Volume 6, Issue 3, March 2016 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 Data Cleansing and Its Tools

[1] Shivangi Rana, [2]Er. Gagan Prakesh Negi, [3]Kapil Kapoor
[1]Research Scholar, CSE Department, [2]Assistant Professor, CSE Department, [3]Associate Professor, ECE Department,
[1, 2, 3]Abhilashi Group of Institutions (School of Engineering) Chail Chowk, Mandi, Himachal Pradesh India

***Abstract:** Data cleansing is about more than good housekeeping, removing duplicate or obsolete data and correcting inaccurate information. In today's climate of data protection and financial pressure on marketing budgets the necessity for cleansed and accurate information is greater than ever. Keeping databases cleansed so that they contain the most accurate and up to date records has always been important. Data cleansing and it's applications can be viewed as one of the emerging and promising technological developments that provide efficient means to access various types of data and information available worldwide. Not only this, these applications also aids in decision making. The paper provides an overview of data cleansing process, data quality, data cleansing methods and comprehensive and theoretical analysis of some data cleansing tools.*

Keywords - ***Data Cleansing, Data Cleansing Tools, Data Quality***

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.**[1]** 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.**[2]** 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. The main objective of data cleaning is to reduce the time and complexity of mining process and increase the quality of datum in data warehouse

II. 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. Data must conform to the set of quality criteria.[3]

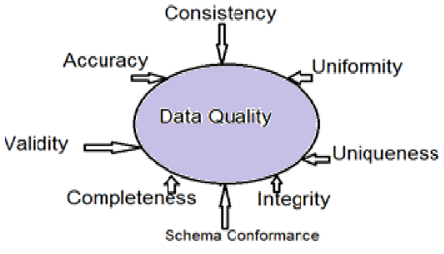

Figure 1: Set of criteria of data quality.

III. 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,[4]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.

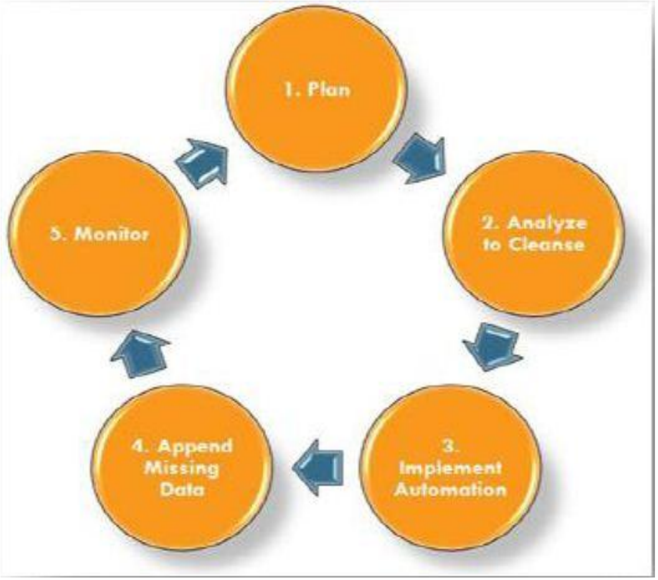

Fig: 2 Data Cleaning Process

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.

IV. DATA CLEANING METHODS

A. *Data cleaning methods for missing value:*

- ***Ignore the tuple*: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
- ***Fill in the missing value manually*: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
- ***Use a global constant to fill in the missing value*: Replace all missing attribute value by the same constant, such as a label like *"Unknown"* or -infinity. Missing values are replaced by, say, *"Unknown,"* then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of *"Unknown."* Hence, although this method is simple, it is not fool proof.
- ***Use the attribute mean to fill in the missing value***:We can use mean of all numbers to fill in the missing value For example, suppose that the average income of customers is $56,000. Use this value to replace the value for *income*.
- ***Use the attribute mean for all samples belonging to the same class as the given tuple:*** For example, if classifying customers according to *credit risk*, replace the missing Value with the average *income* value for customers in the same credit risk category as that of the given tuple.
- ***Use the most probable value to fill in the missing value:***This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.**[5]**

B. Data cleaning methods for Noisy Data

"What is noise?" Noise is a random error or variance in a measured variable. Given a numerical attribute such as, say, *price*, how can we "smooth" out the data to remove the noise? There are following data smoothing techniques:

a. *Binning:* Binning method smooth a sorted data values by consulting its neighbourhood that is values around it. The sorted values are distributed into a number of "buckets" or Bins.Because binning methods consult the neighbourhood of the values, they perform the local smoothing.Figure 3 illustrates some binning techniques.In this example, the data for price are first sorted and then portioned into equal frequency bins of size 3(that is, each bin contains three values). In smoothing by bin means, each value in a bin is replaced by mean value of the bin.For example the mean of the values 4,8 and 15 in Bin1 is 9.Therefore each original values in this bin is replaced by value 9. Similarly, smoothing by bin medians can be employed,in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as bin boundaries.Each bin value is then replaced by the closest boundary value. In the figure 3 the boundaries values of Bin1 are 4 and 15, so 8 is replaced by closest boundary value 4. Similarly in Bin2 the boundary values are 21 and 24,so the middle value 21 will remain same,since it is equivalent to the boundary vale 21. In Bin3 the boundary values are 25 and 34, so the middle value 28 will be replaced by nearest boundary value 25.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure 3: Binning Methods for Data Smoothing

b. Regression: Data can be smoothed by fitting the data to a function, such as with regression. *Linear regression* involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. Figure 4 shows the regression. Here linear regression is shown by two variables X and Y. *Multiple linearregression*is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

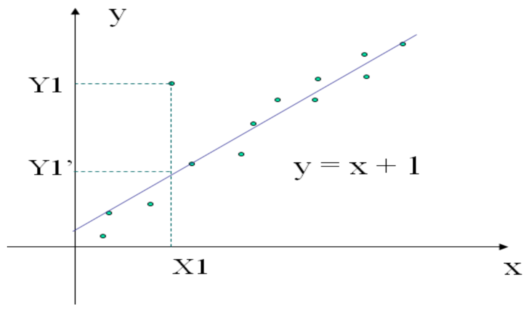

Figure 4: Regression

C. Clustering:

Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers **[5]**. Figure 5 shows clusters.

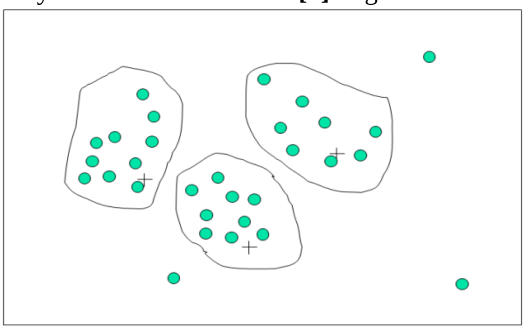

Figure 5: Cluster Analysis

V. A BRIEF OVER VIEW OF DATA CLEANSING TOOLS

Data cleansing is a valuable process that can help companies save time and increase their efficiency. Data cleansing software tools are used by various organisations to remove duplicate data, fix and amend badly-formatted, incorrect and amend incomplete data from marketing lists, databases and CRM's. They can achieve in a short period of time what could take days or weeks for an administrator working manually to fix. This means that companies can save not only time but money by acquiring data cleaning tools.

Within data cleansing, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences

The development and application of data cleansing algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult.

Tools available for data cleansing are briefed as below:

1. Winpure Clean and Match:

WinPureClean & Match is a very powerful software that will ensure clean, duplicate free, professional looking lists and databases, with its amazing suite of list cleaning modules, together with a powerful deduplicatation engine and intelligent merge options. [8]

WinPureClean & Match is an invaluable application that could save your business lots of time and money. With its revolutionary easy-to-use interface, together with its huge array of list cleaning options, it can be used by anyone and with virtually any size of list or database. **WinPureClean & Match** have some extraordinary features. Some of the benefits and features available on **WinPureClean & Match are:-**

Benefits

- **Increase the accuracy** of your customer data; process both business and personal data from any country.
- **Reduce marketing postage costs** by eliminating duplicates from your database, CRM, spreadsheets, etc.
- **Improve virtually any type of list** (contact lists, e-mail lists, prospects, etc.) from a variety of list sources such as Excel, Access and text files generated from Outlook or CRM systems.
- **Improve your company image** by using standardized styles for names and addresses with error-free data.
- **Reduce your team's dependence on the IT staff**—WinPureClean & Match can be used directly by marketing professionals, not just technical IT people.
- **Preventing bouncing emails** with more accurate email addresses.

Features

- **Revolutionary User friendly interface** with extensive help options and tutorials.
- **7 Powerful** List / Data Cleansing Modules.
- **Clean/Match/Merge/Purge** one or two lists.
- **Advanced deduplication engine with phonetic fuzzy matching** (Bob/Robert, Food Limited/Food ltd, Part Street/Park St, etc).
- **Correct invalid email addresses** (WinPure***Clean & Match*** will give suggestions to bad emails).
- **Identify missing data** with scoring system, ideal for fully populating name & address details.
- **Automatically standardize name and addresses** (For example: john mcneal John McNeal).
- Automatically split names, email addresses, telephone numbers and even words into separate columns.
- Automatically merge/concatenate different columns.

2. Rapid Minor:

Rapid Minor is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development. . We can also clean our data using[6] RapidMinor.This tool contain various operators for data cleaning or data cleansing. It released on 2006, latest version available is RapidMinor 6. It can be installed on any operating system that is cross platform, Language independent, Licensed by AGPL proprietary. Rapid minor support about twenty two file format . It easily reads and writes Excel files and different databases. Using RapidMinor we can clean, transform our data. Some of the features and benefits of Rapid Minor are :

Features:

- It represents a new approach to design even very complicated problems by using a modular operator concept which allows design of complex nested operator chains for huge number of learning problems.
- Rapid miner uses XML to describe the operator trees modeling knowledge discovery process.
- It has flexible operators for data input and output file formats.
- It contains more than 100 learning schemes for regression classification and clustering analysis.

- Rapid miner supports about twenty two file formats. [7]
- Rapid Miner has a lot of functionality, is polished and has good connectivity.
- Solid and complete package.
- It easily reads and writes Excel files and different databases.
- You program by piping components together in a graphic ETL work flows.
- If you set up an illegal work flows Rapid Miner suggest Quick Fixes to make it legal.

Benefits:

- Has the full facility for model evaluation using cross validation and independent validation sets.
- Over 1,500methods for data integration, data transformation, analysis and, modelling as well as visualization – no other solution on the market offers more procedures and therefore more possibilities of defining the optimal analysis processes
- .RapidMiner offers numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.

Excel Data Cleaner:

Excel Files Data Cleaning Utility is an useful addin for Excel to Clean and Organize Data. It is fast & Reliable and you can save your precious time & Money. [9] Excel has many features to help you get data in the precise format that you want. Sometimes, the task is straightforward and there is a specific feature that does the job for you. It provides the following features:

Features:

- ***Spell checking:***You can use a spell checker to not only find misspelled words, but to find values that are not used consistently, such as product or company names, by adding those values to a custom dictionary.
- ***Removing duplicate rows***:Duplicate rows are a common problem when you import data. It is a good idea to filter for unique values first to confirm that the results are what you want before you remove duplicate values.
- ***Finding and replacing text*:**You may want to remove a common leading string, such as a label followed by a colon and space, or a suffix, such as a parenthetic phrase at the end of the string that is obsolete or unnecessary. You can do this by finding instances of that text and then replacing it with no text or other text.
- ***Changing the case of text*:**Sometimes text comes in a mixed bag, especially when the case of text is concerned. Using one or more of the three Case functions, you can convert text to lowercase letters, such as e-mail addresses, uppercase letters, such as product codes, or proper case, such as names or book titles.
- Removing spaces and nonprinting characters from text
- Fixing numbers and number signs
- Merging and splitting columns

Benefits:

- Save time by automating your editing
- Eliminate manual errors
- Create "edit rules" that you can apply over and over again
- Remove excess blanks, html "spaces", and non-keyboard characters
- Convert text to any case
- Convert "text" numbers to real numbers

VI. CONCLUSION

After a concise study, we came to a conclusion that Data cleansing is applied with different intensions and within different areas of the data integration and management process. It defined it as the sequence of operations intending to enhance to overall data quality of a data collection. Data cleaning methods and approaches depend upon the type of data which we want to clean and according to that we apply particular methods. This paper also present some of data cleansing tools, these tools provide nice graphical interfaces, focus on the usability and interactivity. They provide flexibility either through visual programming within graphical user interfaces or prototyping by way of scripting languages. In future work features of these tools can be extended further to check other data anomalies and improving the data cleansing results**.** So that data cleansing results can be improved, which help in better decision making.

REFERENCES

[1] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: *Tools for Data Translation and Integration*. In [26]:3-8, 1999.

[2] Li Lee Mong , *Cleansing Data for Mining and Datawarehousing*, school of computing National University of Singapore, 1999 .

[3] Rahm E. & Hai Do Hong, *Data Cleaning: Problems and current approaches*, IEEE Bulletin of the Technical Committee on Data Engineering, 2000
[4] Müller Heiko & Christoph Freytag Johann , *Problems, Methods, and Challenges in Comprehensive Data Cleansing* ,Humboldt-Universität zu Berlin zu Berlin,10099 Berlin, Germany
[5] HansJaiweiand KamberMicheline "Data Mining Concepts and Techniques", Second Edition.
[6] www.rapidminor.com
[7] Mikut Ralf & Reischl Wiley Markus Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011
[8] www.winpure.com
[9] www.exceladin.net

