

Measuring the Effectiveness of Your Data Warehouse

By Sid Adelman, Principal
Sid Adelman & Associates
May 2002

Running a data warehouse without the advantage of metrics is like trying to navigate a ship without a chart, compass, or sextant. Without metrics, we have no way of knowing whether we have delivered a data warehouse that meets user requirements and can be deemed "successful." We would have no idea about response times, machine utilization, availability, user satisfaction, or the quality of the data in the warehouse.

This article suggests metrics for assessing data warehouse success, recommends standards or service level agreements (SLAs), and suggests who should be responsible for measuring. The article addresses who should be responsible for taking action to correct situations that are out of compliance with the standards, and recommends how to represent the results of the measurements to management.

Most projects have explicit or implicit measures of success and most of these can be measured. The measurements of a data warehouse will determine if it was or was not a success.

* TYPES OF METRICS

There are a number of types of metrics that are relevant for understanding how well we are doing.

USAGE: Usage tells us if the data warehouse is being used, to what extent, and by whom. For example, our goal may be to have 90% of the trained users actually using the data warehouse. Our metrics may show that our goal was met. 92% of the users, who were trained on the tools, ran a query or report during the previous month.

Metrics on usage are often an eye opener as you discover that large numbers of intended users do not use the system or use it only sporadically. This type of information can point to deficiencies in training, in poor targeting of intended users, lack of predefined queries and reports, inadequate support, or lack of emphasis by the users' management.

PERFORMANCE: Performance is usually reflected in response time. While I don't recommend an SLA for response time, it is very important to measure how long a user has to wait for his or her answer to come back. It's usually too late once the users start

calling or screaming that response time is terrible---the damage has already been done to the reputation of the warehouse. You will want to know what percentage of the queries ran longer than---for example---ten minutes, how long they actually ran, and you will want to know which departments experienced the long run times.

Measurements sometimes uncover poor performing queries that are the result of users not understanding the ramifications of some of their actions. The appropriate response could be more comprehensive training that focuses on the dos and don'ts of writing queries. The realization that a field is very frequently being summarized would lead to the creation of a summary table. As it becomes apparent that another field is frequently being accessed, or that tables are being joined on a specific key, the DBA would build an index for those fields.

AVAILABILITY: Availability is the percentage of time the system can be accessed by the users during scheduled hours. Scheduled hours may be 24/7 for global systems or a subset such as 18/5. Most organizations do not have the same availability requirements for their data warehouse as they do for their operational systems.

Sometimes operations will not give the data warehouse environment the right level of attention and which results in poor availability. In addition, the ETL process aborting or not completing on time will also impact availability.

Metrics help identify these obstacles, giving staff members the opportunity to improve system performance and, ultimately, meet availability requirements.

RESOURCE UTILIZATION: This would include the number of machine cycles and the accesses to the disk. The information about the percentage of disk utilization should result in a better distribution and partitioning of the data on the array of disks or should point out the necessity of purchasing additional disk space.

USER SATISFACTION: User satisfaction surveys tell us how well the data warehouse meets the users' expectations. Satisfaction surveys should be taken two to four times per year and the results of those surveys should promote action to improve the areas where the users are unhappy. Sometimes these surveys uncover misunderstandings in the way the system was intended to be used. User satisfaction surveys typically include questions on the use of the access and analysis tool, data quality, availability, response time, and support.

DATA QUALITY: The quality of the data can be automatically quantified and should include the percentage of values that are outside of the valid values, the percentage of fields that are missing, data that is the wrong data type, data that is outside of the acceptable ranges, and data that violates business rules. Metrics that have a bearing on the quality of data would include record counts, the number of distinct values, the number of records with null values, the number of records within a certain range, the number of records with a certain distinct value, and the rate of change of the data.

Management is always surprised and dismayed when they discover just how dirty their data is. The data quality metric will serve as a barometer of the constant data quality improvement that should be a part of every data warehouse initiative.

DORMANT DATA: Dormant data is data that is never, ever accessed. Loading this unused data night after night is expensive, consumes disk space, and may reduce the likelihood that the system will be available to the users by 8:00 AM (when they were expecting it). Dormant data is a total waste to the organization and is the albatross that will weigh down the budget, extend ETL time, and tax the skills of the DBA staff.

Dormant data exists either because the requirements gathering process was lax or because the users were unable to sufficiently articulate their requirements resulting in loading useless data for fear that the users may possibly need it. There are tools available to help identify dormant data (Ambeo is one example).

USE OF TOOLS: (which tools are being used and at what level): Since many data warehouse installations have multiple business intelligence tools, this metric should tell us which tools are used and to what extent. Most organizations have anticipated which tools will be used by which departments and by which category of users (e.g., power users, casual users). When the tool use does not match the department and user profile, it's important to understand why. There may be a misunderstanding of the tools and their purpose or the profiling may have been incorrect.

In "The OLAP Report," (www.olapreport.com) a survey found that 30% of data warehouse software was not being used. An organization may be purchasing software when it already has seats available and does not need to purchase any additional software. The organization may be paying maintenance for software on the shelf. In both cases, metrics could identify opportunities to save money.

COSTS: This is what the data warehouse costs, both on initial installation and as an ongoing expense. The value of measuring the costs, and comparing those costs to the anticipated costs (budget), will give the organization the information it needs to better anticipate costs and then to help determine if a project is cost justified before it actually is implemented. Ongoing costs are often grossly underestimated. When anticipating costs, consider the total cost of ownership, including servers, software licenses, support staff and so on.

BENEFITS: Presumably, the warehouse was justified by anticipated benefits and it is important to identify and quantify those benefits. Some of the intangibles do not lend themselves to quantification but it is important to identify them whenever possible. The organization must measure the benefits to determine if they were achieved. This knowledge will help, along with the information gleaned from measuring actual costs, in adding to the enterprise knowledge of anticipating the ROI of each project and in setting priorities.

SECURITY CONFORMANCE: What security violations have been detected, where are the violations coming from, and have there been any breaches? This information will help plug security holes and will also provide a level of comfort to upper management as they worry about security exposures.

In addition to the types of metrics, an effective measurement initiative should include SLAs, assigning responsibility for measurement, determining what means should be used to measure, knowing how the results of the measurements will be used, and defining how these metrics should be reported to management.

*** SERVICE LEVEL AGREEMENTS (SLAs)**

SLAs are written agreements between the business (the folks that will be using the data warehouse) and IT (the people who are responsible for building and providing the data warehouse infrastructure). The SLAs will identify your goals and the metrics will tell you if these goals have been met. Metrics supporting the SLAs may include:

- *Availability
- *Response time
- *Response to problems

SLAs let you know which of the user requirements are being met. SLAs also serve to hold user expectations in check.

*** RESPONSIBILITY FOR MEASUREMENT**

It isn't always clear who should be responsible for measurement. Some organizations have groups dedicated to performance and these are the folks who normally have the primary responsibility. Depending on the specific measurement, the responsibility may fall to the DBA group, the Architecture Group, or to Capacity Planning. Measurement is usually not a full-time job but it is a job that cannot be forgotten or denied. If there are performance problems or availability problems, near real-time awareness is paramount so that the responsible persons can be alerted and effective actions can be taken. Ultimately, the metrics reflect the focus of the organization, e.g. response time, and they should be agreed upon by all partners from the start. At the same time, however, metrics are somewhat fluid and will change as the organization changes.

*** MEANS TO MEASURE**

A number of the data warehouse products have built in the ability to capture and to report on the system. These metrics can be accessed and delivered in a form (with some effort) that is meaningful to both the technical people, to the users, and to management. In addition, there are add-on products that supplement in various areas such as data quality and performance. Quite often, organizations have these means of measuring but are either unaware of their existence, or no one has been assigned to execute the measurements and take action on their results.

*** USE OF MEASUREMENTS**

We need to measure because the data warehouse should always be considered a work in progress, always needing improvements. The process should always be to measure, identify problems and opportunities, and take appropriate action to solve the problems and exploit the opportunities.

Chargebacks are rarely welcomed but for those organizations that do charge back the use of their systems to the departments that employ them, metrics are critical to an equitable distribution of costs. This becomes even more important when money is transferred from one organization to another.

*** REPORTING RESULTS TO MANAGEMENT**

Management wants to know how things are going. They just spent \$5 million for the data warehouse and they want to know if they are getting their money's worth. Are the users using the system, are they happy, and are they achieving the benefits they were expecting?

Management is usually content with monthly reports unless there are serious problems. In which case, management will want to be briefed more frequently on the problems, the steps that are being taken to resolve those problems, and results of the resolutions. Metrics should be reported with just the information that is of interest to each manager. A good approach is to use conventional data warehouse tools accessing a small metrics data mart. Any metrics that represent problems should be highlighted or shown in red. A dashboard is appropriate for metrics of performance and availability.

*** SUMMARY**

Each organization should identify the metrics they will need and use as they continually work to improve their own data warehouse. An understanding of the appropriate metrics, the responsibility for gathering the metrics, and the use of those metrics can make the difference between success and failure of the data warehouse project.

Sid Adelman, President of Sid Adelman and Associates, is a noted consultant in the field of data warehouse project management. He can be reached at sidadelman@aol.com