# Various Data Quality Issues in Data Warehousing

**Sujeet Kumar & Utkarsh Kumar**

Student, Information Technology Engineering, MDU University , Haryana , India

sujeetkumar0110@gmail.com

Student, Information Technology Engineering, MDU University , Haryana , India

utkarsh.kumar10@gmail.com

## ABSTRACT

*In our paper, we will be putting forward some of the complicacies of the data quality problems faced in data warehousing. We will identify some of the reasons for data deficiencies, non-availability or reach ability problems in various stages of data warehousing. 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.*

***KEYWORDS:*** *Data warehouse, data quality, quality parameters*

## 1. INTRODUCTION

Many business organizations, large or small, are implementing data warehouses to collect, store, and process large amount of data. As defined and constructed 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. Many researchers and practitioners believe that a data warehouse architecture can be formally understood as layers of materialized views on top of each other. The success of data warehousing initiatives majorly depends on the quality of the data stored in it. Good quality data ensures user's trust in data warehouse system making it more usable and, optimizes and increases the business benefits gained overall. However, detecting defects and improving data quality comes with a cost and if the targeted quality level is high, the costs often negate (offset) the benefits.

## 2.Quality Parameters and their Metrics

| S.No. | Quality Parameter | Quality Metric |
|---|---|---|
| 1. | Functionality | Number of modules that are not appropriate for the task |
| 2. | Reliability | Number of failures |
| 3. | Usability | Acceptance by users |
| 4. | Efficiency | Performance in terms of response time, processing time, etc |
| 5. | Maintainability | Man hours required to maintain and test the applications |
| 6. | Portability | Number of cases where applications failed to work on new environments |
| 7. | Accessibility | Number of NULL values stored (where they are not expected) |
| 8. | Accuracy | Number of records with accurate values |
| 9. | Consistency | Number of records violating constraints |
| 10. | Security | Number of modules that could not protect the system from unauthorized access |
| 11. | Compliance | Number of modules non-compliant with standards/ conventions/ regulations |
| 12. | Recoverability | Number of times the software was unable to re-establish its level of<br>performance and recover the affected data |
| 13. | Analyzability | Man-hours required for diagnosing defects or failures |
| 14. | Changeability | Man-hours required for removing defects from the system |
| 15. | Testability | Man-hours required for validating the software product |
| 16. | Install ability | Man-hours required for installation of the software in a specified environment |
| 17. | Implementation Efficiency | Number of resources used for the development of software and percentage of used recourses with respect to the originally |

| | | |
|---|---|---|
| | | expected<br>ones |
| 18. | System Availability | Number of cases when relevant information is not available |
| 19. | Currency | Number of pieces of information where transaction time though required was not present |
| 20. | Volatility | Number of pieces of information where valid time though required<br>was not present |
| 21. | Completeness | Number of records with incomplete values |
| 22. | Credibility | Number of records with inaccurate values |
| 23. | Data Interpretability | Number of pieces of information that are not fully described or documented |

## 3. Data 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. Some common quality assurance criteria are

- Completeness
- Consistency
- Validity
- Conformity
- Accuracy
- Integrity

## 4.Data quality control

Data quality control is the process of controlling the usage of data with known amount quality measurement—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. The Data quality control process uses the information from the quality assurance process, then it decides to use the data for analysis or in an application or business process.

## 5. CONCLUSION

Data quality is an important factor in the success of data warehousing projects. Our objective was to put forth such a descriptive classification which covers all the phases of data warehousing which can impact the data quality This paper will be helpful for the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing.

## REFERENCES

[1] McFadden, F. (1996) Data Warehouse for EIS: Some Issues and Impacts, Proc. 29th Annual Hawaii International Conference on System Sciences, Hawaii, (January).

[2] Kimball, R. (1996) The Data Warehousing Toolkit, John Wiley, New York.

[3] Redman, T. (1998) The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM, 41, 79-82.

[4] Wang, Y. and Strong, D.M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 12, 5-34.

[5] Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997) Data Quality in Context, Communications of the ACM, 40:5, 103-110

[6] Singh R., Singh K. A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010.

[7] Wixom, B. H. and Watson, H. J. 2001. An empirical investigation of the factors affecting data warehousing success. MIS Quart. 25, 1, 17–41.

[8] Redman, T. C. 1996. Data Quality for the Information Age. Artech House, Boston, MA.

[9] Shankaranarayanan, G., Ziad, M., and Wang, R. Y. 2003. Managing data quality in dynamic decision making environments: An information product approach. J. Database Manag. 14, 4, 14– 32.

[10] Tayi, G. K. and Ballou, D. P. 1988. An integrated production-inventory model with reprocessing and inspection. Int. J. Prod. Res. 26, 8, 1299–1315

[11] Lee, Y. W., Pipino, L., Strong, D. M., and Wang, R. Y. 2004. Process-embedded data integrity. J. Database Manag. 15, 1, 87–103.

[11] Madnick, S., Wang, R. Y., and Xian, X. 2003. The design and implementation of a corporate house holding knowledge processor to improve data quality. J. Manag. Inform. Syst. 20, 3, 41– 69.

[12] R. Weir, Taoxin Peng, and Kerridge Jon, Best Practice for Implementing a Data warehouse: A Review for Strategic Alignment, DMDW, 2003

[13] Quality Enhancement Provider Handbook, developed jointly by Louisiana Office of Aging and Adult Services and Office for Citizens with Development Disabilities, August 2008. Available at http://new.dhh.louisiana.gov/assets/docs/OCDD/waiver/QEProviderHandbook080108.pdf.

[14] D. H. Besterfield, C. Besterfield-Michna, G. Besterfield and M. Besterfield-Sacre. Total Quality Management. Prentice Hall, 1995.

[15] M. Bouzeghoub, F. Fabret, M. Matulovic, E. Simon. Data Warehouse Refreshment: A Design Perspective from Quality Requirements. Technical Report D8.5, DWQ Consortium (1998). Available at http://www.dbnet.ece.ntua.gr/~dwq/