Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organisations in Australia

Amit Rudra
School of Information Systems
Curtin University of Technology, Perth
AUSTRALIA
RudraA@cbs.curtin.edu.au

Emilie Yeo
School of Information Systems
Curtin University of Technology, Perth
AUSTRALIA
emergper@singnet.com.sg

Abstract
This paper discusses the emergent key issues of data quality in a data warehousing environment. The research study leading to our outcome is described. We investigate the relationship between data quality and data consistency; determine whether data inconsistencies are present in data warehouses and explore how organisations ensure, plan and maintain data quality. Our research outcome - an improved understanding of how organisations can maintain data quality and consistency. We found that the quality of data in a data warehouse could be influenced by factors like: data not fully captured, heterogeneous system integration and lack of policy and planning from management.

1. Introduction
The concept of data warehousing (although, not always under that name) has been in existence for a few years (Porter and Rome, 1994; Warmouth and Yen, 1992). Its drivers, the informational needs of decision makers, have always been there (Shah and Milstein, 1997). A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process to make timely, accurate decisions (Cobb, 1996; McClanahan, 1996; Inmon, 1995). As distinct from an operational database, a data warehouse is managed data situated after and outside the operational systems (Gupta, 1997). Here data is stored and organised for informational and analytical processing over a long historical time (Inmon, 1995). Figure 1 shows an example of a data warehouse. Since a data warehouse contains historical data to supply a time-related view (e.g. for trend analysis), it is mainly intended for analytical applications such as decision support systems and executive information systems (Gallagher, 1995). Of recent, its development has taken off due to the Internet and the fact that information is now almost entirely captured electronically (Raden, 1996; Gupta, 1997; George, 1996). Research shows that implementation of data warehouses in organisations is on the increase. For example, according to a survey commissioned by Tandem on data warehousing and decision support systems, Metrica Consultants found that 85% of the top 1000 European companies either have a data warehouse or plan to create one in the next two years (“Clear Targets Vital for Data Warehousing”, 1996).

Figure 1. Data warehouse as a subject oriented database
The data warehouse allows both simple and complex queries to the database (Lais, 1996). It also provides the tools and procedures that let users manipulate the data to get only the pieces they need in the most useful form to support decision-making (Lais, 1996). Organisations argue that they can use their old programs and databases to create reports and print them out as specified by the users (Bussert, 1997). However, it is a fact that there would be requirements for different types of reports. One cannot satisfy all the needs of the users using the old report generator. The data warehouse allows different combinations of results into reports for efficient query processing (Bussert, 1997; Chaudhuri and Dayal, 1997).

Lately, there has been a rise in expectations about what the data warehouse can do for the organisation (Marshall, 1996). This is an important point because end users often are lured into believing that a data warehouse would bring immense benefits to the organisation. Bell Atlantic faced the problem of data organisation and the solution to that problem was to know what the data is, where it came from, and how significant and accurate it was (Campbell, 1997). A data warehouse cannot support cross-departmental analysis unless the data permits it. To enable such analysis, IT staff may need to model what the departments do and what their data mean. They may even need to study and revamp business processes (Atre, 1997).

This research study aims to provide a real-world analysis of data warehousing and examines how some of the large organisations in Australia plan for information quality in a data warehousing environment. It is expected that organisations considering implementing a data warehouse may benefit from the findings of his study. Moreover, quality issues and the organisation's role in achieving data quality may persuade the whole organisation to participate and take part in the development of a data warehouse.

2. Research Study Objectives

The importance of information and data management has been highlighted by many authors, of which Galliers’ (1993) is an important one outlining the key issues of information systems research. As per the article, data warehousing should belong in both information architecture and data resource management; both of which were the first and second highest ranked 'key information systems management issues’ in the United States. However, in Australia, information architecture was placed third and data resource management a low tenth position (Galliers, 1993). This means that Australian organisations do not view data management as an important issue and hence less attention has been paid to it.

Early research into data warehousing focused on the changing culture of corporate decision making which reflected the need and justification of data warehousing. However, the issues of ensuring information quality in data warehousing have not been fully explored. Current literature focuses on maintaining data quality in the area of general awareness and the usage of new tools to automate processes. For example, English (1996) and Strehlo (1996) both discuss data quality, but their focus was on tools to sustain the data warehouse. Furthermore, English (1998a) considers the costs of low quality data and how it could affect the organisation. In other articles, he has formulated a self-assessment test on quality issues and advocates training as the solution to data quality problems (English, 1998b; English, 1998c). Nonetheless, most authors such as English (1998a), Strehlo (1996) and Cipriano (1995) have not fully covered the various ways data quality can be achieved. Therefore, our research aims to uncover possible solutions to this problem. We found that training is only a part of the solution organisations can utilise to minimise data pollution and there are other measures that help in minimising it. These will be discussed in the next few sections. Before we review the details of our research, we would like to first discuss some terms, like data quality, used in our research.

2.1. Data Quality

Data quality refers to how relevant, precise, useful, in context, understandable and timely data is (Firth, 1997; Barry and Parasuraman, 1997; Miller, 1996). From the research of Delone and McLean (1992), the seven most important items emerging from research were information accuracy, output timeliness, reliability, completeness, relevance, precision, and accuracy. As information quality holds the same qualities as data quality, many authors have used the terms interchangeably (Firth, 1997; Barry and Parasuraman, 1997).

One of the fundamental obstacles in the current data warehousing environment concerns the existence of inconsistent data. Data inconsistencies occur when similar entities appear in multiple systems and there are multiple records of the same entities (Kelly, 1997). To minimise integrity errors and improve information quality, there must be quality control. Quality control demonstrates the following data quality aspects (Clements, 1990):

- Determines the performance of a product.
- Ensures that there are features beyond the primary function of the product.
- Ensures reliability.
- Ensures conformance to standards.
- Ensures durability.
- Ensures maintainability.
- May include aesthetics.
- Influences perceived quality.
It has been found that often, many end-users, including managers are unaware of the quality of data they use in a data warehouse (Lambert, 1996). Data quality in the data warehouse, is generally poor and there are many foreseeable setbacks (such as - economic failure, ineffective planning of business strategies). Organisations with management plans such as Just-In-Time manufacturing would not be able to function properly because of inaccurate data (Lambert, 1996). There have been some surveys of data quality in last few years. For example, Cipriano (1995) found that there is a correlation between quality management practices and quality information flows. Information flows and technologies also contribute towards obtaining high quality performances and low defects. In another research conducted by Forza (1995), quality information can be obtained with quality management practices, quality information systems and quality performance.

Thus, the issue of data quality in a data warehouse is of great importance. Its success depends on two important processes namely, data cleansing and data-quality improvement (English, 1996). Currently, organisations worldwide are focusing on implementation issues, the most crucial being data integrity (Jordan, 1997). Data integrity in a data warehouse is vital to warehousing success as all decision support, data mining, marketing, service and business decisions are dependent on it (Jordan, 1997).

2.2. Data Inconsistency

Data inconsistency occurs when there are different versions of the same data in the database (Awad and Gotterer, 1992). This can be caused by various stages of update or when a change has been made in one file and not in the remaining files. Inconsistencies in stored data are one of the most common sources of errors in a computer application (McFadden and Hoffer, 1991). They lead to inconsistent documents and reports as well as cause poor integrity of the system (Brathwaite, 1985). As depicted in Figure 2, there is a negative relationship between data consistency and data redundancy. This means that at low levels of data redundancy, data consistency is high.

On the other hand, there is a direct relationship between data consistency and data integrity (Figure 3). As Awad and Gotterer (1992:15) suggest, “data which is logically inconsistent lacks integrity because it cannot be depended upon”.

![Figure 2. Relationship between data consistency and data redundancy](image)

![Figure 3. Relationship between data consistency and data integrity](image)

We can achieve data consistency by controlling or eliminating data redundancy. Coupled with good data administration this will promote a high level of data integrity.

3. Research Methodology

We now discuss how we carried out our research. Since there has been little research (in what was being done) in regard to the issue of data pollution in data warehousing in Australia, our aim was to find out how organisations around Australia are handling the data quality issue. Cipriano’s (1995) survey research was about data quality in the United States. Therefore, we thought it would be useful to conduct research on data quality in an Australian perspective. We used Firth’s (1997) causation of data pollution (Table 1) and modified it into a questionnaire for data administrators. The survey consisted of both open and close-ended questions. The respondents were asked to rank, on a five-point scale (low occurrence to high occurrence), some of the issues concerning causes of data pollution. They were also asked to rank the perceived benefits of using a data warehouse. Quantitative analysis was used to analyse the data received. Since the data was derived from a relatively small sample and our research largely exploratory, we avoided regression analysis.
Table 1. Causes of Data Pollution
(Source: Firth, 1997)

<table>
<thead>
<tr>
<th>When/Where</th>
<th>Why</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Conversions, Migrations or Reengineering</td>
<td>Inadequate data quality testing on conversion process. Conversion programs introduce new errors. Reengineering does not consider data context, usage or definitions.</td>
</tr>
<tr>
<td>Heterogeneous System Integration</td>
<td>Data is inconsistent or contradictory across systems. Inadequate data quality testing on integration.</td>
</tr>
<tr>
<td>Post-Integration of Heterogeneous Systems</td>
<td>Data remains inconsistent or contradictory across systems. Subtleties of poor data quality arise as new scenarios develop.</td>
</tr>
<tr>
<td>Production Software</td>
<td>Software requirements were incomplete or errors were introduced in the development Process. Lack of applied software engineering or production controls.</td>
</tr>
<tr>
<td>Database Design</td>
<td>Record and field definitions are too loose, unstructured or are not normalised. Schema lacks sufficient validation, and integrity rules.</td>
</tr>
<tr>
<td>Data Ageing</td>
<td>The company cannot track the age of data, or has no program to update or enrich data.</td>
</tr>
<tr>
<td>Customer (Un-)Response</td>
<td>Data never fully captured. Customer form is badly designed, or no incentive is given to customer to offer response.</td>
</tr>
<tr>
<td>Fraud</td>
<td>Physical and logical system security is lax or compensating controls are absent.</td>
</tr>
<tr>
<td>Systems Internationalisation</td>
<td>Overlapping or inconsistent interpretation or usage of codes, symbols, formats due to national differences.</td>
</tr>
<tr>
<td>Input Error</td>
<td>The system input method is badly designed, or lacks automatic validation. Human errors easily introduced.</td>
</tr>
<tr>
<td>Business Rules</td>
<td>System requirements lack adequate or current reference to business rules for data.</td>
</tr>
<tr>
<td>Policy and Planning</td>
<td>Lack of management attention to data quality management.</td>
</tr>
</tbody>
</table>

In our survey, we felt that it was necessary to ask the data administrators what they thought were benefits which the data warehouse would bring. Their response to this question would tell us how well they understood the business rules of the organisation and implement the data warehouse to suit business needs and wants.

The benefits of the data warehouse as identified by Kelly (1997:195) are:
- Improved customer service.
- Reduced risk.
- Increased opportunity between the organisation and the customer.
- Improved IT system maturity.
- Reduced cost.
- Improved strategic decision making.

Kelly’s ideas were incorporated into the survey as an intensity ranking scale for the respondents to vote upon.

3.1. The Sample

The Data Administrators (of data warehouses) were targeted, because they were considered to have the best feel of the quality of data handled by their data warehouses and possibly, have a more objective view of the issues involved therein. Due to time limitations we decided to limit the sample to a reasonable size. Since there were no list of users of data warehousing technology in Australia available to us (or if it was there it was confidential), we targeted Australia’s top 50 (Thomas,
large businesses (with over 500 employees). Also, as the number of users of data warehousing technology in Australia is relatively few, we utilised the recommendations of some of the data administrators we contacted and an updated list was prepared. The prospective respondents were approached (by telephone, email and fax) to check their readiness to participate in the survey. This took a considerable amount of time. Further, upon inquiry, we found that a number of them (especially, banks and financial institutions) were reluctant to participate in the survey. The reason - their organisations had a non-disclosure policy pertaining to their databases/data warehouses. The final list included 26 data administrators. Apart from 4 of these data administrators, the rest 22 were from the state of Western Australia. Respondents were assured of confidentiality and the organisations ranged from the government sector, mining industry to retail.

4. Findings

Ultimately, 11 responses were received and analysed; reflecting a 42% response rate. Most of these (9) respondents are based in the state of Western Australia. Of the other two, one each was from the states of New South Wales and Victoria. Upon close scrutiny, we found 9 responses to be valid and useful in our analysis. A profile of the respondents was derived from demographic questions: 55% or most of the respondents were from government, the rest being from retail, mining and insurance sectors; all the organisations had a data warehouse and only one used data marts on top of it; the average length of time a data warehouse installed was 37 months; the average number of users of the data warehouse was 36 and it varied from a low 12 to a high of 300 for a large retail giant. The data analysis also shows that there were two distinct categories of data administrators involved in data warehousing. We identified them as category One – data administrators who have been with the organisation less than 6 years, category Two – data administrators who have been with the organisation for more than 6 years. Although category One had less time immersed in the organisation, their views and opinions were by and large similar to those from category Two. Another classification that we used was whether an organisation had an established data warehouse; defined as one, which has been implemented for over a year. We found that 33% of the respondents classified their data warehouse as established, 44% as not established (since most have just begun, in the midst of, or near the end of the implementation phase) and 22% not sure.

In the following sections we list our findings in detail.

4.1. Causes of Data Pollution

As outlined before, the respondents were asked to rank, on a five-point scale (low occurrence to high occurrence), some of the issues concerning causes of data pollution. Table 2 shows our findings. In regard to the question of whether ‘poor systems conversion, migration or re-engineering cause data pollution’ - 83% of all respondents have given it a low ranking. As the mean was a low 1.7 and the standard deviation 0.82, it implies that most respondents agreed that the cause of data pollution is unlikely due to poor systems conversion. Similar results were obtained for factors like: inadequate software production; security and fraud; and 'systems localisation due to inconsistent interpretation or usage of codes, symbols and formats'. The top three reasons for data pollution in the data warehouse seems to be:

- Data never being fully captured.
- Heterogeneous system integration.
- Lack of policy and planning from management.

Some of the other reasons, but to a somewhat lesser extent, contributing to data pollution are: inferior database design and the inability to cope with ageing data.

Table 2. Ranking of causes of data pollution

<table>
<thead>
<tr>
<th>Cause of data pollution</th>
<th>Mean Score</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data never fully captured</td>
<td>3.0</td>
<td>1.41</td>
</tr>
<tr>
<td>Heterogeneous system integration</td>
<td>2.6</td>
<td>.79</td>
</tr>
<tr>
<td>Lack of policy and planning from management</td>
<td>2.6</td>
<td>.98</td>
</tr>
<tr>
<td>Inability to cope with ageing data</td>
<td>2.4</td>
<td>1.13</td>
</tr>
<tr>
<td>Inferior database design</td>
<td>2.4</td>
<td>1.51</td>
</tr>
<tr>
<td>Business rules lack currency</td>
<td>2.3</td>
<td>.76</td>
</tr>
<tr>
<td>Input error</td>
<td>2.3</td>
<td>1.25</td>
</tr>
<tr>
<td>Systems localisation due to inconsistent interpretation or usage of codes, symbols and formats</td>
<td>2.1</td>
<td>.79</td>
</tr>
<tr>
<td>Security and fraud</td>
<td>2.0</td>
<td>1.41</td>
</tr>
<tr>
<td>Inadequate production software</td>
<td>1.9</td>
<td>.69</td>
</tr>
<tr>
<td>Poor system conversions, migration or re-engineering</td>
<td>1.7</td>
<td>.82</td>
</tr>
</tbody>
</table>

4.2. Measurement of Data Quality

In the survey, data quality is measured in the following ways:

- Completeness of data.
- Consistency of data entry.
- Accuracy of data.
- Uniqueness of account numbers (e.g. primary keys, index numbers).
- Durability of the business rules that underpin the data.
This was an open-ended question where the respondents were asked to comment on the above 5 aspects of data quality. Based on the respondents’ comments, the central theme that appears from the analysis is that it is vital to consider the organisation’s business rules. For example, business rules should be formulated in the schema to ensure the consistency of data entry. Following which it should be implemented in the RDBMS to enhance data accuracy. Records should be validated often and metadata made visible to users to maintain ownership of data. To ensure that the business rules are correct, one retail giant commented that there must be communication with all development groups that provide data to the warehouse such that they are always aware of the change in business rules and hence, assess the impact. The development team should also have had experience with the existing systems to further comprehend the business rules.

The source systems header and trailer records should also be analysed, as data quality is dependent on the source systems. Some organisations rely on client feedback to obtain information on the completeness of data. Research has shown that a few organisations check the completeness, consistency and accuracy of data on an ad hoc basis. Reliance on clients is insufficient if used alone. Audits and system validation checks should be used together with feedback to provide the data warehousing team with a clearer picture of the state of data quality.

From the data collected, legacy systems are present in several data warehouses. As there may be flaws and inaccurate data in the legacy systems, these systems are not checked for uniqueness in account numbers. Other modern systems using Oracle, for example, has database constraints employed to ensure unique numbers.

As most of the data warehouses are meant for analysing strategic decisions, it is necessary to ensure that the data meets the users’ needs and expectations. One particular government organisation carries out client surveys and has a customer service council to deal with customer concerns in this aspect.

The data refresh rate ranges from daily to monthly. If the data refresh rate is slow, data quality would be affected, as the figures would no longer be very useful. As one respondent commented “Quality may be achieved, but it is achieved at a high cost”.

The survey showed that the respondents could be classified into two categories according to the length of time they have been employed in their respective organisations. We expected that those respondents who have been employed in the organisation for a longer period of time to know the organisation and its business rules in greater detail. However, we found that the two groups understood the requirements and the focus of the data warehouse. There was no great distinction or gap between the two groups except that the respondents who have had 6 years or greater working experience in the organisation had a more realistic opinion of the purpose of the warehouse.

Other than that, they have largely similar ideas and thoughts with regards to how data quality is measured in their organisation.

4.3. Perceived Benefits of a Data Warehouse

Our survey showed that 86% of the respondents feel that the data warehouse would bring about improved customer service, reduced risks. A further 99% agreed that ultimately, the data warehouse would help improve strategic decision making. As most of the data warehouses were customer-oriented, another major benefit was the increased opportunity between the organisation and the customer (71%). However, there were mixed reactions about whether the data warehouse would bring about IT system maturity (standard deviation = 1.21). Our results showed that although most data administrators agree that the data warehouse would bring about reduced costs in the long run, the set-up costs for this new technology is relatively steep. Further, for an efficient data warehouse development teams, along with data administrators, have to understand the organisations business rules and requirements.

5. Conclusion

This research found that some of the most common ways in which data gets polluted in a data warehouse are as follows:

- Data never fully captured.
- Heterogeneous system integration.
- Lack of policy and planning from management.

In addition, most organisations feel that in order to maintain data quality in a data warehouse, the development team has to understand the organisation’s requirements, hence their business rules. Validation and audit checks should also go hand in hand with end-user feedback to minimise percentage of data pollution in the data warehouse.

We believe our research has highlighted some key issues affecting the quality of data in a data warehousing environment and that it would benefit the practitioners, viz. the data administrators, to keep in mind some of the factors when planning to set up a data warehouse. To the academics we hope that our research has sparked some ideas for future research in this field of study.

A concurrent research we are undertaking is how data warehouse end-users perceive the state of data quality in the database. Future research could be done in this area as a further development on data inconsistencies. For example, the same study could be carried out in five years to find out if the organisations had done anything to
improve on their current situation or whether their data quality problems have become worse.

6. References


Kelly, S., 1997, Data Warehousing in Action, John Wiley and Sons Ltd.


Porter, J.D. and Rome, J.J., 1994, *The Data Warehouse: 2 Years Later...Lessons Learned*, Arizona State University, 1994 CAUSE Annual Conference held in Orlando, FL.


