TDWI

MONOGRAPH SERIES

OCTOBER 2008

Beyond Reporting: Requirements for Large-Scale Analytics

By Wayne W. Eckerson

Director, TDWI Research

The Data Warehousing Institute

tdwi

THE DATA WAREHOUSING INSTITUTE

About TDWI Research

TDWI Research provides research and advisory services to business intelligence and data warehousing (BI/data warehouse) professionals worldwide. Unlike other research or analyst firms, TDWI Research staff focuses exclusively on BI/data warehouse issues and teams up with industry practitioners and TDWI Faculty members to deliver both a broad and deep understanding of the business and technical issues surrounding the deployment of BI/data warehouse solutions. TDWI Research delivers commentary, reports, and inquiry services via TDWI's worldwide Membership program and provides custom research, benchmarking, and strategic planning services to both user and vendor organizations.

About TDWI

The Data Warehousing Institute (TDWI), a division of 1105 Media, Inc., provides in-depth, high-quality education, training, research, and certification for business intelligence (BI) and data warehousing professionals worldwide. TDWI can help your BI team stay abreast of new and emerging trends and techniques and gain the skills they need to deliver effective BI and data warehouse solutions. Through our Membership program and regional chapters, TDWI can also help you and your team establish a network of peers in the industry to whom they can turn for assistance and advice in career development and to solve thorny technical and organizational problems. www.tdwi.org

About Aster Data Systems

Aster Data Systems is a proven innovator in analytic databases for frontline data warehousing—bringing deep insights on massive data analyzed on clusters of commodity hardware. Co-founded by three colleagues in the Stanford Computer Science PhD program, the Aster *n*Cluster database provides patent-pending innovations in performance, availability, and in-database analytics. Aster is headquartered in Redwood City, California, and is backed by Sequoia Capital, Cambrian Ventures, and First-Round Capital. For more information, visit www.asterdata.com or call 650.232.4400.

Executive Summary

To increase the value of their data investments, many organizations are eager to move beyond the delivery of basic reporting capabilities. They want to empower business analysts—who sit at the intersection of data, process, and math—to create data-driven applications that deliver bottom-line insights from massive volumes of data.

There are many challenges to becoming an analytics-driven company. Organizationally, companies must overcome executive perceptions that analytics is too complex, expensive, and abstruse to offer sustainable value. They also need to challenge the traditional approach of running analytic applications on a specialized analytics server or having analysts explore data sets by downloading them to a local server or desktop application.

To apply analytics against large volumes of data, organizations need a purpose-built analytics platform that runs on a massively parallel processing environment and supports custom-built analytical programs that can be invoked via SQL. In addition, the analytic platform should enable IT departments to create sandboxes in the corporate database that give analysts free reign to add their own data and run high-performance analytics without impeding performance of other users on the system.

The Perfect Storm: Analytics Meets Big Data

Beyond Reporting. For the past 5 to 15 years, organizations have been building data warehousing and reporting environments to deliver a consistent set of enterprise information to decision makers. For the most part, these repositories and reports have helped organizations improve the accuracy, comprehensiveness, and freshness of data they use to understand business activity, improve key processes, and optimize performance. Most organizations are ready to move beyond reporting and invest in analytics—a broad set of techniques and technologies that enable organizations to tease patterns and relationships from data.

Analytics uses math and statistical functions to detect patterns in data to assist—and sometimes automate—business decisions.

Business analysts use these techniques to explore data in an ad hoc fashion to understand the cause of a problem or unearth new opportunities. Often, they use complex mathematical and statistical operations to detect patterns and relationships in the data to assist—and sometimes automate—business decision making. Sometimes, the most profitable or useful of these analyses get baked into formal programs called analytic applications. (See Figure 1 for a list of the top categories of analytic applications.)

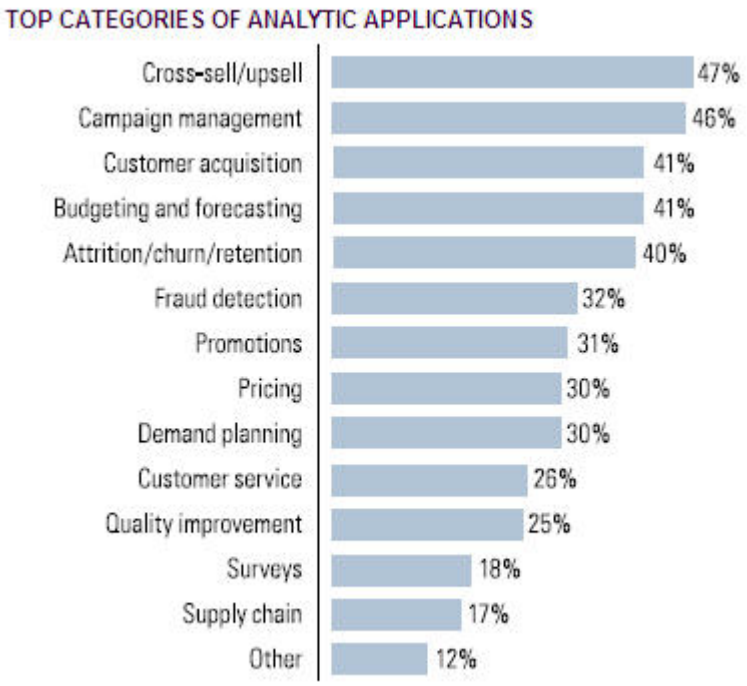

Figure 1. Based on 167 respondents who have implemented analytics. Respondents could select multiple answers. From Wayne Eckerson, *Predictive Analytics: Extending the Value of Your Data Warehousing Investment,* TDWI Best Practices Report, 2006. Available at *www.tdwi.org/research*.

Today, the notion of a Terabyte Club seems quaint... It might be time to start a Petabyte Club.

Data Volumes Explode. At the same time, data volumes to which organizations want to apply analytics are exploding. A decade ago, I participated in the formation of the "Data Warehouse Terabyte Club," which highlighted the few leading-edge organizations whose data warehouses had reached or exceeded a terabyte in size. Today, the notion of a Terabyte Club seems quaint, as many organizations have blown through that threshold. In fact, it might be time to start a Petabyte Club. (A petabyte equals roughly 1,000 terabytes or the equivalent of 250 billion pages of text.) Case in point: Yahoo! announced this spring that it has built the industry's largest analytical database, which will scale to tens of petabytes in 2009.[1]

Deep Analytics. Organizations are creating these outsized data warehouses to perform "deep analytics." For Internet companies such as Yahoo!, the goal is to gain a laser-sharp picture of how people use their Web sites so they can enhance visitor experiences and provide advertisers with more granular target advertising, among other things. Telecommunications companies mine millions of call detail records to better predict churn and fraud; retailers analyze transactions to better understand customer shopping patterns, forecast demand, optimize

[1] From "Yahoo Claims Record with Petabyte Database," *InformationWeek*, May 21, 2008.

merchandising, and increase the lift of promotions. Few industries today are immune from the siren's song of mining big data.

The intersection of analytics and large-scale data warehousing has created the perfect storm.

Perfect Storm. However, the intersection of analytics and large-scale data warehousing has created the perfect storm. Most database management systems were designed to handle online transaction workloads that involve finding and updating one or two records at a time. They were not designed to run complex analytical queries and algorithms against millions of records that often require making multiple passes through the data. Traditional relational database platforms choke on such workloads or, at best, require expensive hardware upgrades and lots of care and feeding that make them cost-prohibitive for all but the biggest companies.

Analytic Platforms. As a result, data-driven companies that want to analyze terabytes or petabytes of data have begun to experiment with new technologies designed from the ground up to handle complex queries against big data. These so-called analytic platforms generally consist of massively parallel (MPP) databases running on either preconfigured or off-the-shelf commodity hardware. Because the systems are purpose-built and designed largely as "plug-and-play" systems, their price-performance and total cost of ownership are jaw-dropping for most grizzled data warehousing veterans who have tried to remediate poor query performance by pitching multimillion data warehousing upgrades to skeptical executives. Many companies are implementing analytic platforms as data marts to off-load more complex processing and in some cases as a complete data warehousing replacement.

Two-Step Process. Providing blindingly fast query performance is not enough. Because many analytic computations are recursive in nature and therefore require multiple passes through the database, most analysts today run SQL queries to create a data set against which they run a procedural program written in Java, C, or another language. Sometimes they run the computation in memory, but in most cases, they export the data to their desktop or a local server where they run the algorithms, and once finished, reload the results back into the data warehouse. This two-step process is time-consuming, expensive, and frustrating.

"I do not like having to switch back and forth between Java and SQL."

"Most of our algorithms rely on set-based operations (i.e., analytical SQL) which returns a result set that we hold in memory and then manipulate procedurally in Java," says an information strategist at a major Internet company. "I do not like having to switch back and forth between Java and SQL—nor do my developers. It is especially frustrating when business logic is spread across the languages. If I can have one language that does it all, it's a big win."

In-Database Analytics. As a result, some vendors of new purpose-built analytical platforms are creating a SQL interface to procedural programs that run natively inside the MPP database. Rather than run procedural computations outside the database, this "in-database analytic" feature enable analysts (not database specialists) to create functions in any language they prefer, insert them directly into the MPP database, and invoke them via SQL calls. This saves time

and money compared to the two-step process. In-database analytics performs all processing within the database in a single pass.

However, we are getting ahead of ourselves. Before we dive into the details of analytic platforms, let's take a step back and look at the landscape for analytics and the reasons for the groundswell of interest in it today. If you're familiar with the basic concepts of analytics, then skip to the final two sections of this report.

Analytics in Context

Business Intelligence. Analytics is a subset of business intelligence, which is a set of processes and tools that enable business users to turn information into knowledge to assist with decisions, enhance planning, and optimize performance. Figure 2 shows a hierarchy of BI capabilities by insights and business value.

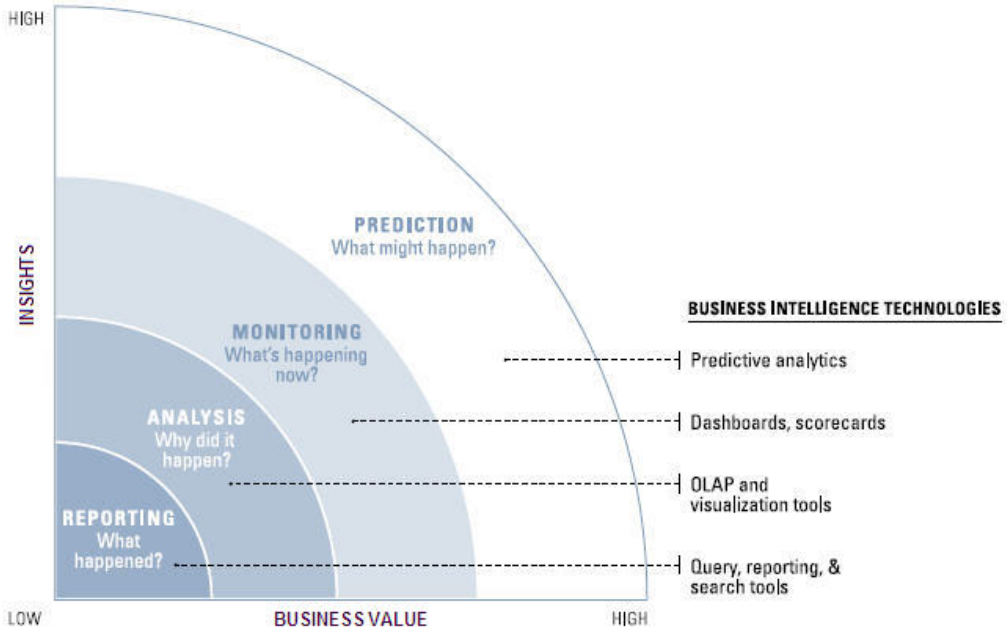

Figure 2. Dimensions of business intelligence. From Wayne Eckerson, *Predictive Analytics: Extending the Value of Your Data Warehousing Investment*, TDWI Best Practices Report, 2006. Available at *www.tdwi.org/research*.

Since the 1990s, most organizations have run reporting and analysis software against historical data to discover what happened and why. Today, organizations are implementing dashboards to identify what is happening now in their businesses so they can make immediate course corrections. These monitoring applications encompass dimensional analysis and reporting to create

"performance dashboards"—layered information delivery systems that parcel out information on a just-in-time basis.[2]

The final frontier for business intelligence is to predict what will happen.

Prediction. The final frontier for business intelligence is to predict what will happen tomorrow, next week, or next month using the past as a guide. While traditional analysis helps business users explore business problems dimensionally (e.g., sales by geography, by product, and by channel), advanced analytics use more complex math and statistics, and sometimes visualization, to detect patterns and anomalies in detailed transactions. Often, analysts turn these patterns into models that can be applied to new transactions to predict behavior or outcomes (for example, "Based on this customer's past purchasing history, this credit card transaction has an 85 percent chance of being fraudulent").

Analytic applications embed analytic models and generate reports or trigger workflows or automated actions.

Analytic Applications. Analytic applications embed these models and run them against new records, generating reports or triggering workflows or automated actions. Analytic applications might analyze customer shopping sequences to identify what types of behavior lead to more or fewer purchases. Other analytic applications might analyze patterns in credit card transactions to detect fraud or analyze telephone calling behavior to predict customer churn or up-sell and cross-sell opportunities. Others might analyze Web traffic to plot common navigation paths through a Web site and their impact on sales and service costs.

Many software vendors now sell functional and industry-specific application packages that embed pre-built analytic routines. These range from front-office applications, such as campaign management, warranty analysis, and Web analytics to back-office applications, such as profitability management, risk management, demand forecasting, and merchandise planning.

The ROI of Analytics. Many studies, including our own, testify to the power of analytics to deliver positive returns on investment and create a competitive advantage.[3] Organizations are using analytics to transform the way they operate and compete. For example, a bank uses analytics to increase campaign response rates sixfold and cut customer acquisition rates in half; a university uses analytics to predict whether a prospective student will enroll, enabling it to better target recruiting and outreach costs; and a hospital uses analytics to improve its ability to identify and treat pediatric brain tumors.

---

[2] See Wayne Eckerson, *Performance Dashboards: Measuring, Monitoring, and Managing Your Business* (John Wiley & Sons, 2005).

[3] A landmark International Data Corporation study shows that business analytics projects have an average ROI of 431% with most generating a payback in less than one year. See "The Financial Impact of Business Analytics," International Data Corp., 2002. Our own data shows organizations average $1.36 million ROI and an 11.2 month payback. See Wayne Eckerson, *Predictive Analytics: Extending the Value of Your Data Warehousing Investment*, TDWI Best Practices Report, 2006. Available at www.tdwi.org/research.

The mainstream business press has caught on to power of analytics to transform organizations and deliver a competitive advantage.

Recently, the mainstream business press has caught on to the power of analytics to transform organizations and deliver a competitive advantage. For example, Stephen Baker, a reporter at *BusinessWeek*, turned a cover-page article titled "Math Will Rock Your World" into a popular book, "The Numerati" (Houghton-Mifflin, 2008). Also, Tom Davenport and Jeanne Harris turned a *Harvard Business Review* article into a best-selling book, "Competing on Analytics: The New Science of Winning" (Harvard Business School Publishing, 2007).[4]

Although analytics delivers the highest degree of business value within the spectrum of BI technologies, it is not yet widely implemented for a variety of reasons. The remainder of this report will examine the challenges of solutions for delivering analytics, especially when applying analytics against large corporate databases.

Challenges and Solutions

There are four major challenges that organizations face when trying to move beyond reporting to analytics:

- Executive perceptions
- Data movement
- SQL expressiveness
- Data scalability

1. Executive Perceptions

Statisticians Not Required. A major obstacle to implementing analytics is perceptions. Many business executives believe that only expensive PhDs with statistical training can create and run analytical computations. Most executives know that PhD statisticians are difficult to find and expensive to retain and require costly specialized analytical programs, workstations, and data management software. Technical executives also worry that statisticians will create and manage shadow data repositories, undermining data consistency throughout the enterprise.

Although this perception was true for years, it is no longer valid today. Business analysts—and even ambitious IT systems analysts—who understand SQL and have a working knowledge of the company's data and business processes can

[4] Other books on analytics circulating in the mainstream business press are Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House, 2007); Ian Ayres, *Super Crunchers: Why Thinking-By-Numbers is the New Way to Be Smart* (Bantam Books, 2007); and Michael Lewis, *Moneyball* (WW Norton & Co. 2003).

perform fairly sophisticated analyses.[5] Although run-of-the-mill analysts may not implement complex machine learning techniques (e.g., neural networks and decision trees), they can still address a wide range of business problems using fairly common algorithms and techniques. (See Sidebar 1, "Analytical Techniques that Don't Require a PhD.")

Recent research suggests that it is not the complexity of the algorithm or model that determines the efficacy of the analysis but rather the number and types of data inputs. According to Anand Rajaraman, a former director of technology at Amazon.com, the real breakthrough in Google's search technology has less to do with its page rank algorithm (which he says is not very unique) than its use of additional data sources as inputs. "The addition of these two additional data sets—hyperlinks and anchortext—took Google's search to the next level," writes Rajaraman in his blog "Datawocky" (http://anand.typepad.com/datawocky).

Analytical Techniques that Don't Require a PhD

Techniques Suitable for "Casual Analysts"

- **Data profiling and transformations—**Functions that analyze row and column attributes and dependencies, change data formats, merge fields, aggregate records, and join rows and columns.
- **Sequential pattern analysis**—Discovers relationships between rows of data. Examples: customer shopping sequences, clickstream sessions, and telephone calling patterns.
- **Time series tracking**—Tracks metrics that represent key behaviors or business strategies. Example: patterns in customer sales that indicate product satisfaction and buying habits.

Techniques Suitable for "More Sophisticated Analysts"

- **Time series forecasting**—Predicts the future value of a measure based on past values. Examples: stock prices, sales revenue
- **Data profiling and transformations**—Uses functions that analyze row and column attributes and dependencies, change data formats, merge fields, aggregate records, and join rows and columns.
- **Bayesian analytics**—Uses simplified probability forecasting.
- **Regression—**Uses regression techniques (linear, logistic, etc.) for prediction (including forecasting time series data), correlation, inference, and hypothesis testing.
- **Classification—**Uses attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest. Examples: credit risk analysis, likelihood to purchase

[5] See Wayne Eckerson, *Predictive Analytics: Extending the Value of Your Data Warehousing Investments*, TDWI Best Practices Report, 2006. Available at www.tdwi.org/research.

- **Clustering or segmentation**—Separates data into homogeneous subgroups based on attributes. Example: customer demographic segmentation
- **Dependency or association analysis**—Describes significant associations between data items. Example: market basket analysis
- **Simulation**—Models a system structure to estimate the impact of management decisions or changes in "nature." Examples: inventory reorder policies, currency hedging
- **Optimization**—Models a system structure in terms of constraints to find the "best possible" solution. Examples: scheduling of shift workers, routing of train cargo, pricing airline seats

Black Box Syndrome. Another problem is that most executives aren't comfortable with the terminology of advanced mathematics and statistics and can't understand analytical techniques, which seem like a black box into which analysts place a hypothesis and data and magically get an answer. As a result, executives often don't trust the results, especially if they run counter to long-held assumptions about how the business operates.

There is no easy way to gain the trust of skeptical executives except through relentless empirical evidence that proves the business validity of the analytics. The recent spate of publicity about analytics and the testimony of leading companies can convince many executives to initiate a test project. Also, lowering the costs of implementing analytic applications can help sway recalcitrant executives.

Analytics aren't useful if workers don't understand their meaning and can't act on the results.

Disseminating the Results. Similarly, analytics aren't useful if workers don't understand their meaning and can't act on the results. To disseminate analytics broadly, organizations need to translate numerical outcomes into business language and actions that executives, managers, and front-line workers can understand. Providing only a numeric value or ratio is not enough. Some organizations are experimenting with ways to embed analytic results into dashboards and reports geared to business users. "We had to simplify the [customer churn] model to make it usable for the sales reps. These folks speak in terms of average order size, not R-squared values," says one business intelligence manager.

2. Data Movement

As mentioned earlier, analysts traditionally download data to specialized analytical platforms, such as Microsoft Excel, SAS, or SPSS, to explore data. One reason they do this is to consolidate data from multiple systems inside and outside the organization. Although a data warehouse presumably performs this task, in reality few data warehouses house the all the data required by analysts. They also download data so they can better profile and manipulate data and put it in a format that is optimal for the type of analysis they are conducting. Finally, most analytical platforms offer built-in support for a variety of analytical

algorithms and techniques and automatically generate code to implement statistical models, among other things.

Off-loading data to a separate application creates problems of its own. Data gets duplicated, manipulated, and enhanced with additional data, creating an analytical silo of information that is not consistent with other views of the organization. Also, when analysts download huge extracts from a data warehouse, they can seriously degrade query performance for other users of the data warehousing system.

The analyst's local server is usually not large enough to handle all the data he or she wants to analyze.

A bigger issue is that the analyst's local server is usually not large enough to handle all the data he or she wants to analyze. In these cases, the analysts must make a Faustian bargain: either they work with the IT department to create and run the analytical calculations in the corporate database (a thought few analysts could stomach until recently) or they work with a subset of data they obtain through sampling or by summarizing the data. Although sampling is often a valid option, most modelers would prefer to work with all the data at a detailed level to optimize model accuracy. Summarization is an option only when performing dimensional analysis using OLAP tools.

To avoid this Faustian bargain, many information architects and analysts are rethinking fundamental approaches to performing analytics. Instead of downloading data to specialized analytical platforms, most now want to run analytics inside corporate databases that contain all detailed transactions. This minimizes or eliminates data movement, improves query performance, and optimizes model accuracy by enabling analytics to run against all data at a detailed level instead of against samples or summaries.

There are design and runtime components to minimize data movement with deep analytics. In the design phase, analysts must explore the data, discover key patterns, and then bake their insights into analytical models or analytic applications. Then, in the runtime phase, analysts execute the analytical models against the entire database.

Sandboxes. To minimize data movement in the design phase, many companies are creating sandboxes within their data warehouses, giving analysts free reign within these partitions to add their own data and integrate it with data in the data warehouse. Analysts should be able to run queries that join data from the warehouse with their own data sets without crashing the system or impeding performance for others. These sandboxes reduce the temptation that analysts will create renegade data marts running on Microsoft Excel, Microsoft Access, or other analytical software.[6]

[6] For more information on analytical sandboxes, see Wayne Eckerson, *Bridging the Divide: Aligning Analytical Modelers and IT Administrators,* TDWI Monograph Series, July 2008. Available at www.tdwi.org/research.

In-Database Analytics. On the runtime side, organizations are eager to embrace in-database analytics, which lets analysts run their models and applications within the data warehouse instead of off-loading data to their desktop or a specialized analytic server. Analytics powerhouse SAS Institute has recognized this demand and begun working with database vendors to translate many of its core analytical functions into SQL extensions that run natively in a relational database. SAS is currently working with Teradata on this score but plans to do the same with other major database vendors in the near future. New database vendors, such as Aster Data Systems, also offer in-database analytics in an MPP environment as a way to differentiate themselves from the competition.

3. SQL Expressiveness

SQL makes it overly complex to execute many types of common business calculations.

Most analytical applications use SQL to query and manipulate data running on relational database management systems. Unfortunately, the SQL paradigm is a set-based, declarative query language designed to retrieve data from record-based tables within relational databases. This approach makes it overly complex to execute many types of business calculations listed in the sidebar "Analytical Techniques that Don't Require a PhD." For example, Internet companies that want to trace the paths that visitors take through their sites in a single session need to apply sequential pattern analysis to the data. The purpose here is to find out how many visitors click on page X followed by pages Y and Z within a two-minute period. This type of analysis requires "row-to-row" queries, which are notoriously difficult in SQL but easily handled in a procedural language.

The lack of procedural expressiveness in SQL poses a problem for analysts.

The lack of procedural expressiveness in SQL poses a problem for analysts. They either must write complex SQL code or circumvent SQL in some way. The most common workaround is to download data to Excel, a desktop database, or a specialized analytical server and perform the calculations there. As previously mentioned, these approaches have performance, accuracy, and scalability implications.

Procedural Extensions. Another common workaround is to leverage the procedural extensions supported by many RDBMS vendors. These extensions enable analysts to write stored procedures (SPs) or user-defined functions (UDFs) that run within in the RDBMS engine and can be invoked via simple SQL calls. Although database vendors offer varying levels of support for UDFs and SPs, they are often single-threaded, black-box functions that operate outside of the purview of the database optimizer. If a memory leak or other error occurs, they can bring down a database or corrupt data records. Because of the delicacy of writing functions directly against a database, many UDFs are written by experienced database specialists and in database- specific languages. In addition, some are not designed to run in a massively parallel database environment. Thus, leading-edge organizations that want to run analytic applications against large volumes of historical transaction data are often out of luck.

New techniques promise to make it easier for analysts to custom code database functions that run in a parallelized database.

MapReduce. Fortunately, there are new techniques that promise to make it easier for analysts to custom code database functions that run in a parallelized database. For example, MapReduce was pioneered by Google to run custom analytics against its massive cloud-computing network so it can better understand Web site activity and user behavior. Aster's version of MapReduce enables analysts to write truly reusable functions in any language they desire—Python, Java, C, C++, Perl—and invoke them with simple SQL calls. These flexible extensions are designed to run in a parallel environment and require no specialized database knowledge other than the MapReduce interface.

4. Data Scalability

The organizational and technical challenges described above have long posed a barrier to the adoption of analytics, especially in the creation of sophisticated analytical models. However, as organizations grow their data warehouses today from tens to hundreds of terabytes (and even petabytes), scaling an analytical environment becomes challenging. As data volumes expand, the technical challenges—data movement, analytical computation, and query performance—are significantly magnified.

MySpace. Take MySpace, for example. MySpace is the third most frequently visited Web site, attracting 120 million unique users and 7 billion page views per day. To get a handle on what customers are doing on its site and how its ad networks are performing, MySpace loads 2 TB of data per day into its 360 TB–capacity data warehouse in one-hour intervals. Rather than summarizing or sampling its data, MySpace runs analytics against the detailed transactions to get a complete picture of customer and ad activity. In essence, it wants to give its analysts ad hoc access to the data warehouse to perform complex analytical operations.

Until recently, running complex ad hoc calculations against terabytes of data was virtually impossible.

Limits of Relational Databases. Until recently, running complex ad hoc calculations against terabytes of data was virtually impossible. Most relational engines are general-purpose engines designed originally to run transactional workloads. To support analytical processing, relational database vendors have added special indexes, functions (including cubes), joins, schemas (such as stars and snowflakes), and aggregation schemes (for example, materialized views). Database administrators have to be proficient in the use of all these techniques to ensure adequate query performance.

Unfortunately, most of these analytical enhancements assume that database administrators can predict the types of queries users will run. Armed with this knowledge, they create indexes and aggregations on the appropriate tables and columns to enhance query response times. However, when a user submits an ad hoc query that doesn't utilize these mechanisms, performance often suffers. Ultimately, general-purpose relational databases can't deliver consistent query performance in a true ad hoc environment.

Historically, it has been costly to use relational databases for large-scale analytics. Indexes consume considerable disk space, sometime six to ten times as much as the raw data itself, so organizations have to purchase more disk and processing power than the raw data suggests. Maintaining all the indexes, partitions, and materialized views increases administrative costs.

Purpose-built Analytical Platforms. Many vendors have recognized the inherent limitations of today's relational databases to support large-scale query processing and analytics. As a result, in the past several years, several new companies, along with some veteran database vendors, have introduced purpose-built analytical platforms. These systems are designed from scratch to support analytical processing and accelerate query performance against large volumes of data. As such, they do not perform well under OLTP (transactional) workloads.

Because they are specialized environments, most analytical platforms offer better price/performance and lower cost of ownership than general-purpose relational databases running the same workloads. Analytical platforms come in a variety of flavors and sizes. As in any early-stage technology market, vendors are taking several approaches to address large-scale query processing. The techniques employed by vendors range from massively parallel processing (MPP) databases and data warehousing appliances to columnar and complex event-driven environments. (See Sidebar 2, "Types of Analytical Platforms.")

Types of Analytical Platforms

The most innovative sector of the business intelligence industry has been among database vendors, both new and old, that have shipped almost two dozen new products in the past year designed to accelerate query performance on large volumes of data. Here is a high-level categorization of these products.

MPP Analytic Databases—Specialized, stand-alone databases designed to run on MPP hardware and accelerate query performance. Examples: Aster *n*Cluster, DATAllegro (now owned by Microsoft), Greenplum 3.2, IBM DB2, Kognitio WX2, Teradata 12.0

Data Warehouse Appliances—A purpose-built machine with preconfigured MPP hardware and software designed for analytical processing. Examples: Dataupia Satori Server; Kickfire Analytic Appliance; Hewlett Packard NeoView; IBM InfoSphere Balanced Warehouse; Netezza Performance Server; Oracle Optimized Warehouse (with various hardware vendors); Teradata 550, 2550, and 5550 machines; Greenplum; and Sun's Data Warehousing Appliance

Columnar Databases—Store data in columns instead of rows, allowing greater compression and faster query performance. Examples: InfoBright Data Warehouse, ParAccel, Sybase IQ, Vertica

Complex Event Processing Systems—A system that captures and analyzes real-time streaming data. Examples: Cognos Now! SeeWhy, Streambase, Syndera, Truviso

Analytical Services—Outsourcing services that provide fast query performance on large volumes of data. Examples: 1010Data, Kognito Data Warehousing as a Service

Requirements for Analytical Platforms

Companies looking to purchase a purpose-built analytical platform have plenty of options. Over the past several years, more than 20 organizations, many of them start-ups, have offered specialized analytical databases. Below are some of the key requirements that organizations should consider when evaluating analytic platforms:

1. MPP with Minimal Overhead. To keep pace with the large volumes of data organizations are collecting without compromising analytical power, organizations should implement an MPP database. To ensure a low total cost of ownership, these platforms should minimize the use of indexes and aggregates that can significantly expand the size of the data warehouse and increase administrative costs.

2. Seamless Scalability and High Availability. When organizations reach the limits of an analytical platform, they should be able to add nodes seamlessly without bringing down the system or reloading the database. Conversely, if a node fails, the system should transparently recover, shifting tasks and data to mirror nodes so users experience no disruption and data is not corrupted. In some systems, the same technique used to support seamless scalability (and avoid forklift migrations) is used to ensure transparent failover and restore, simplifying the environment and administrative tasks.

3. In-database Analytics. The platform should provide a simple interface that makes it easy for analysts to add procedural functions to the MPP database without specialized database knowledge. Analysts should be able to write these functions in a programming language of their choice and call them via SQL. The functions should work natively in an MPP environment and be reusable across a variety of applications. In-database analytics eliminates the need for analysts to download data to a separate analytical server with all the associated problems described earlier in this report.

4. Application Ecosystems. Many analytical platform vendors are encouraging customers and partners to build and share in-database analytical “plug ins” to create an ecosystem of functions that run on the platform. Obviously, the easier it is to create and insert such functions, the quicker vendors can create a rich library of applications to offer prospects and customers. In addition, vendors are courting mainstream business intelligence, ETL, and utility vendors to support their platforms.

5. Mixed Workloads. It’s critical that analytical platforms support mixed workloads so processes don’t interfere with each other. For example, a complex, long-running query or database load shouldn’t impede performance of multiple short tactical queries. Otherwise, the system will choke as the number of users and processes expands. There is no purpose in providing screamingly fast query performance if only one user or application can benefit at a time. In addition,

mixed-workload capabilities and partitions enable organizations to "wall off" portions of an MPP database to create analytical sandboxes for individual analysts.

Recommendations

As organizations seek to gain greater value from their data warehousing investments, they invariably look to push the envelope of their front-end environment and supplement reporting with more powerful analytics. Providing adequate support for the analysts and analytic applications is challenging. Here are several recommendations to optimize your analytics environment.

1. **Create a quick win**. Identify a new application, special project, or department where you can pilot a new approach to rich analytics and create a quick win.
2. **Deploy a purpose-built analytical platform.** Don't struggle to deliver adequate query performance using a general-purpose database. Replace it with a purpose-built analytical platform designed from the ground up to accelerate query performance against large volumes of data. This is game-changing technology that can boost the success of your BI program.
3. **Create in-database analytical functions.** Make sure your analytical platform vendor supports in-database analytical functions so analysts can have their cake and eat it, too. That is, they can run complex analytical functions against large volumes of detailed data without having to off-load the data via sampling or summarization to an outboard analytical server.
4. **Create sandboxes.** Make sure your analytical platform vendor provides robust workload management and partitioning so that you can assign analysts individual sandboxes on the system that allow them to maintain their own data sets, import external data, and query across the entire data warehouse.