International Journal Of Mechanical Engineering And Information Technology
An Official Publication Of IGM Publication

Analysis of Data Quality Issues in Big Data Migration Process

Authors
V.Rathika[1], Dr.L.Arockiam[2]
[1]Research scholar, Dept. of Computer Science Mother Teresa Women's University Kodaikanal, Tamil Nadu, India
[2]Associate Professor, Dept. of Computer Science St. Joseph's College Trichy, Tamil Nadu, India.
Email- *rathirajaphd2013@gmail.com larockiam@yahoo.co.in*

ABSTRACT –
Large volume of data is extracted, transferred, structured and loaded in the process of big data migration. It is not a minor task, because the data stored in the legacy systems should be analyzed, measured, preserved and improved before being brought over to the target systems. Every day we meet large volumes of data and store them in multidimensional form that is in data warehouses. After storing the data will be analyzed and used in accurate decision making in the business enterprises. The quality is the key attitude in each and every analysis and also in computing applications. This paper is going to find all possible data quality issues which are related to data warehouses while migration from source to target.
Keywords: *Data Quality, Data Warehousing, Data Quality Issues*

1. INTRODUCTION

Data migration is very important when a company upgrades its database or software from source to target. The general survey explained about the migration status such as 84% of data migration projects fail to meet expectations, 37% experience budget overruns, and 67% are not delivered on time [1]. Data quality is the key factor which affects the migration status. It must be ensured after the migration process in target systems. It has three factors that are the quality of the data itself, the quality of the application programs and the quality of the database schema.

Today decision making is more essential for all businesses. Inaccurate and incomplete reports lead to wrong decision. As well as redundant data handling leads to economic problems [5]. Poor data quality directs maintenance, repair costs. It creates malfunctioning, operational inefficiency, inefficient business analysis, unavailability of timely data. And also it affects customer satisfaction, reputation or even strategic decisions [14]. Implementation of Data Warehouse is exact solution to complex business intelligent applications. But it fails to meet expectations because of lack of data quality. So, the data quality is a severe issue in implementation and management of data warehouses. Before using data it should be analyzed and cleaned.

2. LITERATURE REVIEW

1. Jaiteg Singh and Kawaljeet Singh (2009) – The data quality was monitored earlier than and later than the induction of automated ETL testing [2].
2. Lixian Xing, Yanhong Li (2010) – They discussed technical issues of data migration [3].
3. Priyanka Paygude, P.R.Devale (2013) – They proposed automated data validation testing approach to find data inconsistencies that have to be checked as a part of data validation process [4].
4. Manjunath T.N, Ravindra S Hegadi Ravikumar G.K (2011) – Discussed and investigated possible set of data quality issues from exhaustive survey. They proposed automating data validation test to

ensure quality assurance and risk management in migration process [6].

5. Manjunath T.N, Ravindra S Hegadi Ravikumar G.K (2012) – They proposed a model to check quality for huge database migration using random sampling [7].
6. Manjunath T.N, Ravindra S Hegadi Ravikumar G.K (2013) – They proposed a mathematical model using deterministic statistical methods to reduce resource utilization and to produce greater data quality [8].
7. Shashikant Patel, Sagar Wakchaure, Mahendra Pingale, Saba Siraj (2014) – They proposed a system architecture to conduct heterogeneous database migration with the features of security, time efficient, flexible and interactive [9].
8. Ranjit Singh, Dr.Kawaljeet Singh (2010) – They focused the number of issues in the stages of data warehouse [10].
9. Priyanka Paygude, P.R.Devale (2013) – They proposed automated data validation architecture to assure data quality in migration projects across the enterprise, multiple platforms, and applications [11].
10. Florian Matthes, Christopher Schulz, Klaus Haller (2011) – They discussed detailed migration process and its various risks. And they fused on types testing which leads successful migration [13].
11. Nikhil Debbarma, Gautam Nath, Hillol Das (2013) – They discussed quality and performance issues in data warehouse [14].
12. Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park (2006) – They introduced a Data Quality Management Maturity Model which improves data quality using data management matures [12].
13. Ramesh Babu Palepu, Dr. K.V.Sambasiva Rao (2012) – They proposed a new model of meta data architecture to find a solution to achieve data quality [13].

3. OBJECTIVES

- This paper focuses about data quality and its important factors.
- It gives overview of data warehouse and its stages.
- It summarizes the data quality issues which are related to the stages of data warehouses.
- It motivates to find better solution to improve accuracy which is the vital component of data quality.

4. GENERAL PHASES OF DATA QUALITY

Data should be comprehensive, understandable, consistent, relevant and timely. Data quality is about bad data that is missing or inaccurate or invalid context. Same data will be estimated to varying degrees of quality according to user's needs (see figure 1).

Data Quality has the 9 dimensions such as definition conformance, completeness, validity, accuracy, precision, non-duplication, derivation integrity, accessibility, and timeliness [6][15].

1. Definition conformance – Definition of chosen object must be having complete details and meaning.

2.Completeness – Data field is having characteristic of all required values.

3. Validity – It refers Data values which include domain values, ranges, reasonability tests, primary key uniqueness, and referential integrity.

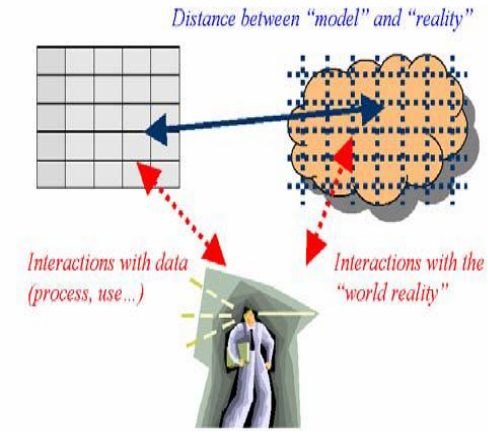

Figure 1-Complexity of Data Quality Objectivity

4. Accuracy – It measures the degree to which data agrees with data contained in an original source.

5. Precision – It refers domain value which specifies business should have correctness as per specifications.

6. Non – Duplication – It refers a one-to-one correlation between records and the real world object or events being represented.

7. Derivation Integrity – It refers the correctness with which two or more pieces of data are combined to create new data.

8. Accessibility – It refers the characteristic of being able to access data on demand.

9. Timeliness – It refers the relative availability of data to support a given process within the timetable required to perform the process.

High quality of data has the benefits such as lower costs, reduced risks, drive efficiency, and improved productivity [8].

5. UNDERSTANDING OF DATA WAREHOUSING

Data warehouse is a collection of technologies which is enabling the knowledge worker to make better and quicker decisions. It is a subject oriented, integrated, time-variant and non-volatile collection of data. It is important strategy to integrate heterogeneous information sources in organization, and to make possible online analytical processing. And it is a copy of transaction data specifically structured for query and analysis. It consists of data sources, staging area, Meta data, highly summarized data and lightly summarized data, raw data, and the details about analysis and reports (see Fig. 2)[10].

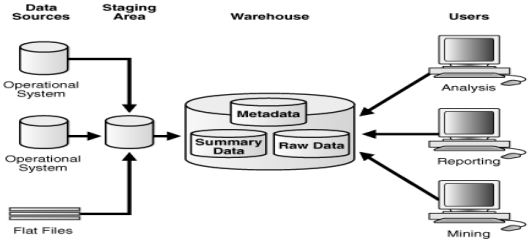

Figure 2- Structure of Data Warehouse

Data Warehouse is having the following stages. All these stages are responsible for data quality problems.

1. Data Sources
2. Data Integration and Data Profiling
3. Data Staging and ETL
4. Database Schema Design

6. FEASIBLE DATA QUALITY ISSUES IN DATA WAREHOUSING

A. Data Source Issues

Different data sources are having different type of problems like legacy data sources are not having metadata, dirty data sources are having data entry error, and part of data comes from text files. The following are the common issues at this stage [6][12].

1. Poor selection of candidate data sources.
2. Poor knowledge of inter dependencies among data sources.
3. Incorrect data formatting.
4. Need of validation routines.
5. Unpredicted changes in source systems.
6. Various sources for same data.
7. Lack of ability to manage with the aged data.
8. Irregular timeliness of data sources.
9. Multiple data sources generate semantic heterogeneity.
10. Missing values in data sources.
11. Unexpected changes in source systems.
12. Deficient of domain level validation in source data
13. Difference between data and Meta data.
14. Unsuited use of special characters.
15. Incorrect data mapping.
16. Lack of domain level validation.

B. Data Profiling Issues

The process of investigating the data available in an existing data source and accumulating statistics and information about that data is called profiling. Profiling can improve accuracy of data. This also is having the following common issues [10][12].

1. Wrong selection of automated tools.
2. Lack of structural analysis

3. Changeable meta data
4. Inadequate data content analysis against external reference data.
5. Inadequate pattern analysis for given fields within each data store.
6. Need of identification of missing data relationships.
7. Inability of evaluation of data structure, data values and data relationships before data integration.

C. Data Staging ETL Issues

Data Staging ETL is having the maximum responsibilities of data quality. It has cleaning process which is executed to improve the accuracy of data warehouse. The following are the common issues at this stage [3][13].

1. Architecture of Data Warehouse affects the quality.
2. Relational and non relational things can affect quality.
3. Insufficient periodical refreshment of integrated data creates quality problems.
4. Some business rules also will affect quality.
5. Incapability of integrating cleansing tasks into visual workflows and diagrams.
6. Incapability of enabling profiling, cleansing and ETL tools to exchange data and Meta data.
7. Insufficient proper functioning of the extraction logic for each source system causes data quality problems.
8. Inadequate error reporting, validation, and metadata updates in ETL process.
9. Incorrect impact analysis of change requests on ETL.
10. Lack of ability to schedule extracts by time, interval, or event.

D. Schema Design Issues

Design of the data warehouse persuades the quality of the analysis that is possible with data in it. So, it is important to give special attention to design issues. It has the following common issues [11][14].

1. Partial and incorrect requirement analysis for schema design.
2. Dimensional modeling can affect quality.
3. Limited and false requirement analysis for schema design.
4. Delayed identification of slowly changing dimensions.
5. Having multi valued dimensions.
6. Inappropriate selection of record granularity.
7. Incorrect identification of dimensions.
8. Lack of ability to support database schema.
9. Need of sufficient validation and integrity rules in schema.

7. CONCLUSION

Data quality is really a key success factor and an enabler of big data migration projects as well as important milestone. The motivation of this research was to focus all possible issues of data quality problems that exist at all the phases of data warehouse. This paper will lead to design best frame work to handle data quality issues to improve accuracy in the big data migration process.

REFERENCES

1. Manjunath T.N., Ravindra S Hegadi, “Data Quality Assessment Model for Data Migration Business Enterprise”, International Journal of Engineering and Technology, Volume 5, No.1, Feb –March 2013, ISSN: 0975-4024.
2. Manjunath T.N., Ravindra S Hegadi, Mohan H.S., “Automated Data Validation for Data Migration Security”, International Journal of Computer Applications, Volume 30, No.6, September 2011, ISSN: 0975-8887.
3. Nitin Anand, Manoj Kumar, “An overview on Data Quality Isuues at Data Staging ETL”, Proceeding of International Conference on Advances in Computer Science and Application, 2013.

4. Priyanka Paygude, P.R.Devale, "Automated Data Validation Testing Tool for Data Migration Quality Assurance", International Journal of Modern Engineering Research, Volume 3, Issue 1, Jan-Feb, 2013, pp.599-603, ISSN: 2249-6645.
5. Priyanka Paygude, P.R.Devale, "Automation of Data Validation Testing for QA in the Project of DB Migration", International Journal of Computer Science Engineering and Inforamtion Technology Reasearch, Volume 3, Issue 3, Aug, 2013, pp-15-22, ISSN: 2249-6831.
6. Manjunath T.N., Ravindra S Hegadi, Ravikumar G.K., "Analysis of Data Quality Aspects in Data Warehouse Systems", International Journal of Computer Science and Information Technologies, Volume 2(1), 2011, 477-485 ISSN: 0975-9646.
7. Manjunath T.N., Ravindra S Hegadi, Archana R.A., "A Study on Sampling Techniques for Data Testing", International Journal of Computer Science and Communication, Volume 3, No.1, January-June 2012, pp.13-16.
8. Manjunath T.N., Ravindra S Hegadi, "Statistical Data Quality Model for Data Migration Bunisess Enterprise", International Journal of Soft Comuting, Volume 8(5), 2013, 340-351, ISSN: 1816-9503.
9. Shashikant Patel, Sagar Wakchaure, Mahendra Pingale, Saba Siraj, "Data Migration System in Heterogeneous Database", International journal of Research in Engineering and Technology, Volume 03, Issue 02, Feb-2014, eISSN: 2319-1163, pISSn:2321-7308.
10. Ranjit Singh, Dr.Kawaljeet Singh, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing", International Journal of Computer Science Issues, Volume 7, Issue 3, No. 2, May -2010, eISSN: 1694-0784, pISSn:1694-0814.
11. Atsa Etoundi Roger, Abessolo Alo'o Ghisiain and Simo Bonaventure Joel. "Migration of Legacy Information System based on Business Process Theory", International Journal of Computer Applications, 33(2), pp. 27–34, November 2011.
12. Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park, "A Data Quality Management Maturity Model", ETRI Journal, Volume 28, Number 2, April 2006.
13. Florian Matthes, Christopher Schulz, Klaus Haller, " Testing & Quality Assurance in Data Migration Projects", 27th IEEE International Conference on Software Maintenance, Williamsburg, VA, September, 2011, 25-30.
14. Nikhil Debbarma, Gautam Nath, Hillol Das, "Analysis of Data Quality and Performance Issues in Data Warehousing and Business Intelligence", International Journal of Computer Applications, Volume 79, No.15, October 2013, ISSN: 0975-8887.
15. Ramesh Babu Palepu, Dr.K.V.Sambasiva Rao, "Meta Data Quality Control Architecture in Data Warehousing", International Jouranl of Computer Science, Engineering and Information Technology, Volume 2, No.4, August 2012.