AN OVERVIEW ON EVOCATIONS OF DATA QUALITY AT ETL STAGE

Sakshi Miglani [1], Dr. Neha Gupta[2]

[1]Research Scholar, [2]Assistant Professor, Faculty of Computer Applications Department, MRIU, Faridabad (India)

ABSTRACT

A data warehouse facilitates the integration of disparate operational databases in an enterprise into a single store. Data quality is one of the most important problems in data management. A database system typically aims to support the creation, maintenance, and use of large amount of data, focusing on the quantity of data. However, real-life data are often dirty: inconsistent, duplicated, inaccurate, incomplete, or stale. However, there is a rapid development and implementation of quality data warehouses specifically that of warehouse data quality issues at various stages of data warehousing. Specifically, problems a during the ETL process, data is extracted from an OLTP databases, transformed to match the data warehouse schema, and loaded into the data warehouse database. The state-of-the-art purpose of the paper is to identify the reasons for data deficiencies, non-availability or reach ability problems at the ETL stage 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. This will help developers & implementers of warehouse to examine and analyse these issues before moving ahead for data integration and data warehouse solutions for quality decision oriented and business intelligence oriented applications.

Keywords: Data Quality (DQ), Data Staging (DS), Data Warehouse(DW)

1. INTRODUCTION

1.1 Understanding Data Quality

Data are of high quality if, "they are fit for their intended uses in operations, decision-making and planning"(J.M.Juran). Furthermore, apart from these definitions as data volume increases, the question of internal consistency within data becomes paramount regardless of fitness for use for any particular external purpose. The one definition of DQ is that it's about bad data – data that is missing or incorrect or invalid in some context. Data quality ensures clear understanding of the meaning, context and intent of the data. Understanding the key DQ dimensions is the first step to data quality improvement.

"The beginning of wisdom is the definition of terms."

Something (data item, record, dataset or database) that can either be measured, or assessedin order to understand the quality of data. In order for the analyst to determine the scope of the underlying root causes and to plan the ways that tools can be used to address data quality issues, it is valuable to understand some common data quality dimensions. Abundant attempts have been made to define data quality and to identify its dimensions.

Dimensions of data quality include accuracy, reliability, importance, consistency, precision, timeliness, fineness;understand ability, conciseness and usefulness.Six primarily data quality dimensions shown in fig-1 are-

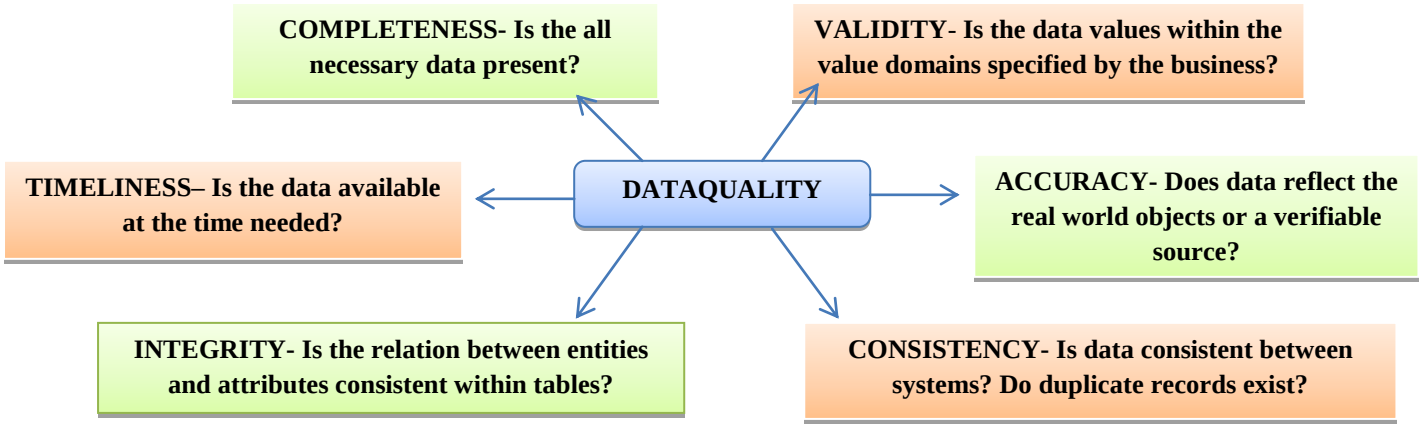

Fig. 1: Data Quality Dimensions

1.2 Data Warehousing

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. As defined by the "father of data warehouse" William H. Inmon, DW 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." Typically the data warehouse is maintained separately from the organizations operational databases, as data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP), applications traditionally supported by the operational databases.

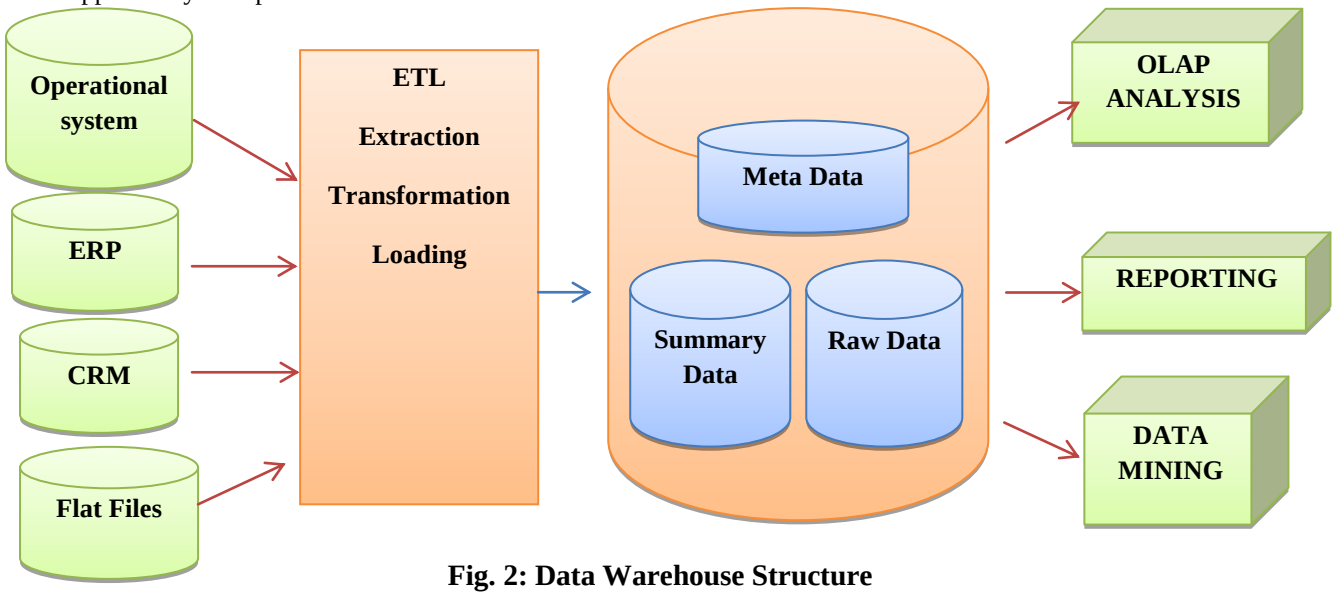

Fig. 2: Data Warehouse Structure

1.3 Phases of DW receptive to DQ problem

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 environment most of which affect is quality to some extent. Despite of 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. 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.

II METHODOLOGY

The study is designed as a literature review of materials published between 1995 to the year 2014 on the topics of DQ and DW. To develop the data quality problems on an ETL phase, the IT implementations infrastructure, data warehousing literature research questionnaires, related to data quality were reviewed.

2.1 Literature Reviewed

- Channah E Naiman& Aris M. Ouksel (1995)-The paper proposed a integration problems which further may have far reaching consequences on data quality.
- Jaideep Srivastava (1999)1 -The principal goal of this paper is to identify the common issues in data integration and data-warehouse creation. Problems arise in populating a warehouse with existing data since it has various types of heterogeneity.
- -Amit Rudra and Emilie Yeo (1999)-The paper concluded that the quality of data in a data warehouse could be influenced by factors like: data not fully captured, heterogeneous system integration and lack of policy and planning from management.
- Scott W. Ambler (2001)-The article explored the wide variety of problems with the legacy data, including data quality, data design, data architecture, and process related issues.
- Won Kim et al (2002)-Paper presented a comprehensive taxonomy of dirty data and explored the impact of dirty data on data mining results.
- Ralaph Kimball (2004)-The data warehouse ETL toolkit.
- AmolSrivastava, MohitBhaduria, HarshaRajwanshi (2008)- “Data warehouse and quality issues”.
- DikshaVerma, Anjali Tyagi, Deepak Sharma (2014)-Data quality problems in data warehousing.

2.2 Phases of ETL

Extraction, Transformation and Loading (ETL) processes are responsible for the operations taking place in the backstage of datawarehouse architecture.

- **Extracts** data from homogeneous or heterogeneous data sources
- **Transforms** the data for storing it in proper format or structure for querying and analysis purpose
- **Loads** it into the final target (operational data store, data mart or data warehouse)

Usually all the three phases execute in parallel since the data extraction takes time, so while the data is beingpulled another transformation process executes, processing the already received dataand prepares the data for loading and as soon as there is some data ready to be loaded in to thetarget, the data loading kicks off without waiting for the completion of the previous phases.

2.2.1 Extraction

Extracting data correctly sets the stage for the success of subsequent processes. During extraction the desired data is identified and extracted from many different sources, including database systems and applications. Common data-source formats include relational databases, XML and flat files or even formats fetched from outside sources by means such as web spidering screen scraping. In general the extraction phase aims to convert the data into a single format appropriate for transformation processing.An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources have the correct/expected values in a given domain (such as default or list of values). If the data fails the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records. In some cases the extraction process itself may have to modify a data validation rule in order to accept the data to flow to the next phase.

2.2.2 Transform

The data transformation stage applies a series of rules or functions to the extracted data from the source to derive the data for loading in to the end target. Some data do not require any transformation at all; known as direct move or pass through data in technical terms. An important function of data transformation is cleansing of data that aims to pass only proper data to the target. When different systems interact with each other based on how thesesystems store data there is a challenge in interfacing/ communicating with each other. Certain character set that may include or available in one system may not be available in other. These cases must be handled correctly or eventually lead to number of data quality related issues.

2.2.3 Load

The load phase loads the data into the end target that may be simple delimited flat file or a data warehouse. Depending on the requirements of the organisations, this process varies widely. Some data warehouses may overwrite existing information with cumulative information, updating extracted data with frequently done on a daily, weekly or monthly basis. Other data warehouses (even other parts of the same data warehouse) may add new data in a historical form at regular intervals, for example hourly.As the load phase interacts with a database schema – as well as in triggers activated upon data load apply(for example uniqueness, referential integrity, mandatory fields), which also contributes to the overall data quality performance of the ETL process.

2.3 Evocations of DQ at ETL stage

In DW, data cleaning is a major part of the so called ETL process. 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 warehouses require extensive support for data cleansing. A data cleaning process is executed in the data staging area in order to improve the accuracy of data warehouse.A staging area or "landing zone", is an intermediate storage area used for processing during the extract, transform and load (ETL) process. The data staging area sits between the data source and data target, which are often data warehouses, data marts or other data repositories. Staging and ETL phase is considered to be most crucial stage of data warehousing where maximum responsibility of data quality effort resides.

Data quality problems at this phase, from a defined literature review are as follows:-

Table 1

S.no	CAUSES OF DQ PROBLEMS AT ETL
1	DW architecture undertaken affects the DQ (Staging, Non Staging Architecture)
2	Type of staging area, relational or non-relational affects the DQ
3	Different business rules of various data sources creates problem of DQ
4	Business rules lack currency contributes to DQ problems [4]
5	The inability to schedule extracts by time, interval, or event cause DQ problems
6	Lack of capturing only changes in source files [5]
7	Lack of periodical refreshing of the integrated data storage (Data Staging area) cause DQ degradation
8	Truncating the DS area cause DQ problems because we can't get the data back to reconcile
9	Disabling data integrity constraints in DS tables cause wrong data and relationships to be extracted and hence cause DQ problems [7]
10	Purging of data from the DW cause DQ problems [5]
11	Hand coded ETL tools used for data warehousing lack in generating single logical meta data store, which leads to poor DQ
12	Lack of centralized metadata repository leads to poor DQ

13	Lack of reflection of rules established for data cleaning, into the metadata causes poor DQ
14	Inappropriate logical data map prepared cause DQ issues
15	Misinterpreting/Wrong implementation of the slowly changing dimensions (SCD) strategy in ETL phase causes massive DQ problems
16	Inconsistent interpretation or usage of codes symbols and formats [4]
17	Improper extraction of data to the required fields causes DQ problems [4]
18	Lack of proper functioning of the extraction logic for each source system (historical and incremental loads) causes DQ problems
19	Unhandled null values in ETL process cause DQ problems
20	Lack of generation of data flow and data lineage documentation by the ETL process causes DQ problems
21	Lack of availability of automated unit testing facility in ETL tools causes DQ problems
22	Lack of error reporting, validation, and metadata updates in ETL process cause DQ problems.
23	Inappropriate handling of rerun strategies during ETL causes DQ problems
24	Inappropriate handling of audit columns such as created date, processed date and updated date in ETL
25	Inappropriate ETL process of update strategy (insert/update/delete) lead to data quality problems
26	Non standardized naming conventions of the ETL processes (Jobs, sessions, Workflows) cause DQ problems
27	Wrong impact analysis of change requests on ETL cause DQ problems
28	Loss of data during the ETL process (rejected records) causes DQ problems. (refused data records in the ETL process)
29	Poor system conversions, migration, reengineering or consolidation contribute to the DQ problems [4] [6]
30	The inability to restart the ETL process from checkpoints without losing data [5]

31	Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects[5]
32	Inability of integrating cleansing tasks into visual workflows and diagrams[5]
33	Inability of enabling profiling, cleansing and ETL tools to exchange data and meta data[5]

III CONCLUSION

Data quality is an increasingly serious issue for large and small organizations. It is central to all data integration initiatives. Before data can be used effectively in a DW or in customer relation management, or business analytics applications, it needs to be analysed 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 on time. In this paper an attempt to collect all the data quality problems at ETL phase is made of data warehousing. Defective data also hampers business decision making and efforts to meet regulatory compliance responsibilities. These causes will really help the data warehouse practioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing.

IV FUTURE WORK

Each item shown in TABLE 1 will be converted in to an item of the research instrument that can be empirically tested by collecting views about the items from the data warehousing practioners, appropriately.

REFERENCES

[1] Channah F. Naiman, Aris M. Ouksel (1995) “A Classification of Semantic Conflicts in Heterogeneous Database Systems”, Journal of Organizational Computing, Vol. 5, 1995

[2] John Hess (1998), “Dealing with Missing Values In The Data Warehouse” A Report of Stonebridge Technologies, Inc (1998)

[3] 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. 118-126

[4] 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

[5] Wayne Eckerson& Colin White (2003) “Evaluating ETL and Data Integration Platforms” TDWI report series

[6] ArkedyMaydanxhik (2007), “Causes of Data Quality Problems”, Data Quality Assessment, Techniques Publications.

[7] Won Kim et al (2002) - "A Taxonomy of Dirty Data " Kluwer Academic Publishers 2002.

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

[9] 7 Sources of Poor Data Quality, https://www.melissadata.com/enews/articles/0 611/2.htm

[10] 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

[11] Markus Helfert, Gregor Zellner, Carlos Sousa, "Data Quality Problems and Proactive Data Quality Management in Data-Warehouse Systems"

[12]Mike(2009" the problem of dirty data" at http://www.articlesbase.com/databasesarticles/the-problem-of-dirty-data-1111299.html

[13] Erhard Rahm & Hong Hai Do (2003) "Data Cleaning: Problems and Current Approaches "

[14]AmolShrivastav, MohitBhaduria, HarshaRajwanshi (2008), " Data Warehouse and Quality Issues", available at http://www.scribd.com/doc/9986531/DataWarehouse-and-Quality-Issues

[15]AhimanikyaSatapathy, "Building an ETL Tool", Sun Microsystems, Available at: http://wiki.openesb.java.net/attach/ETLSE/ETLIntroduction.pdf