International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

Vol.4., Issue.1., 2016
(Jan-Feb)

REVIEW ARTICLE

ISSN: 2321-7758

A DATA CLEANING MODEL FOR DATAWARE HOUSE

VINTI PARMAR[1], ROMIKA YADAV[2], MAMTA SHARMA[3]
[1,2]Research Scholar, Indira Gandhi University, Meerpur, Rewari, India
[3]Jamia Hamdard University, New Delhi, India

ABSTRACT

Data quality is an important factor for the success of data warehousing projects. Improvement in the quality of data is required in data warehouse, because it is used in the process of decision support, which needs accurate data. There are many errors and inconsistencies that occur in the data sets when brought in from several sources. Data cleaning is the activity of identifying and removing or correcting errors in the data. In this paper we propose enhanced algorithm to clean data in the data warehouse, to detect and correct most of the error types and expected problems, such as lexical errors, domain format errors, irregularities, integrity constraint violation, and duplicates.

Keyword-Data cleaning, data quality, data set, data warehouse (DW), Extraction-Transformation-Loading (ETL)

I. INTRODUCTION

Data cleaning is defined as an activity which is performed on the datasets of data warehouse to enhance and maintain the quality and consistency of data. Data cleaning is the process of identifying and removing or correcting errors in the data. It determines and detects the useless, unwanted, corrupt, inconsistent and faulty data and rectify it to enhance the quality of data. Data cleaning is the process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. This process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors [6]. Data quality identified as an 'error-free' approach in the data warehouse. The quality of data needs to be increased by using the data cleaning techniques. Organizations accumulate much data from their businesses that they want to access and analyze as a consolidated whole. However, the data often have inconsistencies in schema, formats, and adherence to constraints, due to many factors like merging from multiple sources and entry errors. Existing data cleaning techniques are used to identify record duplicates, missing values, record and field similarities, and duplicate elimination. The main objective of data cleaning is to reduce the time and complexity of process and increase the quality of datum in the data warehouse. Data quality refers that data is exactly fit for the purpose business use; the features of quality of data as coherency, correctness and accuracy along with the newness and accessibility of data .Improving the quality of data is important in data warehouse, because it is used in the process of decision support, which requires accurate data. Data Warehouse of an enterprise consolidates the data from multiple sources of the organization/enterprise in order to support enterprise wide decision making, reporting, analyzing and planning. The processes performed on data warehouse for above mentioned activities are highly sensitive to quality of data. They depend on the accuracy and consistency of data. Degraded quality of data leads to wrong conclusions of these processes which ultimately lead to wastage of all kinds of resources and assets [1]. Data cleaning is to deal with the dirty data in data warehouse so as to keep high data quality. The principal of data cleaning is to find and rectify the errors and inconsistencies for the data.

II. RELATED WORK

Big The word Data Warehouse (DW) was given by Bill Inmon in 1990, which he defined in the following way: "A warehouse is a subject oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Ralph gives a much simpler definition of a data warehouse as "a copy of transaction data specifically structured for query and analysis" [7]. With this Data warehousing efforts often aim to consolidate data from heterogeneous sources in hoping to provide a unified view of the data that can be used for business decision support, customer relationship management and a large number of other data analysis tasks. Accuracy of such analyses is crucial and relies upon the accuracy of the data loaded into the data warehouse. However, data received at the data warehouse from external sources usually contains errors, e.g. spelling mistakes, inconsistent conventions across data sources, and/or missing fields [2]. Significant amounts of time and money are consequently spent on data cleaning, the task of detecting and correcting errors in data. Following are steps for data warehouse design:

1) Data Acquisition

Data extraction is one of the most time-consuming tasks of DW development. Data consolidated from heterogeneous systems may have problems, and may need to be first transformed and cleaned before loaded into the DW. Data gathered from operational systems may be incorrect, inconsistent, unreadable or incomplete. Data cleaning is an essential task in data warehousing process in order to get correct and qualitative data into the DW. This process contains basically the following tasks [8]:

- Converting data from heterogeneous data sources with various external representations into a common structure suitable for the DW.
- Identifying and eliminating redundant or irrelevant data.
- Transforming data to correct values (e.g., by looking up parameter usage and consolidating these values into a common format).
- Reconciling differences between multiple sources, due to the use of homonyms (same name for different things),synonyms (different names for same things) or different units of measurement.

2). Extraction, Cleansing, and Transformation Tools

The work of extracting data from a source system, cleansing, and transforming the data and loading the consolidated data into a target system can be done either by separate products or by a single integrated solution. Integrated
Solutions fall into one of the following categories [8]:

- Code generators.
- Database data replication tools.
- Dynamic transformation engines.

III. DATA QUALITY: A FAVOURABLE OUTCOME

Large Data quality refers that data is exactly fit for the purpose of business use; that it is consistent, accurate, complete and uniform. Cleaning of data refers to an activity which determines and detects the unwanted, corrupt, inconsistent and faulty data to enhance the quality of data [1]. Companies use up millions of dollars per year to detect the errors in the data. In a study it is found that the combined cost due to bad data to be over US$ 30 billion in year 2006 alone [5]. As business operations trusts more and more on computerized systems, this cost is bound to increase at an high rate. Data quality is the degree to which data meet the specific needs of specific users, which contains several dimensions. The following are characteristics and measures of data quality [7]:

- Definition Conformance: The chosen object is of most important and its definition should have complete details and meaning of the real world object.
- Timeliness: It is the relative availability of data to support a given process within the timetable required to perform the process.
- Precision: The domain value which specifies business should have correct precisions as per specifications.
- Derivation Integrity: It is the correctness with which two or more pieces of data are combined to create new data.
- Completeness (of values): It is the characteristic of having all required values for the data fields.
- Validity (Business rule conformance): It is a measure of degree of conformance of data values to its domain and business rules. This includes Domain values,
- Ranges, reasonability tests, Primary key uniqueness, Referential Integrity.
- Accuracy (to the Source): It is a measure of the degree to which data agrees with data contained in an original source.
- Non-duplication (of occurrences): It is the degree to which there is a one-to-one correlation between records and the real world object or events being represented.
- Accessibility: Is the characteristic of being able to access data on demand.

Defective data lead to breakdowns in the supply chain, poor business decisions, and inferior customer relationship management. Data are the core business asset that needs to be managed if an organization is to generate a return from it. Hao et al. [4] designed the framework as based on rules-base, rule scheduling and log management. Data cleaning process is divided into four parts: data access interface, data quality analysis, data transformation and results assessment. The strong points are: The framework design is unified as all data cleansing process performed at single place. The data access interface provides unified data extraction interface for single source and multi-source data. The process log management records the operation information of whole data cleansing process. The limitations are: The data quality analysis should be done only once and should not be a repetition work. Yu et al. [3] suggested a framework that consists of access to database objects for user model, definition of user model and definition of quality model based on user model. The user model is a data model which is abstracted from the real model in perspective of the user. Arora et al. [1] defines the error of duplicity of data of string type in data warehouse in different data marts. As per Jebamalar et al. [6] the main objective of data cleaning is to reduce the time and complexity of the mining process and increase the quality of datum in the data warehouse.

IV. SOURCES OF ERROR IN DATA

The identification of the sources of data errors can be useful in designing data collection techniques that lessen the introduction of errors, and in developing appropriate post-hoc data cleaning techniques to detect and ameliorate errors.

1) Data entry errors: Data is often corrupted at entry time by typographic errors or misunderstanding of the data source.

2) Measurement errors: In some cases measurements are undertaken by human processes that can have errors in their design (e.g., improper surveys or sampling strategies) and execution (e.g., misuse of instruments).

3) Distillation errors: In many settings, raw data are preprocessed and summarized before they are entered into a database

4) Data integration errors: integration of data from multiple sources can lead to errors. This is because the sources often contain redundant data in different representations.

V. PROPOSED FRAMEWORK DESIGN

In Here we propose a framework for data cleaning. We have attempted to solve all the errors and problems that are expected, such as Lexical Error, Domain Format Error, Irregularities, Integrity Constraint Violation, Duplicates, Missing Value, and Missing Tuple.

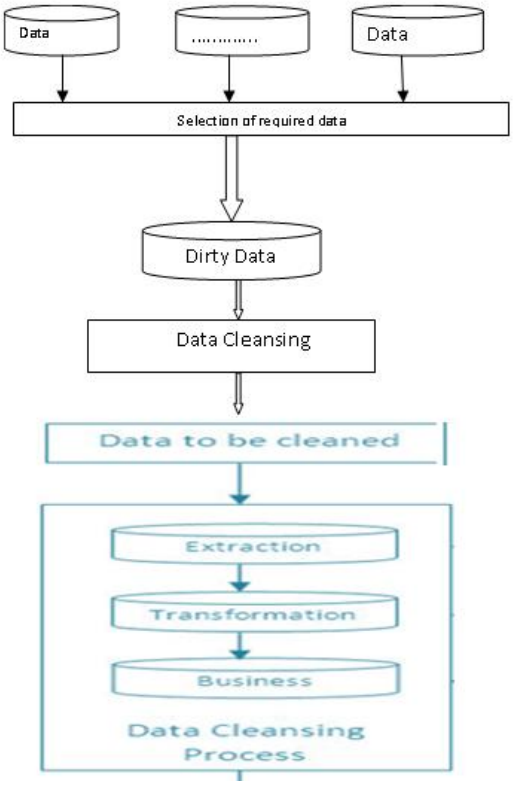

Figure.1 Proposed Component Framework Design

This proposed model can easily be developed in a data warehouse, by the following algorithm:

Input: Sources of Dirty Data

Output: Clean Data

- Select the sources
- For each source:
    - Extract source
    - Select the needed attributes from source(Selection of Attributes)
    - Select the quantitative and discrete attribute only
    - Search and identify the errors and error instances (auditing data)
    - Correct the errors by special algorithms for each instance (the suitable algorithm)
    - Display to user for correction manually
    - Storage in temporary tables
- Combining of all attributes in temporary tables (which needed for the target)
- Check and process to: record redundancy (record duplications)
- Load to the target

VI. CONCLUSION

The most important step in any data processing task is to verify that data values are correct or, at the very least, conform to some a set of rules. Data cleaning are important tasks in data warehousing. Incorrect and misleading data lead to all sorts of unpleasant and unnecessary expenses. In the existing data cleaning techniques, some of the cleaning methods are implemented. But those existing techniques are only good in some parts of cleaning process. For example, duplicate elimination cleaning tools are suited for data elimination process and a similarity-cleaning tool is well suited for field similarity and record similarity. Some approaches enable the declarative specification of a more comprehensive data cleansing processes. In this paper, we focused on errors in quantitative and limited value attributes of huge data. We have proposed comprehensive algorithm for data cleaning for data warehouse.

REFERENCES

[1]. R. Arora, P. Pahwa, S. Bansal, "Alliance Rules of Data Warehouse Cleansing", IEEE , International Conference on Signal ProcessingSystems, Singapore, May 2009, Page(s): 743 – 747.

[2]. S. Chaudhuri, K. Ganjam, V. Ganti, "Data Cleaning in Microsoft SQL Server 2005", In Proceedings of the ACM SIGMOD Conference,Baltimore, MD, 2005.

[3]. Yu Huang, Xiao-yi Zhang, Zhen Yuan, Guo-quan Jiang, "A universal data cleaning framework based on user model", IEEE, ISECS International - Computing, Communication, Control, and Management, Sanya, China, Aug 2009, Page(s): 200 – 202.

[4]. Hao Yan, Xing-chun Diao, Kai-qi Li, "Research on Information Quality Driven Data Cleaning Framework", IEEE, FITME '08. International Seminar - Future Information Technology and Management Engineering, China, Nov 2008, Page(s): 537 – 539.

[5]. Sang-goo Lee, Seoul Nat, Univ Seoul, "Challenges and Opportunities in Information Quality", E-Commerce

Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2007, Tokyo, Jul 2007, Page(s): 481 – 481.

[6]. J. Jebamalar Tamilselvi and Dr. V. Saravanan, “A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse”, ACM, IJCSNS Intenational Journal of Computer Science and Network Security, Vol.8 No.5, May 2008, Page(s): 117 – 121.

[7]. T. Manjunath, S. Ravindra, and G. Ravikumar, “Analysis of data quality aspects in data warehouse systems,” International Journal of Computer Science and Information Technologies, Vol. 2, No. 1, 2010, pp. 477-485.

[8]. B. Pinar, A Comparison of Data Warehouse Design Models, Master Thesis, Atilim University, Jan. 2005.

[9]. Prerna S.Kulkarni, Dr. J.W.Bakal, Survey on Data Cleaning, IJESIT 2014.

[10]. Mortadha M. Hamad, An Enhanced Technique to Clean Data in the Data Warehouse, IEEE,2011. *uring engineering*, vol. 8, (2010), issue 3.