DATA QUALITY PROBLEMS IN DATA WAREHOUSING

Diksha Verma, Anjli Tyagi, Deepak Sharma
Department of Information Technology, Dronacharya college of Engineering

Abstract- **Data quality is critical to data warehouse and business intelligence solutions. Better informed, more reliable decisions come from using the right data quality technology during the process of loading a data warehouse. It is important the data is accurate, complete, and consistent across data sources. Data warehousing is gaining in eminence as organizations become awake of the benefits of decision oriented and business intelligence oriented data basesOver the period of time many researchers have contributed to the data quality issues, but no research has collectively gathered all the causes of data quality problems at all the phases of data warehousing Viz. 1) data sources, 2) data integration & data profiling, 3) Data staging and ETL, 4) data warehouse modeling & schema design. The state-of-the-art purpose of the paper is to identify the reasons for data deficiencies, non-availability or reach ability problems at all the aforementioned stages of data warehousing and to formulate descriptive classification of these causes. We have identified possible set of causes of data quality issues from the extensive literature review and with consultation of the data warehouse practitioners working in renowned IT giants on India. We hope this will help developers & Implementers of warehouse to examine and analyze these issues before moving ahead for data integration and data warehouse solutions for quality decision oriented and business intelligence oriented applications.**

I. INTRODUCTION

Data cleaning, also called *data cleansing* or *scrubbing*, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

Implementing a data warehouse infrastructure to support business intelligence (BI) can be a daunting challenge, as highlighted in recent years by the focus on how often data warehouse implementations fail. Data integration technologies such Informatica PowerCenter has gone a long way towards enabling organizations to successfully deliver their data warehouse projects under budget and with greater business adoption rates. Despite this some data warehouse implementation still fail to meet expectations because of a lack of attention to data quality.

Moving data from its various sources into an easily accessible repository is only part of the challenge in delivering a data warehouse and BI. Without paying attention to the accuracy, consistency and timeliness of the data as part of the data integration lifecycle, BI quickly leads to poor decision-making, increased cost and lost opportunities.

According to industry analyst fi rm Gartner, more than 50 percent of business intelligence and customer relationship management deployments will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues. The impact of poor data quality is far reaching and its affects are both tangible and intangible. If data quality problems are allowed to persist, your executives grow to mistrust the information in the data warehouse and will be reluctant to use it for decision-making.

7 Sources of Poor Data Quality

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. We have seen the rise and assimilation of tools and methodologies that promise to make data cleaner and more complete. Best practices have been developed and discussed in print and online. Data quality is no longer the domain of just the data warehouse. It is accepted as an enterprise responsibility. If we have the tools, experiences, and best practices, why, then, do we continue to struggle with the problem of data quality?

The answer lies in the difficulty of truly understanding what quality data is and in quantifying the cost of bad data. It isn't always understood why or how to correct this problem because poor data quality presents itself in so many ways. We plug one hole in our system, only to find more problems elsewhere. If we can better 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.

Each instance of a quality issue presents challenges in both identifying where problems exist and in quantifying the extent of the problems. Quantifying the issues is important in order to determine where our efforts should be focused first. A large number of missing email addresses may well be alarming but could present little impact if there is no process or plan for communicating by email. It is imperative to understand the business requirements and to match them against the assessment of the problem at hand. Consider the following seven sources of data quality issues.

1. Entry quality: Did the information enter the system correctly at the 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.

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 willful 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.

The cost of entry problems depends on the use. If a phone number or email address is used only for informational purposes, then the cost of its absence is probably low. If instead, a phone number is used for marketing and driving new sales, then opportunity cost may be significant over a major percentage of records.

Addressing data quality at the source can be difficult. If data was sourced from a third party, there is usually little the organization can do. Likewise, applications that provide internal sources of data might be old and too expensive to modify. And there are few incentives for the clerks at the point of entry to obtain, verify, and enter every data point.

Process Quality

Process quality issues usually occur systematically as data is moved through an organization. They may result from a system crash, lost file, or any other technical occurrence that results from integrated systems. These issues are often difficult to identify, especially if the data has made a number of transformations on the way to its destination. Process quality can usually be remedied easily once the source of the problem is identified. Proper checks and quality control at each touchpoint along the path can help ensure that problems are rooted out, but these checks are often absent in legacy processes.

Identification Quality

Identification quality problems result from a failure to recognize the relationship between two objects. For example, two similar products with different SKUs are incorrectly judged to be the same.

Identification quality may have significant associated costs, such as mailing the same household more than once. Data quality processes can largely eliminate this problem by matching records, identifying duplicates and placing a confidence score on the similarity of

records. Ambiguously scored records can be reviewed and judged by a data steward. Still, the results are never perfect, and determining the proper business rules for matching can involve trial and error.

Integration Quality

Integration quality, or quality of completeness, can present big challenges for large organizations. Integration quality problems occur because information is isolated by system or departmental boundaries. It might be important for an auto claims adjuster to know that a customer is also a high-value life insurance customer, but if the auto and life insurance systems are not integrated, that information will not be available.

While the desire to have integrated information may seem obvious, the reality is that it is not always apparent. Business users who are accustomed to working with one set of data may not be aware that other data exists or may not understand its value. Data governance programs that document and promote enterprise data can facilitate the development of data warehousing and master data management systems to address integration issues.
MDM enables the process of identifying records from multiple systems that refer to the same entity. The records are then consolidated into a single master record. The data warehouse allows the transactional details related to that entity to be consolidated so that its behaviors and relationships across systems can be assessed and analyzed.

Usage Quality

Usage quality often presents itself when data warehouse developers lack access to legacy source documentation or subject matter experts. Without adequate guidance, they are left to guess the meaning and use of certain data elements. Another scenario occurs in organizations where users are given the tools to write their own queries or create their own reports. Incorrect usage may be difficult to detect and quantify in cost.

Thorough documentation, robust metadata, and user training are helpful and should be built into any new initiative, but gaining support for a post-implementation metadata project can be difficult. Again, this is where a data governance program should be established and a grassroots effort made to identify and document corporate systems and data definitions. This metadata can be injected into systems and processes as it becomes part of the culture to do so. This may be more effective and realistic than a big-bang approach to metadata.

Aging Quality

The most challenging aspect of aging quality is determining at which point the information is no longer valid. Usually, such decisions are somewhat arbitrary and vary by usage. For example, maintaining a former customer's address for more than five years is probably not useful. If customers haven't been heard from in several years despite marketing efforts, how can we be certain they still live at the same address? At the same time, maintaining customer address information for a homeowner's insurance claim may be necessary and even required by law. Such decisions need to be made by the business owners and the rules should be architected into the solution. Many MDM tools provide a platform for implementing survivorship and aging rules.

Organizational Quality

Organizational quality, like entry quality, is easy to diagnose and sometimes very difficult to address. It shares much in common with process quality and integration quality but is less a technical problem than a systematic one that occurs in large organizations. Organizational issues occur when, for example, marketing tries to "tie" their calculations to finance. Financial reporting systems generally take an account view of information, which may be very different than how the company markets the product or tracks its customers. These business rules may be buried in many layers of code throughout multiple systems. However, the biggest challenge to reconciliation is getting the various departments to agree that their A equals the other's B equals the other's C plus D.

A Strategic Approach

The first step to developing a data strategy is to identify where quality problems exist. These issues are not always apparent, and it is important to develop methods for detection. A thorough approach requires inventorying the system, documenting the business and technical rules that affect data quality, and conducting data profiling and scoring activities that give us insight in the extent of the issues.

After identifying the problem, it is important to assess the business impact and cost to the organization. The downstream effects are not always easy to quantify, especially when it is difficult to detect an issue in the first place. In addition, the cost associated with a

particular issue may be small at a departmental level but much greater when viewed across the entire enterprise. The business impact will drive business involvement and investment in the effort.

Finally, once we understand the issues and their impact on the organization, we can develop a plan of action. Data quality programs are multifaceted. A single tool or project is not the answer. Addressing data quality requires changes in the way we conduct our business and in our technology framework. It requires organizational commitment and long-term vision.

The strategy for addressing data quality issues requires a blend of analysis, technology, and business involvement. When viewed from this perspective, an MDM program is an effective approach. MDM provides the framework for identifying quality problems, cleaning the data, and synchronizing it between systems. However, MDM by itself won't resolve all data quality issues.

An active data governance program empowered by chief executives is essential to making the organizational changes necessary to achieve success. The data governance council should set the standards for quality and ensure that the right systems are in place for measurement. In addition, the company should establish incentives for both users and system developers to maintain the standards.

The end result is an organization where attention to quality and excellence permeate the company. Such an approach to enterprise information quality takes dedication and requires a shift in the organization's mindset. However, the results are both achievable and profitable.

Data Warehousing

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. As defined by the "father of data warehouse", William H. Inmon, a data warehouse is "a collection of Integrated, Subject-Oriented, Non Volatile and Time Variant databases where each unit of data is specific to some period of time. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support" (Inmon, 1996). In the "Data Warehouse Toolkit", Ralph Kimball gives a more concise definition: "a copy of transaction data specifically structured for query and analysis" (Kimball, 1998). Both definitions stress the data warehouse's analysis focus, and highlight the historical
nature of the data found in a data warehouse.

Figure 2: Data Warehousing Structure

Stages of Data Warehousing Susceptible to Data Quality Problems

The purpose of paper here is to formulate a descriptive taxonomy of all the issues at all the stages of Data Warehousing. The phases are:

- Data Source
- Data Integration and Data Profiling
- Data Staging and ETL
- Database Scheme (Modeling)

Quality of data can be compromised depending upon how data is received, entered, integrated, maintained, processed (Extracted, Transformed and Cleansed) and loaded. Data is impacted by numerous processes that bring data into your data environment, most of which affect its quality to some extent. All these phases of data warehousing are responsible for data quality in the data warehouse. Despite all the efforts, there still exists a certain percentage of dirty data. This residual dirty data should be reported, stating the reasons for the failure in data cleansing for the same.

Data quality problems can occur in many different ways [9]. The most common include:

- Poor data handling procedures and processes.
- Failure to stick on to data entry and maintenance procedures.
- Errors in the migration process from one system to another.
- External and third-party data that may not fit with your company data standards or may otherwise be of unconvinced quality. The assumptions undertaken are that data quality issues can arise at any stage of data warehousing viz. in data sources, in data integration & profiling, in data staging, in ETL and database modeling. Following model is depicting the possible stages which are vulnerable of getting data quality problems

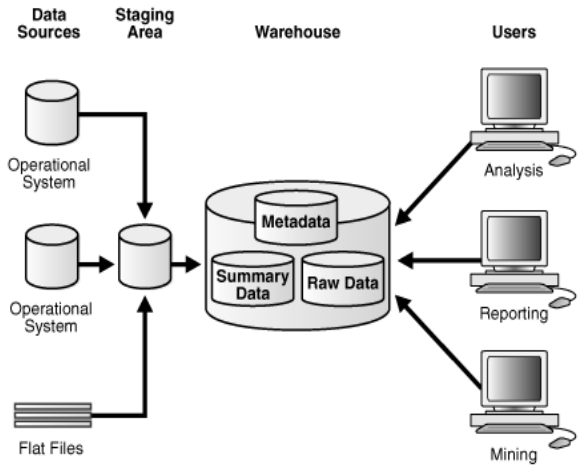

Figure 1: Data Warehousing Structure

II. CONCULSION

Data quality is an increasingly serious issue for organizations large and small. It is central to all data integration initiatives. Before data can be used effectively in a data warehouse, or in customer relationship management, enterprise resource planning or business analytics applications, it needs to be analyzed and cleansed. To ensure high quality data is sustained, organizations need to apply ongoing data cleansing processes and procedures, and to monitor and track data quality levels over time. Otherwise poor data quality will lead to increased costs, breakdowns in the supply chain and inferior customer relationship management. Defective data also hampers business decision making and efforts to meet regulatory compliance responsibilities. The key to successfully addressing data quality is to get business professionals centrally involved in the process. Informatica Data Explorer and Informatica Data Quality are unique, easy-to-use data quality software products specifi cally designed to bridge the gap between and better align IT and the business, providing them with all they need to be able to control data quality processes enterprise-wide in order to reduce costs, increase revenue, and improve business decision-making.

REFERENCES

1 A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing by Ranjit Singh, Dr. Kawaljeet Sing, Research Scholar, University College of Engineering (UCoE), Punjabi University Patiala (Punjab), INDIA

2. 7 Sources of Poor Data Quality, https://www.melissadata.com/enews/articles/0611/2.htm
3. Data Quality in Data Warehouse & Business Intelligence Informatica Data Quality and Informatica Data Explorer
4. Channah F. Naiman, Aris M. Ouksel (1995) "AClassification of Semantic Conflicts in Heterogeneous Database Systems", Journal of Organizational Computing, Vol. 5, 1995
5. John Hess (1998), "Dealing With Missing Values In The Data Warehouse" A Report of Stonebridge Technologies, Inc (1998).
6. Jaideep Srivastava, Ping-Yao Chen (1999) "Warehouse Creation-A Potential Roadblock to Data Warehousing", IEEE Transactions on Knowledge and Data Engineering January/February 1999 (Vol. 11, No. 1) pp.
7. Amit Rudra and Emilie Yeo (1999) "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999
8. Jesús Bisbal et all (1999) "Legacy Information Systems: Issues and Directions", *IEEE Software September/ October 1999*
9. 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, 01 Jul 2001.
10. Scott W. Ambler "The Joy of Legacy Data" available at: http://www.agiledata.org/essays/legacyDatabases2.html