ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Causes of Data Quality Problems in DW and Module of ETL

Monali Kirange [1], Mayuri Kirange[2]

P.G. Student, School of Management Studies, NMU, Jalgaon, India[1]

ME(CSE), Department of Computer Engineering, S.S.G.B.O.T, Bhusaval, India[2]

ABSTRACT: Data warehousing is gaining in eminence as organizations become awake of the benefits of decision oriented and business intelligence oriented data bases. However, there is one key obstacle to the rapid development and implementation of quality data warehouses specifically that of warehouse data quality issues at various stages of data warehousing. Specifically, problems arise in populating a warehouse with quality data. All the causes of data quality problems at all the phases of data warehousing are 1) data sources, 2) data integration & data profiling, 3) Data staging and ETL, 4) data warehouse modeling and schema design. The purpose of this seminar is to identify the reasons for data deficiencies, non-availability or reach ability problems at all the mentioned stages of data warehousing and to formulate descriptive classification of these causes.
ETL is the core process of building data warehouse and is developed by and based on operation system. The data stored in data warehouse is from transaction system and external data sources A metadata management System with good design can highly improve the ETL efficiency. However, the large amount and the wide distribution of metadata in ETL processes cause the mismanagement of metadata To deal with this problem, this seminar suggests intensively manage ETL by metadata repository, logical optimization of ETL processes, modeling it as a state-space search problem, abstract local ontologies from metadata of different data sources, global ontology from data warehouse via OWL DL under the direction of domain ontology.

KEYWORDS: Extract Transform, Load, ETL, Data Warehouse Loading, Operational Data Store, Online Transaction Processing, Data quality.

I. INTRODUCTION

Data Warehouse plays an important part in theprocess of knowledge engineering and decision-making for Enterprise, as a key component of thedata warehouse architecture, the tool that supportdata extraction, transformation, loading (ETL) is acritical success factor for any data warehouse projects Traditional methods of ETL development The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken.
The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. A broader definition is that data quality is achieved when organization uses data that is comprehensive, understandable, consistent, relevant and timely. Understanding the key data quality dimensions is the first step to data quality improvement. To be process able and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness.
To solve heterogeneity problems of different data sources in the processes, this framework models for optimization of the ETL processes by using semantic web technologies and discusses how ontologies are used to support the data integration. A metadata management System with good design can highly improve the ETL efficiency There are different strategies that can be used and they each have different costs and Benefits.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

II. DATA WAREHOUSING AND THE MAIN MODULE OF ETL

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support".Accurate way to design ETL process as in fig. to make it efficient, flexible and maintainable, ETL can be divided into 5 modules: data extraction, data validation, data cleaning, data conversion and data loading.

1] Data extraction

This step requires a great deal, first of all, need to find out which business system does the business data come from, what kind of database management system the business database server runs. Secondly, need to know what kind of table structure the database has and the corresponding meaning of each table structure. Thirdly, need to check whether there exist the manual data and the quantity of the data. Fourthly, define whether there exist the unstructured data. After collecting all of the information, give out he data explanation file.

2] Data validation

Data validation involves a lot of checking work, including the effective value of the property, foreign key checking and so on. As for the low-quantity data, refuse them firstly, and then these data will be stored in order to be fixed in the field of the amendment.

3] Data cleaning

The task of data cleaning is to filter out the undesirable data, and then send them to the business operation department. These undesirable data include: incomplete data, wrong data, duplicated data and so on.

4] Data conversion

From the micro-details perspective, data conversion involves the following types: direct mapping, field operations, character string processing, null value determination, date conversion, date operation, and assemble operations and so on.

5] Data loading

Data will be moved to the center of the target datawarehouse table, it is usually the last step in the process of ETL. As to the best way to load data, the implementation depends on the type of operation and the quantity of data. There are two ways to insert or update data in the database table: SQL insert/update/delete or batch loading application program.Fig. 2. Shows the four staging steps of a data warehouse.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

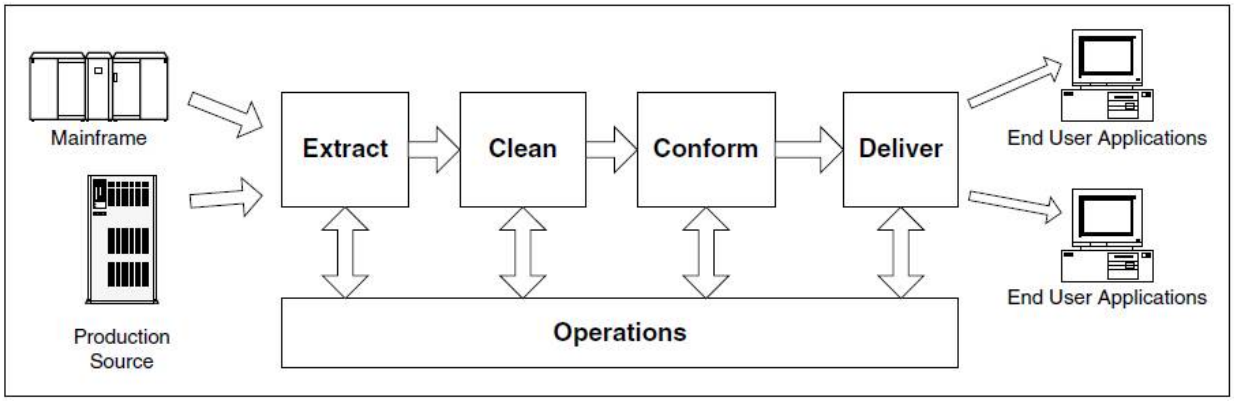

Fig. 2.The Four Staging Steps of a Data Warehouse.

III. LITERATURE REVIEW

ETL is the core process of building data warehouse and is developed by and based on operation system. The data stored in data warehouse is from transaction system and external data sources.
Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. As defined 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. Data Warehouses can contain detailed data, lightly summarized data and highly summarized .Data, all formatted for analysis and decision support" (Inmon, 1996). In the "Data Warehouse Toolkit", Ralph Kimball gives a more concise definition: "a copy of transaction data specifically structured for query and analysis" (Kimball, 1998). Both definitions stress the data warehouse's analysis focus, and highlight the historical Nature of the data found in a data warehouse.
The study is designed as a literature review of materials published between 1992 and 2008 on the topics of data quality and data warehouses. The Figure presents the resulting research model formulated trough extensive literature review. To develop the research model, the IT implementations infrastructure, data warehousing literature, research questionnaires related to data quality were reviewed to identify various reasons of data duality problems at the stages mentioned in the model. Classification of causes of data quality problems so formed will be divided into the factors responsible for data quality at the phases. Later in the next phase of study, it will be converted into survey instrument for the confirmation of these issues from the data warehouse practitioners.

1) Channah E Naiman&Aris M. Ouksel (1995)- the paper proposed a classification of semantic conflicts and highlighted the issue of semantic heterogeneity, schema integration problems which further may have far reaching consequences on data quality . John Hess (1998) the report has highlighted the importance of handling of missing values of the data sources, specially emphasized on missing dimension attribute values.

2) JaideepSrivastava, Ping-Yao Chen (1999) 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.

3) AmitRudra 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.

4) Scott W. Ambler (2001) the article explored the wide variety of problems with the legacy data, including data quality, data design, data architecture, and political/process related issues. The article has provided a brief bifurcation of common issues of legacy data, which contribute to the data quality problems.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

5) Won Kim et al (2002) paper presented a comprehensive taxonomy of dirty data and explored the impact of dirty data on data mining results. A comprehensive classification of dirty data developed for use as a framework for understanding how dirty data arise, manifest themselves, and may be cleansed to ensure proper construction of data warehouses and accurate data analysis.

6) Wane Eckerson & Colin White (2003) report says ETL tools are the traffic cop for business intelligence applications. They control the flow of data between myriad source systems and BI applications. As BI environments expand and grow more complex, ETL tools need to change to keep pace. ETL tools need to evolve from batch-oriented, single-threaded processing that extracts and loads data in bulk to continuous, parallel processes that capture data in near real time. They also need to provide enhanced administration and ensure reliable operations and high availability.

7) Wayne Eckerson (2004) data warehousing projects gloss over the all-important step of scrutinizing source data before designing data models and ETL mappings. The paper presented the reasons for data quality problems out of which most important are 1) Discovering Errors Too Late 2) Unreliable Meta Data. 3) Manual Profiling. 4) Lack of selection of automated profiling tools.

IV. UNDERSTANDING DATA QUALITY

The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. A broader definition is that data quality is achieved when organization uses data that understands the key data quality dimensions is the first step to data quality improvement. To be process able and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understand ability, conciseness and usefulness.
Completeness: deals with to ensure is all the requisite information available? Are some data values missing, or in an unusable state?
Consistency: Do distinct occurrences of the same data instances agree with each other or provide conflicting information. Are values consistent across data sets?
Validity: refers to the correctness and reasonableness of data
Conformity: Are there expectations that data values conform to specified formats? If so, do all the values conform to those formats? Maintaining conformance to specific formats is important.
Accuracy: Do data objects accurately represent the "real-world" values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications.
Integrity: What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication across your system
Accurate way to design ETL process as in fig. to make it efficient, flexible and maintainable, ETL can be divided into 5 **modules:** data extraction, data validation, data cleaning, data conversion and data loading.The ETL process flow involves four major operations: extracting the data from the sources, running it through a set of cleansing and conforming transformation processes, delivering it to the presentation server, and managing the ETL process and back room environment.

STAGES OF DATA WAREHOUSING SUSCEPTIBLE TO DATA QUALITY PROBLEMS

The purpose of this seminar here is to formulate a descriptive taxonomy of all the issues at all the stages of Data Warehousing. The phases are:

- Data Source
- Data Integration and Data Profiling
- Data Staging and ETL

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

- Database Scheme (Modeling)

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 data environment, most of which affect its quality to some extent. All these phases of data warehousing are responsible for data quality in the data warehouse. Despite 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. Fig .2 shows Stages of Data Warehouse Susceptible for DQ Problems as below.

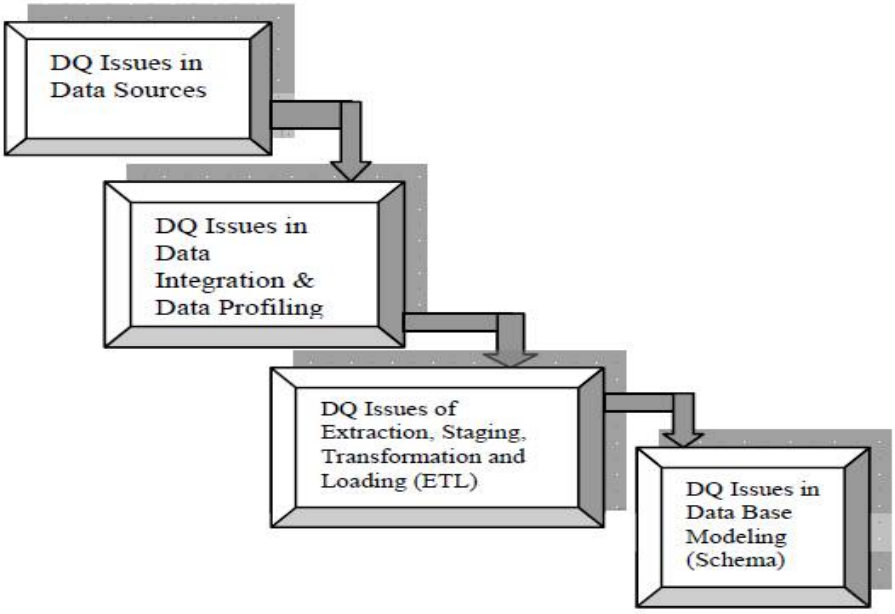

Fig. 2.Stages of Data Warehouse Susceptible for DQProblems

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.
The assumptions undertaken are that data quality issues can arise at any stage of data warehousing viz. in data sources, in data integration & profiling, in data staging, in ETL and database modeling. The model is depicting the possible stages which are vulnerable of getting data quality problems.

V. DATA QUALITY ISSUES

To determine the scope of the underlying root causes of data quality issues and to plan the design the tools which can be used to address data quality issues, it is valuable to understand these common data quality issues.

One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse. A data cleaning process is executed in the data staging area in order to improve the accuracy of the data warehouse.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Sr No	Data Quality issues at Data Staging ETL
1	Data warehouse architecture undertaken affects the data quality
2	Type of staging area, relational or non relational affects the data quality.
3	Different business rules of various data sources creates problem of data quality.
4	Lack of capturing only changes in source files, periodical refreshing of the integrated data Storage
5	periodical refreshing of the integrated data storage
6	Inability to support database schema refactoring cause data quality problems.
7	Lack of centralized metadata repository leads to poor data quality and Lack of reflection of rules established for data cleaning, into the metadata causes poor data quality.
8	Inappropriate logical data map prepared cause data quality issues.
9	Inconsistent interpretation or usage of codes symbols and formats and Improper extraction of data to the required fields causes data quality problems
10	Lack of generation of data flow and data lineage documentation, availability of automated unit testing facility, error reporting, validation, and metadata updates by the ETL process causes data quality problems.
11	Inappropriate ETL process of update strategy leads to data quality problems.
12	Loss of data during the ETL process (rejected records) causes data quality problems.
13	Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects

VI. CONCLUSION

Here mentioned all possible causes of data quality problems that may exist at all the phases of data warehouse. Our objective was to put forth such a descriptive classification Which covers all the phases of data warehousing which can impact the data quality? It has attempted to think on near possible set of causes of data quality problems at all the phases at one attempt. These causes will really help the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing. It would also be helpful for the vendors and those who are involved in development of data quality tools so as to incorporate changes in their tools to overcome the problems highlighted in classifications. It also reviewed the various optimization rules for ETL processes in order to solve the data integration problem of structural and semantic heterogeneity. Utilizing ontology makes ETL more flexible and efficient and the whole processes can be partially automated with the ETL rules in the framework.

REFERENCES

[1] Channah E Naiman&Aris M. Ouksel (1995)-"*Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data* ",The Data warehouse Institute (TDWI) report ,a vailable at www.dw-institute.com .

[2] JaideepSrivastava, Ping-Yao Chen (1999)Why Data Warehouse Projects Fail: Using Schema Examination Tools to Ensure Information Quality, Schema Compliance, and Project Success. Embarcadero Technologies.

[3] AmitRudra and Emilie Yeo (1999), T.; "State-space optimization of ETL workflows", Knowledge and Data Engineering, IEEE Transactions on Volume:17 , Issue: 10 Digital Object Identifier: 10.1109/ TKDE.2005.169 Publication , Page(s): 1404 –1419.

[4] Scott w.amble(2001); "A Framework Model Study for Ontology-Driven ETL Processes", Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on Digital Object Identifier: 10.1109/ WiCom.2008.2651 Publication Year: 2001, Page(s): 1 – 4

[5] Won Kim et al (2002), *"Data Profiling: A Tool Worth Buying (Really!)" ,Information Management Magazine, June 1, 2004 available http://wwwInformationmanagement. com/issues/20040601/1003990- 1.html* Erhard Rahm& Hong Hai Do (2002)

[6] *Wane Eckerson & Colin White (2003) Cleaning: Problems and Current Approaches* ", available at:http://homepages.inf.ed.ac.uk/wenfei/tdd/reading/cleaning.pdf.

[7] Simitsis, A.; Wilkinson, K.; Dayal, U.; Castellanos, M.; "Optimizing ETL workflows for fault-tolerance",Data Engineering (ICDE), 2010 IEEE 26th International Conference on Digital Object Identifier: 10.1109/ICDE.2010.5447816 Publication Year: 2010 ,Page(s): 385 – 396

[8] Wayne W. E. (2004) "Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data ", The Data warehouse Institute (TDWI) report, available at www.dw-institute.com .