2005

Data Quality and the Data Warehouse: A Decision Support System for Allocation of Scarce Resources

M. Pamela Neely
Rochester Institute of Technology, pneely@cob.rit.edu

Data Quality and the Data Warehouse: A Decision Support System for Allocation of Scarce Resources

M. Pamela Neely
Rochester Institute of Technology
pneely@cob.rit.edu

ABSTRACT

This paper describes a decision support system (DSS) for use in allocating scarce resources associated with data quality efforts in the construction of a data warehouse. The DSS is populated with metadata from a data warehouse project, including tags that identify the quality at intersections of data field, data use, and data dimension. The resulting DSS can then be queried by novices on a data warehouse project based on criteria and priorities set by management. The results of an experiment using business students are then presented. It can be shown that, given the proper set of skills, business students, as proxies for novices on a data warehouse development team, can effectively use the tool to analyze and prioritize hundreds of potential fields in a data warehouse project.

Keywords

Data warehouse, data quality, decision support, novice vs. expert, scarce resources

INTRODUCTION

The development of a data warehouse is a costly, time-consuming process (Golfarelli and Rizzi 1999). It is recognized that the quality of data is a critical success factor for a data warehouse (Wixom and Watson 2001), but it must also be recognized that the quality of the data is only relevant within the context of the use of the data (Ballou and Tayi 1989; Redman 1998; Strong, Lee 1997), thus not all available data fields will be migrated to the warehouse. In addition, even within the constraints of data that is fit for the intended use, there will never be enough resources available to bring every available field to a quality level of 100%. Given that numerous sources may be available to populate a data warehouse, it is important that the warehouse developer address quality efforts only to those fields that will best support the goals of the warehouse. A specific project may have hundreds or even thousands of fields available to populate the warehouse, thus it is necessary to have a structure in place for analyzing all of the available fields and then deciding which fields should ultimately be migrated to the warehouse. This structure will act as a decision support system for the warehouse developers.

This paper briefly describes the concept of fitness for use in relation to data quality. An experimental study was conducted utilizing the Data Quality Knowledge Management (DQKM) tool (described in (Neely 2001)), a relational database decision support system intended to be used by novice members of a data warehouse development team. The remaining sections of the paper discuss the research question and the results of the experimental study.

DATA QUALITY AND FITNESS FOR USE

The term "fitness for use", as coined by Juran (1988) originally attempted to describe the many variables that needed to be considered when evaluating the quality of a product. This term has been used frequently in the data quality literature (Ballou and Tayi 1989; Redman 1998; Strong, Lee 1997). Given a data field, the quality of the data in that field will be based on the intersection of the quality dimension being considered (e.g. accuracy, completeness, relevance, etc (Wang and Strong 1996) and how the data will be used (e.g. to answer questions about client demographics, to forecast sales for the next ten years). See Figure 1, which illustrates the relationship among field, dimension, use and measure. This measure, or tag, becomes a part of the metadata (data about the data) and can be used to justify further quality enhancement decisions.

As indicated earlier, the DQKM is a relational database, designed to hold metadata for a data warehouse project. It has several tables that must be populated in order to be useful. Metadata concerning the actual fields available for merging into the data warehouse must be entered. One of the fields in this table is data category. A data category is a classification of a particular data field. For example, you might have two fields from two different data sources. One is called ETHNIC_GRP and the other is called RACE. They both represent data fields that capture the same information- data about ethnicity. By assigning a data category to the fields, it is possible to quickly locate fields from multiple sources that represent data that is

similar in nature. Data categories are specific to a particular project. The individual who is populating the DQKM with metadata can code each available data field with one category from the list. The DQKM used in the experimental study was intended to store metadata associated with a data warehouse designed to support decision making for a state agency responsible for programs provided to the homeless. Thus some of the categories are Assessed Service, Discharge Date, Ethnicity, Facility Demographics, and Pregnant. This allows for very efficient querying, removing the problem of homonyms and synonyms.

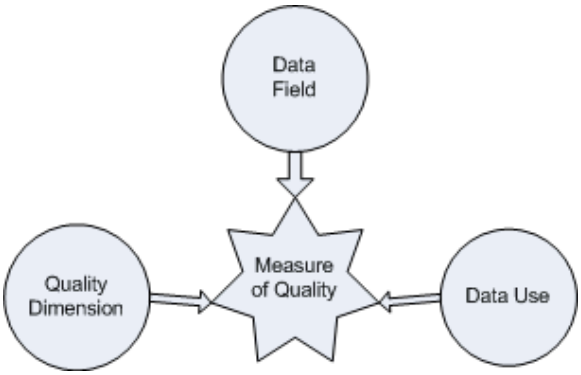

Figure 1- Fitness for Use

Another table lists the uses for the data. Given the assumption that a data warehouse is constructed to address specific needs, those needs can be identified. The DQKM used in the study has uses of client demographics, services assessed to the client (potential services), services actually offered to the client, shelter characteristics, length of stay and identification of a specific shelter. To expand on this, if a field has been identified as having a use of client demographics, it is because the data in the field would help the user of the data warehouse group individuals into different demographic groups. A particular field can address multiple uses. A field that stores gender could be used to group individuals demographically, but it could also be used to determine what shelter characteristics would be needed (i.e. a male only shelter would accept clients with a gender of male.) The uses are defined based on how the data will be used in the data warehouse. Thus, they will also vary from project to project.

The dimension of quality to be considered must also be available in the DQKM. Fifteen dimensions, corresponding to the Wang and Strong work (1996), were included in a separate table. These included such dimensions as accuracy, completeness, relevance and accessibility, among others. These dimensions will not change from project to project.

As stated earlier, the DQKM is a relational database. The table structure has been created by the researcher, but it must be populated with metadata in order for it to be a useful resource. Once the DQKM has been populated with metadata (including measures associated with combinations of data field, data use, and quality dimension), the DQKM becomes a decision support system which can then be used to make resource allocation decisions for quality enhancement.

Thus, it is a two step process. A subject matter expert (SME) will be responsible for determining what the measure is, based on the other three factors. A table in the DQKM stores records for each combination of data field, data dimension and use that has been assigned a measure. These measures are based on the judgment of the SME and are recorded for all instances where the SME has knowledge. No judgment is made at that time to determine whether or not the fields should be evaluated for further enhancement. Once the measures are agreed upon they can be loaded into the DQKM. They are then queried by novices, using preset criteria, to determine which fields should be evaluated for further quality efforts.

To illustrate the complexity of the initial metadata population, consider that the test DQKM had 362 data fields available from 16 source databases. All of these fields are loaded into the DQKM because the tool can also serve as a knowledge management tool to preserve the corporate memory, although not all fields will be migrated to the warehouse. Of these 362 fields, 129 were coded into a specific category, indicating that they would address the questions for which the warehouse was being constructed. Four dimensions were considered important for this project and six uses were identified. Not every available categorized field will be used for all 6 uses. In fact, many of the fields will only address one or two uses. Each combination of field, use and dimension that is identified becomes a record in the DQKM table. In the case of the

experimental project, the DQKM was populated with 690 records showing the measure, or tag, at a particular junction of field, use and dimension.

However, once this process is completed, a tool has been created that captures this knowledge of the SME. At this point, judgments can be made as to where to allocate resources for further enhancement. As noted earlier, the SME simply assigns a measurement of quality based on their knowledge of quality dimensions and how the data will be used in the warehouse. Assuming that measurements are assigned in terms of percentages, the SME could just as easily assign a measure of 40% as 90%. At this point in the process, the data must be sifted to determine where to focus the quality enhancement efforts.

Research Question

Given the complexity described above, it is obvious that the number of fields available for a data warehouse project could quickly become impossible to analyze without the support of some type of tool. Although the assignment of measure is subjective, based on the evaluation of the SME, it is necessary to condense these measures into an objective, finite set of fields that should be considered for improved quality. There is a high cost associated with determining the measures. It would be beneficial if we could then assign less costly individuals to the task of ferreting out the precise fields to improve in quality. Thus, the following research question is asked:

1. With the use of a decision support system, can subject matter novices make resource allocation decisions for data quality efforts in a data warehouse project?

Research Methodology

An experimental procedure was developed and pilot tested with a goal of having novices on a data warehouse development team extract data fields from the DQKM that met minimum criteria for further evaluation. On the assumption that data of a given (poor) quality is not cost effective to enhance, subjects were provided with a set of minimum criteria on which a given data field, for a given use and dimension, were to be evaluated. (See Table 1 for the criteria used in the study.) Each dimension and use was paired with a data category such as Facility Demographics or Pregnant. They were asked to evaluate a total of five data categories. The output from each query of the database would be only those fields in a given category, on a given dimension and for a given use that were above the minimum criteria percentage. In other words, the measure, or tag, must be greater than the minimum criteria percentage. If the specific data field met the minimum criteria, it would be evaluated further for quality enhancement.

As the queries are processed, new records are appended to a table in the database. The procedure asked the participants to export this table to an Excel worksheet and analyze the data based on the following priorities (in order of priority):

1. The more uses that a data field addresses the higher priority it should be given for enhancement
2. Data used for length of stay (one of the uses) has a higher priority than the other uses
3. Data that is closer to a 100% measure will require fewer resources and thus, should be given higher priority

Dimension	Use	Criteria
Complete	Client Demographics, Services Assessed, Services Offered	50%
Complete	Shelter Characteristics, Length of Stay, Specific Shelter ID	60%
Accurate	Client Demographics	75%
Accurate	Services Assessed, Services Offered	50%
Accurate	Shelter Characteristics, Length of Stay, Specific Shelter ID	60%
Reputable	Client Demographics	40%
Reputable	Services Assessed, Services Offered, Shelter Characteristics	55%
Reputable	Specific Shelter ID	80%
Accessible	All Uses	50%

Table 1- Criteria for Further Consideration of Quality Enhancement

The participants were instructed that they had a budget of $100,000 for quality enhancement. Each data field had a cost to produce associated with it that was exported to the Excel worksheet. The cost to enhance a particular field on a particular dimension could be extrapolated from the cost to produce. A particular data field would have been repeated in the export if the data field addressed multiple dimensions or uses. Working within the given priorities, particuipants were to decide which

specific fields should be considered for quality enhancement. Other than the priorities, directions for deciding which specific fields to enhance were left intentionally vague. In particular, they were not told what formula to use to arrive at cost to enhance, and they were not told how to organize the worksheet to sort the data by priority. It is expected that novices would have sufficient analytical skills to sort the data and prioritize the fields to most effectively use the $100,000 budget within the given constraints.

The participants were students in a junior level database management systems class. 27 students completed the exercise. All but 3 of the students were MIS majors. All were business students. Over half of the students had been on co-op, meaning that they had a paid six-month assignment working in the field of MIS. All of the students had taken a course on Excel and Access within the previous two years.

They were provided with the DQKM, a copy of the procedure, and copies of the lists of data categories, data uses, and dimensions. The procedure was conducted during a class period and thus they had 1 hour and 50 minutes to complete it. Although this seems like an unnecessary time constraint all of the students completed the exercise to the best of their ability within that time period. Additional time was not requested by any of the participants. Step 1 of the procedure required the participants to familiarize themselves with the metadata available to populate the data warehouse. Step 2 involved querying the database multiple times based on the criteria given above in Table 1. The results of this step should have been 135 data fields that met the minimum criteria for further quality efforts. Step 3 was the exportation of the fields meeting minimum criteria to an Excel spreadsheet where the students were asked to prioritize and identify fields that should be enhanced for quality. The priorities were indicated earlier in this section.

The results of this step should have been 84 fields targeted for quality enhancement. The final step of the project was to report these fields, along with the appropriate data source, back to management for their evaluation. At that point, management can make a decision as to whether or not the results fit with their expectations. Participants were allowed to request clarification on procedure points and discuss methods of analysis with each other, much as they would in a project environment on a data warehouse development team.

Results

As stated earlier, 27 students completed the exercise. All of the students were able to extract records meeting minimum requirements and export the data to Excel.

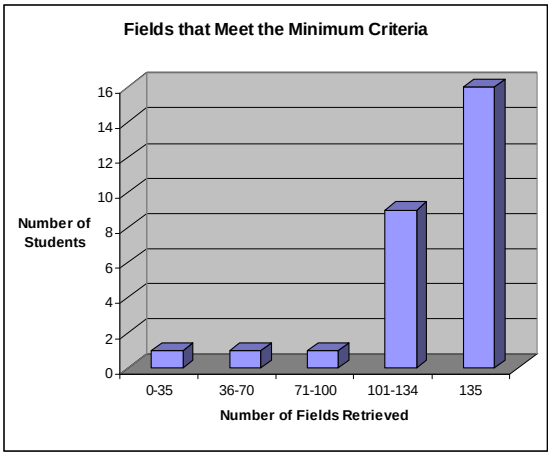

Figure 2- Records that Meet Minimum Criteria

As can be seen in Figure 2, 16 students extracted exactly the right records from the DQKM. Of the 11 students who retrieved less than 135 records, the average number of missing records was 13. In addition, 5 students retrieved records that did not meet the minimum criteria for further evaluation. Extraction of the records from the DQKM was primarily a mechanical process, consisting of running a query multiple times with the correct inputs of data category, use, dimension, and criteria percentage. However, as there was no "undo" button on the DQKM, it was important for participants to pay close attention to detail.

In analyzing the results of the prioritization it is difficult to come up with summaries. Obviously, the records that are extracted will impact the results of the prioritization. Additionally, students made three different assumptions regarding the third priority, "Data that is closer to 100% in a given dimension will require fewer resources and thus, should be given higher priority." The next section will discuss these three assumptions.

Assumptions

Participant assumptions fell into three groups for calculating cost to enhance. Each assumption will result in a different total cost to enhance (all records) and the number of records that can be enhanced within the given budget.

Some participants assumed that the measure should not be a part of the formula when calculating the cost to enhance. Thus, they believed that the cost to enhance was the same whether the measure was 65%, 90% or 100%. If a given field, dimension and use record had a cost to produce of $2,400, then that would be the cost to enhance. With this assumption, 27 records can be enhanced with $2,200 left over. Total cost to enhance all 135 records would be $8,474,700 with this assumption.Table 2 shows the data for participants using this assumption. The number of records to be enhanced, as identified by the student, ranged from 22 – 29. They identified 20 – 27 valid records as identified by the solution to the exercise. Total cost to enhance was generally very close to the exercise solution. Variation was dependent on records extracted in the previous step.

Some participants argued that if a record were already at 65% quality it would only cost 35% of the cost to produce to enhance the quality to 100%. Although this is not totally true, since it will probably cost more to enhance the last fraction of the data, it is probably a good approximation, particularly given that no data field would need to be enhanced to 100%. This group used a formula that multiplied the cost to produce times 1 minus the measure. With this assumption 84 records could be enhanced within the $100,000 budget. Total cost to enhance would be $618,860. Table 3 shows the data for this assumption. Although the total cost to enhance is consistent within this group, the records identified for further enhancement vary much more widely than the previous two groups. The number of records identified to enhance was between 30 and 94. Valid records were 19 – 81.

Other participants factored in that records with a measure of 100% would need no further enhancement. The cost to enhance for those records would then be zero. All other records would have a cost to enhance that equaled the cost to produce. With this assumption 27 records could be enhanced with $1,300 left over. The reason that there is less left over with this assumption than with the first assumption is that 2 records in the first assumption costing $4,800 would not need to be enhanced and additional records costing $5,700 could be enhanced, Total cost to enhance with this assumption would be $2,924,800. Table 4 shows the data for this assumption. Participants identified between 26 and 30 records, of which between 24 and 27 were valid records. Total cost to produce varied only when the records extracted was incorrect.

Table 5 has the data for 10 participants that could not be grouped into the previous assumptions.

DISCUSSION

The results of the experiment indicate that with the proper analytical skills novices can make an important contribution to the overall process of evaluating records for quality enhancement and allocating scarce resources to ensure that the resources will be aimed at records which will most contribute to the success of the data warehouse.

Over half of the participants were able to properly extract records from the DQKM that met minimum requirements for further evaluation. As stated earlier, analysis of the prioritization was affected by the actual records extracted. However, it should be noted that within a particular assumption, results were generally close to the expected results. This is particularly true of the participants who did not factor the measure into the calculation of cost to enhance, or only considered the measure when it was at 100%.

Those participants who factored in the measure had consistent Total Cost, but varied widely on the number of records. The participant on the low end of the scale identified a field costing $62,500 to enhance, thus using a significant portion of the budget. A primary problem for this group appeared to be the way they sorted the data. The first two groups didn't get beyond the second priority- all records in the solution had multiple uses. In this assumption they needed to include some (but not all) of the records that addressed only one use. They did not prioritize within this group on the 2nd and 3rd priority properly, thus causing variation in the results

Table 2- Assumption: Measure not considered

# Records	$ Spent	Valid Records	Total Cost
28	$100,000	25	$8,744,100
29	$ 97,800	24	$8,474,700
27	$ 97,800	25	$8,474,700
29	$ 97,800	27	$8,474,700
22	$ 79,200	20	$6,347,800

Table 3- Assumption: Cost to Produce * (1- measure)

# Records	$ Spent	Valid Records	Total Cost
94	$ 79,320	19	$ 618,860
89	$ 99,558	81	$ 618,860
53	$ 99,997	50	$ 618,860
84	$ 99,256	84	$ 618,860
49	$ 99,925	42	$ 613,433
30	$ 99,978	28	$ 617,906

Table 4- Assumption:
No cost for 100%, no other measure considered

# Records	$ Spent	Valid Records	Total Cost
26	$ 97,900	25	$2,924,800
26	$ 97,900	26	$2,924,800
26	$ 97,900	26	$2,924,800
30	$ 99,000	27	$2,894,500
29	$ 97,800	26	$2,924,800
27	$ 97,800	24	$2,924,800

Table 5- Miscellaneous

# Records	$ Spent	Valid Records	Total Cost
0	$ -	0	$ -
0	$ -	0	$ -
27	$ -	24	$ -
30	$ 93,901	4	$ -
122	$ 78,106	0	$ 78,106
6	$ -	0	$ -
0	$ -	0	$ 147,500
0	$ -	0	$ 607,863
0	$ -	0	$ 618,860
0	$ -	0	$ 605,897

As noted in Table 5, ten of the students had results that would not be acceptable. Codifying the assumptions for the other groups would have eliminated many of the discrepancies seen in the results. Clearly, this group had several problems. Some did not calculate a cost to enhance at all. Others calculated a cost but never prioritized the records. A final subsection of this group had such poor results in the extraction process that the analysis was virtually meaningless. Inability to interpret the instructions and inattention to detail contributed to the unacceptable results.

The researcher had previously conducted the experiment with information systems seniors in a school of computer science. Although the research conditions were similar in both experiments, conclusions from that experiment indicated that students could not serve as proxies for novices in a business team (see (Neely 2002) for discussion). However, the current group of students has several skills that were lacking in the previous group. Based on these additional skills, the results of the experiment indicate that students can be used as proxies for novices, and that novices can effectively perform the required analysis.

A major advantage of the current participant group is that they are college of business students. They have all had accounting, finance and data analysis classes, in addition to the class in Excel and Access. They are accustomed to focusing on the details and were familiar with the techniques of sorting and formula building in Excel. The previous group was unable to do any of the analytical work in Excel.

A second characteristic that would seem to work in favor of this group of students is work experience in the field. As indicated earlier, about half of the group had already been on co-op. Although they were earlier in their educational career than the previous group, they were beyond the previous group in terms of actual work experience.

Finally, the stakes for this group of students was higher than for the previous group. The exercise had been conducted in a colleague's class in data quality during the previous experiment. The students did not have an attachment to the researcher or a stake in the outcome, as the other professor was not grading it. This time the students were in the database class taught by the researcher and participation points were awarded for completion of the exercise. This parallels a business environment where the employee would want to do a good job to keep their job and possibly earn recognition in the form of a promotion or raise.

LIMITATIONS

Although a case can be made that students are adequate proxies for novices on a data warehouse team, it is a limitation of this study that the experiment was not conducted with actual novices in a business environment. Ideally, this tool would become a part of the methodology of a development process and the procedure would be prepared by the manager in charge of quality enhancement.

CONCLUSION

As a decision support system the DQKM can play an important role in the development of a data warehouse. Individuals knowledgeable of the source and the warehouse must populate the DQKM with metadata. However, once the DQKM has the records that contain measures or tags of quality based on the field, use and dimension, the DQKM can then be assigned to a novice for the final analysis. The novice must have certain skills such as attention to detail, analytical skills, and spreadsheet skills. The analysis consists of extraction of records meeting certain minimum criteria and then prioritization given priorities assigned by management. Once management receives the report the analysis can be redone with different criteria and priorities if the results are not in line with what was expected, or other questions arise.

REFERENCES

1. Ballou, D. P. and Tayi, G. K. (1989) Methodology for Allocating Resources for Data Quality Enhancement, *Communications of the ACM*, 32, 3, 320-329.
2. Golfarelli, M. and Rizzi, S. (1999) Designing the Data Warehouse: Key Steps and Crucial Issues, *Journal of Computer Science and Information Management*, 2, 3.
3. Juran, J. M., *Juran's Quality Control Handbook*. 4 ed, ed. F.M. Gryna. 1988, New York: Mc-Graw-Hill, Inc.
4. Neely, M. P. *A Proposed Framework for the Analysis of Source Data in a Data Warehouse*. in *Proceedings of The Conference on Information Quality*. 2001. MIT, Cambridge, MA.
5. Neely, M. P. *Data Quality Knowledge Management: A Tool for the Collection and Organization Of Metadata In A Data Warehouse*. in *Proceedings of the Ninth Americas Conference on Information Systems*. 2002. Dallas, TX.
6. Redman, T. C. (1998) The Impact of Poor Data Quality on the Typical Enterprise, *Communications of the ACM*, 41, 2, 79-82.
7. Strong, D. M., Lee, Y. W., and Wang, R. L. (1997) Data Quality in Context, *Communications of the ACM*, 40, 5, 103-110.
8. Wang, R. Y. and Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems (JMIS)*, 12, 4, 5-34.
9. Wixom, B. H. and Watson, H. J. (2001) An Empirical Investigation of the Factors Affecting Data Warehousing Success, *MIS Quarterly*, 25, 1, 17-38.