# The Quality Data Warehouse:
# Solving Problems for the Enterprise

**Bradley W. Klenz, SAS Institute Inc., Cary NC**
**Donna O. Fulenwider, SAS Institute Inc., Cary NC**

## ABSTRACT

Enterprise quality improvement is quite different from quality improvement at the process level. Enterprise quality improvement requires information from many departments within an organization, such as production, quality assurance, engineering, customer service, and purchasing. Typically, these groups collect large amounts of data from disparate and disconnected systems. The systems include a variety of Statistical Process Control (SPC) systems implemented on different production lines at different plants, as well as data collected in Laboratory Information Management Systems (LIMS), Manufacturing Execution Systems (MES), Enterprise Resource Planning (ERP), and Manufacturing Resource Planning (MRP) systems. The disparity and disconnection of these systems poses a major problem for the implementation of enterprise quality improvement.

Data warehousing is a technology that is designed to facilitate analytical processing on data from disparate sources. This technology is being employed successfully today in many industries such as finance and retail where it aids, for example, in the understanding of customers and their buying habits. Now, forward thinking companies in the manufacturing industry are beginning to use data warehousing to address enterprise level quality. Transactional systems exist to do the job of gathering and storing the detail data, whereas the data warehouse exists to serve the decision-making needs of the enterprise. To achieve enterprise quality improvement, a Quality Data Warehouse must be built to serve the analytical needs of the manufacturing enterprise.

This paper discusses some of the emerging business problems in quality improvement and how the field has evolved to the current situation. Data warehousing is presented as a key technology for continued quality improvement. The presentation starts with what information is delivered with a data warehouse and how that information is delivered. Next, steps are laid out for building a Quality Data Warehouse, including appropriate sources of data and who needs to be involved in the data warehousing project. Finally, some important additional points on exploiting the Quality Data Warehouse are covered.

## EMERGING BUSINESS PROBLEMS FOR QUALITY IMPROVEMENT

Think about what it would take to answer questions like the following:

- You collect data throughout the production process. You have information on raw materials going into the process and measurable quality characteristics on the outgoing product. How can you relate the information that you have at the two ends of your production process?

- You have warranty claim data available from the corporate office. You would like to relate this data back to the production conditions that existed at your plant. The warranty data is tagged with one set of identifiers, the process measurement data is tagged with different identifiers. Some data from shipping records can tie them together. How do you combine this data and analyze it to find process improvements that would reduce warranty claims?

- You have multiple plants that are producing the same or similar products. Although you have corporate guidelines on operations, each plant has evolved differently. Now, the differences make it difficult to use data across plants. How can this data be combined so they are available for analysts to learn from the cross-plant variation?

In many organizations, the procedure to answer these questions goes like this:

1. Pick up the phone and call your IT contact to submit a request for data. This data could be in many different places: a legacy system (custom export program), an enterprise data warehouse (SAS®), an open database (Oracle®), a proprietary database (extract to ASCII text files), a spreadsheet (Excel®), or even on paper.

2. Identify problems in the data and clean the data. This would include conversions to standard units of measure, removing outliers, or standardizing corrective action terms.

3. Submit a request for programming resources to write a unique and custom application to clean the data and combine appropriately.

4. Run the program or tool to do the desired analysis.

5. Distribute the results.

Have a new question? Go back to step 1.

## EVOLUTION OF THE CURRENT SITUATION

Most manufacturing companies have both operational and quality improvement systems at work in many departments. Various quality improvement systems are in place at different locations within the company. On one production line a particular SPC system is used because of the high data rate generated by a check-weighing device. A different SPC system is used earlier in the production line due to the variety of instruments needed for measurements. A LIMS package is used to record tests run by QA analysts in the laboratory. Figure 1 shows how various systems might be deployed within a typical company. These systems serve a particular need and are part of the total quality picture at the company; however, the data are owned by each department or group.

Corporate Office
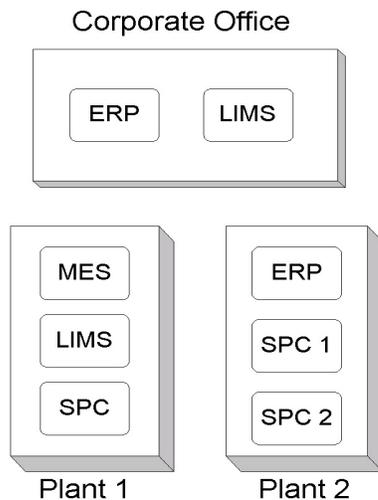


**Figure 1: The Current Situation**

These systems serve the transactional needs within each department of the company. At the time they were implemented, collecting data was the most pressing need. Current transactional systems have added open architectures in an attempt to address data sharing requirements. While this makes it technically feasible to combine data from the different sources, the reality is that no one system combines the data for analysis needs. Within each

transactional system, the data are collected in different formats. The open architectures allow access to the underlying data, but terminology and business rules are imbedded in the systems. For quality improvement, the business rules must be included in the analysis as well. Complicating matters is the fact that the operating systems where the data reside might be UNIX®, AS/400®, or DOS/Windows®.

So what about the quality needs for the entire organization? Individual systems have a wealth of detail data. To make enterprise level quality decisions, data from these individual systems must be combined into meaningful information. Systems have come online over time. Slightly different requirements were used in choosing each system. On top of that, different technologies were available as individual plants went into production; thus, there is a combination of old and new technology meshed together. Collectively, millions of dollars have been spent to get the production facility to the level of automation that exists today, and it is prohibitive to think of dismantling the existing structure to obtain all new technology.

## TRANSACTIONAL NEEDS VS. ANALYTICAL NEEDS

Transaction systems are better understood at manufacturing plants. Systems have been in place to automate production and control individual machines. Online Analytical Processing (OLAP) has data needs that are different from transactional systems. Online Transaction Processing (OLTP) systems are optimized to provide quick data collection with feedback that is limited to what is needed for direct machine control. SPC systems that are tied into OLTP are specialized to standard control charts for a single process. OLAP needs access to current data, but not to the extent of OLTP systems. OLAP will also sacrifice response time to get the information required, whereas OLTP must respond quickly so the production line will not be disrupted. IT and operations are usually happy to off-load data from transactional systems so that OLAP can be managed separately from OLTP systems. A main point of the Quality Data Warehouse is that you don't **monitor** real-time processes from the data warehouse, you determine **what to monitor** from the data warehouse.

Note that this does not preclude the data warehouse from supporting real-time feedback. For real-time feedback to be meaningful, a historical context is needed. The data warehouse provides this context. Real-time systems can use historical information from the data warehouse along with current process measurements to provide feedback that allows real-

time decision support.  McClellan (1997) discusses the future of MES and how data warehouses and decision support systems will need to be integrated with the operational environment.

**DO ENTERPRISE RESOURCE PLANNING SYSTEMS SOLVE THE PROBLEM?**

Enterprise Resource Planning (ERP) systems are currently being installed at many companies.  These installations promise much better integration between functional business units.  This will greatly improve the data sharing issues needed for enterprise level data analysis.  The current ERP systems are very good at integrating business transactions.

However, ERP systems are not designed to provide the statistical analysis needed for decision support. Data are stored in a transaction optimal format. Some vendors are developing data warehouse capabilities to address analysis needs.  Currently these ventures are not mature.  The quality engineer should carefully evaluate how much work is needed to integrate other systems (for example, shop floor control systems) that are not part of the ERP package.  In addition, the statistical analysis tools provided by ERP vendors are not robust.  This is especially true in the area of statistical quality improvement.

**WHAT INFORMATION DOES A DATA WAREHOUSE DELIVER?**

A data warehouse delivers information across the enterprise; one version of the truth.  This allows meaningful comparisons between plants, production lines, and products.  The data become information meaningful for all levels of decision-makers within

the company.  For the IT staff, data are in a clean, consistent, and documented format.  For the engineer or analyst, data are convenient, in a common format, and if desired, exportable to other common formats.   The key difference in the data warehouse is that the data are geared toward analysis, not transactions.

With data from each stage of the production process readily available, it becomes possible to explore relationships across the production line.  You can determine how variation in earlier stages of the process affects later stages of the process.  This can lead to batches being scrapped much earlier in the production process, where the loss is less costly. You can also determine which factors are most influential on the final product quality in order to facilitate monitoring of those factors much earlier in the process.

The Quality Data Warehouse enables decisions to be made closer to the production line.  Historical data can provide a short list of previous problems that are consistent with current operating conditions. Not only are problems known, but also the corrective actions taken and how they worked.  In this manner, guided and informed decisions can be made.

**HOW IS THE INFORMATION DELIVERED?**

The data warehouse is the foundation of analytical decision support.  From the data warehouse, a number of data marts and info marts are populated to support the various analytical needs.  Figure 2 provides a high-level diagram of how these components are related.
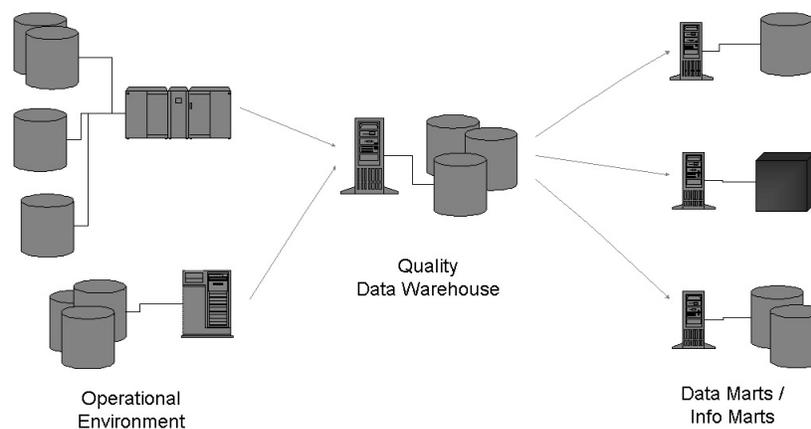


**Figure 2: The Quality Data Warehouse**

A *data mart* is a data file (or files) that have clean data in a format that allows efficient data analysis. Any business rules, such as how to assign materials to production batches, have been applied in creating the data file. The file is ready for analysis, and the engineer does not need to spend time on manipulating the data to the needed format.

An *info mart* is not only the data for analysis, but also reports, charts, and user interfaces. Info marts typically enable a group of users to access information in a format that is conducive to their decision support needs. The info mart might be printed for distribution or published on a corporate intranet. In some cases, a custom 'fat client' interface may be written to navigate the info mart. The interface could also enable business users to act on their decisions from within the interface.

To better understand data marts and info marts, consider the following examples:

Comparing setup results between different plants: A food processing company has multiple production lines that allow products from several different recipes to be produced, based on current demand. Variations in ingredients for different recipes affect settings such as flow rate and dryer temperature. An info mart with a web interface could enable shop floor personnel, at any plant, to review past runs with similar characteristics. Various settings would be displayed along with the measured results. The interface would also enable the operator to input the settings chosen for the current run, so that information would be available for future analysis.

Electrical test data analysis: In semiconductor and electrical component production, electrical testers produce a large volume of data for each component tested. Engineers need a central database with the test results and tag data to identify lot, wafer, and die. The data mart would have the data with the appropriate tag values. An interface to the data mart would allow subsetting for the desired tag values. This data would be in a format that could be analyzed easily using advanced statistical tools.

Support for Six Sigma projects: Six Sigma projects leaders need to know what measurement data are

available that would be relevant to their project. Identifying currently available data prevents duplication of effort. Metadata (explained in more detail later) enables project managers to search the data warehouse to see what measurement data have already been collected. Data within the data warehouse have already been cleaned and appropriate business rules, such as grouping of assignable cause terms, have been applied. The data warehouse enables the project manager to export data to whatever analysis environment is needed. The results of Six Sigma projects can be added to the data warehouse to allow the process improvements to become part of the standard operating procedure.

Other data mart and info mart subjects include enterprise-level SPC, cost of poor quality analysis, total cost of manufacturing, and supplier quality analysis.

**BUILDING THE QUALITY DATA WAREHOUSE**

For most companies, the initial steps to constructing a Quality Data Warehouse are already in place. Data collection is automated and routinely recorded. Some data transfer and conversion techniques have been established. Departments collaborate to the extent that is possible to provide reports and data for interdepartmental analysis needs.

The Quality Data Warehouse does not replace the enterprise data warehouse. If a company has an enterprise data warehouse, the Quality Data Warehouse should be incorporated within it. The data marts and info marts are fed from the enterprise data warehouse. Likewise, once the Quality Data Warehouse is built, it can expand to support other analysis areas.

**STARTING WITH THE DATA MODEL**

The Quality Data Warehouse starts with a data model. A logical data model describes the relationships in the current transactional data. It provides a high-level view of the data without distractions from implementation details. These relationships rarely change, so the model is fairly stable. Figure 3 is a segment of such a data model.
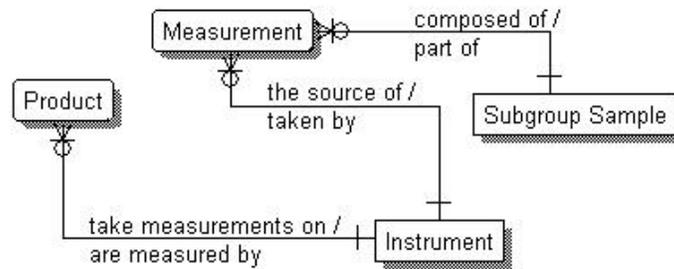
**Figure 3: Logical Data Model**

Logical data models do not contain derived data. For quality improvement, most analysis is built off derived data. For example, the subgroup mean, range, and standard deviation are derived from the subgroup measurements. A data warehouse data model encompasses the derived data. Silverston, Inmon, and Graziano (1997) provide a discussion of the appropriateness of derived data at this point. Also discussed are the criteria for deciding what derived data should be included in the data warehouse data model. The data warehouse data model is the most useful model to the quality improvement analyst. Figure 4 shows the previous logical data model segment with derived data added.

**SOURCES FOR DATA TO THE QUALITY DATA WAREHOUSE**

Transactional systems provide the majority of the data for the Quality Data Warehouse. The formats of the data in the respective systems are (and should be) tailored to the transactional needs addressed by each particular system. Example

sources for the Quality Data Warehouse are as follows:

- process measurement data from the SPC systems. This data would also contain out-of-control flags detected on the shop floor. Cause indications and corrective actions entered by operators come from this data source.

- production scheduling data from a MRP or ERP system

- materials data from a Supply Chain system, including quality characteristics supplied by vendors

- production execution data from an MES

- QA lab data from an LIMS

- customer complaint data from the call center

- warranty claim data
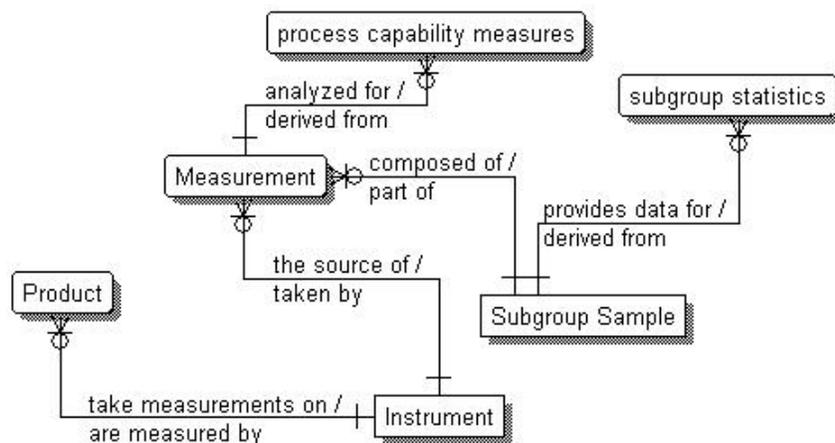


**Figure 4: Data Model with Derived Data**

**WHO NEEDS TO BE INVOLVED?**

Typically, a data warehouse project involves a number of departments. The purpose of the Quality Data Warehouse is to address the analytical needs of the quality organization within the company. Therefore the quality division is the driver for the project, on both the corporate level and the plant level. Product engineering has an interest in the Quality Data Warehouse as well, for answering questions related to product design and production. Plant operations is involved since their transactional systems are the source of much of the data. IT support is required to implement and update the data warehouse. Executive management should be the sponsor of the project and ultimately is the recipient of improved decision support information from all levels of the organization. The successful data warehouse project stresses the enterprise benefits of the warehouse.

Quality management and quality engineers have a critical role in the data warehouse project. IT personnel will have resources that are available for the general process of building and maintaining a data warehouse. The responsibility of the quality organization is to supply the subject knowledge. This subject knowledge is what differentiates the Quality Data Warehouse from a standard business data warehouse. Requirements must be specified to ensure that the data are combined correctly for analysis; otherwise, the warehouse may fail to provide accurate information. Examples of typical failures to be aware of include

- improper data aggregation
- incorrect subgrouping
- failure to allow for multiple sources of variation
- insufficient structure for traceability
- no support for measures of uncertainty

For additional information on successful data warehouse project management, refer to Inmon (1996).

**METADATA**

Another benefit of the Quality Data Warehouse is the metadata, or the 'data about data'. Metadata provides information on where data values came from and how they were derived. Metadata tells what data are available, and how current they are. For data warehouse administrators, metadata provides technical information on problems encountered in the data loading process. Metadata can be used to document the quality of a process.

Metadata also tells how the data are linked. A variety of identifiers (or tags) are placed on data

values. For analysis, these identifiers must be connected to determine relationships. Building the metadata provides these links. One example would be a link between the subgroup statistics and the control limits that were in effect when the subgroup sample was taken. In many cases the links can be built automatically when the data warehouse is populated. If manual input of key fields is required, the data warehouse can identify what values are missing. A data entry and notification system can be put in place to facilitate quick and accurate manual input.

**DATA WAREHOUSE BASICS**

Similar to how transactional systems are chosen to address operational needs, the data warehouse is modeled to address analytical needs. The physical implementations of data warehouses are modeled by defining *facts* and *dimensions*.

*Facts* are the analysis variables in the model. They are the input values used to generate the statistics of interest. It is important to note that the fact values themselves are usually a statistical result. It is important to determine the granularity needed to support the analysis desired from the data warehouse. Some example facts for the Quality Data Warehouse are

- subgroup statistics (instead of individual data measurements)
- process status (in-control or not)
- process statistics, such as $C_p$ and $C_{pk}$
- attribute data, such as defect counts
- control limits, specification limits, and target values

Note how the level of granularity depends on the fact. The subgroup mean and range are calculated for a particular sample, where as $C_p$ and $C_{pk}$ pertain to a whole process (typically over a specified phase of production).

*Dimensions* are the categorical variables in the model. They are chosen to support the classifications that are needed for analysis. Dimensions in the Quality Data Warehouse would include

- product identifiers, including lot or batch
- process identifiers, such as machine and measurement characteristics
- time intervals (shift, day, month)
- supplier identifiers
- process change or shift identifiers
- defect categories

The facts and dimensions define the logical data warehouse. Once the data warehouse data model is established, the physical data model can be determined. The physical data model defines the actual storage architecture of the data warehouse. Design constructs for the physical data model include the following:

- summary tables: These are typical relational tables that contain summarized data.

- multidimensional tables: Multidimensional tables contain summary statistics that can be accessed at various crossings. They allow the 'slice and dice' type of ad hoc analysis. They are also called multidimensional databases or cubes.

- star schema: This structure has a central fact table with keys to several dimension tables. The dimension tables contain the identifiers to break down categories. They provide optimal access to the fact values by allowing subsetting to be done on the dimension tables and then access to the fact table for analysis.

- snowflake schema: The snowflake structure is a derivation of the star schema. This structure is used when a dimension table has subcategories that can be broken down to allow more efficient access.

For descriptions of data warehouse constructs, refer to Kimball (1996) and Welbrock (1998).

Populating the data warehouse is performed on a scheduled basis. The frequency of the schedule is determined by the timeliness of data needed for analysis. The timing of the extract from transactional data is determined by the operational needs of the transaction system. Frequently, data warehouses are updated on a nightly or weekly basis. It is possible to update portions of the data warehouse on different schedules.

The basic implementation of populating and updating the data warehouse follows these steps:

- data access and extract. Data are brought in from Online Transaction Processing (OLTP) and external sources.
- transformation. The umbrella term for the tasks that prepare data for loading into the warehouse.
- loading the physical data warehouse.

**HISTORICAL DATA REPOSITORY**

OLTP systems must provide quick response to the shop floor. To provide this response, historical data are only available for a short time. OLAP needs access to data for a longer time frame, although it may not be required to access data to the same level of detail that is provided from the transaction system. The Quality Data Warehouse provides an efficient storage mechanism for historical data that are needed for analysis.

**DELIVERING THE INFORMATION AND MAKING BETTER DECISIONS**

Delivering information effectively and efficiently is facilitated through the use of data warehouse technology. A growing list of analysis and exploitation tools are available, ranging from basic quality improvement techniques to reliability analysis and simple predictive modeling to sophisticated knowledge discovery methods such as neural networks and decision trees. Online Analytical Processing (OLAP) and decision support systems enable ad hoc data analysis to be performed by a diverse group of employees within a company. New technology in the form of intranet and web-based interfaces can be used to distribute results to a large number of people at many required locations.

**CONCLUSION**

Analytical needs for enterprise-level quality improvement are not being met. A large obstacle to meeting these needs is data accessibility. Data are available in various transactional and SPC systems throughout the organization, but analytical processing directly on these sources is not desirable. Data must be collected from these various systems and transformed to match structures that are needed for analysis. Transactional processing requires quick response time to transaction requests, which may be hampered by analytical applications that run directly against transactional data. Data mining techniques are being used in other industries to discover previously unknown relationships in the data, and there is a great deal of interest in applying these techniques in the manufacturing quality area as well. Before data mining methods can be applied, a data warehouse with analysis-ready, clean data must exist. A subject-oriented Quality Data Warehouse is needed to serve the decision-making needs of the enterprise.

**REFERENCES**

Inmon, W. H. (1996), *Building the Data Warehouse,* New York, NY: John Wiley and Sons, Inc.

Kimball, R. (1996), *Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses,* New York, NY: John Wiley and Sons, Inc.

McClellan, M. (1997), *Applying Manufacturing Execution Systems,* Boca Raton, FL: St. Lucie Press.

Silverston, L., Inmon, W. H. and Graziano, K. (1997), *The Data Model Resource Book:  A Library of Logical Data Models and Data Warehouse Designs,* New York, NY: John Wiley and Sons, Inc.

Welbrock, P. (1998), *Strategic Data Warehousing Principles Using SAS Software,* Cary, NC: SAS Institute Inc.

**CONTACT**

Brad Klenz
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Email: brklen@wnt.sas.com