

Data Quality in Health Care Data Warehouse Environments

Robert L. Leitheiser
University of Wisconsin – Whitewater

Abstract

Data quality has become increasingly important to many firms as they build data warehouses and focus more on customer relationship management. This is especially true in the health care field where cost pressures and the desire to improve patient care 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 allow practitioners and researchers to focus on processes that generate data quality problems. The paper also describes how the model was used in a real world health care organization and what implications there are for practitioners and researchers.

1. Introduction

For health care organizations, data is central to both effective health care and to financial survival. Data about the effectiveness of treatments, the accuracy of diagnoses, and the practices of health care providers is crucial to organizations that strive to maintain and improve health care delivery. Hospitals, clinics, and other health care facilities are also under increasing pressure to hold down costs that is being driven by managed care organizations and increasingly stingy state and local governments. Good cost, charge and payment data is essential to keeping costs down and remaining competitive. The health care industry is unique in that it needs to bring together efforts to improve the quality of individuals' health with the effort to cut costs to employers and governments. These are two very different kinds of customers who have seemingly opposing goals. To meet these goals, health care organizations are bringing together, often for the first time, data from their clinical information systems with data from their financial systems. This integration is costly and time consuming. It also poses special problems for data quality.

The purpose of this paper is to examine the issues health care organizations face in trying to deliver high quality information to clinical and financial end-users in an environment with many diverse source systems and organizational units with different business rules. Specifically, the paper examines an architected data

environment that includes various source systems, data marts, and data warehouses [7]. This paper reviews existing literature 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 and applied to a mid-sized hospital based health care organization. Recommendations are made for applying the model to other health care organizations and for further research.

2. Defining Data Quality

Significant effort has gone into defining what is meant by "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.

More expansive views are now widely held. For example, Strong, et.al., [20] take a consumer focused view that quality data is "data that is fit for use by data consumers." A data quality problem, then, occurs when there is a difficulty with one or more of the data's quality dimensions that makes the data unfit for use. Strong, et.al., get to this view by treating data processing as a data manufacturing system. In a data manufacturing system there are three important roles: data producers, data custodians, and data consumers. Data producers are people, groups, or systems that generate data. Data custodians are people who provide and manage computing resources for storing and processing data. Finally, data consumers are people, groups, or systems that use data. As the final consumer, data users are critical in defining data quality.

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.

Wang and Strong [27] take an empirical approach to defining data quality. The authors use a marketing research survey to gather data on perceived data quality attributes and combine them into dimensions and categories. The final results of their two-phase study are presented in Table 1.

Table 1: Perceived Data Quality Dimensions	
Information Quality Category	Information Quality Dimensions
Intrinsic	Accuracy, Objectivity, Believability, Reputation
Accessibility	Access, Security
Contextual	Relevancy, Value-Added, Timeliness, Completeness, Amount of Data
Representational	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation
From Wang, Richard Y., and Strong, Diane M. [27]	

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.

3. Importance of Data Quality

Many of the things organizations are trying to do are affected by data quality. For example, the recent emphasis on Customer Relationship Management is dependent on high quality customer data [19]. In the

health care industry improved data quality is linked to better management of health plans [6], improved accounts receivable practices [11, 21], and better management of health care supplies [23].

While these benefits are all financial they are also potentially significant clinical benefits to improved data quality. Recently, the Institute of Medicine [8] shocked the public with a report that estimated that 98,000 people die every year from medical errors. Some of the errors are the result of missing or bad information about drugs, orders and treatments. While the focus of this paper is on data quality for decision support applications rather than for on-line patient management, there are also potential health consequences for bad data being used for decision making about cost-effective treatment plans for certain types of diagnoses.

In this more competitive environment, hospitals are turning more and more to customer relationship management to help them be more competitive. Khalil and Harcar [12] describe a private hospital in Louisville, Kentucky where patients can be tracked from the point of initial contact through all subsequent interactions with the hospital. The database also provides the information needed to implement the hospital's targeted direct mail programs. They conclude that "effective use of relationship marketing strategy requires excellent data quality."

More generically, Redman [18] has evaluated the impacts of poor data quality on organizations. He reports that "Unless an enterprise has made extraordinary efforts, it should expect data (field) error rates of approximately 1-5%. (Error rate = number of erred fields/number of total fields.)" Poor data quality has adverse affects at the Operational, Tactical, and Strategic levels of an organization. At the operational level, poor data reduces customer satisfaction, increases operational costs, and reduces employee job satisfaction. At a tactical level, poor data quality leads to poor decision making, more difficult data warehouse implementation, more challenging reengineering efforts (which are based on data analyses), and an increase in organizational dissention. Finally, at the strategic level, poor data quality negatively impacts on strategic decision making, on strategy execution, on issues of data ownership, and on the ability to focus management's attention on truly important business issues. While the author does not have hard numbers on these costs, he does make persuasive logical arguments on their importance.

IT executives appear to be buying these arguments. Wallace [24] reports in a recent survey of 300 IT executives that "81% of survey respondents ranked improving customer data quality the most important post-year 2000 technology priority." Other high data priorities were to improve data mining (#6) and data warehousing (#7). About 70% of respondents said they expected to

significantly increase spending for improving customer data quality in the next 12 months. This was the second highest spending category increase. As a result of this spending, the worldwide market for products and services for data cleansing, translation, mapping, and analysis is estimated to be over \$100 million in 1998 and could grow to \$150 million by 2001 (from the Aberdeen Group).

While data quality is important for all systems, it is especially important, and visible, in architected data warehouse environments [4, 10, 5]. A 1997 survey by the Data Warehousing Institute of 320 warehouse managers found data quality to be an especially critical issue for data warehouse success. [13]

4. Challenges to Data Quality

Understanding the meaning and importance of data quality is important, but the major challenge faced by organizations is to ensure and improve the quality of data in information systems.

“Now, no serious information system has data quality of 100%. The real concern with data quality is to ensure not that the data quality is perfect, but that the quality of the data in our information systems is accurate enough, timely enough, and consistent enough for the organization to survive and make reasonable decisions.” [16]

In order for this to happen, Orr [16] says two things must happen: (1) a comparison must be done between system data and the real world, and (2) any differences must be corrected. While this sounds obvious and simplistic he goes on to add an important condition to achieving high data quality: the data must be used. The more its used, the longer its used, and the more stringently its used, the better the quality of the data will be. Users will find problems and report them. Systems personnel will be focusing on the data that is being used. When data is not used it goes bad. Orr proposes “Use-based data quality programs” based on finding innovative, systematic ways to ensure that critical data is used through the use of audits, system redesign, training, and continuous measurement.

Tayi and Ballou [22] identify difficulties in ensuring data quality. First, the wide variety of structured and unstructured types of data makes ensuring data difficult. Second, data quality is often not given a high priority and is cut during budget squeezes. Third, there are a wide variety of dimensions on which data quality can be affected; e.g., timeliness, accuracy, completeness, etc. Fourth, there often is no widely accepted means of determining the seriousness of data deficiencies. And finally, it may be difficult to define levels of data quality that are appropriate to the organization.

Burch [2] takes a more technical view in describing four major categories of data contaminants that make data warehouse data quality assurances tricky:

- Multiple internal and external data sources
- Free-form text containing multiple hidden or inaccessible business
- Unexpected data values in fields
- Spelling variations for the same real world object

Burch proposes data-reengineering which involves four phases: 1) data investigation, 2) data standardization, 3) data consolidation and enrichment, and 4) data "survivorship." In the last phase the final content and format of data to go into the data warehouse is determined.

Ballou and Tayi [1] further point out that data from different sources may contain both semantic differences and syntactic inconsistencies. Moreover, the desired data may simply not have been gathered. Semantic differences in health care may be differences between the how patients are classified in the hospital setting and in the clinical setting. A syntactic difference would be how gender is coded in a particular system (e.g., 1 vs. "F").

Strong, et.al., [20] studied 42 data quality projects in three different organizations including two in the health care field (a hospital and an HMO). Some of their findings, based on the data categories identified in Wang and Strong [27] are:

1. Intrinsic data quality problems often result from mismatches among sources of the same data, and the introduction of judgment or subjectivity into the data production process,
2. Accessibility data quality problems were associated with technical accessibility, data-representation issues, and data-volume, and
3. Contextual data quality problems often resulted missing data, inadequately defined or measured data, and data that could not be appropriately aggregated.

In summary, there are many ways that data can go bad. Data quality assurance must consider the sources of poor data quality identified by these studies.

5. 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.

Measurement is central to many data quality approaches. Oman and Ayers [15] demonstrate how well it works with a data quality improvement program for a large Federal Government database. For the combined user organizations the % correct value improved from about 60% to over 80%.

In contrast to the manufacturing approach, Kaplan, et. al. [9] describe 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.

Also in the model development category is the work by Ballou and Tayi [1] who propose an integer-programming model to identify data quality projects that would increase the utility of the data in a data warehouse. The model systematically incorporates trade-offs, including the value of obtaining data sets not currently available and evaluating gains from reducing the quantity of stored data. To work, the data warehouse manager must:

- Identify supported organizational activities,
- Determine data needed for those activities,
- Evaluate data quality for each set of data needed,
- Define potential projects for enhancing quality of data warehouse data,
- Estimate impact on data quality for each project, and
- Define the change in warehouse utility caused by each project.

The model is formulated in the paper but apparently not explicitly applied to a real-world situation.

The approach taken for this paper is perhaps closest to that used by English [3] where the author proposes a parallel process for (1) cleansing data and (2) data quality improvement [see Figure 1]. There is an initial development process that includes Data Warehouse Design, Initial Clean Up of Data, an Initial Mapping of Source to Target Systems, and the Initial Load of the warehouse. Data quality must be considered in those processes. There is also an ongoing operational process that involves a continuous Data Quality Improvement Program, a continuing Assessment of Data Quality, and an effort to Improve Defects in the current Process.

English discusses how important the data cleansing step is and describes automated tools to assist with the data cleansing phase.

In summary, data quality has been defined in terms of both its relative fitness and an empirically derived framework of dimensions. The importance of data quality to all organizations, and especially to health care organizations with data warehouses, has been demonstrated. A review of data quality efforts has produced a set of issues to consider and a number of high level multi-step processes to go through. Also reviewed were formal models for describing some aspects of the data quality problem. This paper will extend existing work by describing data quality issues from the perspective of an architected health care data environment. The approach is related to that of English [3] in that it focuses on the developmental and operational processes of such an environment but will be much more explicit in describing the processes and how they lead to data quality challenges and opportunities.

6. Methodology

The methodology used in this study had three steps. The first step was to develop a generic process model that would provide insight into where data quality problems are introduced into the development and operation of an architected data warehouse environment. The second step was to apply that model to the health care field and address the specific needs of that industry. The last step was to introduce the model into an actual health care organization to test its validity and usefulness.

7. 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 warehouse(s),
- Datamarts,
- End-user analysis tools, and
- Transformation/Translation/Transportation 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 health care field these systems tend to be concerned with registration/scheduling, financial transactions, and clinical/health information about current patients.

Data warehouses are "subject oriented, integrated, non-volatile, and time variant collections of data in support of management's decisions." [7] In a health care organization it is the place where registration/scheduling, financial, and clinical 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 health care organizations 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 data 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. To validate the model and this approach, it was applied to an actual health care organization.

8. Application of Model

The generic process model described above and in Figure 2 was applied to the architected data environment of a midsized health care organization located in the Midwestern region of the United States. This organization has historically been hospital based but is expanding into clinics, managed care, and other services in order to offer complete and integrated care to its patients. The data warehouse project is in its third year and is operational with clinical, financial, and registration data.

After tiring of fighting data quality fires the organization decided to invest some effort into developing a systematic process for detecting and addressing quality problems. The first step in this process was to use the elements of the generic process model [Figure 2] to describe the specific architected data warehouse environment in the organization. The result is illustrated in Figure 3. At the time of this study there were three principal source systems: a clinical system for managing patients in the hospital, a registration system for registering and scheduling patients, and a financial system for billing and managing payors.

The organization had one large data warehouse that had enterprise-wide scope (i.e., the Enterprise Data

Warehouse or EDW) but at the time of this study was principally used for supporting hospital-based decision making and reporting.

Several approaches were taken for the Source to Target Transformation process depending on the characteristics of the source system. The principal tool for loading the EDW was provided by the clinical source system vendor and was required because of the nature of the proprietary data storage used in that source system. The registration source system also used proprietary storage requiring contracting with the vendor to write the extraction programs. The financial system used a standard relational database product for data storage facilitating the extraction process. This variety of source systems and the use of proprietary data storage make data quality assurance for the Source to Target Transformation process in this organization a difficult challenge. Most of the quality problems that have been identified to date come from this part of the overall process.

At various times in the history of the EDW there have been different numbers of datamarts. Each specialized datamart presented its own quality challenges but since they were all developed using standard relational database products they were relatively easy to manage.

The organization selected a powerful report writing tool as its standard end-user interface to the EDW and datamarts. This product allowed the creation of local data cubes that could be refreshed on demand or according to a predetermined schedule. End-users have found the complexity of the tool to be challenging and to date most of the reports have been developed by the EDW team. There is a goal to have more end-users use the tool themselves which increases the chance of errors in creating the data cube and writing the reports.

For each step in the process model [Figure 3] where data problems could occur, an analysis was done to identify specific issues and their causes. Following from English [3], both developmental and operational processes were considered. For example, data quality problems could be introduced by end-users when they are using the end-user reporting tool. During development of the report an end-user could define a data cube that includes more data than they desire (e.g., includes patients of a different type than they expect). Reports produced from this cube would then be in error. An on-going operations example would be when a report is printed before its data cube was refreshed. This would also result in data that was not what was expected. The list of development and operations data quality issues and causes was quite lengthy and detailed.

Identifying data quality issues and their causes is only the first part of the program. The organization also needed to find ways to address the issues and to make logical decisions on where to put its data quality assurance efforts.

To help identify opportunities for addressing the data quality issues, the organization developed a 2x2 grid of general approaches that can be used for data quality assurance. One dimension of the grid was a separation of Development processes from Operations processes. The other dimension consisted of ways to Avoid Quality Problems and ways to Detect & Repair Quality Problems. Stated differently, one element of the dimension was being “proactive” and one was “reactive.” The grid that resulted is show in Table 2.

	Avoid Quality Problems	Detect & Repair Quality Problems
Development	communication training support for users support for IS documentation	user quality reviews IS testing audits
Operations	communication documentation	monitoring user quality reviews audits

The responses in the grid were then applied to different parts of the model to generate specific programs that could improve quality. For example, to avoid quality problems that result from the incorrect use of the analytical tool client, training was put into place for end-users that included specific content related to quality reports and queries. Metrics have also been defined for different elements of the process model in order to catch developmental and operational problems at an earlier point in time.

In assessing the effectiveness of this approach it would be nice to be able to point to specific quality metrics and show their improvement. In fact, changing organizational factors (e.g., the major implementation of a new registration source system and the efforts required to address Year 2000 problems) have created a variety of new situations and conditions, which were not anticipated in the analysis. What can be stated with confidence is that the use of the model and the analysis have raised the level of appreciation and understanding of data quality. Data quality problems are identified sooner and remedies are more quickly formulated. Data quality metrics have been revised and improved and more are being put into place. Overall, in this one health care organization at least, the firm feels it has a program in place to insure the value of its data warehouse resources.

9. 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 health care 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 identify 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. Research can also help by linking specific health care outcomes to improvements in data quality. In the end, health care organizations and consumers are most concerned about how health care can be improved and costs reduced through the use of information technology. 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*; Englewood; Oct 1997, S5-S8.

11. Kenyon, William W. **Analysis of the collection cycle**, *Journal of Health Care Finance*; Gaithersburg; Fall 1993, 20, 1, 10.
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, 89-101.
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.
19. Reiner, David. **Distilling the data stream**, *Banking Strategies*; Chicago; Jul/Aug 1999, 75, 4, 6-14.
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.
26. Wang, Richard Y. **A product perspective on total data quality management**, *Association for Computing Machinery. Communications of the ACM*; New York; Feb 1998, 41, 2, 58-65.
27. Wang, Richard Y., and Strong, Diane M. **Beyond accuracy: What data quality means to data consumers**, *Journal of Management Information Systems*; Armonk; Spring 1996, 12, 4, 5.

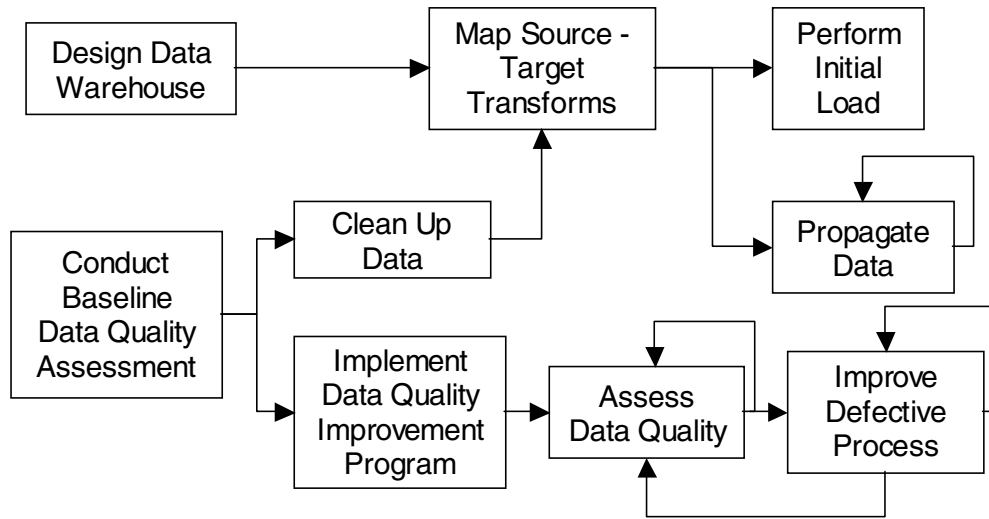


Figure 1: Parallel Data Quality Processes (From English, Larry P. Help for data-quality problems, *Informationweek*; Manhasset; Oct. 7, 1996, 600, 53-62.)

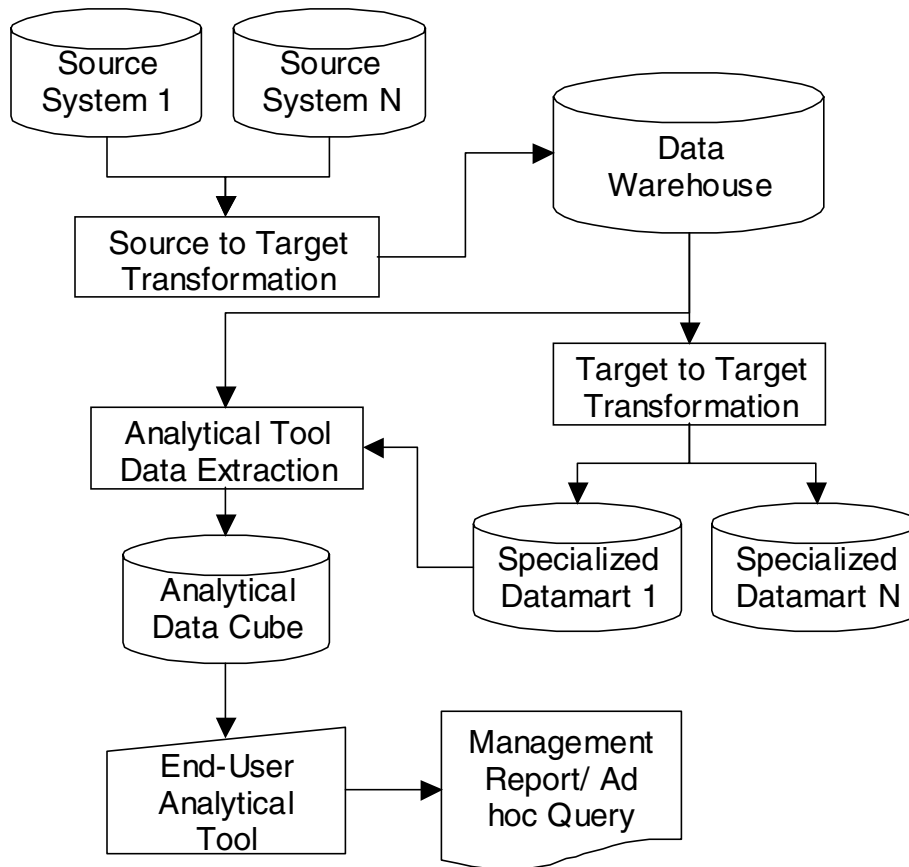


Figure 2: Process Model of Architected Data Environment

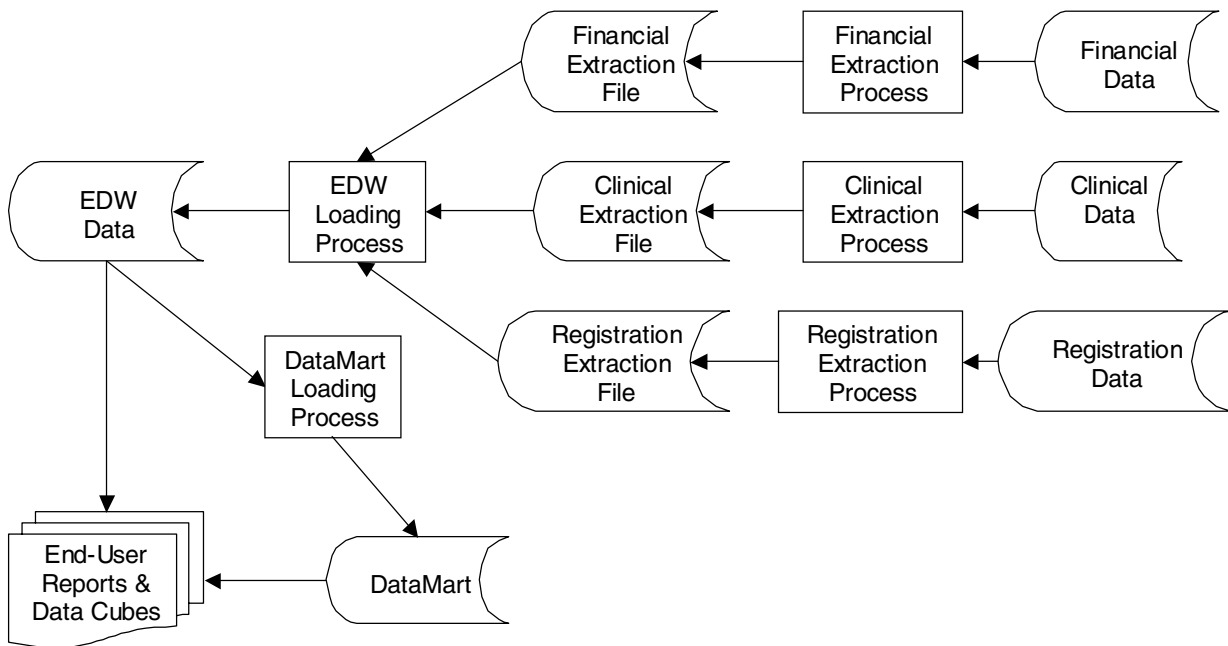


Figure 3: Process Model Applied to Example Health Care Organization