Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Study over Problems and Approaches of Data Cleansing/Cleaning

Nidhi Choudhary

Department of Computer Science,

UPTU, India

Abstract— Data cleaning is mostly required when integrating heterogeneous (similar) data sources and should be addressed together with schema-related data transformations. This paper identifies the data quality problems that are inscribed by data cleaning and provides an overview of the main approaches as a solution. Also in data warehousing, the data cleaning is an important part of the commonly-known ETL process and is BI Tool can be integrated with ETL Tool or vice-versa. The current techniques that supports the data cleaning are also discussed.

Key Words:- Data Cleaning, Heterogeneous Data, Schema-related Data Transformation, Data Warehousing, ETL Process, BI Tool, ETL Tool.

I. INTRODUCTION

Data scrubbing, also called **data cleansing**, is the process of amending or removing data in database that is incorrect, incomplete, improperly formatted, or duplicated. An organization in a data-intensive field like insurance, retailing, banking, tele-communications, or transportation might use a data scrubbing tool to systematically examine data for flaws by using rules, alogrithms, and look-up tables. Basically, a database scrubbing tool includes programs that are capable of correcting a number of specific type of mistakes, such as adding missing zip codes or finding duplicate records. Using a data scrubbing tool can save a database administrator a significant amount of time and can be less costly than fixing errors manually.

Data cleansing, **data cleaning** or **data scrubbing** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data.

Fig: 1 Elements of Data Cleaning

II. PROCESS OF DATA CLEANING

A simple, five-step data cleansing process that can help to target the areas where the data is weak and needs more attention. From the first planning stage up to the last step of monitoring the cleansed data, the process will help the team zone in on dupes and other problems within the data. The most important thing to remember about the five step process, is that it's a on going circle. Therefore, it can start with small and make incremental changes, repeating the process several times so as to continue improving data quality.

Fig: 2 Data Cleaning Process

1. Plan

First of all, there is need to identify the set of data that is critical for making your working efforts the best they can possibly be. When looking at data, it should focus on high priority data, and start small. The fields that are need to identify will be unique to the business and what information is specifically looking for.

2. Analyze to Cleanse

After an idea of the priority data that is desired, it's important to go through the data that is already exist in order to see what is missing, what can be thrown out, and what, if any, are gaps between them. There is a need to identify a set of resources to handle and manually cleanse exceptions to the rules. The amount of manual intervention is directly correlated to the amount of acceptable levels of data quality. Once a list of rules or standards get builds, it'll be much easier to actually begin cleansing.

3. Implement Automation

Once the cleansing begins, it should begin to standardize and cleanse the flow of new data as it enters the system by creating scripts or workflows. These can be run in real-time or in batch (daily, weekly, monthly) depending on how much data has been taken for working. These routines can be applied to new data, or to previously keyed-in data.

4. Append Missing Data

This is important especially for records that cannot be automatically corrected. For examples, emails, phone numbers, industry, company size, etc. It's important to identify the correct way of getting a hold of the missing data, whether it's from 3rd party append sites, reaching out to the contacts or just via Google.

5. Monitor

You will want to set up a periodic review so that you can monitor issues before they become a major problem. You should be monitoring your database on a whole.

At the end, it is to bring the whole process full circle. Again revisit your plans from the first step and re-evaluate. Can the priorities be changed? Do the rules implemented still fit into the overall business strategy? Pinpointing these necessary changes will equip towards the work through the cycle; make changes that benefit the process and conduct periodic reviews to make sure that the data cleansing is running with smoothness and accuracy.

III. ETL PROCESS

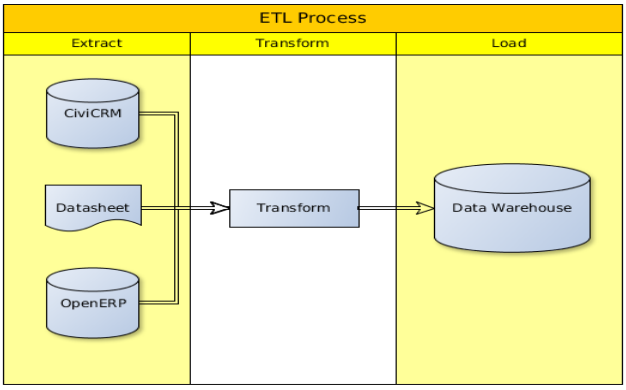

Fig: 3 Working of ETL

The working of the ETL Tool consist of three major components which are as follows:
1. EXTRACTION
2. TRANSFORMATION
3. LOADING

1. EXTRACTION: The integration 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. Data is extracted from heterogeneous 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 several ways to perform are:
i. Update notification
ii. Incremental extract
iii. Full extract

2. 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. The several ways to perform is:
i. Data Smoothing
ii. Data Aggregation
iii. Data Generalization
iv. Normalization
v. Attribute Construction

3. LOADING: Loading data to the target multidimensional structure is the final step in ETL. In this step extracted and transformed data is written into dimensional structures actually is accessed by the end user and application systems. Loading includes both dimensional tables and fact tables.

IV. BI TOOL

Business intelligence (BI) is regarding to create a value for the organizations depends on data or, more precisely, on facts. While it looks like another buzzword to describe what successful entrepreneurs have been doing for years, that is, using business common way. From a modern business-value perspective, corporations use BI to develop decision-making capabilities for managerial processes (e.g., planning, budgeting, controlling, assessing, measuring, and monitoring) and to ensure vital information is explored in a appropriate manner. Computer systems are the equipments that help us do work better, faster, and with more reliability and effectivity.
Business intelligence systems, is also known as EIS[Executive Information Systems], or Decision Support Systems, are a non-transactional IT system used to support business decision making and resolve management issues, generally used by executives and managers. Almost all people agrees that OLAP and data warehouse systems are a vital and essential part of business intelligence systems. Many business intelligence systems were in the structure of a data warehouse systems.

DIFFERENCE B/W ETL AND BI TOOL
The study over both the tools in data cleaning came to a following comparison:

ETL TOOL	BI TOOL
1. It helps to abstract and store the data.	It helps to control the data for decision-making and forecasting etc.
2. It facilitates the data to store in meaningful form.	While the BI tool supports in order to query the data for reporting and predicting
3. It helps in loading the data into the target tables which involves lots of transformation to execute the business logics and standards.	It also helps in analyzing the market conditions, customer's preferences, company capabilities etc.
4. It is a function in Data Warehousing.	Whereas BI is part of MIS[Management Information System]
5. ETL Tools incorporate the BI Tools.	But BI Tools encompasses OLAP[Online Analytical Processing]

Fig:4 Concept of BI Tool

V. SIGNIFICANCE OF DATA QUALITY

Data Quality - in it's simplest explanation -is ***a measurement of the value of a specific set of data, utilized in a specific manner, towards specific goals*** – then the levels of Data Quality attainable are relationally knotted to the specificities of the data within. Data Quality - by its close connection with the true value (vs. observed value) and applicability of a company's data – is an interior element of ROI[Region of Interest] determination and feasibility of the multiple uses to which the data is assigned; *i.e. marketing, business intelligence, and so forth*. Data Quality, in its most fundamental definition, is a metric by which the value of your data to your organization can be measured. Data Quality though, is also an actionable philosophy, as Data Quality can be manipulated, whereby it increases or decreases the value of the data upon which it acts respectively.

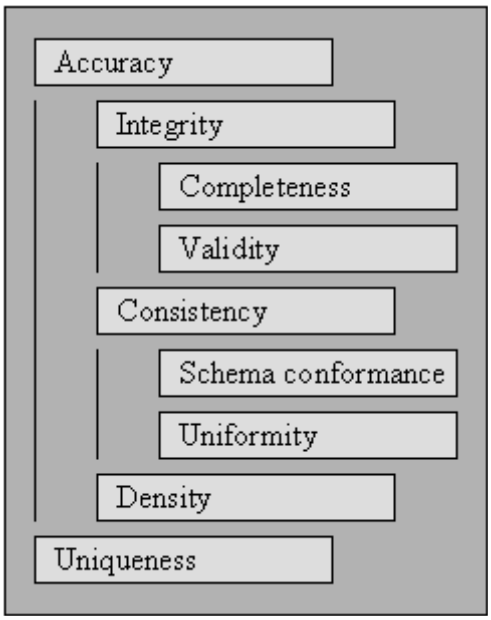

Figure 5: Hierarchy of data quality criteria

VI. CHALLENGES AND PROBLEMS OF DATA CLEANING

1. Conflict Resolution and Error Correction

The most challenging problem within data cleansing remains the correction of values toeliminate domain format errors, constraint violations, duplicates and invalid tuples. In many cases the available information and knowledge is insufficient to determine the correct modificationof tuples to remove these anomalies. This deletion of tuples leads to a loss of information if the tuple is not invalid as a whole. This loss of information can be avoided by keeping the tuple in the data collection and mask the erroneous values until appropriate information for error correction is available. The data management system is then responsible for enabling the user to include and exclude erroneous tuples in processing and analysis where this is desired. The ability of managing alternative values allows to defer the error correction until one of the alternatives is selectedas the right correction. It is a technical challenge to manage the large amount of different logical versions and still enable high performance in accessing and processing them.

2. Preservation of Cleansed Data

Cleansing data is a time consuming and expensive task. After having performed data cleansing and achieved a data collection free of errors one does not want to perform the whole datacleansing processOnly the part of the cleansing process should be re-performed that is affected by the changed value. This affection can be determined by analysing the cleansing lineage. After one of the values in the data collection has changed, the cleansing workflow has to be repeated for those tuples that contains the changed value as part of their cleansing lineage.The broad definition of require the collection and management of a large amount of additional meta-data to keep track of cleansing lineage. The additional information resulting from the initial workflow execution has to be collected in order to be able to speed-up ensuing cleansing workflow executions.

3. Data Cleansing in Practically[Virtually] Integrated Environments

The problems mentioned in the preceding section intensify when performing data cleansing in environments of virtually integrated sources, like IBM's DiscoveryLink [HSKK+01]. In these environments it is often impossible to propagate corrections to the sources because of their autonomy. Therefore, cleansing of data has to be performed every time the data is accessed. This considerably decreases the response time. By collecting and managing appropriate metadata like cleansing lineage and performed operations in a data cleansing middleware the performance could be increased considerably. The middleware should only collect as much data as necessary but still enable fast cleansing of data.

4. Data Cleansing Mechanism

In many cases it will not be possible to describe the whole data cleansing graph in advance.This makes data cleansing an iterative, interactive and explorative task. The whole data cleansing process is more the result of flexible workflow

execution. Process specification, execution and documentation should be done within a data cleansing framework which in turn is closely coupled with other data processing activities like transformation, integration, and maintenance activities. The framework is a collection of methods for error detection and elimination as well as methods for auditing data and specifying the cleansing task using appropriate user interfaces. The requirements for a data cleansing framework are (this might overlap with requirements for other data integration and management tasks):

• The ability to uniformly describe the structure and access the contend of data sources possibly heterogeneous in the data model used.

• Abstract definition of possible error types so the user can be guided within the data auditing task. For each error type there need to be methods that can be configured and used for error detection and elimination. Machine learning techniques can be applied within the data cleansing framework.

• For data cleansing a convenient user interface is needed. This has to be closely coupled with languages used for the specification of other activities in maintaining and processing data.

• The documentation of the performed operations as well as the collection of additional meta-data allows the verification of correctness for each of the tuples. This include the cleansing lineage.

VII. APPROACHES FOR DATA CLEANING/CLEANSING

1.Data analysis: With a view to find out the nature of errors and irregularity which are suppose to be eliminated, a descriptive analysis of data is necessitate. Besides the hand operated investigation of the data and data samples resolution programs are also used to get metadata regarding the features of data and detect the data quality issues.

2.Definition of transformation workflow and mapping rules: Based upon the number of data sources, their level of heterogeneity and the "uncleanliness" of the data, a huge number of data transformation and cleaning steps may have to be implemented. Sometimes, a schema translation is used to draw sources to a day-today data model; in data warehousing, generally a parallel representation is used. Recent data cleaning steps could correct single-source instance problems and manage the data for integration. Previous steps deals with schema/data integration and cleansing multi-source instance problems, e.g., duplications. In data warehousing, the control and data flow for such transformations and cleaning steps are specified within a workflow that describes the ETL process (Fig. 3).

3.Verification: The accuracy and effectiveness of a transformation workflow and the transformation should be tested and estimated, e.g. a sample of the source data, to refine the explanations if required. Various iterations of the analysis, design and affirmation steps may be required, for e.g., since few errors only become recognizable after applying few transformations.

4.Transformation: Implementation of the transformation steps either by running the ETL workflow for loading and refreshing a datawarehouse else throughout answering queries on multiple sources.

5.Backflow of cleaned data: Later,when errors are removed, the cleansed data should also replace the unclean data in the genuine sources in result to provide legacy applications the improved data too and to avoid redoing the cleaning work for future data extractions. The transformation process definetly required a huge amount of metadata, such as schemas, instance-level Data properties, transformation mappings, workflow definitions, etc. For regularity, flexibility and ease of reuse, this metadata should be managed in a DBMS-based repository. In order to support the data quality, detailed information about the transformation procedure is to be recorded, both in the repository or in the transformed instances, in specific information about the completeness and originality of source data and lineage information about the origin of transformed objects and the transformation applied to them.

LIMITATIONS

The limitations of the study are as follows:

1.The study has been done on narrow basis.

2.This research has been conducted on one or two IT organisations.

VIII. CONCLUSION

After a concise study, we came to a conclusion that Data cleaning is not only useful for data warehousing but also it is beneficial for query processing on heterogeneous data sources like in web-based information systems [Web-Designing].

- The Web Based environment acquire much more restrictive performance constraints for data cleaning that need to be considered in the design of suitable approaches.
- Moreover, data cleaning for semi-structured data like XML based, is of great importance which gives the reduced structural constraints and the speedily increasing amount of XML data.

- Data cleaning process helps in :
  i. Poor data accuracy
  ii. Manual data entry error
  iii.CRM/ERP Integration effectiveness
  iv.Migration of legacy system & database
  v.Limited customer insight
  vi.Acceleration of data dependant projects

- This also has been found that the ETL Tools like:
  i. Oracle Warehouse Builder
  ii. IBM Infosphere Information Server
  iii.SAS Data Integration Studio
  iv. SQL Server Integration Services
  v. Oracle Data Integrated etc. are effectivelty used in organizations.

- In Data Warehousing, the data cleaning process is far more effective in conflict detection[resolution].
- There is difference in ETL Tool and BI Tool on a basis of their working as BI is widely viewed in management context and ETL is viewed in as a complete organization system.
- More work is needed on the design and execution of the best language approach for supporting both schema and data transformations.

REFEERENCES

[1] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: *Tools for Data Translation and Integration*. In [26]:3-8, 1999.
[2] Bernstein, P.A.; Dayal, U.: *An Overview of Repository Technology*. Proc. 20th VLDB, 1994.
[3] Cohen, W.: *Integration of Heterogeneous Databases without Common Domains Using Queries Based Textual Similarity*. Proc. ACM SIGMOD Conf. on Data Management, 1998.
[4] Doan, A.H.; Domingos, P.; Levy, A.Y.: *Learning Source Description for Data Integration*. Proc. 3rd Intl. Workshop The Web and Databases (WebDB), 2000.
[5] Hernandez, M.A.; Stolfo, S.J.: *Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem*. Data Mining and Knowledge Discovery 2(1):9-37, 1998.
[6] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*. Proc. 10th Intl. Conf.Database and Expert Systems Applications (DEXA), 1999.
[7] Milo, T.; Zohar, S.: *Using Schema Matching to Simplify Heterogeneous Data Translation*. Proc. 24th VLDB,1998.
[8] Quass, D.: *A Framework for Research in Data Cleaning.* Unpublished Manuscript. Brigham Young Univ., 1999

WEBSITES:

1. http://link.springer.com/chapter/10.1007%2F978-0-85729-320-6_91#page-1
2. http://slide-share.com Bouayad Mehdi & Bouzoubaa Marouane

