Issues of Data Quality in Data Warehouses

Jyoti Sheoran
Research Scholar, Deptt. Of CSE
RKGITW, Ghaziabad

Abstract
In recent years, corporate scandals, regulatory changes, and the collapse of major financial institutions have brought much warranted attention to the quality of enterprise information. If we understand the underlying sources of quality issues, then we can develop a plan of action to address the problem that is both proactive and strategic. The relationship between data quality and data consistency is investigated in this research paper. Poor data quality is costly as it adds expenses and lowers user satisfaction so Data Profiling could be enforced. It was observed that quality of data in a data warehouse is affected by factors like: data not fully captured, lack of planning, data aging.

Keywords
Aging, data consistency, data profiling, data quality, data warehouse.

1. INTRODUCTION

1.1 Data warehousing

Concept of data warehousing has been in existence since a long time. (Shah and Milstein, 1997) its drivers, the informational needs of decision makers, have always been there. William H. Inmon defined Data warehouse as a collection of subject-oriented, integrated, non volatile and time variant databases in which each unit of data is specific to a certain period of time. The data warehouses can contain detailed data, lightly summarized data and highly summarized data, all of which are formatted for analysis and decision support (Inmon, 1995). Clearly distinct from an operational database, a data warehouse is managed data that is situated after and outside the operational systems (Gupta, 1997). Here the data is stored and organised for informational and analytical processing over a long historical period (Inmon, 1995). Since the data warehouse contains historical data to provide for a time-related view (for e.g. trend analysis), it is intended, mainly for analytical applications such as executive information systems and decision support systems (Gallagher, 1995). Figure 1 shows an example of a data warehouse.

Research has shown that the implementation of data warehouses in various organisations is on an increase. Data warehouse allows both simple as well as complex queries to the database (Lias, 1996). It provides different tools and procedures that allow users to manipulate the data to get only the information they need in the most useful form to support decision making.

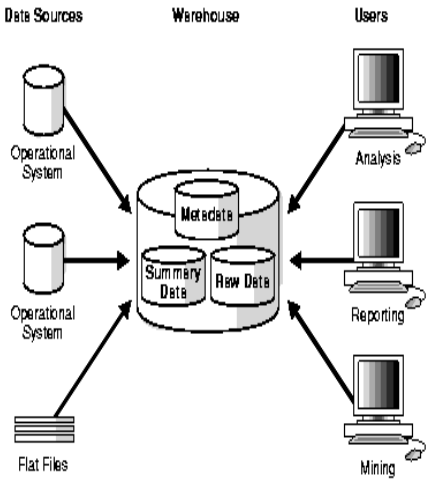

Figure 1. Data Warehouse

Lately, there has been a rise in the expectations about what a data warehouse can do for the organisations (Marshall, 1996). This is important as the end users are lured into believing that a data warehouse could bring immense benefits to the organisation.

This research study basically aims to provide a real-world analysis of data warehousing and to examine how some large organisations plan to achieve information quality in a data warehousing environment. Moreover, quality issues and organisation's role in achieving data quality may even persuade the whole organisation to participate and take part in the development of a data warehouse.

1.2 Data quality

The existence of data all alone cannot ensure that all the management functions and decisions can be undertaken smoothly. Data quality refers as to how relevant, precise, useful, in context, understandable and timely the data is (Firth, 1997; Miller, 1992). A broader definition is that data quality is achieved when an organisation uses data that is comprehensive, understandable, timely, relevant and consistent.

The first step to data quality improvement is the understanding of the key data quality dimension. Data has to satisfy a set of quality criteria to be process able and interpretable in an effective and efficient manner. The most important dimensions emerging from the research were information accuracy, completeness, output timeliness, reliability, relevance, precision and understandability.

In current data warehousing environment one of the fundamental obstacles concerns the existence of inconsistent data. Data inconsistencies mainly occur when similar entities appear in multiple systems and there exist multiple records of the same entities (Kelly, 1997). Thus to minimise integrity errors and also to improve information quality, there must be

quality control. To accommodate rapidly changing business needs organizations also require high quality information. A data quality program that utilizes proven data quality techniques can greatly improve strategic value and success of data warehouse. The quality control demonstrates the following data quality aspects (Clements, 1990):

- Determines the performance of a product.
- May include aesthetics.
- Ensures that there are features beyond the primary function of the product.
- Ensures durability.
- Ensures maintainability.
- Influences perceived quality.
- Ensures reliability.
- Ensures conformance to standards.
-

It has been observed that often, many end-users, including the managers are unaware of the quality of the data that they use in a data warehouse (Lambert, 1996). Poor data quality can often lead to many foreseeable setbacks (which could be-ineffective planning of business strategies or economic failure). Due to inaccurate data, organisations with the management plans such as Just-In-Time manufacturing approach would not be able to function properly (Lambert, 1996).

Thus, issue of data quality in a data warehouses is of great importance. Its success majorly depends on two important processes namely- data quality improvement and data cleaning. Organisations are presently focusing on various implementation issues, the most crucial one being data integrity. Data integrity in a data warehouse is vital to data warehousing success as all the decision support, marketing, data mining, service and business decisions are dependent on it.

2. RESEARCH METHODOLOGY

2.1 Technique

The study is basically designed as a literature review of materials that were published between 1992 and 2010 on topics of data warehouses and data quality. To carry out this work the data warehousing literature, the IT implementations infrastructures related to data quality were reviewed to recognise the various reasons of data duality problems. The classification of major causes of data quality problems so formed will be then divided into the various factors that are responsible for data quality degradation.

2.2 Literature reviewed

John Hess (1998) this report has highlighted the importance of handling the missing values in the data sources, it specially emphasized on the missing dimension of attribute values.

Amit Rudra and Emilie Yeo (1999) this paper concluded that the quality of data warehouse could be largely influenced by factors such as: data not fully captured, Lack of planning from the management and heterogeneous system integration.

Scott W. Ambler (2001) this article explored the wide variety of problems accompanied by the legacy data as, data design, data quality, data architecture and certain process related issues. The article has briefly bifurcated some of the common issues of legacy data which contribute to problems data quality.

Wayne Eckerson (2004) the report says data warehousing projects gloss all the important step of scrutinizing source data before designing data models and ETL mappings. The paper presented some important reasons for the problems of data quality. These were:

1. Manual profiling.
2. Discovering data errors too late.
3. Lack of selection of automated profiling tools.
4. Unreliable Meta data.

2.3 Classification of quality issues of data

A major cause of data warehousing and business intelligence project failure is to attain the wrong or poor quality data. Different sources of data have different types of problems associated with it like data acquired from legacy data sources do not even have the metadata to describe them. A source that offers any kind of unsecured access can become unreliable and ultimately contributes to poor data quality. The characteristics that are responsible for poor data quality can be described as follows:

1. Entry quality: Relates to whether the information enters the system correctly at the point of origin.

2. Process quality: Was the integrity of the information maintained during processing through the system?

3. Identification quality: Are two similar objects identified correctly to be the same or different?

4. Integration quality: Is all the known information about an object integrated to the point of providing an accurate representation of the object?

5. Usage quality: Is the information used and interpreted correctly at the point of access?

6. Aging quality: Has enough time passed that the validity of the information can no longer be trusted?

7. Organizational quality: Can the same information be reconciled between two systems based on the way the organization constructs and views the data?
A plan of action must account for each of these sources of error. Each case differs in its ease of detection and ease of correction. An examination of each of these sources reveals a varying amount of costs associated with each and inconsistent amounts of difficulty to address the problem.

8. Entry Quality: Entry quality is probably the easiest problem to identify but is often the most difficult to correct. Entry issues are usually caused by a person entering data into a system. The problem may be a typo or a wilful decision, such as providing a dummy phone number or address. Identifying these outliers or missing data is easily accomplished with profiling tools or simple queries.

3. CONCLUSION

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

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

In addition to it, most of the organisations feel that in order to maintain data quality in a data warehouse, the development team must understand the organisations requirement, hence their business rules. Validation and audit should go hand in hand for end-user feedback to minimise the percentage of data pollution in the data warehouse.

Hopefully this research has highlighted some of the key issues affecting data quality in a data warehousing environment and that it would also benefit the practitioners viz. The data administrators, in order to keep in mind some of the factors while planning to set up a data warehouse.

4. Future Work

As a further development, research could be done in this area on data inconsistencies. For example, the same study could be carried out in next five years so as to find out if the organisations have done anything in order to improve their current situation or whether their problems of data quality have further become worse.

5. References

[1] Scott W. Ambler (2001) Challenges with legacy data: knowing your data enemy is the first step in overcoming it, practice Leader, Agile Development, Rational Methods Group, IBM,01Jul 2001

[2] IJCSI International Journal of Computer Science issues, Vol 7,Issue 3, No 2, May 2010

[3] John Hess (1998), Dealing With Missing Values in the Data Warehouse, a report of Stonebridge Technologies, Inc(1998)

[4] Atre,S., 1997, Achieving Unity of data, Computerworld, Sep 15, 1997.vol. 31 n37, pp.79(2)

[5] Clerar Targets Vital for Data Warehousing, 1996, Insurance Systems Bulletin, Oct 1996, Vol 12(4), pp.6

[6] Clements, R.B., 1990, Creating and Assuring Quality, ASQC Quality press

[7] David Loshin, "The data quality business case: projecting return on investment", Information White paper. Available at: http://www.melissadata.com/enews/articles/1007/2.htm

[8] Won Kim et al (2002)- "A Taxonomy of Dirty data" Kluwer Academic Publishers 2002

[9] Matteo Golfarelli (2009) Survey on Temporal Data Warehousing, International Journal of data warehousing and mining

[10] Wrembel, R.& Mendelzon, A. (2001).Metadata Management in a Multiversion Data Warehouse. Journal of Data Semantics,8, 118-157

[11] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992

[12] Kimball, R. The data Warehouse Toolkit. John Wiley, 1996

[13] Wuisdom, J."Research Problems in Data Warehousing" 4th International CIKM Conf. 1995

[14] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In OSDI, 2004

[15] Gupta, A., & Mumick, I.S. (1995). Maintenance of materialized views: problems, techniques and applications. Data Engineering Bulletin, 18(2), 3-18