Ensuring Data Quality in Data Warehouse Environments through Attribute-based Metadata and Cost Evaluation*

Yu-Chi Chu† Shan-Shan Yang‡ Chen-Chau Yang†

†Department of Electronic Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan, Republic of China
<ycchu@sun.epa.gov.tw ccyang@et.ntust.edu.tw>

‡Program Management & Coordination Dept.
Telecommunication Laboratories
Chunghwa Telecom Co., Ltd.
Taipei, Taiwan, Republic of China

Abstract

Data quality will be a significant issue as data warehousing becomes more and more popular. On the basis of the consideration of quality assurance (QA), we present an attribute-based metadata model which plays as an explanation for identifying data quality. The data stored in warehouses therefore can be classified into attribute data and their quality metadata. A four-phase process is introduced for data quality management during the life cycle of data warehouses. Overall data quality conditions can be identified and related information can be provided for determining whether the data meet "fit to use" principle and whether they need to be improved. Furthermore, we use cost/benefit evaluation model to find out the non-quality data and set the priority for improvement under limited resources. Our approach allows system developer to document relevant quality data as the metadata which may be associated with the life cycle of data warehouses. The quality metadata not only can enrich the interpretation of attribute data, but also can provide the diagnostic information to figure out the reasons and the sources of errors. In addition, the cost/benefit evaluation model we developed may provide a foundation for the quantitative analysis of data quality.

1 Introduction

In recent years, data warehouse systems (DWS) have attracted a great deal of attention both of practitioners and academician. A significant characteristic of data warehouses is the prominent roles of historical data. Most operational systems focus on current data, yet data warehouse systems (informational systems) often involve temporal comparisons for supporting decision making. Generally, the information stored in the warehouse can be structured and organized in a form that makes it easy to use for applications. A data warehouse can therefore be seen as a set of *materialized views* defined over the remote sources, and warehoused data is usually used for decision making, rather than for operations.

Data warehousing efforts have to address several potential problems[13]. For example, data from different sources may have serious semantic differences, and is likely to contain syntactic inconsistencies. Moreover, the desired data may simply not have been gathered. Therefore, data warehousing projects may not succeed for various reasons, such as poor system architecture or unacceptable query performance, but nothing is more certain to yield failure than lack of concern for the issue of data quality. Non-quality data will lead either to wrong decisions being made, or knowledge workers losing confidence in the data warehouse. Unfortunately, most of the research work over the past few years on data warehousing are restricted to quantitatively selecting view sets for materialization[5, 13].

There are some studies have been done on the quality issues of information systems and data warehouse environments [1, 6, 11, 12]. Wang et al. propose a framework of data quality analysis, based on the ISO 9000 standard [12]. This framework reviews a significant part of the literature on data quality, while only the research and development aspects of data quality seem to be relevant to the cause of data warehouse quality design. In [11], an attribute-based model is presented and can be used to incorporate into quality aspects of data products. The quality data and attribute data are simply hard to be separated for accessing or evaluating since they focus on the environment of traditional databases. Moreover, the query languages for standard databases such as SQL may need to be modified for fitting a new schema architecture. Jarke et al. presented an approach to exploring the architecture and quality in data warehouses based upon extended repository [6]. This approach extends the Goal-Question-Metric approach from software engineering to capture some quality dimension in data warehousing architecture. The quality issues discussed in [6] are focused on the quality of the *design and implementation of data warehouses,* rather than the quality

*This work was supported in part by National Science Council, R.O.C. grant #NSC88-2213-E011-004

of the data stored in data warehouses.

On the basis of the consideration of quality assurance (QA), we propose a methodology for data quality assurance in data warehouse environments. Our methodology not only concerns the processes of improving data quality, but also takes into account the cost of data quality improvement. Since metadata plays an important role through out the life cycle of data warehouse, we adopt an attribute-based metadata which plays as an explanation for identifying the data quality. In terms of the information provided by quality metadata, we can identify the diversity of poor data which are not fit to use. Furthermore, we use a cost/benefit model to find out the most fatally poor data and set the priority for improvement under limited resources.

This paper is organized as follows. Section 2 presents a framework for data quality assurance and how the data quality can be represented as a hierarchy structure, as well as a four-phase process for data quality management (DQM) with detailed description. The proposed attribute-based metadata model is given in section 3. Section 4 describes the evaluation of data quality based on the cost/benefit model. Section 5 contains concluding remarks and future research directions.

2 Framework of Data Quality Assurance

Data quality has two distinct aspects: one involves the objective "correctness" of data (such as accuracy and consistency), and the other involves the appropriateness of data for some intended purpose. Data producers and users generally assume that the purpose of data quality assurance is to provide the best data possible. But this obscures the need to evaluate data. The implication is that if a data set is the best available and is as good as it can be made, then there is no other options but to use it. In this case, there is no point in worrying about just how good as it can be made. The flaw in this is that merely saying that a data set is as good as it can be made does not tell us *how* good it is or whether it is *any* good at all. What may be considered good data in one case may not be sufficient in another case. For example, analysis of the financial position of a firm may require data in units of thousands of dollars, while audit requires precision to the cent. Therefore, the term "data quality" may best be defined as "fit to use," which implies the concept of data quality is relative[10].

2.1 Data quality hierarchy

On the basis of the goal of "fit to use," the data quality can be classified to four dimensions, and each dimension might be composed of several "data quality factors." Moreover, each data quality factor may have some "data quality indicators." Therefore, data quality issues can be formed as a hierarchy structure investigating the relationship between each level in a systematic understanding[1, 11]. Figure 1 shows the hierarchy structure of data quality. We briefly explain the four dimensions of data quality as follow.

- **Accessibility** From the user's point of view, it is hampered that a DWS should provide an efficient mechanism for accessing the data in data warehouse under certain consideration of security. The mechanism should be able to reduce the efforts of searching in a large and poorly structured information space, as well as avoiding interference of data analysis with operational data processing. When the amount of data in the data warehouse becomes huge, the factor of performance should be taken into account for evaluating the balance between access efficiency and system loading.
- **Interpretability** It remains difficult for DWS users to interpret the data because the semantics of data description languages for data warehouse schemata is weak, fails to take domain-specific aspects into account, and is usually not formally defined and therefore hardly computer-supported. The data interpretability dimension concerns with data description, such as data layout for legacy systems and external data, table description for relational databases, primary and foreign keys, aliases, defaults, domains, explanation of coded values, etc.
- **Contextual** We adopt the amount of information, relevancy, and timeliness as three factors to evaluate the data quality of contextual. A great deal of information might help the process of decision making, but obviously it might also cause the degradation of system performance and waste of resources. Thus, the relevance between user's need and the data in the warehouse should be evaluated during the construction of data warehouse. With regard to the factor of timeliness, we can evaluate it by examining two indicators: *non-volatile*, which means the use of data is independent on temporal relationship, and *current* means dependent on temporal relationship.
- **Believability** Since most the DWS users often do not know the credibility of the source and the accuracy of the data, the believability of data is hampered. We can evaluate the believability of data in terms of the completeness, consistence, accurate, and credibility[2, 4]. The completeness means the percentage of the real-world information entered in the sources and/or the warehouse. For example, completeness could rate the extent to which a string describing an address did actually fit in the size of the attribute which represents the address. The accuracy stands for the correctness of the data entry process which happened at the sources. The consistency represents the logical harmony of the in-

formation, both in syntactic and semantic aspects. The credibility describes the trustworthiness of the sources that provided the information.

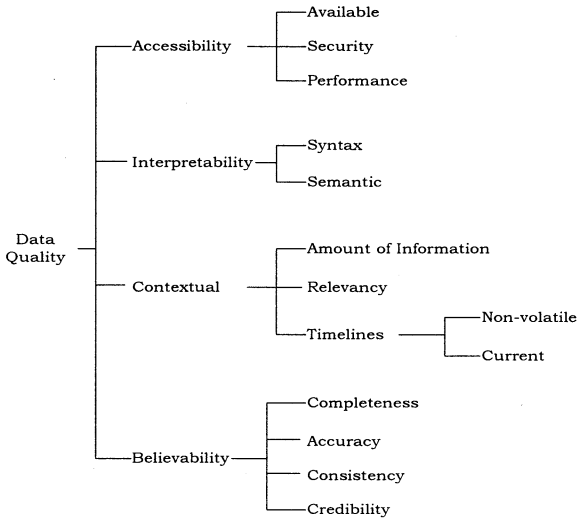

Figure 1: Data quality hierarchy

2.2 Process of data quality management

The quality of data stored in the warehouse is obviously not a process by itself; yet it is influenced by all the processes which take place in the warehouse environment. Thus, quality considerations have accompanied data warehouse research from the beginning. Generally, data stored in the warehouse come from various sources including internal databases and external data resources. If there are quality problems in those sources, these problems will obviously be transformed to the warehouse. This causes an unpredictable situation when a warehouse was applied for decision making. Furthermore, even if there are no quality problems in those data sources, data errors or the degradation of data quality might occur during the processes of data integration and transformation when constructing and maintaining data warehouses. Hence, there is a need to develop a mechanism or operational procedure that can be used to enhance the data quality during the life cycle of data warehouse.

We propose a systematic process for data quality assurance shown as figure 2. The process covers the life cycle of data warehouse development and consists a number of phases that can be viewed as the activities for data quality management: (1) analyzing data quality requirement, (2) constructing attribute-based metadata for data quality interpretation, (3) identifying data quality, (4) performing cost/benefit evaluation. Once accomplishing these activities, we may determine which data items need to be improved to meet the goal of " fit to use." We briefly describe the basic concepts and functionalities of each phase as follows.

Phase 1: Analyzing data quality requirement. This phase is somewhat similar to logical design of conventional database systems, whereas the system designer have to figure out the semantic ambiguity and syntactic inconsistencies from various sources. Data issues and quality issues should be taken into account during this phase. For example, what kinds of data quality factors and how many factors should be involved to meet user's needs. Do these factors meet "fit to use" principle? The results of this phase will be the specifications of quality requirement which can be understood by both users and system designers.

Phase 2: Constructing attribute-based metadata. Data warehouse systems usually have multidimension schema to store integrated data from diversity sources. For ensuring the data quality, we may add an extra dimension which is dedicated to the description of data quality correspondent with specific attribute. Moreover, the quality data can be combined with attribute data to simplify the description. Although the description of data quality will cause the overhead of storage resources, we believe the benefits from good data quality can cover the cost of storage space.

Phase 3: Identifying data quality. Since data warehouse may support decision making to the users at different levels in the organization, the requirements of data quality differ from various points of view. This is consistent with the principle of "fit to use." For example, concerning the data factor of timeliness, some users may need the data only last year, and others may need the data within the past decade for detail analysis. We should identify the data that fail to meet the quality requirements, and find the reason why they are not qualified.

Phase 4: Performing cost/benefit evaluation. Once the unqualified data are identified, we have to figure out how to improve the quality of those data. On the practice perspective, we need to take the cost issues into account to see how many efforts should be payed to improve the quality of unqualified data. In fact, it is impractical to achieve a flawless state of data quality. Therefore, we have to evaluate a balance condition between the cost and benefit before we decide how to improve the quality of unqualified data.

After finishing above process, the results may provide the system designer with a helpful support to adopt appropriate strategies for data quality assurance. For the data that meet the requirements of data quality, we may

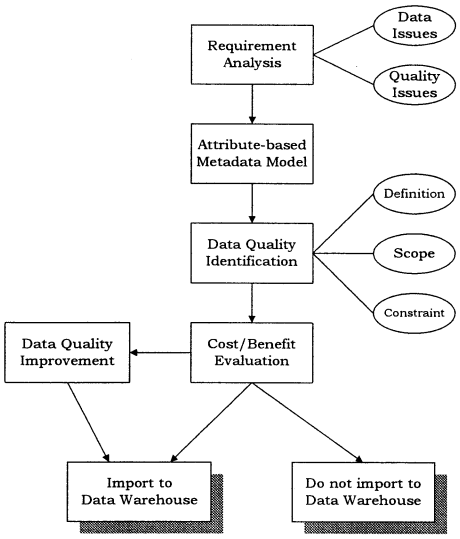

Figure 2: Process of data quality management

import them to the warehouse immediately. The unqualified data can be divided into two aspects. The first portion will be imported to the warehouse after we improve their quality for the cost of improvement is under consideration. Another portion represents that unqualified data do not meet the balance between cost and benefit. It is a tradeoff issue whether such data should be imported to the warehouse. We suggest that system designer and users might make decisions based on the subjective deliberation.

3 Attribute-based Metadata Model

Metadata may describe or be a *summary* of the information content of the data described in an intensional manner. There are many kinds of metadata that can be associated with a data warehouse, including metadata to improve or restrict access to data, to facilitate sharing and interoperability, to characterize and index data, etc. Metadata may also be used to define the user's expectations of data quality and to describe the conditions of data quality at data warehouses. Thus, the metadata may play as a *data quality profile*, which can be easily extended as required[9].

When a data set is obtained from a data source or intermediate data center, the objective aspects of its quality (accuracy, consistency, etc.) may already have been determined. Various agents and users themselves may combine, aggregate, filter, edit and modify data from different sources in order to prepare a data set for a specific use. In addition to recording information about data values themselves and evaluating the quality of these values, it is important to record information about the processes that affect data, both to ensure that data quality is not corrupted and to allow improving these processes. The next two subsections present an overview of attribute-based metadata model which can be used to evaluating, assessing, maintaining and improving data quality.

3.1 Metadata to support data quality

The attribute-based data quality model for information systems was proposed in [7, 11]. The approach adds an indicator to specify the quality requirement for each data attribute. Those indicators can be viewed as the goal that data quality should achieve during the life cycle of information systems. Once the data field associates with its quality indicators, the storage structure for the data field will be adapted to a pair structure as $<$ `Attribute, Quality_key` $>$. For accommodating the pair structure, the typical languages for database manipulation and query such as SQL should be modified to fit the requirement. One of the shortcomings of this mechanism is that it causes overheads both in the aspects of storage resource and system performance.

We adapt attribute-based data quality model in [11] with more flexible and simplified considerations. From the standpoint of data quality management, we propose an attribute-based metadata that adds quality description to an attribute to play as the quality explanation in warehouse environments. Since most data warehouse systems have multi-dimension characteristics for storing and retrieving data, we may use one of dimensions dedicated to specifying quality metadata. Therefore, the data quality can be corresponded to specific data attribute by an internal linkage. For example, table 1 shows an extended table in which the original data table is $<$ `ID, Addr, Owner, Profit_est.` $>$. After analyzing the requirement of data quality, we add three items of quality metadata, $<$ `Entry_date, Evaluator, Entry` $>$ for the attribute of `Profit_est`. These metadata represent the date of data being input, who evaluated the profit, as well as who punched the data.

3.2 Specifying quality with metadata

For ensuring proper association of data attribute and data quality in the warehouse, the results obtained from the phase of requirement analysis of data quality will provide a helpful support during the implementation. Because each data attribute may have several quality metadata items, we need to develop a mechanism for investigating their association relationship within the schemas of data warehouse systems. If a data attribute connects with quality metadata, then a symbol of "Δ" will be marked to identify all the data belong to the attribute can be associated with its quality metadata.

The overall picture of the implementation forms a multi-level framework. The quality metadata in the same level are viewed as a *quality schema*. The primary index in the schema is *quality index* which connects the

Table 1: Attributed-based metadata model with quality description

ID	Addr.	Owner	Profit_est.	Entry_date	Evaluator	Entry
B001	Taipei	David	$1,000	1998-03-21	Monica	Mary
B002	Taichung	Mike	$3,000	1998-03-21	Bill	John
⋮	⋮	⋮	⋮	⋮	⋮	⋮
data attribute				**quality metadata**		

attribute data and quality metadata. For example, if we concern the data quality of profit estimation such as who made the estimation and when the data were imported, we may mark the symbol of "Δ" to the attribute of <Profit_est.>. This will form a quality schema shown as table 2.

Table 2: Extended attribute with quality description

ID	Addr.	Owner	Profit_est. Δ
B001	Taipei	David	$1,000
B002	Taichung	Mike	$3,000
⋮	⋮	⋮	⋮

The attribute <Profit_est. Δ> in table 2 can be further extended with its quality metadata to form a new quality schema including quality describtion shown as table 3; the attribute <Source> in table 3 can make a further describtion redagding the quality of data sources shown as table 4. In the example, we may find that the profit estimation for certian stores are depended on the *evaluator's quality*. The decision maker may judge the beleviability of the data in the warehouse based on who generate the data and when the data were imported. A detailed illustraction of the implementation of quality metadata can be found in [14].

Table 3: Level one of quality metadata

Profit_est. Δ	Source Δ	Entry
B001 Δ	Monica	Mary
B002 Δ	Bill	John
⋮	⋮	⋮

Table 4: Level two of data quality metadata

Source Δ	Evaluator	Entry_date
B001 Δ	Monica	1999-03-21
B002 Δ	Bill	1999-03-21
⋮	⋮	⋮

Our approach for constructing the attribute-based metadata deffers from the approach in [11] in two ways. First, we use an extra dimension of data warehouse for storing quality information instead of using a pair structure which might cause the diffculty in implementation. Second, our approach may benefit the process of the query of data quality, and users are able to query the data quality without the modefication of query language. Therefore, the performance for assuring data quality in the warehouse can be overall improved.

Furthermore, when data quality issues are associated with the life cycle of data warehouses, we need to take the problem of integration into account. Concerning the inconsistency situation during the process of data access, we have to modify quality metadata simultaneously when corresponding attribute data were modified or deleted.

4 Identifying and Evaluating Quality

It is a subjective issue when we evaluate the data quality under the considerations of cost and benefits. Before we transform the data that extracted from various sources into data warehouse, we need to consider not only from user's point of view, but also the internal and external situation. For example, apart from the product issues, a commercial subject data warehouse has to consider the factor of market change as well. The process of data quality evaluation therefore compose a circularity characteristic. In this section, we present the procedures of data quality idenfication and cost/benefit evaluation. We first introduce how to establish data quality constraints in conjunction with attribute-based metadata, and how the errors and anomalies can be detected. Second, we apply a cost/benefit evaluation model to find out the priority of non-quality data that should be improved.

4.1 Data quality idenfication

The procedures of data quality idenfication on data warehouse are more complicate than traditional databases. However, accoding to the survey in [3, 8], almost 80% of data quality problems cuased by 20% defects. Hence, we may refer such situation as Pareto principle, which implies that we should chase up the problems and causes that have the biggest impact on quality and cost.

On the basis of quality requirements that defined in terms of data quality metadata, we can establish a list of

data quality constraints based on the quality deminsion of contextual introduced in section 2. For example, some data items might have the consideration of expiration. If the the profit estimation in table 1 was made two years ago, it might not reflect current situation at all.

By using data quality constraints, data error and anomaly may be detected based on rules that are constructed by the user. If the rules in a constraint are completely specified, then for each record that satisfies the IF condition, we may assign the quality condition a certain state to corresponding attribute data. For example, the following constraint #1 states that attribute data `Entry_date` has a expiration constraint; constraint #2 states that if attribute data `Profit_est.` is greater than $3000, and metadata shows that the data is from source of "`Monica`", such data may be assigned to a condition of anomaly because it contradicts user's expectations.

```
%Constraint #1:
    IF (M.Entry_date<1999-03-31)
    THEN Condition = expired
%Constraint #2:
    IF (A.Profit_est.>3000) AND
        (M.Source="Monica")
    THEN Condition = anomaly
```

For the sake of balance between cost and benefit, we need to rank the data quality indiators based on cureent condition of data quality. The purpose of ranking is to assist warehouse's managers to figure out the most critical part of non-quality that affects data quality in data warehouses. We apply Pareto chart to prioritizing non-quality data accoding to the constraints sitting above. The goal of Pareto chart is to identify the most important issues or defects that need to be addressed. Ranking criteria can be classified by different subjects and concerns. For example, the rate of data errors, the cost caused by defected data and how many resources are required to fix non-quality data.

The outcomes of data quality idenfication provide a foundation for further cost/benefit evaluation. We are currently implementing a set of tools for assisting warehouse's manager to generate and maintain data quality constraints, as well as for detecting the errors and anomalies. These tools will consist of a group of graphic user interfaces and middleware to access attribute data and quality metadata in warehouses. A more detailed description of the implementation of data quality identification can be found in [14].

4.2 Cost/benefit evaluation

One of the important issues for evaluating data quality is how to quantifying each indicator of data quality. The cost/benefit evaluation model we developed is based on an assumption of the relationship between elapased time and data quality. We believe that the degradation of data quality in the data warehouse is highly related to the temporal aspect. Therefore, the relationship between time and data quality may provide a foundation for the quantitative analysis of data quality.

We define the degradation of data quality as a function of time, denoted as $Q(t)$, representing the data quality in data warehouse at a certain time point t. In addition, the overall data quality is composed of several indicators of data quality; thus $Q(t)$ can be defined as

$$Q(t) = \sum_{i=0}^{n} Q_i(t) \tag{1}$$

where $Q_i(t)$ represents the degradation of each indicator of data quality. According to equation (1), we may not be able to find out the priority between each indicator, and all indicators are in the same weight of cost and importance. Yet if we view the degradation as the proportions of non-quality data in warehouses, $Q(t)$ can be quantified to a real number between $0 \sim 1$, then equation (1) can be modified as follows.

$$Q(t) = \sum_{i=1}^{n} \frac{\rho_i}{W} \tag{2}$$

where W represents the amount of attribute data, and ρ stands for the number of non-quality data.

We propose that the total cost of data quality should include the *lost cost* and the *improvement cost*. In accordance with these two issues, we can determine what kinds of data item in the data warehouse should be modified and improved. We explain the lost cost and the improvement cost as follows.

Lost cost means the cost caused by the non-quality data. The cost may be expressed in terms of lost funding, lost production, lost assets or legal liability. Generally, lost cost is dependent on the degradation of data quality.

Improvement cost means the cost that must be spent for improving data quality to a certain level. It is dependent on the number of data quality indicators which need to be improved or modified.

Let non-quality data exist from $t = t_0$, the improvement of non-quality start at $t = t_1$ and finish at $t = t_2$. Then, the lost cost caused by non-quality data is $Q(t_2)$, and improvement cost is $Q(t_2 - t_1) + C$, where C is the cost of time-independent issues such as material resources. Moreover, we found that $\frac{dQ}{dt}$ is more convenient for evaluation than $Q(t)$, because $\frac{dQ}{dt}$ means the proportions of non-quality data within a time unit, and appropriately represents the degradation of data quality in warehouses. We therefore construct the evaluation model based upon the following criteria, i.e. modeling $\frac{dQ}{dt} \sim t$ fot its distributed relationship.

1. Let the degradation of data quality $\frac{dQ}{dt}$ be the scale of non-quality data in data warehouse within a time unit.
2. The lost cost is directly proportional to $\frac{dQ}{dt}$ with two coefficients C_1, C_2.
3. $\frac{dQ}{dt}$ is directly proportional to time t, and there is a coefficient β to identify the rate of the degradation of data quality .
4. When $t \geq t_1$, the rate of the degradation of data quality becomes $\beta - \lambda x$, where λ is the average rate of the improvement. In an ideal condition, the assumption should satisfy $\beta < \lambda x$.
5. On the basis of the temporal relationship, the improvement cost of data quality can be classified as follows.
   - **time-dependent**: let C_2 be the coefficient of improvement cost within a time unit, then the improvement cost of each quality indiator is $C_2(t_2 - t_1)$, for instance, the input of manpower can be viewed as a time-dependent cost.
   - **time-independent**: let C_3 be the coefficient of improvement cost, for instance, the input of material resources is a time-independent cost.
6. We classify non-quality data based on each data quality indicator, therefore $\frac{dQ}{dt}$ can be determined by $\sum \frac{dQ_i}{dt} (i = 1, 2, \ldots, n)$, and we obtain

$$\begin{aligned} \frac{dQ(t)}{dt} &= \frac{d\sum_{i=1}^{n} Q_i(t)}{dt} \\ &= \sum_{i=1}^{n} \frac{dQ_i}{dt} \end{aligned} \tag{3}$$

We construct the model based upon $\frac{dQ}{dt} \sim t$ relationship shown as figure 3 using the assumptions mentioned above. During $t_0 \leq t \leq t_1$, the degradation rate of data quality, denoted as β, will be the linear ratio along with time. As $t = t_1$, the amount of inadequate data become $\frac{dQ(t=t_1)}{dt} = q$.

When $t_0 \leq t \leq t_2$, it is the time period that occurs inadequate data, and the degradation of data quality is $Q(t_2) = \int_0^{t_2} \frac{dQ}{dt} dt$, i.e. the triangular area in figure 3 when $0 \leq t \leq t_2$. We compute lost cost **LC** based on the assumption 2 introduced earlier:

$$\mathbf{LC} = C_1 Q(t_2) = C_1 \int_0^{t_2} \frac{dQ}{dt} dt = \frac{1}{2} C_1 q t_2 \tag{4}$$

Let $\frac{q}{t_2 - t_1} = \lambda x - \beta$, we obtain

$$\mathbf{LC} = \frac{1}{2} C_1 q t_1 + \frac{C_1 q^2}{2(\lambda x - \beta)} \tag{5}$$

During $t_1 \leq t \leq t_2$, it is the time period that modified inadequate data; the degradation of data quality

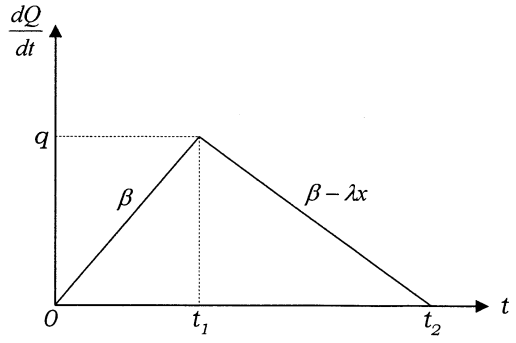

Figure 3: $\frac{dQ}{dt} \sim t$ relationship

becomes $Q(t_2 - t_1) = \int_{t_1}^{t_2} \frac{dQ}{dt} dt$, i.e. the triangular area in figure 3 when $t_1 \leq t \leq t_2$. Based on assumption 5 introduced earlier, we obtain the improvement cost as follows.

$$\mathbf{IC} = C_2 x(t_2 - t_1) + C_3 x = \frac{C_2 q x}{\lambda x - \beta} + C_3 x \tag{6}$$

Accordingly, plus lost cost and improvement cost, we have

$$\mathbf{TC} = \frac{1}{2} C_1 q t_1 + C_3 x + \frac{C_1 q^2}{2(\lambda x - \beta)} + \frac{C_2 q x}{\lambda x - \beta} \tag{7}$$

For determining how may quality indicators should be improved, we need to reduce the total cost. Let $\frac{dC}{dx} = 0$, we can compute minimum total cost using C's first order derivative.

$$\mathbf{x} = \sqrt{\frac{C_1 \lambda q^2 + 2 C_2 \beta q}{2 C_3 \lambda^2}} + \frac{\beta}{\lambda} \tag{8}$$

The $\mathbf{x}$ in equation 8 plays a fundamental role in result of the cost/benefit modeling. It represents that the amount of quality indicators we may handle with the minimum cost. We observe that $\mathbf{x}$ is composed of two parts where $\frac{\beta}{\lambda}$ represents the cost for improving inadequate data, i.e. that β is the degradation rate of data quality and λ is the average improvement rate of each data quality indicator. It is obviously that the slope $\beta - \lambda x$ will be negative and has a chance to cross with axis t in figure 3 only if $x > \frac{\beta}{\lambda}$. Another part for determining the amount of quality indicators is related to each parameter when we adopt the model. When the average improvement rate λ and the coefficient for improvement cost C_3 are increasing, the amount of improved quality indicators decreased. Moreover, when the degradation rate of data quality β, degradation condition of data quality as the improvement started q, and coefficient for lost cost C_1 are increasing, the amount of improved quality indicators increasing as well. There are some de facto

factors have to be considered as we adopt the model to evaluate the improvement plan for data quality. In general, C_1, C_2, C_3 can be viewed as the constants, β and q can be obtained by examination, and λ can be computed by experienced data warehouse managers. We may gradually rectify the parameters in equation 8 for adapting the model close to de facto distribution.

5 Conclusions

We have described an attribute-based metadata model which separates data in the warehouse into two aspects, attribute data and quality data. We also explore how quality data can be represented as metadata and how it can be accessed in data warehousing architecture. Thus, the data quality issues can be effectively managed and assured. Concerning the degradation of data quality in warehouse, we propose a cost/benefit model to perform the evaluation and find out what kinds of data items should be modified or improved. The main contributions in this paper can be summarized as follows.

- Attribute-based metadata model can enhance the data interpretation and assist to deal with the semantic conflict in data warehousing architecture. As time goes on, the users of warehouse may face the problem of interpretability, because they do not know how the data are transformed into the warehouse. Our approach allows the system developer to document the related quality data as the metadata which may be associated with the life cycle of data warehouse.
- Quality requirement can be formally and clearly defined in terms of attribute-based metadata and that provides the diagnostic information to figure out the sources of data error.
- Overall data quality conditions can be identified and relevant information can be provided for determining whether the data meet "fit to use" principle and whether they need to be improved.
- Users may filter the data that retrieved from the warehouse based upon various quality requirements. On the basis of constraints of cost and timing consideration, we may then figure out what kinds of data should be preferentially modified or improved in order to achieve maximin benefit of data quality.

We believe that data quality will be a significant issue as data warehousing becomes more and more popular. There is obviously a great deal of work remained to be done for obtaining better data quality to support decision making. One direction of current work will continue to expand the cost/benefit model for more detail evaluation and analysis of data quality. For example, we may further analyze of $\lambda(q)$ to explore the relationship between λ and the degree of data degradation q. We will use a real-world example to justify the model proposed in section 4. In addition, AI technique for data quality definition, and Machine Learning to enhance the capability of non-quality data detection and identification may also be taken into account for future work.

References

[1] DWQ Project. http://www.dbnet.ece.ntua.gr/~dwq/.

[2] Ballou, D. P., and H. L. Pazer, "Cost/Quality Tradeoffs for Control Procedures in Information Systems." *International Journal of Management Science* 15(6), pp. 509-521, 1987.

[3] Barquin, R., and H. Edelstein, *Building, Using, and Managing the Data Warehouse.* PTR Publishing, 1997.

[4] Huh, Y. U. et al., "Data Quality," *Information and Software Technology.* 32(8), pp. 559-565, 1990.

[5] Inmon, W. H., *Building the data warehouse 2/e.* Wiley Publishing, 1996

[6] Jarke, M., M. A. Jeusfeld, C. Quix, and P. Vassiliadis, "Architecture and quality in data warehouse: an extended repository approach." *Information Systems.* 24(3), pp.229-253, 1999.

[7] Kimball, R., *The data warehouse toolkit.* Wiley Publishing, 1996

[8] Parsaye, K., and M. Chignell, *Intelligent Database Tools and Applications: Hyperinformation Access, Data Quality, Visualization, Automatic Discovery.* Wiley Publishing, 1993.

[9] Rothenberg, J., "Metadata to support data quality and longevity." *Proc. of 1st IEEE Metadata Conference,* 1996.

[10] Tayi, G.-K. , and D. Ballou, "Examining data quality." *CACM.* 41(2), pp. 54-57, 1998.

[11] Wang, R. Y., M. P. Reddy, and H. B. Kon, "Toward Quality Data: an Attribute-based Approach." *Decision Support Systems.* vol. 13, pp.349-372, 1995.

[12] Wang, R. Y., V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research." *IEEE Trans. Knowledge and Data Engineering.* 7(4), pp.623-640, 1995.

[13] Widom, J., "Research Problems in data warehousing." *Proc. of 4th Int'l Conference on Information and Knowledge Management (CIKM),* 1995.

[14] Yang, S. S., *An Evaluation Mechanism for Data Quality Assurance in Data Warehouse Environments.* Master thesis, National Taiwan University of Science and Technology, 1999. (in Chinese)