DOI: 10.15415/jotitt.2014.22021

A Review of Contemporary Data Quality Issues in Data Warehouse ETL Environment

RUPALI GILL[1], JAITEG SINGH[2]

[1]Assistant Professor, School of Computer Sciences, [2]Associate Professor, School of Computer Applications, CU, Punjab, India

Email: rupali.gill@chitkara.edu.in

Received: October 7, 2014| Revised: November 15, 2014| Accepted: December 19, 2014

Published online: December 30, 2014

Abstract: In today's scenario, Extraction–transformation– loading (ETL) tools have become important pieces of software responsible for integrating heterogeneous information from several sources. The task of carrying out the ETL process is potentially a complex, hard and time consuming. Organisations now –a-days are concerned about vast qualities of data. The data quality is concerned with technical issues in data warehouse environment. Research in last few decades has laid more stress on data quality issues in a data warehouse ETL process. The data quality can be ensured cleaning the data prior to loading the data into a warehouse. Since the data is collected from various sources, it comes in various formats. The standardization of formats and cleaning such data becomes the need of clean data warehouse environment. Data quality attributes like accuracy, correctness, consistency, timeliness are required for a Knowledge discovery process.

The present state -of –the- art purpose of the research work is to deal on data quality issues at all the aforementioned stages of data warehousing 1) Data sources, 2) Data integration 3) Data staging, 4) Data warehouse modelling and schematic design and to formulate descriptive classification of these causes. The discovered knowledge is used to repair the data deficiencies.

This work proposes a framework for quality of extraction transformation and loading of data into a warehouse.

General Terms: Data warehousing, data cleansing, quality data, dirty data

Keywords: Data inconsistency, identification of errors, organization growth, ETL, data quality

Journal on Today's Ideas – Tomorrow's Technologies, Vol. 2, No. 2, December 2014 pp. 153–160

I. INTRODUCTION

Business today forces the enterprises to run different but coexisting information systems. Correct information is the most imperative resource in many enterprises for the business success. Decision support and business intelligence systems are used to mine the data in order to obtain knowledge that supports the decision-taking process affecting the future of a given organization. A data warehouse is a class of relational database that is designed for analytical processing rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. The challenge in data warehouse environments is to integrate, rearrange and consolidate large volumes of data over many systems, to provide a unified information base for business intelligence.

The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. In ETL data is *extracted* from different data sources, and then propagated to the DSA (Data Staging Area) where it is *transformed* and cleansed before being *loaded* to the data warehouse. Source, staging area, and target environments may have many different data structure formats as flat files, XML data sets, relational tables, non-relational sources, web log sources, legacy systems, and spreadsheets. To work in an operational environment several quality issues have been seen in an ETL environment. Cleansing data of errors in structure and content is important for data warehousing and integration. Current solutions for data cleaning involve many iterations of data "auditing" to find errors, and long-running transformations to fix them. There are various approaches used in cleaning data in manufacturing industries, schools/colleges/universities, organizations and many more. Users need to endure long waits, and often write complex transformation scripts.

II. RELATED WORK

E. Rahm and H. Hai Do [16] classify data quality problems that can be addressed by data cleaning routines and provides an overview of the main solution approaches. The article also presents contemporary tool support for data cleaning process.

Alkis Simitsis [15] in this paper presented a conceptual and a logical model for ETL processes. The article provides a methodology for the transition from the conceptual to the logical model.

Muller and Freytag [14] classified quality problems into syntactical anomalies which concern data formats and values for data representation

(e.g. lexical errors, domain format errors and irregularities). The authors also discussed the Semantic anomaly and coverage anomaly which hinders data collection from being comprehensive and non-redundant representation of the world such as integrity constraints, contradictions, duplicates and invalid tuples. In the work of Müller and Freytag no appropriate framework was designed to support the cleansing process.

Singh and Singh in [10], highlights major quality issues in the field of a data warehouse in his review has collected various issues in data ware house process. The author has classified various causes of data ware house process which may help the practitioners to proceed towards their research.

Rahul K. Pandey [2] has tried to gather various sources of data quality problems at various stages of an ETL process. The researcher has classified the problems as problems at data sources, data profiling problems , staging problems at ETL, problems at data modelling.

Panos Vassiliadis et.al.[11] in his research identified generic properties that characterize ETL activities. The researcher provided a taxonomy that characterizes ETL activities in terms of the relationship of their input to their output and the proposed taxonomy can be used in the construction of larger modules which can be used for the composition and optimization of ETL workflows.

Ahmed Kabiri [8] suggests that ETL is a critical layer in DW setting. It is widely recognized that building ETL processes is expensive regarding time, money and effort. The researcher has highlighted (1) the review of open source and commercial ETL tools, along with some ETL prototypes coming from academic world, secondly (2) the modeling and design works in ETL field. Also, (3) ETL maintenance we approach (4) review works in connection with optimization and incremental ETL.

Sakshi Agarwal [7]reveals various reasons of data quality problems in DW ETL The researcher has identified the possible set of causes of data quality issues which would be beneficial for DW practitioners, researchers and implementers.

K.Srikanth et al. [4] discusses issues related to Slowly Changing Dimensions - SCD type 2 will store the entire history in the dimension table. In SCD type 2 effective date, the dimension table will have Start_Date and End_Date as the fields.The implementation of the same has been done in Informatica. The considered example for research is the Employee dimension.

Jasna Rodi´c et al. [13] have proposed various rules that can be used in data warehouse process. The researchers have generated metadata tables for these tables that store information about the rules. The information about the rules violations is stored to provide analysis of such data. This could give a

significant insight into our source systems. Entire data quality process will be integrated into ETL process in order to achieve load of data warehouse that is as automated, as correct and as quick as possible.

The published work by Singh and Singh [12] substantiates that very diminutive information available on the quality assurance of ETL routines. The proposed methodology helped to solve the common errors. The researcher suggested that by automating very basic quality checks for data quality management have given satisfactory results but still there is a need to study the scope of automated testing in extraction, transformation and loading routines independently.

Chinta et al. [9] gave a data cleaning framework to provide robust data quality. The authors have worked upon missing values and dummy values using the Indiasoft data set.

Satkaur et al. [6] provided a guide viable ETL solution. The work gave the step by step guide to ETL prototype.

Sujatha R.[5] in her research explores designed a framework for non-parametric iterative imputation based mixed kernel estimation in both mixture and clustered data sets. The research has implemented a framework to fill in incomplete instances.

The research by J. Anitha[3] has covered all the major aspects of ETL usge and can be used to compare and evaluate various ETL tools. The implementation of SCD Type - 2 has been done to show comparison.

The work by P. Saravanan [1] provided a integrated unit for imputing missing values for the right attribute. The kernel based iterative non-parametric estimators work for both continuous and discrete values.

III. DISCUSSIONS AND OBJECTIVES

The previous comprehensions has given a detailed idea is various data quality issues in data warehouse environment. The data quality issues of

1. naming conflicts
2. structural conflicts
3. date formats
4. missing values
5. changing dimensions

are taken into consideration by various authors and have been implemented through various tools. But no single framework has provided a solution to all the above problems at a single place. Moreover, the frameworks implemented which covers all the issues are implemented through paid tools. So we propose a framework to implement the above issues by hand-coding.

IV. PROPOSED WORK

While developing a new dimensional data warehouse or replacing an existing environment, the ETL (extract, transform, load) implementation effort is inevitably on the critical path. Difficult data sources, unclear requirements, data quality problems, changing scope, and other unforeseen problems often conspire to put the squeeze on the ETL development team. It simply may not be possible to fully deliver on the project team's original commitments.

Data Quality is not only a challenge in an idealized mono-cultural environment**,** but it is inflamed to epic proportions in a ETL environment. The data quality issues namely:

- Quality of data like dirty values, date formats, naming conventions
- How source and target data structures can be mapped like mapping various databases
- How "complex" are the data relationships like when to consider a fact table as dimension table

complicates the data warehouse process and hamper the implementation of Data warehouse ETL process in industry. Taking into consideration the above issues we propose a data quality service routine as follows:

The proposed service handles various quality issues of the ETL at the following stages**:**

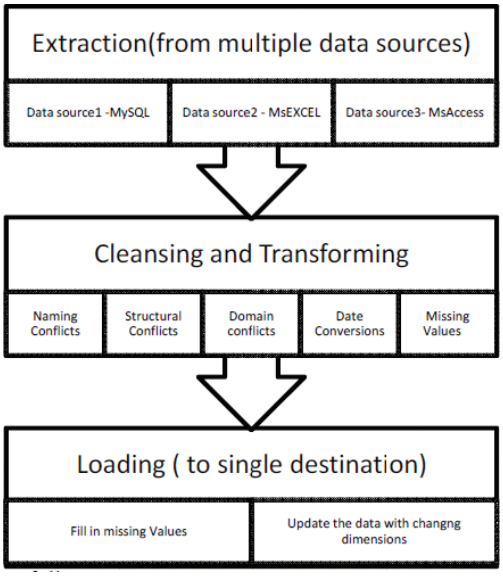

Figure 1: ETL Stage flow Diagram

1. Extraction
2. Transformation
3. Loading

1. Extraction: The amalgamation of all of the disparate systems across the enterprise is the real challenge to getting the data warehouse to a state where it is usable. This step consolidates the data from different data sources. Flat files and relational databases are the most common data sources. The main objective of extract the data is to retrieve all the required data from the source system with as little resources as possible. It is also known as Data discovery phase. The validated data from extraction is backed up and archived at the staging area.

2. Cleansing and Transformation: It applies a set of rules to transform the data from the source to the target. This includes converting the measured data to the same dimension using the same units so that they can be later joined. This step involves resolving

1. Naming Conflicts
2. Structural Conflicts
3. Applying Domain Constraint Checks
4. Handling Missing Values

3.Loading: Loading data to the target data source structure is the final step in ETL. In this step extracted and transformed data is written into dimensional

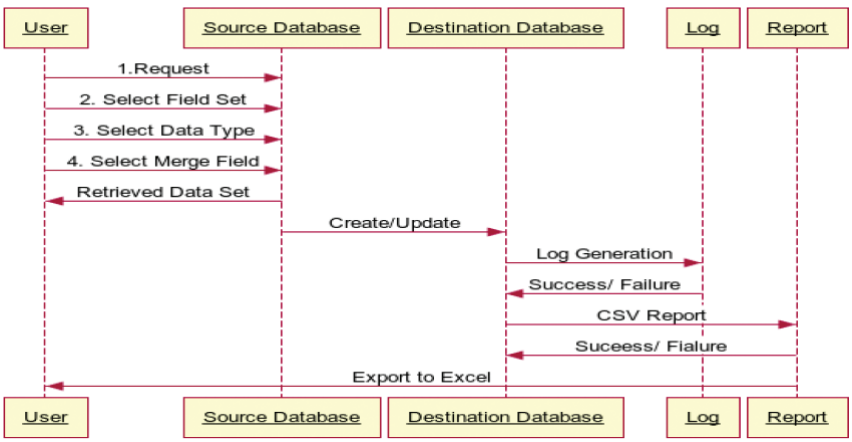

Figure 2: Sequence Diagram of Proposed ETL Quality Workflow

structures actually is accessed by the end user and application systems. Loading includes both dimensional tables and fact tables.

V. CONCLUSION

Data quality has become a major concern activity performed by most organizations that have data warehouses. Every organization needs quality data to improve on its services it renders to its customers. In view of this a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of data cleansing. The papers reviewed in this report looked at critical aspects of data quality and the various types of data that could be cleansed.

FUTURE WORK

In the future work we propose to implement the above mentioned System Routine according to following sequence diagram.

REFERENCES

[1] Chinta Someswara Rao, J Rajanikanth, V Chandra Sekhar, Bhadri Raju MSVS (2012) "Data Cleaning Framework for Robust Data Quality in Enterprise Data Warehouse" , IJCST e- ISSN: 0976-8491 p. ISSN : 2229-4333, Vol. 3, Issue 3, pp. 36-41

[2] K. Srikanth , N.V.E.S Murthy, J. Anitha (2013) " Data Warehousing Concept Using etl Process For SCD Type-3" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Issn: 2276-6856, Vol. 2, Issue 5, pp. 142-145.

[3] Kabiri A.; Chiadmi D. (2013) "Survey on ETL Processes", Journal of Theoretical and Applied Information Technology. Vol. 54, No. 2

[4] Pandey K.Rahul (2014). Data Quality in Data warehouse: problems and solution.IOSR-Journal of Computer Engineering, Volume 16, Issue 1, pp. 18-24.

[5] Rahm, E., Do, H.H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bull. Vol 23 No. 4, pp. 3-13

[6] Rodi´c J.; Baranovi´c M. (2009) "Generating Data Quality Rules and Integration into ETL Process", DOLAP'09 ACM

[7] Sakshi Agarwal '*Reasons of Data Quality Problems in Data Warehousing*' International Journal of Computer, Information Technology & Bioinformatics (ijcitb) issn: 2278-7593, Volume-1, Issue-4 ieee & ieee Computational Intelligence Society, 2013.

[8] Saravanan p. (2014) "An Iterative Estimator for Predicting the Heterogeneous Data Sets", Weekly Science Research Journal issn: 2321-7871, Volume 1, Issue 27, pp. 1-15.

[9] Satkaur; Mehta a.(2013) "a Review Paper on scope of etl in retail domain", International Journal of Advanced Research in Computer Science and Software Engineering 3(5), ijarcsse, pp. 1209-1213.

[10] Singh J.; Singh K. (2009) "Statistically Analyzing the Impact of Automated ETL Testing on the Data Quality of a Data Warehouse", International Journal of Computer and Electrical Engineering, Vol. 1, No. 4.

[11] Singh R.; Singh K. (2009). A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing International Journal of Computer and Electrical Engineering, Vol. 1, No. 4

[12] Srikanth k.; Murthy n.v.e.s.; Anitha j. (2013) "Data Warehousing Concept Using etl Process for scd Type-2", American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936, Volume-2, Issue-4, pp. 86-91' 2013.

[13] Sujatha.R (2013) "Enhancing Iterative Non-Parametric Algorithm for Calculating Missing Values of Heterogeneous Datasets by Clustering", International Journal of Scientific and Research Publication issn: 2250-3153, Volume 3, Issue 3, pp. 1-4.

[14] Vassiliadis P.; Simitsis A.; Baikousi E. (2009) "A Taxonomy of ETL Activities" DOLAP '09 Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP, pp 25-32.

[15] Heiko Muller, Johann-Christoph Freytag. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing, pp. 21.

[16] Vassiliadis P.; Simitsis A.; Skiadopoulos S.(2002) "Conceptual Modeling for ETL Processes", Proceedings of the ACM tenth international workshop on Data warehousing and OLAP, pp. 14-21.