Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Analysis of Data Cleansing Tools

Shivangi Rana
Research Scholar
CSE Department
Abhilashi Group of Institutions
(School of Engg.)
Chail Chowk, Mandi, India

Er. Gagan Prakesh Negi
Assistant Professor
CSE Department
Abhilashi Group of Institutions
(School of Engg.)
Chail Chowk, Mandi, India

Kapil Kapoor
Associate Professor
ECE Department
Abhilashi Group of Institutions
(School of Engg.)
Chail Chowk, Mandi, India

Abstract: ***What should be kept in mind is that data cleansing is not an easy process. Not only is it time-consuming and requires a considerable amount of work, but also the expense of it is significant. This may be the reason why some organizations underestimate the importance of data cleansing, which can lead to numerous business failures as well as adverse effects caused by inaccurate or inconsistent databases. Problematic data can lead users to distrust the very applications they rely on to make marketing and sales decisions. The only way to reverse this situation is to "clean" your data. But writing code to do that can be time-consuming and costly. Fortunately, there are a number of data quality methods that will clean your data for you. This paper presents a comparison and analysis of data cleansing tools and features and benefits of each tool. Additionally it present comparative analysis of data cleansing tools and determine the best one.***

Keywords: ***Data, Data cleansing, Data Quality, Databases.***

I. INTRODUCTION

Data cleansing, as the term suggests, is exactly that—a database cleaning process that involves the removal and/or correction of "dirty data" from said database. When data is stored using any type of process, certain errors are inevitable. Once these errors enter the system, irregularities are bound to happen and "dirty data" (i.e., "*data that is incorrect, out-of-date, redundant, incomplete, or formatted incorrectly")* is born—potentially threatening to pollute a clean database. In essence, the goal of data cleansing is to minimize these errors and prevent or eradicate dirty data. In an attempt to keep all data as useful and as up-to-date as possible, the process of data cleansing usually involves a read-through of a set of records to verify the accuracy of each. Also referred to as *data scrubbing*, this process is very important in maintaining a smooth workflow for data-dependent businesses. It is a valuable process that allows companies to save time and money, and at the same time increase the efficiency of their transactions. Data cleaning, data cleansing or data scrubbing as it may be called is the automatic batch process of correcting or removing data from your database or mailing files that is incorrect, inaccurate, out-of-date, incomplete, or wrongly formatted.

If some of the clients within a database do not have accurate phone numbers, for example, employees cannot easily contact them. If clients' email addresses are not formatted correctly, as another example, an automated email system would be unable to send out the latest coupons and special deals.

As such, data cleansing is crucial to organizations that deal with a large amount of data: banks, government offices, SMEs, and many other types of data-heavy businesses. Data management professionals also encourage these firms to actively invest in cleansing tools that prevent any sort of decline in data efficiency caused by a mismanaged database or partner with a company that offers outsourced research services. Incorrect or inconsistent data can create a number of problems which lead to the drawing of false conclusions. Therefore data cleaning can be an important element in some data analysis situations. However, data cleaning is not without risks and problems including the loss of important information or valid data.

There are a large variety of tools available that can be used to support data cleaning. Additionally, many statistical programs have data validation built in, which can pick up some errors automatically.

II. DATA CLEANING

Manual Data Cleansing

If done manually, the process involves a person deliberately combing through a pile of data to correct typos and spelling errors, properly label and file all mislabeled data, and carefully supplying missing entries in incomplete files. This manual process would also entail the eradication of out-of-date records so that they do not disrupt the current workflow or occupy space that can otherwise be allotted to new and relevant data.

Electronic Data Cleansing

In more complicated scenarios that involve complex operations, data cleansing is usually performed by computer applications that are programmed to follow a set of rules and procedures that are initially determined by the user. These

rules are often based on a specific purpose or set of purposes, e.g. to delete records that have not been accessed or updated within a certain time frame, to perform spell checks, or to get rid of duplicate copies. More sophisticated programs are even capable of filling in missing data or change them based on a certain preset.

Data cleansing software tools are often used by various organizations to fix and improve badly formatted data from marketing lists and CRMs. Through this, they are able to quickly achieve results that could otherwise take days or weeks should the process be carried out manually. Needless to say, companies can save not only time but money by investing in data cleaning tools.

Managing big data has never been easier—thanks to the ever-evolving digital technology we have this day and age. Businesses need to know how to properly handle, analyze, encode and store the raw data they've gathered to convert it into valuable information vital to their operation. Leave all your big data concerns to us and we will address them through our data management services, allowing you to focus on other important business matters.

III. IMPORTANCE OF DATA CLEANSING SOFTWARE

In this day and age, many businesses seek to cut down costs and adjust their spending habits, which is why their interest in acquiring new software is quite low. Still, in case you really seek to adjust spending habits for your company, an investment you need to consider is data cleansing software, as reliable data cleansing tools can save your business thousands of dollars yearly.

A problem that companies have to confront is linked to duplicated data that can generate useless spending for your company. The purpose of data cleansing software is to eliminate such duplicated data and avoid errors like misguided funds. According to studies, companies end up losing over 6 million dollars yearly only because they do not employ proper data cleansing tools. By employing reliable data cleansing software, the quality of stored data is greatly improved. What can happen when you are not using data cleansing tools is to increase the volume of duplicated data, despite only slight variations being made between an old version and a new one.

There are many ways in which you can lose money because of duplicated data. For instance, the same marketing materials being sent to the same people will be nothing by wasted money. Your clients' interest will decline and, in the end, they will just ignore any emails and notifications coming from you. Wasting money is not good for your business. By appealing to reliable data cleansing software, you will ensure that your hard earned profits will not be melted away by poor marketing decisions, taken as a result of duplicated data. What you need are data cleansing tools that can help you keep your data up to date.

The main purpose of data cleansing software is to remove all the duplicated records and to identify abnormal data records that must be removed manually. Also, good data cleansing tools can remove invalid records and leave only valid data to be used for company activities and decisions. Also, you will be able to use data cleansing software in order to increase the accuracy of the data generated by your company. Accurate data will eliminate the possible mistakes and errors in the future. You will never take decisions based on erroneous information, and you will not end up losing money because of duplicated data.

Another aspect that data cleansing software can help you with is to keep your business on track, by supervising regular data entry activities. These ongoing tasks must be monitored all the time, if you want to eliminate the risk generated by erroneous data. Maintaining your databases clean of unnecessary data will definitely contribute to the success of your business, and will also save you a lot of money.

IV. DATA CLEANSING TOOLS

Every good business decision is backed by reliable & accurate data. Data quality issues generally arise when anomalies are found in the database. The problem occurs when you integrate data from different sources into one single data source. Object Identity problem is the main trouble with data and it affects the quality of the database. Data cleansing tools & data cleansing techniques help to tackle this problem effectively. Most of the organizations these days rely heavily on the information stored in their database. Duplicate elimination, data warehouse construction activities can be effectively achieved by using a good data cleaning tool. These errors occur due to mistyping of the words or when the data is collected from various sources. This at times leads to duplicate entries and brings down the inconsistencies within the database. There are powerful, easy-to-use and affordable data cleaning tools that will help to tackle the problem of bad quality data.

A. Rapid Miner:

It is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, data cleaning, 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 and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures. We can also clean our data using RapidMiner. This tool contains various operators for data cleaning or data cleansing. It released on 2006, latest version available is RapidMiner 6. It can be installed on any operating system that is cross platform, Language independent, licensed by AGPL proprietary. Rapid Miner support about twenty two file format. It easily reads and writes Excel files and different databases. Using Rapid Miner we can clean, transform our data.The operators for data cleaning are:-

- Attribute name and Role Modification Operator
- Type conversion

- Attribute set reduction & transformation
- Value Modification
- Outlier Detection
- Filtering
- Sorting
- Rotation
- Aggregation
- Set operation

Using 'Attribute name and role modification' operator we can change the name and role of operator, 'Type conversion' operator can change the data type of attribute to another data type. 'Attribute set reduction & transformation 'further consist of 19 'Generation', 7 'Transformation' and 14 'Selection' operators. 'Value Modification' operator consist of 'Numerical Value Modification',' Date Value Modification and ' Nominal Value Modification' operator, using these operator we can modify the values. 'Outlier Detection' it provides further operators 'Detect Outlier (Distance)','Detect Outlier (Densities)','Detect Outlier (LOF that is Local Outlier Factor)', 'Detect Outlier (COF that is Class Outlier Factor)'.Using 'Filtering' operator we can remove the duplicates. Data can be sorted in ascending or descending order by using 'Sorting' operator.
Transpose, Pivoting, De-Pivot can be implemented on the data using 'Rotation' Operator. 'Aggregation' operator can aggregate the data. 'Set Operations' provides Append, Join, Set Minus, Intersect, Union, Superset and Cartesian operators.

1) Technical specification:

- Released on 2006
- Latest version available is Rapid miner 6.
- Licensed by AGPL Proprietary
- Cross platform i.e. can be installed on any operating system
- Language Independent
- Can be downloaded from www.rapidminer.com.

2.) Features:

- Rapid miner supports about twenty two file formats.
- 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.
- 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

3.) 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, modeling 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.

B. Winpure Clean and Match:

WinPure Clean & Match is the complete list cleaning, data cleansing and data deduplication software, all rolled into one powerful easy-to-use application. Clean your mailing lists, marketing databases, spreadsheets and emails with its five unique list/data cleaning modules, then perform a powerful data deduplication and merge/purge on one or two lists to ensure they do not contain any duplicates. **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. WinPure Clean & Match is split into 3 sections (**DATA, CLEAN, MATCH**). The **DATA** section allows you to import one or two lists, the **CLEAN** section contains 7 separate list/data cleansing modules, and the **MATCH** module is for matching/deduplicating one or two lists. **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.

1) Technical specification:

- Released on 2009
- Latest version available is Winpure clean and match(2012) 6.2.8
- Licensed by shareware
- Used with Windows XP-SP2 / 2003 / Vista / Windows 7 &8
- Can be downloaded from www.winpure.com.

2.) Features:

- Revolutionary User friendly interface with extensive help options and tutorials.
- 7 Powerful and unique List / Data Cleansing Modules.
- Clean/Match/Merge/Purge one or two lists.
- Advanced deduplication engine (Bob/Robert, Food Limited/Food ltd, Part Street/Park St, etc) find duplicates you never thought existed.
- Correct invalid email addresses (WinPure Clean & Match will give suggestions to bad emails).
- Identify missing data with scoring system and graphs, ideal for fully populating name & address details.
- Automatically standardize name and addresses (eg. john mcneal > John McNeal, Ibm > IBM, etc)
- Identify missing data, ideal for fully populating name & address details

3.) Benefits

Improve the Performance of Your Marketing Efforts. By maintaining clean customer lists, WinPure Clean & Match will help you:

- Increase the accuracy of your customer data: business or consumer, local or international.
- Reduce marketing postage costs by eliminating duplicates from your database using advanced deduplication and intelligent merge/purge.
- 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 clean, correct, fully populated and duplicate-free data.

V. COMPARATIVE STUDY OF TOOLS

After the study of Data Cleansing Tools we have analyzed features and technical specification of these tools and also the modules which are used by these tools for data cleaning. Rapid Miner and Winpure Clean&Match have different modules for performing different functions according to the requirement of the customer.Analytical study was made by taking into account technical specifications and feature.

Table 1: Technical Overview of Data Cleansing Tools

S.N	Tool Name	Release Date	Release date/ Latest version	License	Operating System	Website
1	Rapid Miner	2006	21November,2013 /6.0	AGPL Proprietary	Cross platform	www.rapidminer.com
2	Winpure Clean & Match	2009	Winpure Clean&Match (2012)6.2.8	Shareware	Windows XP-SP2 / 2003 / Vista / Windows 7 &8	www.winpure.com

The table shown gives the technical overview of the tools which includes name of the tool, description of release date, latest version release date, license, operating system and official website.

Table 2: Analytics of feature of Data Cleansing Tools

S.N	Tool Name	Type	Features
1	Rapid Miner	Statistical analysis, data mining, predictive analytics, Data Cleaning	•More than 20 new functions for analysis and data handling, including multiple new aggregation functions •File operators to operate directly from Rapid Miner •A macro viewer that shows macros and their values in real time during process execution • Intuitive GUI
2	Winpure Clean&Mach	Data Cleaning, Data Matching/ Deduplication, Data Merging	•Revolutionary User friendly interface •7 Powerful List / Data Cleansing Modules. •Advanced deduplication engine with phonetic fuzzy matching •Correct invalid email addresses •Clean/Match/Merge/Purge one or two lists

The given table describes the basic features and functionality provided by these tools

VI. RESULTS AND DISCUSSIONS

Table. 3 represent the excel files ,cleaned using data cleaning tools , Customer Data file is cleaned using "RapidMinor", it contain 7 % missing values and 21 duplicate records. NGO Data file is cleaned using "Winpure clean & Match", it contains 25% missing values and 10 duplicate records.

Table 3 Files used and Properties

File Name	No. of records	No. of Fields	Missing Value	Duplicate Records
Customer Data	840	13	7%	21
NGO Data	410	9	25%	10

The experiment was performed on two Data Cleaning tools. Data anomalies are detected using these tools that are duplicate data, missing values, Illegal values etc. Two tools Rapid Minor and Winpure clean & Match are used for data cleaning. Two Excel dataset are tested on these tools. Comparison of results is done. Data of the files mentioned in the table is imported in to the data cleansing tool and different modules of the tool are used to clean the dirty data , missing values, duplicate record are eliminated using these tools

Table 4 The results of comparison between two frameworks

Tools / Problems	Rapid Minor	Winpure Clean & Match
Availability	Desktop	Desktop
Missing Values	Yes	Yes
Duplication	Yes	Yes uses the matching
Illegal Values Elimination	No	Yes
Varying values representation	No	Yes
Misspelling	No	No
Merge	Yes	Yes
File Format	CSV, Database, Excel, Access, binary, XML	Text files, Excel , commercial DBMS,
Ease of Use	Moderate	High

Table 4 shows the results of comparison between the two above frameworks from several aspects, the most significant (Missing values, Availability, Duplication, Illegal values elimination, Misspellings, Varying value representations, File formats, Ease of usage). "Rapid Miner" deals with missing values by using its 'Replace Missing Value' operator. This operator replaces missing values in the selected attributes by a specified replacement. Missing values can be replaced by the minimum, maximum or average value of that attribute. "WinPure Clean & Match 2012" deals with missing values by using another table as a master to fill the missing values. Missing values of a single table handling is not supported. Additionally, "WinPure Clean & Match 2012" deals with Illegal values elimination by Range Constraints, Regular expression patterns and Unique Constraints

VII. CONCLUSION

Data cleansing has become a major activity performed by most organizations that have data warehouses. Every organization needs quality data to improve on its services it renders to its customers. The development and application of data cleansing techniques requires the use of right choice of software tools. This paper provides information about two data cleansing frameworks, how they deal with dirty data and use its different modules for this purpose. From the above comparative study we have concluded that both the tools have their own specific features but out of those Winpure clean and match is widely used for data cleansing purpose. This work can be a helping hand to provide an insight in future to develop an application with more efficiency and availability i.e. a tool can be designed which instead of supporting a specific area can be extended to more fields.

REFERENCES

[1] Rahm E. & Hai Do Hong, *Data Cleaning: Problems and current approaches*, IEEE Bulletin of the Technical Committee on Data Engineering, 2000

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

[3] Tamraparni Dasu and Theodore Johnson. Exploratory Data Mining and Data Cleaning. Wiley,2003.

[4] Cohen, W.: *Integration of Heterogeneous Databases without Common Domains Using Queries Based Textual Similarity*. Proc. ACM SIGMOD Conf. on Data Management, 1998

[5] Ibrahim Housien Hamed, Zuping Zhang & Qays Abdulhadi Zainab,*A comparison study Of Data Scrubbing algorithm and framework in Data Warehousing*,,*International Journal of Computer Applications (0975 – 8887) April 2013*

[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] Mikut Ralf & Reischl Wiley Markus Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011

[8] www.rapidminor.com

[9] www.winpure.com

