An Important Data Quality Tools for Data Warehouse: Case Study

Rajesh Singh
Asst. Prof., Dept. of CSE, B.S.A.College of Engg. & Technology, Mathura, U.P.

Abstract: Ensuring Data Quality for an institute or group data repository various data quality tools are used that focus on this issue. The scope of these tools is moving from specific applications to a more worldwide point of view so as to make sure data quality at every level. A more organized structure is needed to assist managers to decide these tools so that that the data repositories or data warehouses may be maintained in a very competent way. Data quality tools are used in data warehousing to ready the data and ensure that clean data populates the warehouse, therefore enhancing usability of the warehouse. This research paper focuses on the various data quality tools which have been used and implemented successfully in the preparation of different types of data.

Keywords- Data Quality, Decision Making, Data Warehouse, Business Value

1.1 Introduction

As we know that the data warehouse is a subject oriented, integrated, time variant and non volatile collection of data in support of management's decision making process.

It is a central repository for all or significant parts of the data that an enterprise's various business systems collect.

The data stored in the warehouse is uploaded from the operational systems such as marketing, sales, etc. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

In other words Data quality refers to the level of quality of Data. There are many definitions of data quality but data is generally considered high quality if, "they are fit for their intended uses in operations, decision making and planning according to J. M. Juran [5]. Alternatively data is deemed of high quality if it correctly represents the real world construct to which it refers. Furthermore, apart from these definitions, as data volume increases the question of internal consistency within data becomes significant, regardless of fitness for use for any particular external purpose. The people's views on data quality can often be in disagreement even when discussing the same set of data used for the same purpose.

1.2 Data Quality

Data quality is an essential characteristic that determines the reliability of data for making decisions. High quality data should be as follows also represented in Fig-1

• Complete: It means that all relevant data such as accounts, addresses and relationships for a given customer is linked should be complete.

• Accurate: Means those common data problems like misspellings, typos, and random abbreviations have been cleaned up. If Data is accurate then mining operation can be applied easily to extract the reliable information.

• Available: This phenomenon describes that required data is accessible on demand; users do not need to search manually for the information. Data should be available 24 by 7 for users.

• Timely: Up to date information is readily available to support decisions. As we also know that Data Warehouse keeps Timely data by which relevant information may be extracted

• Validity: Data resides in Data Warehouse should have some validity. It should be resourced from some valid platform.

• Reliability: Data resides in Data Warehouse should be reliable, means it should be available each and every time whenever required for decision making.

• Integrity: Data resides in Data Warehouse is in the form of heterogeneous, multidimensional, so that should be integrated with each other for smooth functioning of different operations.

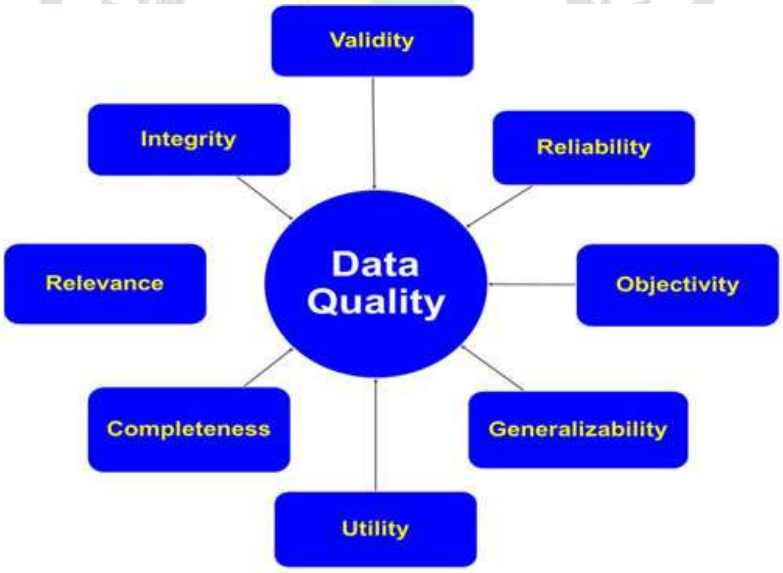

Fig-1- Different Data qualities

1.3 Business value of data quality

1. Data quality associated problems cost companies millions of dollars annually because of lost revenue opportunities, failure to meet regulatory compliance or failure to address customer issues in a timely manner. Poor data quality is often cited as a reason for failure of critical information intensive projects. By implementing a data quality program, organizations can:
2. Deliver high quality data for a range of enterprise initiatives including business intelligence, applications consolidation and retirement, and master data management
3. Reduce time and cost to implement CRM, data warehouse/BI, data governance, and other strategic IT initiatives and maximize the return on investments
4. Construct consolidated customer and household views, enabling more effective cross selling, up selling, and customer retention
5. Help improve customer service and identify a company's most profitable customers
6. Provide business intelligence on individuals and organizations for research, fraud detection, and planning
7. Reduce the time required for data cleansing saving on average 5 million hours, for an average company with 6.2 million records

2.1History

Before the rise of the economical computer data storage, massive mainframe computers were used to sustain name and address data for delivery services. This was so that mail could be properly routed to its goal. The mainframes used business rules to correct common misspellings and typographical errors in name and address data, as well as to track customers who had moved, died, gone to prison, married, divorced, or experienced other life-changing events. Government agencies began to make postal data available to a few service companies to cross reference customer data with the National Change of Address registry (NCOA). This technology saved large companies millions of dollars in comparison to manually correction of customer data. Large companies saved on postage as bills and direct marketing materials made their way to the intended customer more accurately. Initially sold as a service, data quality moved inside the walls of corporations, as low-cost and powerful server technology became available.

2.2 Overview

There are a number of theoretical frameworks for understanding data quality. A systems theoretical approach influenced by American pragmatism expands the definition of data quality to include information quality, and emphasizes the inclusiveness of the fundamental dimensions of accuracy and precision on the basis of the theory of science. One framework, dubbed "Zero Defect Data" adapts the principles of statistical process control to data quality. Another framework seeks to integrate the product viewpoint and the service perspective .Another framework is based in semiotics to evaluate the quality of the form, meaning and use of the data Price and Shanks, 2004. One highly theoretical approach analyzes the ontological nature of information systems to define data quality rigorously Wand and Wang, 1996.

3.1Data Quality Assurance

Data quality assurance is the process of profiling the data to discover inconsistencies and other anomalies in the data as well as performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality.

These activities can be undertaken as part of data warehousing or as part of the database administration of an existing piece of applications software.

3.2 Data quality control is the process of controlling the usage of data with known quality measurements for an application or a process. This process is usually done after a Data Quality Assurance (QA) process, which consists of discovery of data inconsistency and correction.

Data QA processes provides following information to Data Quality Control (QC):

- Severity of inconsistency
- Incompleteness
- Accuracy
- Precision
- Missing / Unknown

The Data QC process uses the information from the QA process to decide to use the data for analysis or in an application or business process. For example, if a Data QC process finds that the data contains too many errors or inconsistencies, then it prevents that data from being used for its intended process which could cause disruption. For example, providing invalid measurements from several sensors to the automatic pilot feature on an aircraft could cause it to crash. Thus, establishing data QC process provides the protection of usage of data control and establishes safe information usage

3.3 Use of Data Quality in Data Warehouse

Data Quality (DQ) is a niche area required for the integrity of the data management by covering gaps of data issues. This is one of the key functions that aid data governance by monitoring data to find exceptions undiscovered by current data management operations. Data Quality checks may be defined at attribute level to have full control on its remediation steps.

DQ checks and business rules may easily overlap if an organization is not attentive of its DQ scope. Business teams should understand the DQ scope thoroughly in order to avoid overlap. Data quality checks are redundant if business logic covers the same functionality and fulfills the same purpose as DQ. The DQ scope of an organization should be defined in DQ strategy and well implemented. Some data quality checks may be translated into business rules after repeated instances of exceptions in the past.

Below are a few areas of data flows that may need perennial DQ checks: Completeness and precision DQ checks on all data may be performed at the point of entry for each mandatory attribute from each source system. Few attribute values are created way after the initial creation of the transaction; in such cases, administering these checks becomes tricky and should be done immediately after the defined event of that attribute's source and the transaction's other core attribute conditions are met.

All data having attributes referring to Reference Data in the organization may be validated against the set of well-defined valid values of Reference Data to discover new or discrepant values through the validity DQ check. Results may be used to update Reference Data administered under Master Data Management (MDM).

All data sourced from a third party to organization's internal teams may undergo accuracy (DQ) check against the third party data. These DQ check results are valuable when administered on data that made multiple hops after the point of entry of that data but before that data becomes authorized or stored for enterprise intelligence.

All data columns that refer to Master Data may be validated for its consistency check. A DQ check administered on the data at the point of entry discovers new data for the MDM process, but a DQ check administered after the point of entry discovers the failure of consistency.

As data transforms, multiple timestamps and the positions of that timestamps are captured and may be compared against each other and its leeway to validate its value, decay, operational significance against a defined SLA means service level agreement. This timeliness DQ check can be utilized to decrease data value decay rate and optimize the policies of data movement timeline.

3.4 Data Quality Tools

The market for data quality tools has become highly visible in recent years as more organizations understand the impact of poor quality data and seek solutions for improvement. Traditionally aligned with cleansing of customer data in support of CRM related activities, the tools have expanded well beyond such capabilities, and forward thinking organizations are recognizing the relevance of these tools in other data domains. Product data often driven by MDM initiatives and financial data are two such areas in which demand for the tools is quickly building.

Data quality tools are used to address various aspects of the data quality problem:

1.Parsing and standardization - Decomposition of text fields into component parts and formatting of values into consistent layouts based on industry standards, local standards for example, postal authority standards for address data, user defined business rules, and knowledge bases of values and patterns

2. Generalized “cleansing” -Modification of data values to meet domain restrictions, integrity constraints or other business rules that define sufficient data quality for the organization

3. Matching -Identification, linking or merging related entries within or across sets of data

4. Profiling - Analysis of data to capture statistics or metadata that provide insight into the quality of the data and aid in the identification of data quality issues

5. Monitoring -Deployment of controls to ensure ongoing conformance of data to business rules that define data quality for the organization

6. Enrichment- Enhancing the value of internally held data by appending related attributes from external sources for example, consumer demographic attributes or geographic descriptors.

3.5 Data Quality Assessment Methodology

A data quality assessment methodology is defined as the process of evaluating if a piece of data meets the information consumers need in a specific use case [1]. In a comprehensive survey [2], it was observed that in the 30 identified approaches; there were no standardized set of steps that were followed to assess the quality of a dataset. Inspired from the methodology proposed in [3] and the lack of a standardized methodology in LD, we propose a methodology consisting of three phases and nine steps. In particular, from each of the 30 approaches, we extracted the common steps that were proposed to assess the quality of a dataset. We then adapted and revised these steps to propose a data quality assessment methodology for LD as depicted in Figure- 2

Our methodology thus consists of the following phases and steps:

1. Phase I: Requirements Analysis

(a) Step I: Use Case Analysis

2. Phase II: Quality Assessment

(a)Step II: Identification of quality issues

(b)Step III: Statistical and Low level Analysis

(c)Step IV: Advanced Analysis

3. Phase III: Quality Improvement

(a)Step V: Root Cause Analysis

(b)Step VI: Fixing Quality Problem

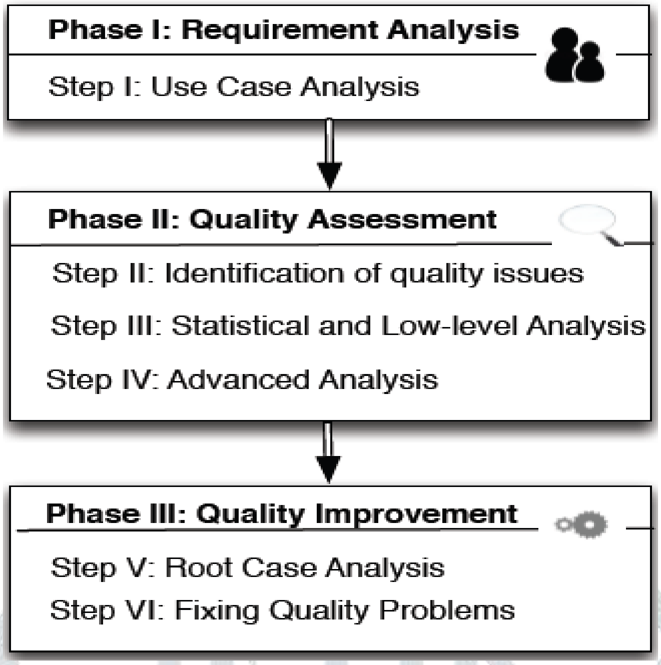
Phase I: Requirement Analysis
Step I: Use Case Analysis
Phase II: Quality Assessment
Step II: Identification of quality issues
Step III: Statistical and Low-level Analysis
Step IV: Advanced Analysis
Phase III: Quality Improvement
Step V: Root Case Analysis
Step VI: Fixing Quality Problems

Figure 2- Steps in data quality assessment [4]

This section describe each of the steps in detail along with the list of data quality dimensions from the 18 dimensions identified in [2] that are applicable for each step.

1 Phase I: Requirements analysis

The multi dimensional nature of data quality makes it dependent on a number of factors that can be determined by analyzing the user's requirements. Thus, the use case in question is highly important when assessing the quality of a dataset. This requirement analysis phase thus includes the gathering of requirements and subsequent analysis of the requirements based on the use case

2 Phases II: Data Quality Assessment

In the previous phase, we identified the user requirements for her dataset with the particular use case she has in mind. This second phase involves the actual quality assessment based on the requirements. In particular, amongst the set of dimensions and metrics discussed in [2], the most relevant ones are selected. Thereafter, a quantitative evaluation of the quality of the dataset is performed using the metrics specific for each selected dimension. Thus, this phase consists of three steps: (II) Identification of quality issues (III) Statistical and Low-level analysis and (IV) Advanced analysis

3 Phases III: Quality Improvement

This phase focuses towards improving the quality of the datasets based on the analysis performed in Phase II focusing on the use case identified in Phase I. This phase consists of two steps: Root Cause Analysis and Fixing Quality Problems

4. Conclusion

In this paper, we have introduced a data quality assessment methodology consisting of three phases and six steps. This methodology is generic enough to be applied to any use case.

In order to validate its usability, we plan to apply it to specific use cases to assess the feasibility and effectiveness of the methodology. This validity will also help us measure its applicability in various domains. Moreover, we plan to build a tool based on this methodology so as to assist users to assess the quality of any linked dataset.

5. References

1.Data Warehouse Implementation of Examination Databases” MA Butt, SMK Quadri, M Zaman, International Journal of Computer Applications 44 (5), 18-23, 2012

2.A.Zaveri,A.Rula,A.Maurino,R.Pietrobon,J.Lehmann,andS.Auer.QualityAssessment Methodologies for Linked Data: A Survey. Under review, available at http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-pen-data.

3. C. Batini and M. Scannapieco Data Quality: Concepts, Methodologies and Techniques
(Data-Centric Systems and Applications)Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

4. Anisa Rula, Amrapali Zaveri, Methodology for Assessment of Linked Data Quality, LDQ 2014, 1st Workshop on Linked Data Quality Sept. 2, 2014, Leipzig, Germany

5. "Data Quality: High-impact Strategies - J. M. Juran, What You Need to Know:

6. “Toxic Data”, Haggerty, N., DM Review Magazine, June 199

7. Horowitz, A. (1998). “Ensuring the Integrity of Your Data”, Beyond Computing, May1998

8.“Star Schema Implementation for Automation of Examination Records”, International , Conference on Computer Science, Computer Engineering and Applied Computing Las Vegas, USA, July 16-19, 2012 ISBN 1-60132-050-7O’

9. Neill, P. (1998). “It’s a Dirty Job: Cleaning Data in the Warehouse”, Gartner Group, January 12, 1998

10. Lehmann, R. Cornelissen, and A. Zaveri. Test driven evaluation of linked data quality. In
WWW, pages 747{758, 2014