Data Bryte:
A Proposed Data Warehouse Cleansing Framework

Steven D. Mohan, Doctoral Candidate
University of Phoenix
fusion@rmi.net
Mary Jane Willshire, Ph.D.
Software Engineering Management Associates (SEMA)
Mjwillshir@aol.com
Charles Schroeder, Ph.D.
Colorado Technical University
cshroed@cos.colotechu.edu

Abstract

Data cleansing is not considered a "sexy" proposition, yet this scrubbing consumes 60--80% of the effort required to build a functional and effective data warehouse. A new cleansing methodology is described that offers a coherent and focused approach for attaining improved information quality. Initial quantitative research results indicate that cleaner data can be obtained.

Introduction

Modern corporations amass valuable "oceans" of data daily. Within the telecommunications industry, for example, companies often accrue more than one half a Terabyte (1,000,000,000,000) a day in just one of the hundreds of databases that serve both residential and commercial markets. Typically these data come from multiple sources, both internal and external to the corporation [6]. In order to avoid being drowned during analysis in this sea of data, the data are migrated, aggregated and summarized into data warehouses. This achieves two purposes. First, the productivity of the current operational transaction processing systems is not disturbed by analysis queries. Second, the data can be logically consolidated into coherent, searchable entities, usually by product/offering, time, classification, category, territory, trend/value, customer, etc.

Unfortunately, the bulk of the data contain a significant number of errors [4, 5, 8, 9, 10, 11, 12, 13]. The data residing in the operational transaction databases, though perhaps of sufficient quality for transaction processing, contain a number of duplications, inconsistencies, errors, missing data, and other issues that make them unsuitable for a decision support data warehouse, without considerable "cleansing" of the data [3, 6].

Research has demonstrated that 60-80% of the effort of building a usable data warehouse is consumed in data scrubbing [14] . Yet there is very little published research that addresses the actual scrubbing of the data. As pointed out by Allen [1], this scrubbing phase consumes a lot of effort and is not a "sexy" proposition. The fact remains that the data need to be **systematically** scrubbed (validated) in order to build data warehouses and data marts with an acceptable level of data quality that can be searched coherently.

The research has attacked various parts and pieces of the topic of data cleansing in general, yet none of the work has organized the theories and partial ameliorations into a balanced, solution-centric data scrubbing architecture. Though a careful management of the relationship between basic accounts and aggregated accounts has been recommended by Wang [15], no one had undertaken to develop an entire data cleansing engineering program from start to finish. In fact, the warehouse developer has been left to attack the data cleansing issues in an intuitive manner, at best [7]. What we propose to show is that a standards-based, data entity cleansing paradigm integrated with a model-based, combined logical data element quality paradigm (that meets the sufficiency requirement of data cleanliness) can be developed that can be used to define, analyze, and provide guidance to improve data quality.

For the purposes of our research, we define a standard to consist of a single (or set of) criteria and/or a clearly specified format against which an atomic data entity can be compared. As a result of that comparison, the absence or presence of errors (and the number of errors) can be determined. An example of the criteria would be the comparison of a corporation's customer address database atomic entities with the United States Post Office Addressing/formatting standard for postal addresses.

However, comparing data to standards only identifies part of the possible range of errors. Errors are typically discovered by comparing values with standards, or "real-world" values, or by checking for inconsistencies with other data [4]. In order to identify "cleansed" atomic entity data which are inconsistent with real world data or are inconsistent with other data contained in a specified database, a data warehouse or data mart model was developed.

This paper presents a progress report on the first phase of this research effort to design, build and verify a data cleansing framework, called the Data Bryte Framework, for improving data quality in a data warehouse for industrial use. The framework is described below. We invite suggestions from the data quality community on this effort.

Research Method

The research to develop, apply, and prove the viability of Data Bryte, a standards and model-based data cleansing framework consists of six steps, each of which is described in more detail in the following paragraphs. The first step develops and describes the general framework for Data Bryte that is utilized to define, analyze, and provide guidance to improve data cleansing quality. The second step is to define measures of success (i.e. generate measurable thresholds) in developing and applying Data Bryte. The third step is to tailor Data Bryte to an existing and/or new database, then apply the tailored Data Bryte to the subject database. Describing and concisely analyzing the results of applying Data Bryte will comprise the fourth step, while drawing conclusions from the application of Data Bryte would be the fifth step. The latter two steps will determine the viability of Data Bryte. Recommendations will be formulated and presented in step six. Each of these steps and current status of the research is discussed in more detail below.

Step 1 Describe the Standards/Model-Based Data Cleansing Framework (Data Bryte)

In an effort to accomplish the first research step of defining Data Bryte; concepts, paradigms and ideas were obtained from the data cleansing and data quality community. These were generalized to form a sub-framework that could be shaped to individual domain or project requirements. In developing the general architecture of Data Bryte, the data cleansing function was so constructed as to scrub data consistently within the context of specific systems (whether automated or not), their functions, the data at hand, and most importantly, the users of the data. As noted by [2] the cleanliness of the data are fundamentally determined by the users and their analysis methods.

Step 2 Define Measures of Success

The second step in this research is to define measures of success. The criteria used to judge whether or not a standards/model-based combined data cleansing methodology will be successful, are clearly articulated in the execution of this step. These criteria include:

1. The general framework is successfully used as a basis to define data cleansing attributes specific to a particular data environment and users' needs.

2. The general framework is successfully tailored to facilitate the measure and analysis of the cleansing of a specific set or sets of data.

3. The general framework provided guidance in identifying methods to improve the quality of data in a selected case or cases.

Fundamentally, quantitative evidence is being researched and presented to clearly substantiate the findings. A database was constructed and benchmarked that was free of errors. The schema of the database was based upon proprietary business data and operational data structures, though the actual data and structures in the constructed database was non-attributable data. The constructed database was then intentionally corrupted. The data in the constructed database was then reduced and the errors in given fields were characterized by both central tendencies (mean, median, mode) and by dispersion and skewness (range, variance, and standard deviation) for grouped data. The result was corrupted data that matched (within a percent) the corruption found in the proprietary business databases.

The ratio scale was utilized to characterize the errors occurring within the data fields. This permitted the determination of whether or not significant improvement occurred utilizing Data Bryte in the cleansing of the data, as well as the (numerical) quality of any processes affecting the quality of the cleansing. These quantitative measurements were also utilized to support any findings.

Step 3 Verify Data Bryte

The third step of this research is to execute Data Bryte using selected portions of the constructed database as test cases; this step will serve to verify Data Bryte. The constructed database first consisted of a number of "fact" tables that comprised the fundamental entities that a nominal corporate transaction processing system would utilize in its day-to-day business operations. The tables consisted of generated (as opposed to actual industry) data in the format shown in Figure 1. Once the entity tables were constructed, the data contained therein was rolled up or "aggregated" into combined logical entity tables of the type normally searched/mined by corporations in a quest for both improved customer service/loyalty and improved profits. The tables consisted of generated (as opposed to actual industry) data in the format shown in Figure 2.

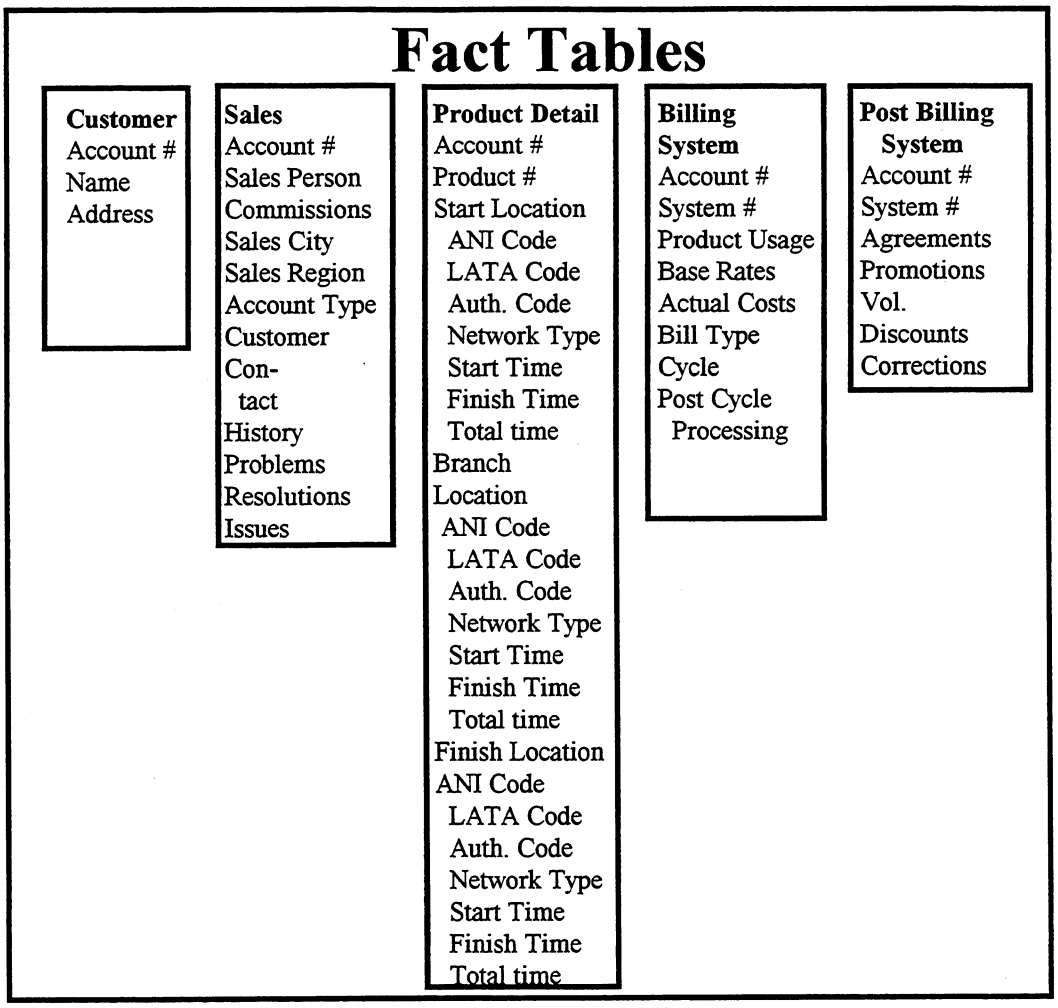

Figure 1 Fact Tables

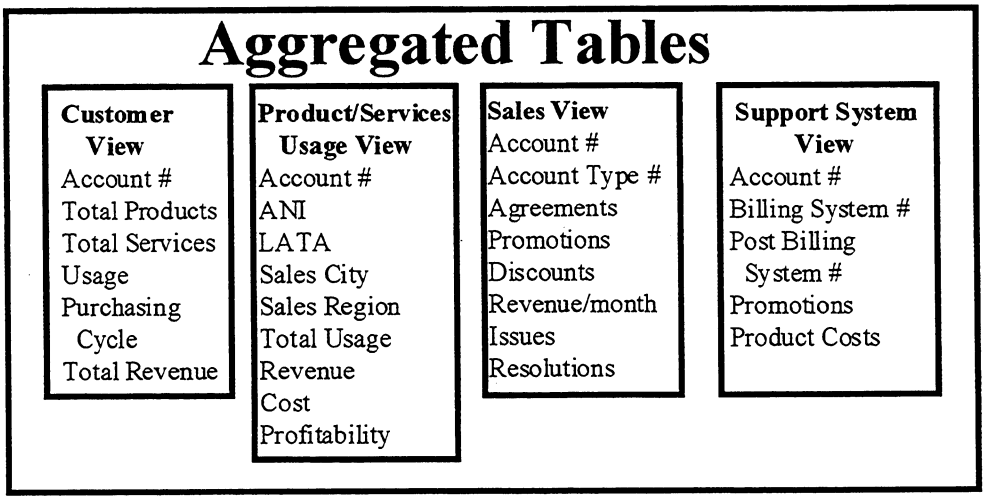

Figure 2 Aggregated Tables

Initially the database was constructed at a 100% cleanliness level. Once the context was modeled, Data Bryte was tailored to the individual functions and data that resided within the

constructed database. The database was then corrupted. The types of errors injected and the relative error rates closely matched (within one percent) the types of errors and rates found in industry proprietary databases. Hence the effectiveness of Data Bryte can be evaluated against problems extant in industry.

After applying Data Bryte cleansing tool against the constructed (corrupted) database, the resulting data fields were reduced for errors. Again, the data in the (resulting) constructed database were then reduced and the errors in given fields characterized.

Step 4 Validate Data Bryte

In step four of the research, the results of applying each of Data Bryte's activities will be presented and analyzed in detail; the impacts and results of the methods selected for each task will be determined and discussed. In sum, the validation of Data Bryte will take place in this step, where it will be demonstrated that Data Bryte is a viable framework for defining, analyzing, and improving data quality through an integrated, standards/model-based cleansing process.

Comparison of data quality before and after will be performed. The generated histograms depicting both central and variance tendencies for selected data fields will be compared and contrasted to identify significant changes. First, the errors in given fields identified by first applying a model-based cleansing process against the enriched entity data sets, will be analyzed against the baseline. Second, the errors identified by applying a standards system against the atomic entities will be analyzed against the baseline. Third, errors identified by applying the combined standards and model-based process against the entire constructed database will be analyzed against the baseline.

Those elements of the data cleansing process that were either poorly or strongly effective will be statistically evaluated and analyzed. Again, the data in the (resulting) constructed database will then be reduced and the errors in given fields will be characterized. As applicable, any significant external data influences will be identified.

Finally, in order for the process to be validated the null hypothesis, that is, that the changes in the identified errors were due to truly random processes rather than the effects of the cleansing

process, must be clearly rejected. A significance level of 5% will be utilized. Any attempts at remedies will be documented, along with the results of their implementations.

Step 5 Draw Conclusions

Any resulting conclusions will be drawn in step five of this method. Initially, a determination will be made as to whether the success criteria for developing Data Bryte were met. The analysis of the quantitative and qualitative measurements will provide the basis for all conclusions drawn.

Step 6 Make Recommendations

The final step of this method will be to offer recommendations indicated as a result of executing the previous five steps. Recommendations will be made concerning the general usage of Data Bryte as well as reasonable extensions that could be made to this approach.

Status

Steps one through three of the research method (description, definition of success measures, and verification) have been completed. The validation (step four) of the research method (by the comparison of data quality before and after the use of Data Bryte) is on-going at this time.

In the actual revenue data itself there is a less than a 1% error rate. This is because a) the data undergoes several cleansing (and auditing) operations before an invoice is ever generated and b) in those instances where a bill is incorrect, the customer (quickly) objects, and the cause of the error is determined and fixed before the next month's bill. Ergo, the feedback system is very strong.

In the customer data there is a 5+% error rate. This is because a) the data are entered manually via call center personnel talking to a customer over a phone line and b) a policy that nothing (not even a screen edit) will slow down the capture (and subsequent processing) of a customer order. Errors resulting in a bad invoice or sending the bill to the wrong place get caught fairly soon. The rest may never get caught.

In the aggregation of atomic revenue and product/service data into logical entities that could be analyzed, initially it was functionaly impossible to determine error rates. This is because a) there is no underlying semantic definition of what constitutes a product or service and b) there are

no rules (or process model) on how such objects are to be aggregated. As a result, "minor" code or data representation changes (in seven billion plus call records per month) can disproportionately skew the trend curves presented each month.

The semantic model and aggregation rules are currently under construction. As past data are mapped to the model, it is likely that substantial reporting errors will be illuminated. Analysis of root cause and effect factors is also on-going.

Conclusion

Telecommunications call records and product/service information provides a rich source of data with which to evaluate a coherent data standards/data model cleansing approach. Data errors range from a low of one percent for the financial data, five percent plus for the customer data, and an (initially) undefined large number of errors for the aggregation data. The Data Bryte coherent data cleansing framework is currently undergoing quantitative analysis to determine viability in terms of thresholds and generalizability to other information domains. Initial data analysis strongly suggests that Data Bryte is a viable data cleansing methodology. Upon final validation, additional research will be undertaken to further optimize the methodology.

References

[1] Allen, S. "Name & Address Data Quality," presented at Conference On Information quality, Cambridge, Massachusetts, 1996.

[2] Ballou, D. P., and Tayi, G. K., "Managerial Issues in Data Quality," presented at Conference on Information Quality, Cambridge, Massachusetts, 1996.

[3] Brown, S. M., "Preparing Data for the Data Warehouse," presented at Conference on Information Quality, Cambridge, Massachusetts, 1997.

[4] Caby, E. C., Pautke, R. W., and Redman, T. C., "Strategies for Improving Data Quality," *Data Quality*, vol. 1, pp. 4-12, 1995.

[5] Devlin, B., *Data Warehouse: From Architecture to Implementation*. Reading, Massachusetts: Addison-Wesley, 1997.

[6] Gardyn, E., "A Data Quality Handbook for a Data Warehouse," presented at Conference On Information Quality, Cambridge, Massachusetts, 1997.

[7] Jarke, M., and Vassiliou, Y., "Data Warehouse Quality: A review of the DWQ Project," presented at Conference on Information Quaility, Cambridge, Massachusetts, 1997.

[8] Klein, B. D., and Rossin, D. F., "A Preliminary Analysis of Data Quality in Neural Networks," presented at Conference on Information Quality, Cambridge Massachusetts, 1997.

[9] Lawrence, J. P., and Pearl, D. K., "Data Quality Issues in the Ranking of U.S. Colleges," *Data Quality*, vol. 2, 1996.

[10] Mandke, V. V., and Nayar, M., "Information Integrity -- A Structure for its Definition," presented at Conference on Information Quality, Cambridge, Massachusetts, 1997.

[11] Mathieu, R. G., and Khalil, O., "Teaching Data Quality in the Undergraduate Database Course," presented at Conference on Information Quality, Cambridge, Massachusetts, 1997.

[12] Orr, K., "Data Quality and Systems Theory," presented at 1996 Conference on Information Quality, Cambridge, Massachusetts, 1996.

[13] Redman, T. C. "Data Quality for the Information Age," Norwood, Massachusetts: Artech House 1996

[14] Rosenthal, A., and Dell, P. "Propagating Integrity Information in Multi-Tiered Database Systems," presented at Conference on Information Quality, Cambridge, Massachusetts, 1997.

[15] Wang, R. Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, vol. 41 #2, pp. 58-65, February 1998