March 2008

The Ever-Evolving Data Warehouse:

Dealing with Changes and Pressures for BI Today

Claudia Imhoff, Ph.D.

Introduction

Business Intelligence (BI) and data warehousing have undergone a significant evolution in the past 10 years. There was a time when BI was a standalone, isolated world of time-dated, historical snapshots of data for analytics. BI implementers extracted all the data they wanted from the operational environment, and then disappeared with it into a world of their own.

Tactical and strategic analyses were performed on daily, weekly, and perhaps monthly snapshots of data. Query performance was important but not a critical factor in the environment's success. The audience was limited to business analysts and highly technical statisticians and researchers. The most difficult problem faced by the implementers was the ability to give these users flexibility in terms of handling unplanned or unusual queries.

Then along came the notion of operational BI. While it seemed innocent enough – just another form of BI – operational BI has turned out to be a major disruptive force in BI environments today. Perhaps it can best be summed up in one characteristic: Operational BI must actively support operational decision making.

What does this mean to your existing data warehouse environment? Plenty! Operational BI has put significant and increasing pressures on the BI environment to meet these new business requirements. This white paper discusses the pressures and their impact on data warehouse environments. We end the paper will suggestions on how to make your infrastructure more "dynamic" and future-proof.

Operational BI Pressures on Today's BI Environments

There are five major categories of pressures that we will discuss.

Data Currency and Volumes

The first major pressure from operational BI users comes in the form of the currency of the data they use. To make good operational decisions, the data must be much fresher or current than that normally found in traditional data warehouses. Detailed *intraday* snapshots of data are trickle fed into data warehouses or operational data stores, allowing operational personnel the ability

to analyze events occurring during the day of the event. As you can imagine, this means that the volume of data being stored increases substantially. Data warehouses now contain tens of terabytes to hundreds of terabytes (even petabytes!) to support all forms of BI now. And it should be noted that the granularity of the data for operational BI must be at the lowest level of detail.

All this means that, not only does the data warehouse infrastructure have to handle faster, more frequent loads of data, but seamless scalability is mandatory – whether it is for processing the data, storing the increased volumes, or maintaining the integrity of the environment (backups, failovers, etc.)

Performance

While performance in traditional BI environments has always been important, it is now a critical success factor for operational form of BI. The mere name, operational BI, should conjure up visions of users requiring sub-second response times – enough to make the stoutest BI implementers' hearts quiver!

Now add to this the fact that the environment must still support the more traditional BI users (tactical and strategic) with appropriate response times. This one facet of the evolving data warehouse environment – handling a mixed workload environment – has stumped many database vendors. A mixed workload environment means the technology must have the ability to prioritize queries, not only according to their importance to the enterprise but also their response requirements.

Number of users

Before the advent of operational BI, most decision support environments appealed to a relatively small number of users – perhaps less than 10% of the entire workforce. With a focus on operational decisions, the new form of BI attracts a great deal more users, maybe even *all* of the business users.

These new users may have very different interface requirements. If they are front line personnel, they are familiar with drop down menus, rigid edit checking, in-line help functions, and so on. These are unlike anything traditional BI implementers have had to deal with before. It means that BI implementers must rethink how BI is delivered, what the interface looks like, how support should work, and where and when the users will use this capability.

And here is a catch – not all the new users are human! Operational BI must also interface with operational applications as well. The operational application may call a BI application to perform a specific analytic and, upon receiving the results, the operational application then moves on to the next step in the process. This interconnectivity between operational BI and operational systems is major driving force for the need for scalability and availability of our BI environments.

Different data types

Another aspect of bringing operational BI into your data warehouse environment is a change in the type of data that is needed. We have seen an increasing interest in unstructured or semi-structured data in support of operational decision-making. Most data warehouses are focused on integrating only structured data, and may not be capable of switching gears to handle the unstructured data.

Unstructured data requires retooling the ETL processes to handle this new form of data. In addition, the BI component must be able to manage the display characteristics needed for this vast amount of new data.

Finally the data warehouse itself has another new source of high volumes of data – again assaulting the scalability of the underlying technology. Between the need for intraday snapshots and the vast amount of unstructured data needed, you can see how data warehouses could reach unheard of sizes.

Support for increasingly complex and dynamic queries

The final pressure on today's data warehouse environment comes from the increasingly innovative analytics performed by business users. We have seen the evolution of BI utilization in most organizations go from the creation of simple reports, to time series comparisons, to complex models of fraud detection and risk mitigation, to intricate predictive analyses of customer and market behaviors. Thanks to operational BI, the time frame of the data for these analytics also changed. In addition to the change in data snapshot frequency, operational systems are now being fitted with streaming and embedded analytics performing complicated analytics to help front line workers make better decisions throughout the day. The interface with operational systems has truly changed the face of a BI environment forever.

Impacts on Data Warehouse Environments

The impact on BI environments of speeding up the overall analytical process can be profound. In our practice, we frequently perform assessments on existing data warehouse and BI environments to determine where and how implementing operational BI capabilities will affect the existing infrastructure. We look for weaknesses in the overall process of data integration, storage, access, and disaster recovery. It is typical that any weaknesses discovered in the overall processes will be exaggerated when they are accelerated. For example, if the data integration processes are not performing efficiently, this weakness will be magnified significantly when these processes are sped up.

To help you perform your own internal assessment, here are the areas you should focus on to ensure you are ready for operational BI:

- **Scalability** – As mentioned earlier, the ability of the data warehouse to scale from 1 to 10 TB, from 10 to 100 TB, and from 100 to 500 TB (or more) is required these days. Keep in mind that scalability is critical to ensure that calls from operational applications can be handled with ease as well. Along with scalability, you may also want to look for data that can be either archived or perhaps deleted. Check for forgotten summarized or derived fields, unneeded or unused data elements, and unnecessary redundancy introduced and forgotten. Confirm that you can effectively manage the growing size by maintaining the relevant data for business queries and analytics online while ensuring timely access to any archived data.
- **Performance** – The biggest performance hurdle is determining how your database will handle the mixed workload expected with the advent of operational BI. Make sure your environment can provide for dynamic BI mixed workloads, BI and transactional workloads, and a combination of data warehouse and data mart workloads on the same system. Your infrastructure will have to deal with issues such as varying levels of query concurrency, resource adjustments to exploit the available capacity and handle different work priorities, protection against "killer" queries, maintenance of consistent response times regardless of the workload, and dynamic resource allocation to react to loads. In addition, determine if you have a smart archiving scheme and appropriate underlying technology so that it can be returned with confidence.
- **Continuous availability** – Due to the increasing dependence of the operational personnel and operational applications on the new BI capabilities, your environment will have to provide extremely high

availability and an environment that is stable for the most mission-critical applications. The data warehouse technology cannot "go down"; therefore, you must determine whether it can mimic the availability expected of operational systems, that is, 99.999% availability.

- **Ability to combine different sources of data** – Ultimately, the business users want the ability to look at all sorts of data from myriad sources seamlessly. They want to combine operational BI results (e.g., the identification of potentially fraudulent transactions) with a current view of the situation from the operational system. They may combine strategic BI results (e.g., customer lifetime value score or next-best-product-offering) with customer information from the customer care system. When assessing your infrastructure's ability to handle the demands of operational BI, make sure you understand and can support this ultimate goal of data "on demand" with your data management suite of technologies.

"Future-proofing" Your Data Warehouse Infrastructure

Changing an existing data warehouse infrastructure can be painful and frustrating. Painful because you may have to retool the entire underlying technologies if they cannot support the new demands placed upon them. Frustrating because you may not know or correctly predict in which direction the business community needs or technology will move.

How can you reduce the risk of making a wrong technological decision? Here are a couple of tips in selecting the best technology for the long haul. They help mitigate the risk that you are on the wrong path to evolve to future, yet-unknown needs and requirements.

- **Maintain an open environment** – We always hear about the need for non-proprietary technologies and infrastructures in BI. A good bet for the future of your environment is to make sure you maintain open components. While proprietary technologies may be useful for specific problems today, they limit your ability to grow with your users. Open technologies give you the ability to swap out one component easily and replace it with newer and more suitable open technology for tomorrow.

- **Ensure that you build flexibility into your infrastructure and database design** – Flexibility comes in many forms. We have discussed the need for seamless scalability, linear performance, archive

capabilities that enhance performance and reduce costs, and bullet-proof availability. The ability to add capacity easily, to increase the numbers of users with minimal effort, and to extend the functionality from the individual through the team to the entire organization are measures of an environment's flexibility. BI must also be responsive not only to the different needs of the business community but also be able to handle all the information – both structured and unstructured – that the users need to make decisions.

Summary

What is needed to ensure sustainability of the modern data warehouse infrastructure? The following attributes: scalability, performance, accessibility, and availability of the data, and the ability to deliver analytic results, operational events, even external information through an open, on-demand data management architecture. We are fortunate that we have such platforms available to grow and evolve with the needs of data warehousing and BI. One such example is IBM with its introduction of the InfoSphere family of products, particularly the InfoSphere Warehouse.

The data warehouse environments of the future must have these characteristics in place now to ensure their ability to support not only today's requirements but to guarantee agility in supporting future, yet-unknown, innovations in this constantly evolving environment.