INCREMENTAL DATA QUALITY IN THE DATA WAREHOUSE

Karsten Boye Rasmussen
Institute of Organization and Management, SDU
University of Southern Denmark, Campusvej 55, DK-5230 Odense M.

Keywords: Data Warehouse, Data Quality, Business Knowledge, Metadata, Knowledge Management

Abstract: The data warehouse is the cornerstone for the production of business knowledge in the organization. The foundation of the quality of the business knowledge is the quality of the data in the data warehouse. Determination of dimensions of data quality in the data warehouse has been obtained through the intuitive, the empirical and the ontological approaches. The first point of this working paper is that data quality is not a static measure and that awareness of the data quality dimensions is a prerequisite to improve the data quality. The second point is that selection is the cornerstone of data quality in the data warehouse in relation to the quality dimensions. Thirdly, that post-load improvement of the data quality is obtainable. Metadata can be added incrementally containing information on the use of data – including the users' selections within the data warehouse.

1 INTRODUCTION

Improvements of data quality in the data warehouse are well described in books and articles on the construction of the data warehouse and the processes of Extract, Transform, and Load. This paper will look into the potentials of improvement of the data quality after the load of data into the data warehouse. The proposition is that the actual utilization of the data warehouse delivers the vehicle for quality improvement.

"The data warehouse provides access to consistent organizational data that can be combined for query, analysis, and presentation of published data with a quality that will act as a driver of business reengineering" (Kimball 1996). This demonstrates the opening of a wider use of the data warehouse; but decision support prevails (Inmon 1996). The underlying assumption is that quality action builds upon quality decisions that stems from quality data

This poster paper starts with a compact overview of approaches to data quality in order to show the dimensions of data quality. Secondly, the dimensions will be seen from the viewpoint of the user (the action and decision support perspective) which will bring forth common characteristics. Thirdly, this will open an opportunity for improvement of data quality through the use of data.

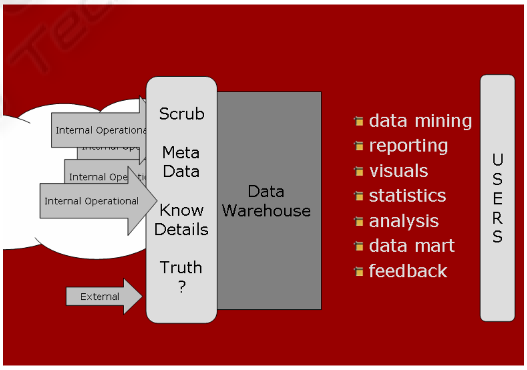

Figure 1: Data warehouse

2 WAREHOUSE QUALITY

The paper is following a conceptual division of approaches to data quality into: 1) the intuitive approach2) the empirical approach, and 3) the theoretical approach (Wang & Strong 1996c).

Boye Rasmussen K. (2004).
INCREMENTAL DATA QUALITY IN THE DATA WAREHOUSE.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 634-637
DOI: 10.5220/0002600806340637

2.3 Intuitive data quality

The intuitive definition of data quality is "fitness for use" (Bruckner & Schiefer 2000;Wang & Strong 1996a) for the "data consumer" (Strong *et al.* 1997). This demonstrates relativity and subjectivity. As what can be interpreted as a reaction to the relativity the intuitive approach to data quality are often primarily focused on metrics and figures: firstly, metrics to describe the extension of the data quality problem; secondly, metrics of a guess or estimate of the (financial) effect of poor data quality; and lastly the proportion of errors in the data that are causing these problems.

The relativity of data quality is important as the rationale for the establishment of the data warehouse exactly is to bring the same data into many different contexts (applications) utilized by many different users (Tayi & Ballou 1998).

The weakness of the intuitive approach is that there is no stated and clear definition of the concept "data quality"; however some quality dimensions are identified: accuracy, currentness, completeness, and consistency (Fox *et al.* 1994).

2.4 Empirical data quality

The user perspective is underlying the intuitive approach but is made explicit when Wang and Strong (1996b) are pursuing a methodological well-based exploratory empirical study of data quality from a user perspective by applying marketing methodology and viewing data as a product and the user as a consumer. The obtained many quality descriptors were processed by use of factor analysis and grouped into four target categories: Intrinsic, Contextual, Representational, and Accessability.

The concept of dimensions implies unsubstitutability. This is demonstrated by the conspicuous ineptness of assertions like "The data are absolutely fitting for the task, but they are not accessible", or "The data arrived in time, but they are impossible to understand". All dimensions have to be present – and can be so in varying degrees - or the data will be "unfit for use".

2.5 Ontological data quality

The structure and categories within the area of data quality are not guaranteed to arise from the intuitive or the empirical approach. A theoretical approach from a systems-design viewpoint is done by Wand and Wang (1996) who build their argumentation on the view that the information system (IS) delivers a representation of the real world system (RW). From the information system the user makes an inferred interpretation of the real world, but is also capable of making a direct observation of the real world. The two views of the real world can lead to deficiencies of data and "inconformity" between the two views. The mapping between the information system and the real world system leads to three categories of defectiveness: Incomplete, Ambiguous, and Meaningless. In its simple forms the extremes implies that the RW has states not found in the IS (incomplete) or the IS has states not existing in the RW (meaningless). Ambiguity arises when a state in the IS is covering more than one state in the RW. Ambiguity precludes the inverse mapping from the information system to the real world.

3 QUALITY DECISIONS

With the determination of both the empirical and the theoretical developed dimensions it is fruitful to return to the original starting point that data quality should improve our acting. "A good decision is an action we take that is logically consistent with the alternatives we perceive, the information we have, and the preferences we feel" (Howard 1988).

The dimensions of data quality are in the ontological approach deducted to data being incomplete, ambiguous, and meaningless while the empirical findings isolated the groups of intrinsic, contextual, representational, and accessible.

The data warehouse is a collection of data for use in many applications and by many users. The fact that most of these applications and users are unknown when the system is designed – as well as when data are extracted-transformed-loaded into the data warehouse - accentuates that the development of the data warehouse must assure extreme flexibility to accommodate changes. The quality of data is embedded not in the data itself, and not in the system, but in the users use of data: "what may be considered good data in one case (for a specific application or user) may not be sufficient in another case" (Wand & Wang 1996).

3.1 Incrementing quality by use

On the other hand the proposition in this paper is that data quality is balanced. It is neither objective nor solely a subjective undertaking. Enhancements are obtained by the users in their use of the data, but the value of the enhancement lies in the distribution of this knowledge in the organization. This is not a change in the data, but a change in the perception of data, and this is made explicit when stored as metadata.

A more detailed journey would look at all the data quality dimensions and at data at three levels: On one level the object is a data file, typically visualized as a relational table with rows and attributes. On the next level the object is the attribute as found in the theoretical potential of the attribute, the domain. The third level is the specific datum of a specific row of a specific attribute buried in the data file.

Examples can be given on how these levels affect the quality dimensions. For the sake of brevity only a few dimensions will be touched in this presentation.

3.2 Incompleteness

Incompleteness of data implies that we know about the real world states and that we are in vain searching for corresponding states in the system. A user with a singular system view will not have the capacity to judge the data as incomplete. The more knowledgeable user must pass on his knowledge by a description of incompleteness. Incompleteness on the file level is demonstrated when the number of rows does not mach the number obtained from the real world. However, we have to take into account that a distribution can be purposefully wrong (as in a stratified sample). Incompleteness is related to the file level description of data and demands metadata on the selection procedures and description of procedures (e.g. weighting for stratification).

Incompleteness on the domain is demonstrated when the distribution of an attribute does not match the distribution of the real world. The non-match to the real world is also remarkably demonstrated when certain known states of the attribute are not found in the system at all. This can point in two directions. Firstly, the data file can be incomplete because rows are missing or secondly, if rows are complete the reason is that a state in the system does not distinguish between two or more states in the real world. (This is also Ambiguity at the domain level).

Ambiguity is also sometimes found to be created on purpose. The design involves decisions on which distinctions (attributes) are relevant. If the color of a product is disregarded in the design of the data warehouse, we have chosen not to use this as a variable in the system view. This implies to users that color is of no importance. A reconstruction will demand great persuasive efforts from the users toward the designers of the data warehouse. Parallel to the selection of rows the quality of the data warehouse also depends upon a selection of attributes.

3.3 Meaning

When states in the system cannot be related to any states in the real world the system is without meaning. Meaninglessness cannot be envisioned to be created on purpose. Meaning is contextual and the lack of meaning is typically the lack of context i.e. the lack of metadata describing the attribute. At the file level we move from the row to the column dimension. The attribute can be without meaning (like an unintelligible description: "Anno Nutrical Excerpt Range 5 Years Hourly Measured"). Without further information we must discard the use of this variable.

At the domain level meaning can easily vanish from the attribute. If an attribute is supposed to store the age of the customer and one customer is coded "Bright Blue" this implies that the data is not verified against the legitimate values for the attribute. The datum is meaningless, but the attribute itself is loosing credibility. Metadata describing the procedures to validate the data will imply that meaninglessness is less inclined to occur.

3.4 Contextual

The empirical contextual dimensions include value-added, relevancy, timeliness, and appropriate amount of data - apart from completeness that has already been discussed above. The correct context for these dimensions is the actual use and the judgment by the user. Although the judgments must inherently be subjective they can be lifted to generality as the subject's judgment is made within the frames of the company. Other employees within the same company are expected to make the same judgment or at least have interest in the judgments made by colleagues. This consensus can ease the operational measurement because the use of the data then signifies a contextual fit. It is expected that logging of the specific use of the data warehouse tables (views), rows, and columns can construct a valid index for the contextual value of the data. The data warehouse is a machine powered by usage. Without users the data warehouse is of "no use". Data without users become obsolete. In popular terms: "Use it or loose it" (Orr 1998) because data systems can suffer of "atrophy". The concrete logging as a foundation for "knowledge of use" is an example of the data warehouse "taking its own medicine" of obtaining knowledge of their customers though the analysis of behavior.

4 INCREMENTAL CHANGE

The dynamic user addendum to the data warehouse is regarded as an incremental change because each user is only adding small portions. However, the many users are aggregated to be delivering a significant change of the data warehouse.

It is unthinkable that a user of a data warehouse would be permitted to change the data of the data warehouse. However, it is totally acceptable that a user will be allowed to add information to the metadata of the data warehouse. This will address the problem of lost knowledge: "Users working with a particular data set come to know and internalize its deficiencies and idiosyncrasies. This knowledge is lost when data are made available to other parties" (Ballou & Tayi 1999). The challenge of the data warehouse is to make this knowledge persistent and develop a facility for transforming this knowledge and the "knowledge of use" into metadata and thus implement "use of knowledge". The loaded data and metadata are normally viewed as consolidated and never to be changed. However, the rationality of this view only relates to the data itself not to the metadata.

5 CONCLUSION

The result of this paper is: Firstly, the nature of data quality has been exemplified in data quality dimensions. Secondly, attention has been drawn to the fact that the quality of data in the data warehouse is closely related to the selection procedures. Thirdly, data quality is not static but can be dynamically improved through the use of the data.

In the constant iterative development it is important technically to secure that information on data use is collected, stored, and disseminated. Organizationally the knowledge of the users must be received, processed, and added to the metadata.

REFERENCES

Ballou,D.P. & Tayi,G.K. (1999) Enhancing data quality in data warehouse environments. *Communications of the ACM* 42, 73-78.

Bruckner,R.M. & Schiefer,J. (2000) Using portfolio theory for automatically processing information about data quality in data warehouse environments. *Advances in Information Systems, Proceedings* 1909, 34-43.

Fox,C., Levitin,A.V., & Redman,T.C. (1994) The notion of data and its quality dimension. *Information Processing & Management* Vol. 30, 9-19.

Howard,R.A. (1988) Decision-Analysis - Practice and Promise. *Management Science* 34, 679-695.

Inmon W.H. (1996) *Building the Data Warehouse (2.ed.)*. John Wiley & Sons.

Kimball R. (1996) *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley & Sons.

Orr,K. (1998) Data quality and systems theory. *Communications of the ACM* 41, 66-71.

Strong,D.M., Lee,Y.W., & Wang,R.Y. (1997) Data quality in context. *Communications of the ACM* 40, 103-110.

Tayi,G.K. & Ballou,D.P. (1998) Examining data quality. *Communications of the ACM* 41, 54-57.

Wand,Y. & Wang,R.Y. (1996) Anchoring data quality dimensions in ontological foundations. *Communications of the ACM* 39, 86-95.

Wang,R.Y. & Strong,D.M. (1996c) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 5-24.

Wang,R.Y. & Strong,D.M. (1996b) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 5-24.

Wang,R.Y. & Strong,D.M. (1996a) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 5-24.