© 2016 IJEDR | Volume 4, Issue 1 | ISSN: 2321-9939

A Brief Survey on Issues & Approaches of Data Cleaning

Ayanka Ganguly
M.E. Student,
Dept. of Computer Engineering,
SAL Institute of Technology & Engineering Research, Ahmedabad-380052, Gujarat, India.

Abstract - In a data warehouse management process cleaning raw data is an essential part. Data cleaning is a process of detecting and correcting inaccurate data. Before loading data into the data warehouse data needs to be cleaned thoroughly which in turn will also improve the quality of data too. Hence for accurate data, improving the data quality is necessary. Data may include text errors, quantitative errors or even duplication of the data. Many researchers have proposed various algorithms till now in order to clean the data which removes errors and inconsistencies from the data. In this paper, I have represented the major issues, challenges and approaches for cleaning data. Various tools for cleaning data have also been described.

***Index Terms* - data cleaning, data quality, tools, challenges and approaches.**

I. INTRODUCTION

Data cleaning or cleansing is a process to determine inaccurate, incomplete, or unreasonable data and then improving the quality of data through correction of detected errors and omissions [2,1]. Data cleansing is much more than simply updating a record with good data. Crucial data cleansing involves decomposing and reassembling the data. Data cleaning is performed by domain expert because it has the authority and knowledge in identifying and eliminating of anomalies. Anomaly is a property of data values it may causes the errors in measurements, lazy input habits, omission of data and redundancies. Anomalies are basically classified into three types:[7]

Syntactic - describes characteristic values and format.

Semantic - hides data collection from a comprehensive and non- redundant representation.

Coverage anomalies - reduce the amount of entities and their properties [10]

The general methods for data cleaning are[9]:

- Defining and determining the error types;
- Find and identify error instances;
- Correction of identified errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

Need for data cleaning

Cleaning of data is needed in order to improve the quality of data by reducing errors in data and make them "fit for use" by users for their documentation and presentation work. There may be different types of known and unknown errors in the data and are to be expected.

However, correcting these errors by various techniques may be time consuming process but it cannot be ignored. Thus it is not only important to delete the errors and clean the data but also the changes made in the document should be traced. The best way for retaining the corrected information is to add corrections of the database in separate field or fields so that incase of any failure in the system we can retrieve the original data from the database.

II. PRINCIPLES OF DATA CLEANING

a) *Planning is Essential:* A good planning is required for a better data management policy. Also a Strategy to implement data cleaning into an organization environment will improve the overall quality of the organization's data and its reputation among users and suppliers will be maintained.[3]

b) *Organizing data improves efficiency:* Organizing the data before data checking, validating and correction can improve the efficiency and reduce the time and cost of data cleaning. For example, by sorting data of any particular field, increase in efficiency can be achieved through checking all records related to the particular location at the same time, rather than going back and forth to key references[3]. Spelling errors in other fields may also be detected in this way.

c) *Prevention is better than cure:* It is more efficient and way cheaper to prevent an error rather than finding it later and correcting it. It is also important that during the detection of errors the feedback mechanisms must ensure that the errors does not occur again[3]. A good database design must ensure that the entities such as Names, Gender, Phone no, should be entered once and at the time of data entry it should be verified.

d) *Documentation:* For a good quality of data maintaining documentation is very much important. Without proper documentation, it may be difficult for users to understand how the data needs to be used and also for the curator to know what and by whom the data quality checks have been done. Generally there are two types of documentation: First, is to check for each and every record that what checks and changes have been made and by whom. The second is the metadata which maintains records of every information at the dataset level.

e) *Accountability, Transparency and Audit-ability:* These are the key elements for cleaning data. Unorganized and unplanned data cleaning can give unproductive and inaccurate results. Various kinds of accountability for cleaning data needs to be adopted or established by using the data quality policies and strategies. To improve the quality of data and to understand how to use the cleaned data, the process for data cleaning needs to be transparent and it should have a well formatted documentation with proper auditing trails in order to reduce duplication of records and to ensure that the errors never re-occur once they have been solved.

f) *Prioritization reduces Duplication:* By organizing and sorting of records, prioritization helps in improving efficiency and reduces overall cost. Due to this a large amount of data can be Cleaned at a lower cost. For example, before working on the more difficult records, they can be examined using batch processing or automated methods. By concentrating on the data which are more useful to user, there may be a chance of various kinds of errors being detected and corrected. Thus due to the immediate use of this data the quality of data will improve by providing greater incentives for data suppliers and users and it also improves client/suppliers relationships and reputations.

III. Data Quality Criteria

The data has to satisfy a set of quality criteria's in order to process and making it easy to understand in an effective and efficient manner. Satisfying those criteria's data of that type is said to be of a quality data. Many techniques and approaches have been made to define data quality and to identify its characteristics. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness[7]. During the optimization of data cleaning the quality criteria can be used by allotting priorities to each criteria which may influence some another criteria when various methods are executed for data cleaning[2].

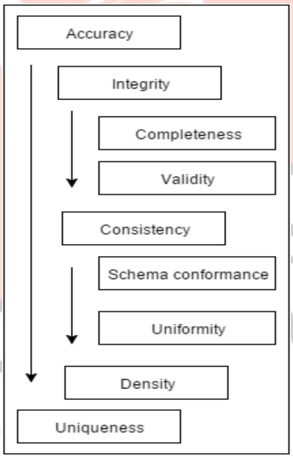

Fig: Data Quality Criteria[2]

Accuracy: it is the combination of integrity, consistency and density. When all these criteria's are accomplished the data is said to have a high quality.

Integrity: it is the combination for both completeness and validity. The inability to link related records together may actually introduce duplication across your systems[2].

Completeness: it is defined as the quotient of entities from M being represented by a tuple in r and the overall number of entities in M[2]. This kind of completeness may lead to data integration problem. Completeness can be achieved not just by deleting the tuples(if they are representations of entities from M) but by correcting the tuples which contains various anomalies.

Validity: to what extend the data is precise and understandable is referred as Validity.

Consistency: Do distinct occurrences of the same data instances agree with each other or provide conflicting information. Are values consistent across data sets?[7]

Uniformity: the data after cleansing process should not posses any inconsistencies and irregularities, i.e., the values within each attribute must be properly utilized. Uniformity is the quotient of attributes not containing irregularities in their values and n, the total number of attributes in r.[2]

Density: it is quotient of missing values in the tuples of r and the number of total values that ought be known because they exist for a represented entity.[2]

Uniqueness: it is the quotient of tuples representing the same entity in the mini-world and the total number of tuples in r[2]. A collection that is unique does not contain duplicates.

Generally, the quality of data is determined on how the data is received, integrated, processed, maintained and loaded in data warehouse which is done by the ETL tool(Extraction, Transformation & Loading). If all these steps are performed sequentially then we can achieve a better quality of data for data warehouse. Sometimes despite going through these phases a small percentage of error still exits. This residual of unclean data should be identified and reported so that the cause for failure is understood.

Data quality problems can occur in many different ways. The most common include[7]:

- 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.

IV. Process For Data Cleaning

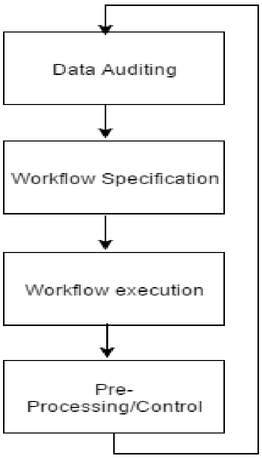

Fig: data cleaning process[2]

a) Data Auditing: It is the first step in data cleaning process where the data is audited to find different types of anomalies within it. Using Statistical methods data is audited and then parsing is used to detect syntactical irregularities. The instance analysis of individual attributes (data profiling) and the whole data collection (data mining) derives information such as minimal and maximal length, value range, frequency of values, variance, uniqueness, occurrence of null values, typical string patterns as well as patterns specific in the complete data collection. Once the entire data is audited the obtained results is beneficial for integrity constraints and domain formats. For future corrections in the irregularities auditing data in initial phase is must.

b) Workflow Specification: Detecting and removing various inconsistencies from data can be done using a sequence of operations on the data. This is called as data cleaning workflow. It is specified after auditing the data to gain information about the existing anomalies in the data collection at hand[2].The data cleansing process insist in the specification of the cleaning workflow where the elimination of all inconsistencies from the data are done. After cleaning process if one needs to modify the erroneous data , the type and the cause of anomalies should be know. During the cleaning process syntax errors are solved first because at the time of cleaning other types of errors and irregularities are detected and eliminated which is additionally hindered by syntax errors. Otherwise there is no specific order to remove the anomalies during the data cleansing workflow.

c) Workflow Execution: After a proper verification and specification of its accuracy the data cleaning workflow is executed. It should be efficient enough to be implemented on large data sets as well. This is often a trade-off because the execution of a data cleansing operation can be quite computing intensive, especially if a comprehensive and 100% complete elimination of anomalies is desired[2]. Therefore, we need heuristic methods in order to achieve a higher accuracy even at a low execution speed.

d) Post-Processing & Controlling: After execution of the cleaning workflow, the results are verified again to check the level of its accuracy with the specified operations. It may happen that during the cleaning process few tuples remain uncorrected, so they are corrected manually. Because of this a whole new cleaning cycle starts again, starting with the auditing phase followed by workflow specification and execution to remove the additional and remaining inconsistencies and further to clean the data by automatic processing.

V. Tools For Data Cleaning

a) AJAX: AJAX is an extensible and flexible framework attempting to separate the logical and physical levels of data cleansing. The logical level supports the design of the data cleansing workflow and specification of cleansing operations performed, while the physical level regards their implementation[2]. AJAX major concern is transforming existing data from one or more data collections into a target schema and eliminating duplicates within this process. For this purpose a declarative language based on a set of five transformation operations is defined. The transformations are mapping, view, matching, clustering, and merging.

b) **FraQL:** FraQL is another declarative language supporting the specification of a data cleansing process. The language is an extension to SQL based on an object-relational data model. It supports the specification of schema transformations as well as data transformations This can be done using user-defined functions. The implementation of the user defined function has to be done for the domain specific requirements within the individual data cleansing process. FraQL supports identification and elimination of duplicates. Also supported by FraQL is filling in missing values, and eliminating invalid tuples by detection and removal of outliers, and noise in data[2].

c) **Potter's Wheel:** Potter's Wheel is an interactive data cleansing system that integrates data transformation and error detection using spreadsheet-like interface. Error detection for the whole data collection is done automatically in the background. Potter's Wheel allows users to define custom domains, and corresponding algorithms to enforce domain format constraints. Potter's Wheel lets users specify the desired results on example values, and automatically infers regular expressions describing the domain format[2]. Therefore, the user does not have to specify them in advance.

d) **ARKTOS:** ARKTOS is a framework capable of modelling and executing the Extraction- Transformation-Load process (ETL process) for data warehouse creation. Six types of errors can be considered within an ETL process specified and executed in the ARKTOS framework. PRIMARY KEY VIOLATION, UNIQUENESS VIOLATION and REFERENCE VIOLATION are special cases of integrity constraint violations. The error type NULL EXISTENCE is concerned with the elimination of missing values. The remaining error types are DOMAIN MISMATCH and FORMAT MISMATCH referring to lexical and domain format errors[2].

e) **IntelliClean:** IntelliClean is a rule based approach to data cleansing with the main focus on duplicate elimination. The proposed framework consists of three stages. In the Pre- Processing stage syntactical errors are eliminated and the values are standardized in format and consistency of used abbreviations. The Processing stage represents the evaluation of cleansing rules on the conditioned data items that specify actions to be taken under certain circumstances[2]. During the first two stages of the data cleansing process the actions taken are logged providing documentation of the performed operations. In the human verification and validation stage these logs are investigated to verify and possibly correct the performed actions.

VI. Issues & Challenges To Clean Data

a) **Conflict Resolution and Error Correction**

One of the most challenging issues for data cleaning is the correction of values which must eliminate constraint violations, redundancies, domain format errors and invalid tuples. Due to insufficient availability of information and knowledge it is difficult to determine the correct modification of tuples which will remove these anomalies. By deleting these tuples it may lead to loss of information if the tuple is not invalid as a whole[2]. To avoid this loss of information mask the tuples with erroneous values and keep the tuple in the data collection until a proper information is available for error correction.

b) **Preservation of Cleansed Data**

Data cleaning is a time consuming process and requires more effort. Once we achieve a data collection free of errors by performing data cleaning process one would not want to repeat the entire cleaning process again, instead only the part that needs to be changed should go through the cleaning process again. The values which have been changed in the data collection, the cleansing process has to be done only for those tuples which contains the changed value which follows a sequential flow.

c) **Data Cleansing in Practically[Virtually] Integrated Environments**

The issues mentioned in the preceding section escalates while performing data cleaning in of virtually integrated environment sources, like IBM's Discovery Link [HSKK+01]. Propagation for corrections to the sources is often impossible in these environments as they are autonomous in nature. Therefore, whenever the data needs to be accessed the cleansing process needs to be performed, which leads to considerably decrease in the response time.

d) **Data Cleansing Framework**

Sometimes it may not be possible to depict the whole data cleansing process graph in advance, This will make the cleaning process more complex, iterative and primary task to perform. The whole data cleansing process is more the result of flexible workflow execution[2]. Process specification, execution and documentation of data cleaning within a framework should be co-related with other data processing activities such as transformation, integration, and maintenance activities. A framework consists of various methods for detecting and eliminating errors and also for auditing and monitoring data once the cleaning process has been performed.

VII. Approaches For Data Cleaning

The following are general approaches for data cleaning:

a) *Data analysis*: A complete data analysis is required to identify which kinds of errors, irregularities and inconsistencies from the data needs to be removed. Not only manual inspection on data or data samples should be done but also analysis programs should be used to get the properties of data of metadata and detect data quality issues.

b) *Definition of transformation workflow and mapping rules*: A data transformation process may consist of number of cleaning steps and each step performs schema/data and instance-related transformations(mappings). A data transformation and cleaning system can generate a transformation code only when the required transformation is specified in an appropriate language ,e.g., GUI, which will minimize the amount of self programming.

c) *Verification*: The transformation definitions and the accuracy and effectiveness of transformation workflow should be estimated, tested and evaluated, e.g. , on a sample of source data, in order to improve the definitions if required. Iterative steps needs to be performed in the analysis, design and verification phase, e.g., even after applying some transformations few errors may be noticeable.

d) *Transformation*: The transformation steps can be executed either by running the ETL process which is used for extracting, transforming and loading in data warehouse or during

e) Execution of the transformation steps either by running the ETL workflow for loading and refreshing a data warehouse or during firing queries on multiple sources.

f) *Backflow of cleaned data*: Once the errors, such as single-source errors, are removed the cleaned data should replace the unclean data in the original sources which will give legacy applications for the new data and will avoid of cleaning back the data for future data extractions.

VIII. Current Algorithms For Data Cleaning

Various researchers have proposed different types of techniques to clean the data. Some of them are mentioned below :

PNRS Algorithm: C. Varol et al. [12] have proposed PNRS algorithm which stands for Personal Name Recognizing Strategy. It corrects the phonetic and typographical errors present in the raw data, using standard dictionaries. It has two algorithms Near Miss Strategy and Phonetic Algorithm which correct words using standard Dictionaries:

(i) *Near Miss Strategy* – Two words are considered "near" if they can be made identical –
o By inserting a blank space
o By interchanging 2 letters
o By changing/adding/deleting a letter

If a valid word is generated using this technique, it is added to temporary suggestion list, which can be reviewed and corrected in the original data automatically or with some manual intervention.

(ii) *Phonetic Algorithm* - Phonetic Algorithm uses a rough approximation of how each word sounds. This is important as "near miss" doesn't provide us with the best list of suggestions when a word is truly mis-spelled. This compares the phonetic code of the mis-spelled word to all the words in the word list. If the phonetic code matches, then the word is added to the temporary suggestion list, which can be reviewed and corrected in the original data automatically or with some manual intervention.

Transitive Closure: Transitive Closure algorithm for data cleaning has been proposed by W.N. Li, et al [13]. This algorithm preprocesses the data to categorize millions and billions of records into groups of related data. The ETL tool using following algorithms processes the individual groups for data cleaning which involves

- Identifying and removal of redundancies - This is especially valuable when we are migrating data from different source systems where we might store same data in different formats
- Filling the blank cells.
- Establishment of "group" relationship between different records leading to faster querying.

In this technique the records are matched on the basis of matching of the keys (keys are selected attributes of the data). Each key is matched one after the other, so as to obtain related group of records. These groups can further be analyzed and corrected. Blanks can be filled, and redundancies can be removed.

Semantic Data Matching: Semantic Data Matching algorithm for data cleaning have been proposed by Russell Deaton , et al[14]which eliminates all the records in the dataset which have the same meaning but have different names. Using this method a more accurate result can be obtained.

Enhanced Technique: Enhanced Technique algorithm for data cleaning have been proposed by Dr. Mortadha M, et al[18] which offers the user interaction by selecting the rules and any sources and the desired targets. Each step from the algorithm is well suited for different purposes. 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.

IX. Conclusion

Data cleaning is primary task in a data warehouse management system. This paper reflects a survey work that aimed the issues and challenges in cleaning data. In order to achieve a clean data, improving the quality of data is also emphasized. Existing data cleansing approaches mainly focuses on the transformation of data and the elimination of duplicate values. In this paper we have also discussed the current algorithms that have been proposed by various researchers.

References

[1] Erhard Rahm, Hong Hai Do, "*Data Cleaning: Problems and Current Approaches*", IEEE, 2000.

[2] Heiko Müller, Johann-Christoph Freytag, "*Problems, Methods and Challenges in Comprehensive Data Cleansing*", Humboldt University Berlin, 2003.

[3] Arthur D. Chapman, "*Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*", version 1.0, Global Biodiversity Information Facility, Copenhagen, 2005.

[4] Nidhi Choudhary, "*A Study over Problems and Approaches of Data Cleansing/Cleaning*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Jonathan I. Maletic, Andrian Marcus, "*Data Cleansing: A Prelude to Knowledge Discovery*", Springer US, 2010.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Paulo Oliveira, Fátima Rodrigues, Pedro Henriques, Helena Galhardas, "*A Taxonomy of Data Quality Problems*", Journal of Data and Information Quality - JDIQ, 2005.

[7] Ranjit Singh, Dr. Kawaljeet Singh, "*A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing*", International Journal of Computer Science Issues IJCSI, Vol. 7, Issue 3, No 2, May 2010.

[8] Surajit Chaudhuri, Umeshwar Dayal, "*An Overview of Data Warehousing and OLAP Technology*", ACM SIGMOD Record, Volume 26 Issue 1, March 1997.

[9] Jonathan I. Maletic, Andrian Marcus , "*Data Cleansing: Beyond Integrity Analysis*", 2000.

[10] Kalaivany Natarajan, Jiuyong Li, Andy Koronios, "*Data Mining Techniques For Data Cleaning*", Proceedings of the 4th World Congress on Engineering Asset Management, Athens, Greece, 2009.

[11] Manjunath T.N, Ravindra S Hegadi, Ravikumar G.K, "*Analysis of Data Quality Aspects in Data WareHouse Systems*", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 2 (1) , 2010.

[12] Cihan Varol, Coskun Bayrak, Rick Wagner and Dana Goff, "*Application of the Near Miss Strategy and Edit Distance to Handle Dirty Data*", Springer, 2010.

[13] Johnson Zhang, Roopa Bheemavaram, Wing Ning Li, "*Transitive Closure of Data Records :Application and Computation*", Conference on Applied Research in Information Technology ALAR, 2006.

[14] Russell Deaton, Thao Doan, and Tom Schweiger, "*Semantic Data Matching: Principles and Performance*", Springer, 2010.

[15] Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma, “*HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses*”, Information Retrieval & Knowledge Management (CAMP), 2012 International Conference, 13-15 March 2012.

[16] Prerna S. Kulkarni, J.W. Bakal ,"*Hybrid Approaches for Data Cleaning in Data Warehouse*"**,** International Journal of Computer Applications (0975 – 8887) Volume 88 – No.18, February 2014.

[17] Ashwini M.Save,Seema Kolkur, "*Hybrid Technique for Data Cleaning*", International Journal of Computer Applications, 2014.

[18] Dr. Mortadha M. Hamad, Alaa Abdulkhar Jihad ,"*An Enhanced Technique to Clean Data in the Data Warehouse*", Developments in E-systems Engineering (DeSE), 2011.