International Journal of Recent Engineering Science (IJRES),
ISSN: 2349-7157, *volume1 issue5 September to October 2014*

Data Quality in Manufacturing Data Warehouse

Dr.Sreenath, S.Karunagaran

[1]Assistant Professor, [2]PG Student,
Department of Mechanical Engineering,
Acharya Institute of Technology, Bangalore, India

Abstract

Data quality has become integral to many organizations as they build data warehouses and focus more on customer relationship management (CRM). This is especially true in the Manufacturing domain where cost pressures and the desire to improve quality of customer experience drive efforts to integrate and clean organizational data. This paper reviews earlier work on data quality and extends it by providing a process model of architected data environments. This model allows practitioners and researchers to focus on processes that generate data quality problems. The paper describes how the model was used in a real world scenario and subsequent implications for practitioners and researchers.

1. Introduction

For manufacturing units, data is central to both effective quality control and to financial survival. Data about the quality of products, the rate of defect, and the process of manufacturing is crucial to organizations that strive to maintain and improve manufacturing process. Every manufacturing unit is striving to reduce the cost of production and also improve the overall quality of the product to gain an edge in the competitive market. Cutting the cost of production as well as improving the quality of the manufactured item is a major challenge. To meet these seemingly imposing challenges, manufacturing industry has begun to integrate data from its management information systems and its financial information systems. This integration is time consuming and costly and also poses challenges with respect to the quality of data.

The purpose of this paper is to examine the issues manufacturing organizations face in trying to deliver high quality information to operations and financial end-users in an environment with many diverse source systems and organizational units with different business rules. Specifically, the paper examines a data environment that includes various source systems, data marts, and data warehouses [7]. This paper reviews existing research on definitions of data quality, the importance of data quality, methods used to insure data quality, and processes that affect data quality. A model for understanding data quality issues in this environment is developed. Recommendations are made for applying the model to the manufacturing units.

2. Data Quality

Many authors have put in significant effort in defining the term *Data Quality*. This is not just an academic exercise because the definition will be tied to specific dimensions and measures in order to support data quality improvement efforts. Traditional information processing thinking about data quality was concerned with accuracy, precision, and timeliness. Levitin and Redman [14] assert that two important considerations for data quality are insuring that data models are clearly defined and that data values are accurate.

Jarke, et.al., has tried to elaborate the concept of data quality by making data quality a function of various factors namely, interpretability, usefulness, accessibility, believability. The said work is organized in a hierarchical tree as shown in figure1.

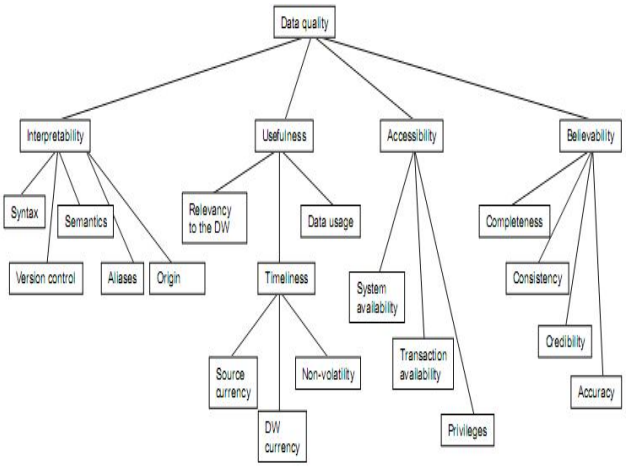


Figure 1: Quality Factors for Data Warehouse.

Wang and Strong [27] adopt an empirical approach in defining data quality. The authors used a marketing research survey to gather data on perceived data quality

attributes and combine them into various dimensions and categories.

Wand and Wang [25] take a theoretical approach to defining quality data. They use a set of assumptions, definitions, and postulates to derive data quality dimensions using a design-based perspective. Their overall premise is that data in an information system should reflect aspects of the real world system. Data deficiencies can be identified where the mapping between the information system state and the real world state break down. Design deficiencies consist of incomplete representation, ambiguous representation, and meaningless states. Operation deficiencies, e.g., garbling, result when real world states are not mapped to information system states properly at operation time. Finally, decomposition-related deficiencies result when properly mapped individual state elements don't map properly when combined at a higher level. As a result of these deficiencies, the data in the information system can be incomplete, ambiguous, meaningless or incorrect.

Strong, et.al., [20] take a consumer focused view that quality data is "data that is fit for use by data consumers." This paper will use the dimensions of Wang and Strong but will also add a relativistic perspective to the view. "The term "data quality" can best be defined as "fitness for use," which implies the concept of data quality is relative. Thus data with quality considered appropriate for one use may not possess sufficient quality for another use. The trend toward multiple uses of data, exemplified by the popularity of data warehouses, has highlighted the need to address data quality concerns." [22] For our purposes, "fitness for use" for the wide variety of users in an integrated health care organization will be the primary determinant of data quality. Underlying this fitness will be the dimensions of information quality given in Table 1.

Information Quality Category	**Information Quality Dimensions**
Intrinsic	Believability, Reputation, Accuracy, Objectivity.
Accessibility	Access, Security
Contextual	Timeliness, Completeness, Amount of Data
Representational	Interpretability, Ease Of Understanding, Concise representation, Consistent representation.

Table 1: Data Quality Dimensions [27].

It is important to observe that in any typical manufacturing process, quality is injected in every process from the very start. For example, a casting metal foundry technician systematically monitors a melt to assess that the required composition of metal compounds like carbon, iron, nickel, chromium, zinc etc. are in place prior to pouring to ensure the quality of the desired metal casting. Similarly, it is imperative that data quality needs to be injected in every phase of information system design and implementation with due diligence to governance, monitoring and auditing, among other things.

3. Issues to be addressed

Before we proceed with the task of carrying out an assessment of data quality, there are several issues which need to be tackled. For example,

- Business requirements documentation is non-existent, not maintained upon change, too high-level, lacks integrated enterprise viewpoint, and lacks supporting business processes.
- Business rules are buried in program code which results in higher maintenance costs, dependency on specialized skills, and a lack of awareness.
- Undocumented definitions and missing semantics.
- Inability to audit and monitor changes to the architecture and contained data.

The issues mentioned are only a tip of the iceberg. The list demonstrates the need to address these issues throughout the development lifecycle of systems as shown in figure 2.

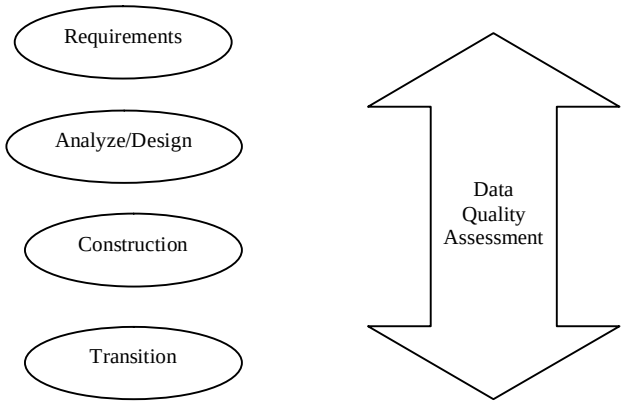


Figure 2: Generic SDLC and Quality Assessment.

4. Data Quality Processes

A systematic approach to data quality would involve the implementation of a quality assurance process. Several authors have provided descriptions of processes to ensure and improve data quality. Redman [17] described three strategies for improving data quality:

1. Identify the problem,
2. Treat data as an asset, and
3. Implement more advanced quality systems.

To implement the first element of the strategy, identify data problems; consider whether the organization is extensively inspecting and correcting data, has significant redundant data, has enough quality data to support key strategic initiatives and re-engineering efforts, and has data users and managers who are frustrated with current data quality. Implementing strategy element 2, treating data as an asset, involves inventorying data, recognition of the value of the processes that create data, assignment of responsibility for data quality, establishing a customer supplier relationships for data, finally, investing in the quality of the asset. Implementing more advanced quality systems, the final strategy element, requires (1) defining processes for error detection and correction, (2) implementing process management to discover and eliminate root causes of data problems, and (3) redesigning processes to make them less error-prone.

Wang [26] proposed a five step approach based on the Total Quality Management concepts developed in the manufacturing realm. This approach is called Total Data Quality Management (TQDM). The approach consists of four components:

1. Definition: identify important data quality dimensions,
2. Measurement: define metrics for data quality,
3. Analysis: determines causes for data quality problems and the effects of poor quality, and
4. Improvement: take action based on analysis to Improve data quality.

In contrast to the manufacturing approach, Kaplan, et. al. [9] describes the assessment of data quality in accounting information systems and attempt to generalize to other type of systems. Graphical Models are used to represent the accounting systems and incorporate representations of forms, processes, transformation points, control procedures, general ledger accounts, and relationships. Based on interviews with professional accounting information system auditors, the authors laid out a four step Accounting Information System Data Quality Assessment Process. The steps include:

1. Develop a minimal set of target error classes,
2. Select the minimum set of controls to test for Reliability for each path in the graphical model,
3. Establish the minimal level of testing needed for each control to ensure the target error classes are detectable at the desired level of assurance, and
4. Run tests to insure that controls are working at the target assurance levels.

The authors develop a model to determine the smallest set of controls that need to be tested given a graphical description of the accounting information system and the set of target error classes.

5. Process Model and Data Quality

This section first introduces the generic process model (see Figure 2) and relates it to the health care field. After introducing the model, the paper describes its use in identifying data quality opportunities/problems that should be considered in a data quality assurance process. An architected data warehouse environment includes the following components (see Figure 2):

- Data Source Systems,
- Data Warehouses,
- Data Marts,
- End user analysis Tools
- Transformation/Translation Tools.

Data source systems are those systems that capture the original data. They are often Online Transaction Processing Systems (OLTP) and are primarily concerned with representing the current state of an organization and with processing the organization's transactions. In the Manufacturing field these systems tend to be concerned with Process scheduling, financial transactions, and inventory information about current stock. Data warehouses are "subject oriented, integrated, nonvolatile, and time variant collections of data in support of management's decisions." [7].In a manufacturing organization it is the place where Process scheduling, financial and inventory data come together to support analyses and decisions that combine those subject areas. A datamart, in this context, is a subset/aggregation of the data warehouse that is designed to support a specific organizational unit or a specific organizational process. In some contexts, datamarts are small specialized data warehouses that exist without depending on an enterprise level data warehouse for data. We will not consider that situation here. In this analysis, datamarts are created for

performance and user interface reasons. They contain data that is redundant (i.e., is duplicated or can be derived from data warehouse data). End-user data analysis tools are client programs that allow end-users to access and analyze data from the data warehouse and/or datamarts. These tools could be thin clients (e.g., web browsers) that display the results of simple queries or they could be sophisticated fat clients that download a subset of data (a data cube or a pivot table) and then analyze that local data in powerful ways.

Connecting the other elements together are hardware/software technologies that extract, transform, translate, cleanse, monitor, and transport data. These tools range from low level input/output tools to meta-level tools that manage the entire architected data environment.

The starting point of the data quality analysis is to create a customized process model for the target organization. Different manufacturing units will have different numbers of source systems, different numbers and levels of datamarts (or no datamarts at all), different numbers and types of end-user tools, and different cleansing, transformation, and transportation needs. The generic model is flexible enough to accommodate those differences and still guide the data quality analysis that follows.

The main contribution of this paper is to describe how the process model of an architected data warehouse environment [Figure 2] can be used to guide data quality assurance. The basic approach is to start at the end of the process (i.e., Management Report/Ad hoc Query) and work back to the beginning (i.e., Source Systems).

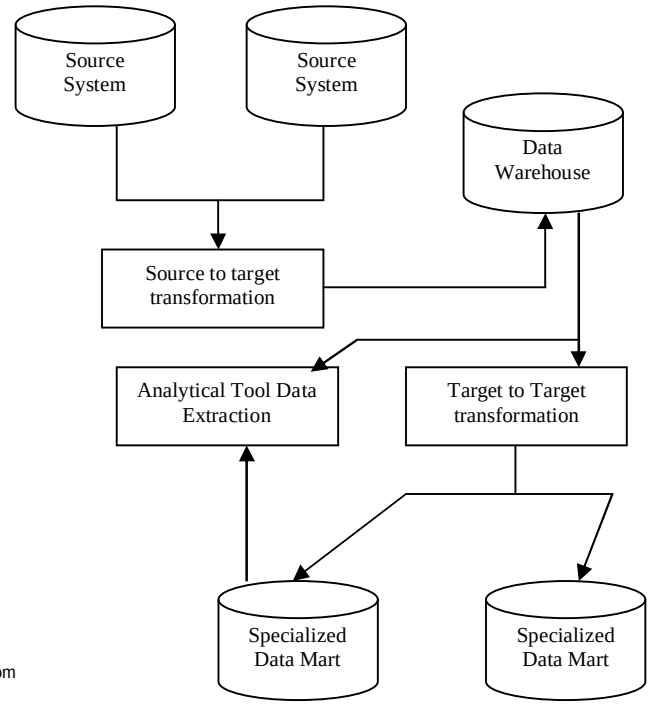
The goal of the entire data warehouse effort is to produce management reports/ad hoc queries that are "fit for use" by the decision makers in an organization. Working backward from this final product, we can assess the quality constraints that determine the success of these reports. The quality of the data in the reports/queries is determined by (1) the quality of the data in the local analytical data cube (or file or pivot table), and (2) by the quality of the report/query specification. Most analytical tool clients have sophisticated report writing and analysis capabilities that can produce inaccurate and misleading results even if the local data set contains no errors. Formulas, constants, and data can be hidden or confusing making it hard for end-users to detect errors or faulty assumptions. This part of data quality assurance is often ignored or left up to end-users.

Moving upstream from the analytical tool client, the quality of the data in the local analytical data cube is dependent on (1) the quality of the data in the datamart (or data warehouse), and (2) the quality of the transformation process that loads the data cube. If the process that loads the local data cube has errors, then those errors will be transferred to the reports and queries. Local data cubes may be considered the responsibility of end-users and therefore may also be left out of standard data quality assurance programs.

Moving further along the data stream, the quality of data in the datamart is dependent on the quality of data in the data warehouse and the target-to-target extraction/transfer processes that load the datamart from the warehouse. This process is made more complicated if there are differences in operating systems, database management systems, and hardware platforms between the datamarts and data warehouse. Problems in the Target-to-Target Transformation process result from bugs introduced during the development of the process to operational failures that happen during normal load processes.

Moving further, the quality of data in the data warehouse is dependent on the quality of data in the various source systems and on the quality of the extraction, cleansing, transformation, and transfer processes that make up the source-to-target transformation. The quality of data in the source systems is determined by many factors including data entry controls, edit controls, system changes, down time procedures, hardware/software bugs, etc. The quality of the source-to-target transformation is critical to the loading of quality date into the data warehouse. This process is made more difficult as the number and variety of the source systems increases. As with the Target-to- Target loading, problems could be introduced during development, maintenance/upgrades, or operations.

The result of using the process model as the basis for analyzing data quality in the data warehouse environment is that the sources of data quality problems are systematically identified and steps can be taken to monitor and improve this quality.



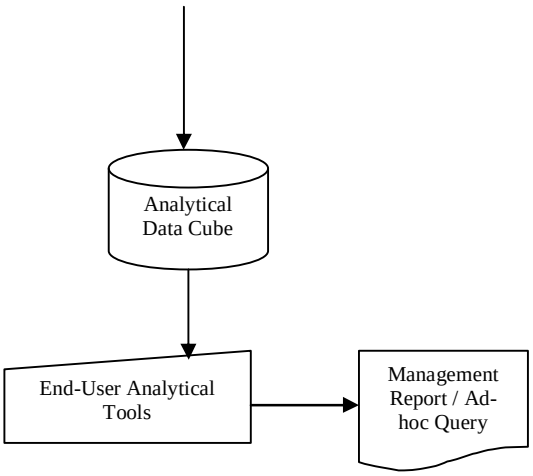


Figure 2: Process Model

8. Conclusions and Recommendations

The model presented in this paper contributes to the literature on data quality by detailing a process model that can be applied to the manufacturing and other industries to help in the assurance of quality data for decision making. Other organizations can apply the generic model [Figure 2] to their situations and identify components that affect the quality of data in their reports and queries. They can apply the Development/Operations grid [Table 2] to those components to help them generate ideas for maintaining and improving data quality in their firms.

Research is needed to identity which specific response to development and operational data quality problems works best at which step in the data warehouse process. One of the challenges the health care organization that was studied faced was how to convert the lists of issues and causes that came out of the analysis into specific priorities and programs. Researchers can help practice by finding better ways to do this linkage. A significant contribution can be made by improving decision making through the use of better organizational data.

References

1. Ballou, Donald P., Tayi, Giri Kumar. Enhancing data quality in data warehouse environments, *Association for Computing Machinery. Communications of the ACM*; New York; Jan 1999, 42, 1, 73-78.

2. Burch, George. Clean data, *Manufacturing Systems*; Wheaton; Apr 1997, 15, 4, 104-108.

3. English, Larry P. Help for data-quality problems, *Informationweek*; Manhasset; Oct. 7, 1996, 600, 53-62.

4. Foley, John. Data warehouse pitfalls, *Informationweek*; Manhasset; May 19, 1997, 631, 93-96.

5. Francett, Barbara. Marts keep data on the move, *Software Magazine*; Englewood; Mar 1997, 17, 3, 55-60.

6. Henderson, Mary, Integrated health care management through comprehensive info, *Benefits Quarterly*; Brookfield; Second Quarter 1995, 11, 2, 48.

7. Inman, W. H. Building the Data Warehouse, 2nd Edition. John Wiley & Sons. New York, 1996.

8. Institute of Medicine. Press Release: Preventing Death and Injury From Medical Errors Requires Dramatic, System-Wide Changes , Nov. 29, 1999.

9. Kaplan, David, Krishnan, Ramayya, Padman, Rema, and Peters, James. Assessing data quality in accounting information systems, *Association for Computing Machinery. Communications of the ACM*; New York; Feb 1998, 41, 2, 72-78.

10. Kay, Emily. Dirty Data challenges warehouses, Software Magazine; Eaglewood: Oct 1997.

11. Kenyon, William W. Analysis of the collection cycle, *Journal of Health Care Finance*; Gaithersburg; Fall 1993, 20, 1.

12. Khalil, Omar E M, Harcar, Talha D. Relationship marketing and data quality management, *S.A.M. Advanced Management Journal*; Cincinnati; Spring 1999, 64, 2, 26-33.

13. Krill, Paul. Data warehouses have need for clean data, *InfoWorld*; Framingham; Mar 16, 1998, 20, 11, 27.

14. Levitin, Anany V., and Redman, Thomas C. Data as a resource: Properties, implications, and prescriptions, *Sloan Management Review*; Cambridge; Fall 1998, 40, 1, 89101.

15. Oman, Ray C., and Ayers, Tyrone B. Improving Data Quality, Journal of Systems Management, May 1988, 31-35 .

16. Orr, Ken. Data quality and systems theory, *Association for Computing Machinery. Communications of the ACM*; New York; Feb 1998; 41, 2, 66-71.

17. Redman, Thomas C. Improve data quality for competitive advantage, *Sloan Management Review*; Cambridge; Winter 1995, 36, 2, 99;

18. Redman, Thomas C. The impact of poor data quality on the typical enterprise, *Association for Computing Machinery. Communications of the ACM*; New York; Feb 1998, 41, 2, 79-82.

20. Strong, Diane M., Lee, Yang W., and Wang, Richard Y. Data quality in context *Association for Computing Machinery. Communications of the ACM*; New York; May 1997, 40, 5, 103-110.

21. Tarplee, Sue, and Cassidy, Bonnie. Medical record department's leadership role in receivables management, *Journal of Health Care Finance*; Gaithersburg; Fall 1993, 20, 1, 41.

22. Tayi, Giri Kumar, and Ballou, Donald P. Examining data quality, *Association for Computing Machinery. Communications of the ACM*; New York; Feb 1998, 41, 2, 54-57.

23. Walera, Edward J., and Button, Charlie. Using a supply usage relational database to reduce costs, *Healthcare Financial Management*; Westchester; Sep 1997, 51, 9, 35-38.

24. Wallace, Bob. Data quality moves to the forefront, *Informationweek*; Manhasset; Sep 20, 1999, 738, 52-67.

25. Wand, Yair, and Wang, Richard Y. Anchoring data quality dimensions in ontological foundations, *Association for Computing Machinery. Communications of the ACM*; New York; Nov 1996, 39, 11, 86-95.