Raab associates, inc.
345 Millwood Rd •Chappaqua, NY 10514
914.241.2117 • www.raabassociates.com

Breaking the Business Intelligence Log Jam

David M. Raab
Raab Associates Inc.
May 2009

BI Thought Leader
Sponsored by

PARACCEL™

Table of Contents

About Raab Associates Inc.

Since 1987, Raab Associates Inc. has provided independent consulting on marketing technology and analysis to major firms in retail, communications, financial services, hospitality, and technology industries. David M. Raab has written hundreds of columns for publications including *DM News*, *DM Review* (now *Information Management*) and *Relationship Marketing Report* and spoken to audiences around the world. He is author of the *Raab Guide to Demand Generation Systems* and *The Marketing Performance Measurement Toolkit*.

Contact:
Raab Associates Inc.
345 Millwood Road
Chappaqua, NY 10514
www.raabassociatesinc.com
info@raabassociatesinc.com

Welcome to the BI Log Jam

Labor is by far the largest cost in any business intelligence project: nearly 60% according to some estimates.[1] But the actual "keyboard time" to configure a system is minimal. All the work goes into deciding what that configuration should be.

This work is the time spent understanding user needs and available data, designing solutions that combine the two, testing and reviewing results with users, and then repeating the process until the users are satisfied with the results. Developers have long recognized that the key to keeping down project costs is minimizing the changes made after delivering the initial result. Modifying, retesting and re-reviewing a fully elaborated existing system can be much more effort than building it the first time.

Yet changes are inevitable. Users and analysts often don't understand their true requirements until they see the first concrete results of the project. Moreover, those requirements will change over time in response to new information, some provided by the BI system itself, and to changes in the underlying business.

Required changes pile up at a steady or accelerating rate...the inevitable result is a growing backlog of unmet business needs.

In other words: required changes pile up at a steady or accelerating rate, while the effort to make each change grows with system complexity. The inevitable result is a growing backlog of unmet business needs. This is the BI Log Jam.

Why Change is a Problem

But what, exactly, makes change so difficult for business intelligence systems? It's true that change is costly for any type of process. Yet it's especially painful for business intelligence systems because even small alterations often require substantial effort to implement. This is because the most common approach used for enterprise business intelligence systems, multi-dimensional data models, must aggregate data into cubes tailored to answer specific questions. If a new question requires new data elements or different dimensions, the entire cube generation process must be modified, starting with extraction of the input data and continuing through aggregation, metadata creation, and user interface adjustments. This applies even if the input data has already been collected in a data warehouse. If a new source must be added to the warehouse itself, even more work is needed.

The data cubes are necessary because querying the raw data directly would often take too long, either because of the sheer amount of detail or because it would require complicated queries that are difficult to write and time-consuming to execute. Data cubes provide a huge improvement over direct queries in response speed and ease of use. In most cases, this value of this improvement greatly outweighs the costs for cube design and update processing. Yet those costs can still be substantial. As we've already seen, the bulk of those costs are for labor.

To put it another way: since hardware and software costs are largely fixed once the system has been deployed, labor efficiency is by far the largest variable affecting the total operating cost. So using labor efficiently – specifically, reducing the cost of changes needed to meet user needs – is the key to success.

Efforts to Minimize Change

Developers have traditionally sought to reduce change costs by reducing the need for changes. Common approaches include:

- *Do a better job of understanding user needs.* This is the simplest solution of all: it boils down to "get it right the first time". But there's more here than motivational wall posters. Companies can make substantive investments in hiring business analysts, formal development methodologies, use case and scenario development, business / IT organizational alignment, process engineering, and other approaches. These are important and effective, but even organizations that do an excellent job of understanding user needs often find themselves with growing backlogs as needs change over time. Although better analysis can reduce rework during the initial development cycle, it cannot eliminate changes completely.

- *Work more efficiently.* This makes another excellent wall poster, perhaps reinforced by the empty cubicle of a co-worker removed during a recent downsizing. Again, companies can also take more concrete steps to avoid rework such as employing agile development techniques that closely involve end-users in the process. These methods have proven effective but take expert management, essentially requiring technicians who have acquired one demanding skill set (BI system design) to master yet another (close cooperation with end-users). Sadly, few people have the talents needed to excel at both.

Reducing the cost of change is the key to success.

- *Use more powerful hardware.* Increased processing power lets companies build larger cubes that can satisfy more user requirements without being modified. More generally, it let BI systems give adequate performance even when they have not been fine-tuned for efficiency. The general idea is to substitute relatively low-cost hardware for high-cost labor. Although this works to some extent, some hardware does not scale linearly – meaning that additional effective capacity is quite expensive. In addition, the performance penalty for poorly tuned designs is often so high that it only a massive hardware increase can compensate. Nor can hardware eliminate the need for substantive changes such as adding a new data or source or changing an aggregation level.

- *Reduce demand.* Detailed project requests, formal cost justifications, charge-backs for IT resources, and similar policies are legitimate tools to optimize use of corporate resources. But they also have the not-entirely-unintended or unwelcome effect of discouraging change requests. The problem, of course, is that needs don't vanish simply because users don't file a request. Instead, they'll often turn to "shadow IT" systems such as Excel spreadsheets that actually cost the company more money in the long run than changing the BI system itself. Or, they wait until a problem becomes urgent and must be resolved through an expensive crash project or inefficient workaround.

Although these and other methods can reduce the need for change, they cannot eliminate it entirely. Users will never fully understand their requirements, analysts will never fully understand users, and requirements will never stop changing over time. Breaking the BI log jam ultimately requires not simply reducing the need for change, but making change itself

[1]Market Demand for Business Intelligence and Performance Management (BI/PM), AMR Research, 2007

easier to accomplish.

A New Hope

Happily, a new generation of technologies does make change easier. These technologies allow vastly more efficient access to raw data, eliminating or sharply reducing the need for predefined data cubes. This means that new reports can be created or new data sources added with a minimum of effort by technical staff. In fact, skilled users, such as departmental business analysts, can often meet new needs without any help from specialists in the IT department. The net result is to greatly reduce the cost of change while doing a better job of satisfying business requirements. Truly, everyone wins.

Three technologies characterize these new systems.

- *In-memory data storage.* In-memory storage removes the bottleneck of reading data from disk drives. Data transfer rates have improved much more slowly in recent years than processor speeds, memory size and storage density. This has had little impact on transaction processing systems, which typically select small sets of data they can locate in advance using indexes. But it is a major issue for BI systems, which typically scan many or all records in a large data set. In fact, a primary reason that data cubes have remained so central to BI processing is that they greatly reduce the amount of data that must be transferred from disk to satisfy user requests. When data is preloaded into memory, on-the-fly aggregation taking advantage of high-speed processors can give performance that would otherwise have required a pre-calculated data cube.

- *Columnar databases.* These systems organize data so that each column (that is, all instances of a given data element such as “city”, “first name” or “account number”) is stored and accessed independently. This contrasts with traditional row-oriented databases, which are optimized to read all values for a single row (for example, an entire customer record). The row-oriented approach makes sense for transaction processing, which typically does access multiple fields on a single record. But BI queries often need to read all instances of just a few data elements. For them, a columnar approach is vastly more efficient. Columnar approaches also allow greater data compression since each column can be compressed using the technique best suited to its particular data type. This makes it even easier to fit the data into memory, making columnar designs a particularly good complement to in-memory technology.

Breaking the BI log jam ultimately requires not simply reducing the need for change, but making change itself easier to accomplish....

...Happily, a new generation of technologies does make change easier.

- *Shared-nothing or MMP (massively-multi-processor) architectures.* These systems divide the data and processing among nodes that each contains its own disk storage, memory and processors. This allows all nodes to work simultaneously on a single problem without the bottlenecks that might appear if all data had to read through a single disk array, accessed in a shared memory space, or run on one processor. MPP systems do require some coordination to spread the data across different nodes so each can work on a piece of a problem, to share data across nodes when necessary, and to aggregate results. Doing this well is quite difficult, but several vendors have developed their own approach. These differ considerably in the details: some use proprietary hardware, some are purely software based, and some use commodity hardware in tightly

controlled configurations. What they share are abilities to process massive data volumes at much higher speeds and much lower costs than conventional systems.

Each of these technologies has been available for years. But early implementations were often very expensive, supported limited data volumes, had excessive data load times, or were incompatible with the industry-standard SQL query language. This made them difficult to integrate with existing company infrastructures and added significantly to their support cost.

But today's versions of these systems have been carefully designed to overcome these obstacles. Apart from a few proprietary "appliances", most run on standard hardware. They use standard operating systems and SQL-compatible databases. Data volume and load time must be examined carefully on a product-by-product basis, but in general the shared-nothing architectures allow them to scale very well. Some of today's systems could actually be used for transaction processing, but most are designed solely for business intelligence applications.

Benefits

These technologies reduce business intelligence change costs in several ways:

- *Adding new data.* New fields and tables can be copied directly into the system in their native data structures or with minor modifications. They can then be queried independently or linked with other data using standard SQL. There is no need to restructure the data for performance, create indexes, or reload existing data cubes. The ability to add data without significant technical effort lets users easily examine new sources and load files they may use only once.

- *Handling unplanned queries*. Users can query detailed source data directly, rather than being limited to predefined questions defined by data cubes. This allows them to execute unplanned queries without waiting for IT to design and load a new cube or modify an existing one. The high performance provided by the new technologies allows these detail-level queries to run quickly enough that cubes are not necessary.

- *Supporting new applications.* Because data is stored in conventional tables, rather than special-purpose analytical structures, it is easily accessible to third-party applications such as predictive modeling or visualization software. This means these applications can be added to the BI infrastructure quickly and with minimal integration cost. If the application itself requires a particular data structure, the new technologies can execute the necessary transformations much faster than conventional databases.

What it all boils down to is this: users can do more for themselves, and IT can complete the remaining work much more quickly. The result is a huge improvement in BI effectiveness and huge reduction in BI change costs.

Limitations

Exciting as these new technologies are, they cannot by themselves solve all business intelligence problems. Companies will still need disciplined development processes to uncover and meet user requirements. They will still need data quality and master data management programs to properly integrate data from different sources. They will still need

to find tools that make data accessible to end-users and to make sure they are trained to use them effectively. In short, although these technologies can substantially reduce some of the labor associated with business intelligence, they do not change its fundamental nature.

What Next?

In-memory data storage, columnar databases, and shared-nothing architecture can be found separately or in different combinations in different products. From a technologist's perspective, systems like ParAccel are the most exciting because they combine all three. But even when just one or two are present, they offer the potential to significantly change the economics of business intelligence by the reducing costs and speeding delivery of solutions that meet a broad range of company needs.

The question facing IT departments is how to take advantage of this opportunity. Fortunately, since the current implementations of these technologies are highly compatible with existing infrastructures, they are easy to test and deploy incrementally. Many vendors offer trial or proof of concept programs that let qualified prospects see what the systems can do before making an actual purchase. Some offer cloud-based deployments that make it even easier to work with the system on a trial basis.

The BI log jam has been building for years. You won't clear it by poking with a stick.

One point to bear in mind is that these systems are radically different from conventional technologies. This means their advantages become most apparent when you use them for tasks that conventional systems simply cannot handle. Reproducing an existing cube-based environment in one of the new systems may not be significantly easier, while building a new project that doesn't use cubes at all will highlight the real differences in development speed and labor. Similarly, starting with a small test of a low-volume application may hide the system's power: it is probably more useful to try a project that you know is much too large for your current technology to support. This is the exact opposite of the "start small" approach IT departments typically use to test new technologies, but it's necessary to demonstrate what makes changing to these systems worth the investment.

The radical nature of these systems is exactly what makes them important. The BI log jam has been building for years. You won't clear it by poking with a stick. But careful planning and the right equipment can clear channels for information to flow more freely than ever before.

BI Thought Leader Collection

www.paraccel.com/data_warehouse_resources/business_intelligence_thought_leader

About ParAccel

ParAccel, Inc. is the proven leader in scalable analytic performance and price-performance. The ParAccel Analytic Database™ is a new generation, MPP-Columnar DBMS that is delivering breakthrough analytic performance and price-performance in customer environments. Available as software or a virtual or packaged data warehouse appliance on standard hardware from all major vendors, it can be implemented stand-alone, or as a drop-in accelerator to extend an existing SQL Server or Oracle investment. ParAccel's management team includes technical founders and industry veterans from noted data management companies Netezza, Oracle, Teradata, Gupta, SenSage, PointBase, and IBM. ParAccel helps leading companies like Autometrics, Merkle and TRX to extend their analytic performance advantage to further differentiate their products and services. ParAccel is based in California with offices in Cupertino and San Diego. For more information please contact us at info@paraccel.com or 866-903-0335, or visit us at www.paraccel.com.