or thousands of processors. Parallel execution greatly accelerates the processing time for analytics calculations, providing very significant performance gains for faster results.

Finally, by leveraging In-Database processing, SAS analytics can be applied to where the data resides; thus, eliminating the process of copying the data over the network for processing on separate servers. For enormous million and billion row datasets, In-Database processing is an efficient approach to address super-large and complex calculations, without resorting to sampling or analysis on smaller dataset.

The ability to streamline the end-to-end analytic process (i.e., data preparation, data exploration, modeling and scoring) and parallelize processing opens the door to resolve new classes of business problems that have been previously unimaginable or technically infeasible.
A top 10 multi-national bank headquartered in Europe with major international presence and focus on Asia was unable to transform its overwhelming amount of data into a competitive advantage. High data management costs and inefficient processes resulted in models that targeted the wrong mortgage prospects.

As the financial sector struggled, the bank acknowledged the need to update the way it targeted its customers. It needed a more agile data model that better reflected the changing market for mortgage loans. However, moving those large volumes of data for processing made it impractical to develop new scoring models. The bank adopted In-Database processing and experienced dramatically improved results and cost savings. Three months after implementing In-Database processing, the bank achieved $6 million in savings, a 55% increase in productivity and 55% reduction in costs. Other benefits included:

- Reducing rescoring process times from a month to hours and/or seconds.
- Enabling rescoring for more than 10 million customers to go from a monthly to daily process.
- Automated notification when the scoring results change over time, indicating model decay.

Using In-Database processing enabled the bank to execute complex analytic computations closer to the data. Now processing times are much faster and organizations require less server and network infrastructure to move data between storage and memory.

As you can see, In-Database processing accelerates enterprise analytics performance, improves the manageability of data and significantly increases processing scalability.

**FIGURE 1: Architecture of combined Grid Computing and In-Database processing solution**

**EMBRACING THE COMPLETE PACKAGE**

The combination of Grid Computing and In-Database processing enables organizations to streamline their entire analytical business process. Where enterprises were previously only able to analyze a limited set of data they can now analyze all available data, including the inherent complexity of the data relationships. By enabling rapid and frequent testing and validation of models against new data, poorly performing models can be more quickly replaced. Analysts can build better, more robust models and move them into production faster with increased confidence levels.

Combining Grid Computing and In-Database processing also provides complete governance of both the data and
Complementary Innovation: Grid Computing and In-Database Processing in Action

A major American financial services company had a goal to create a scalable and highly-available business analytics environment and deploy new models faster to respond rapidly to changing market conditions. To better support their directly issued consumer credit cards, they needed to refresh their data management and analysis system. In order to solve their scalability issues and install fault-tolerant architecture while providing faster tools for their analysts, they chose to institute a phased implementation of the SAS Grid and In-Database solutions.

The organization’s business analytics environment supported analysts from the risk, marketing, and payment services divisions. The organization’s existing arrangement, an off-site hosted solution, not only did not scale well but was also a single-point-of-failure. Implementing Grid Computing provided scalability and high availability.

The IT department is now able to update software and restart hardware without any interruptions to business users. In addition, the company is realizing documentable savings through workload balancing designed to maximize throughput without allowing a single department to capitalize all of the resources.

Building on existing investments in both SAS and Teradata technology, another goal is to minimize data movement and redundancy, while accelerating time-to-value for analytics. Understandably, the firm needed to capitalize on its ability to proactively scale the environment, while also keeping infrastructure costs to a minimum. The company co-located the grid with the data warehouse platform, which consistently yields productivity improvements including completing a comprehensive marketing model in three and a half hours rather than nine hours.

The financial services organization is now in the process of extending their highly available SAS Grid environment to include the SAS In-Database processing analytics platform. Once implemented, they will have a dependable, integrated environment for predictive analytics. This setup enables the company to enhance the performance of model scoring and allows more analytic models to be processed faster.

For some situations, large companies may have dozens or hundreds of statistical analysts, all independently competing for time on the High Performance Computing tier to process their analytics jobs. The parallel nature of the grid environment enables companies to cost effectively expand their pool of computing tier resources, significantly increasing their ability to run more jobs quicker, in parallel.

In other situations, large companies have enormous data stores to analyze, typically with millions or billions of records. For these big data cases, the utilization of In-Database processing analytics empowers users to run their calculations much faster by pushing them to the database tier. This minimizes large, slow data transfers over the network and leverages the parallelization of the database platforms.

There are many situations where both of the above apply — companies with hundreds of analysts, which have enormous data stores to analyze. In these cases, these users will benefit from both Grid Computing and the systems within the business analytics environment. Improved workload management and prioritization enable fine-grained control of business analytics job placement and performance.

When an organization decides to implement a business analytics environment that combines Grid Computing and In-Database processing, they often choose to deploy the technologies in a phased approach. Enterprises can implement either technology independently to steadily improve the business analytics environment without dramatically impacting business critical processes and minimizing risk. Which technology should be implemented first depends on the needs of the organization.
In-Database processing technology. The combination of these complementary technologies provides companies with enhanced access to business analytic computing resources, as well as faster execution of big-data jobs, running in the database. Analysts are able to run more jobs, and get the individual results back quickly. By adopting both technologies, IT and business organizations are able to gain the following benefits:

**IT:**
- **Optimize resources** — leveraging the software and hardware to analyze more data and provide faster results for any data intensive analytical jobs
- **Support incremental growth** — as organizations collect more data, leverage the power of In-Database processing and Grid Computing to grow infrastructure and costs in a controlled manner
- **Lower costs** — having efficient processes lower IT costs to manage the system and data
- **Minimize data movement** — eliminate data duplication, maintenance and management
- **Guarantee uptime and continuity** — with faster throughput and better use of the system

**Business:**
- **Better decisions** — the ability to manage and analyze more data and related data fields will provide more insight into difficult and complex decisions
- **Decrease time to results** — offering faster results for business executives to make informed decisions
- **Maximize analyst productivity** — significantly increase the number and quality of models built by analysts
- **Solve big data problems** — the ability to address problems once thought to be unsolvable.

**AN INNOVATIVE, STRATEGIC APPROACH**

Grid Computing and In-Database processing technologies drive the vision of the analytic enterprise. These technologies enable organizations to operate more strategically and bring innovation to the forefront of the business.

One of the key benefits of a combined environment is the ability to handle and manage data and user growth in an efficient and effective manner. With Grid Computing, servers are always available for immediate access to computing resources. With In-Database processing, the duration of complex calculations is minimized. With enterprise data warehouses, data is structured in a reusable, easy-to-access manner, shortening the end-to-end analytic processes significantly. Together, these efficiencies enable analysts to iterate more frequently, experiment more, share insight, and attempt to solve never-before-possible business problems.

Innovations like In-Database processing have helped lower the cost of experimentation tremendously. Michael Schrage, a research fellow at the MIT Center for Digital Business, contends that, “thanks to IT, the cost of experimentation in business has collapsed.”

Business Analytics expert Tom Davenport says this is exactly what constitutes the growing trend of “Agile Analytics” — an iterative, multi-pass approach to analytics.

For one financial services firm (see sidebar on page 7), the combined solution allowed them to initiate a cultural change in the relationship between business users and IT. By bringing together its analytic tools with an enterprise data warehouse, the firm aimed to give analysts in its customer information and decision-management group the ability to generate new insights and understanding of critical market and consumer trends.

The resulting leading-edge decision-support capabilities helped users identify innovative new revenue and customer service opportunities. This integrated platform has resulted in significant savings through cost takeout, revenue generation, enhanced productivity and reduced IT run rates. All of which help keep the firm at the top of its industry.

**CALL TO ACTION**

In the 2010 “State of the CIO” survey by IDG’s CIO magazine, technology leaders showed a strong and increasing focus on business gains. Long-term strategic thinking, planning, collaboration and influence continued to top the list of IT business priorities. But evidence of a significant change was noted: the biggest increases came in the shift of CIOs to begin looking externally to the customer and commercial opportunities and aligning with business goals.

For forward-thinking CIOs and business unit heads, business analytics can play a key role in achieving their goals. Informed decision-making starts with good data. It is IT’s job to leverage their unique technology insights to make that data actionable, available, secure and of the highest quality.

As high value, centrally managed data and relevant insights are incorporated into the different types of business decisions, the demand for consistent data governance, utilization of consistent business terms and applications, and transparency are applied to business processes.
A combined Grid Computing and In-Database processing offering goes a long way toward easing the administrative overhead involved with operationalizing complex analysis quickly and efficiently. It goes beyond that. The complementary usage enables the business to think differently. IT can concentrate more on data generation, distribution and management, rather than on supporting one-off analyst queries and sample data requests. Business and IT can work jointly to identify business needs and implement appropriate High Performance Computing solutions quickly.

CONCLUSION
High Performance Computing enables organizations to drive proactive, evidence-based business decisions and agile strategies to anticipate and manage change in a volatile market. The combination of Grid Computing and In-Database processing and provides documented business results:

- Rapid return on investment.
- Lower total cost of ownership.
- Reduced complexity.
- More efficient IT resources.
- Optimized use of an organization’s most precious resource: its data.

Business Analytics is a growth industry. A recent Network World magazine survey (June 2010) reported that spending on BI and analytics is on the upswing, with 57% of respondents anticipating a bigger investment in the next year.3 And with cloud-based business analytics solutions on the horizon, the time is now for progressive companies to embrace a strategic outlook that allows the business to go beyond spreadsheets and query-wait-and-repeat scenarios.

By combining SAS Business Analytics with technologies from Teradata and Platform Computing, there is a real opportunity for forward-thinking organizations to take a revolutionary step that taps existing infrastructure, data and technology in a new and productive way. Ultimately, with Grid Computing and In-Database processing, businesses can leverage their most valuable internal resource — data — to make better decisions more quickly and gain market share.


SAS® Grid Computing
A flexible, centrally managed computing environment for high availability and faster processing
Read more

SAS In-Database and Teradata
The strength of the Teradata platform combined with the powerful analytics of SAS creates a unique value proposition for customers. Read more

ADDITIONAL READING
Raising the Bar on Business Analytics: Innovation Powered by Grid
Investment in business analytics solutions can improve competitiveness
Read the IDC White Paper

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.
® indicates USA registration. Other brand and product names are trademarks of their respective companies.
All rights reserved. 000000_000000.00.09
Teradata and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide.