Data Conditioning and Comparison of Data Cleansing tools

Research Article

Jahnavi Prathipati
George Mason University, VA
Sooraj Cheeti
George Mason University, VA
Sai Hemanth Nirujogi
George Mason University, VA
Siva Swetha Yalamanchili
George Mason University, VA

***Abstract*—With ever increasing technologies in the world, people will be more concerned about the data that grows with it. Data conditioning plays a crucial part in analytics. Data analysis requires clean and error-free data. An entire organization's marketing and sales decisions depend on the data used for analysis. Incorrect or problematic data might lead to false predictions and decisions. The only way to alter this situation is to clean the data. Data cleansing is not an easy process, it is a time consuming and requires a lot of work. This paper presents and studies all the data conditioning tools in the market and compares the features and benefits of each tool to determine the best one.**

Keywords*—Data Conditioning, Data Cleaning, Data cleansing tools

I. Introduction

Data conditioning is the use of data optimization and data management techniques to filter the raw data and securing it for storage in a system. It helps an organization's data center to improve system performance and utilization which in turn lowers the operating costs and capital expenditures.

Data cleaning is also referred to as Data cleansing and Data scrubbing. Data cleaning is the process of altering the data either by correcting or removal of noise from the data to make sure that the data on the whole is appropriate. There are a lot of ways to clean the data. It can be done manually or electronically. Regardless of the method chosen the main criteria of data cleaning is to achieve high quality of data.

High quality data should pass a set of quality criteria which include Validity, Accuracy, Completeness, Consistency, and Uniformity. Data quality is the key to success. Data readiness is used to describe the status of the data. It qualifies the readiness of the data we are going to use. To provide a sense of state of our data, Data readiness levels are used. [1]

The world of business depends on the data which ever shape or form it is in. The data can either be the information about all the customers or regarding the sales or something else. If this data is used correctly it could be a great help for the organization to improve the business and make profits. To make sure of this, the companies rely on data cleaning which helps them to save time, increase their efficiency and make more money.

II. Data Quality

A. Definition

The quality of data is generally considered high when the data is suitable for business operations, customer management, marketing, decision making and planning in an organization.

B. Dimensions of Data Quality

Most of the organizations select their dimensions or attributes based on their requirements and level of risk etc. Each dimension is assigned with different weights based on the importance of that dimension for the organization.

The six main important dimensions of data quality are: [2]

1. *Completeness* of data is the percentage of stored data to the expected attributes of data provided.
2. *Uniqueness* of the data is achieved when there are no duplication or repeatedness is identified.
3. *Accuracy* is the condition of quality of being error-free, correct and exact from the real world object.
4. *Timeliness* is the degree to which the data reflects reality from the required event of time.
5. *Validity* is the correctness and reasonableness of the data.

6. *Consistency* of the data represents how the process of keeping data is uniform over its life cycle.

Figure-1: Dimensions of Data Quality

III. Data Quality Problems

Poor data quality can lead to wastage of time and resources. It results in poor decision making and marketing. The data quality problems are categorized into two parts. They can be single source problems or multisource problems.

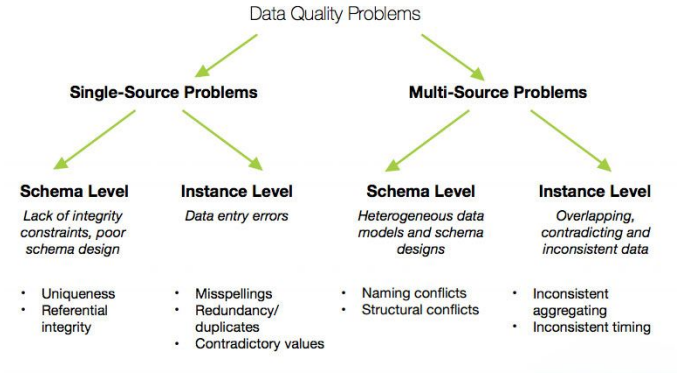

Figure-2: Data Quality Problems [4]

C. Types of Data Quality Problems [3]

- *Single Source Problems*
  - *Schema level problems* that occur due to lack of application specific integrity constraints inside standardization of database. For example, values that are not in the domain range, violation of uniqueness of the data, not defining the reference fields etc.
  - *Instance level problems* that occur due to inconsistencies, errors and null values in the contents of the dataset. Instance level problems are not visible to schema level problems. For example, misspellings, missing or null values, duplicate values, duplicate contacts with different value, etc.
- *Multi Source Problems* are occured when many datasets merge together in a warehouse, the need for data cleaning is also increased. This happens due to the datasets mostly contain the same data with different attributes and subsets.
  - *Schema and instance* problems are recorded, it means that data is affected with inaccuracies along with some structural issues. The schema problems are mostly the cause of instance problems, which means the structure of the data is corrupted.

D. Data Quality Assurance

This has 3 main activities

- *Data Profiling:* This process involves exploring the data, identify the data quality issues and summarize the data in table row counts, null values etc.
- *Data Cleansing:* It involves fixing the data issues like misspellings, typos, incorrect formats, etc.
- *Data Monitoring:* The data is maintained in a clean state. Users check the business rules first and submits the data quality issues to the data store.

IV. Data Cleaning Process

The data cleaning process circles in 4 steps: [5]

- *Data Auditing*: The first step is to audit the data to find the types of anomalies inside it. Statistical methods and parsing of data are done to detect anomalies. The results are used to specify the integrity constraints and domain formats. Each and every constraint is checked to identify possible violating tuple.

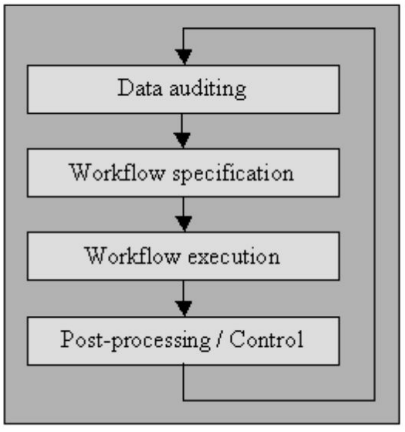

Figure-3: Data Cleaning Process

- *Workflow Specification:* Common order problems are eliminated by operating multiple time over the data. This is called data cleansing workflow. It is specified to obtain the information about the existing anomalies that are in the data. The specification of cleansing workflow is applied to the raw and unstructured data automatically removing the anomalies remains as the main challenge in Data cleansing.
- *Workflow Execution:* The cleansing workflow is executed automatically after the specification and verification of the accuracy of data.
- *Post-Processing/Control:* The results are verified once again for accuracy after executing the cleansing workflow.

E. Methods of Data Cleaning

Data cleaning can be done in one of the two methods

- *Manual Data Cleaning*: This method can be considered as an old hat as this is one of the redundant methods. Manual data cleaning is labor intensive as it required a person to go through all the data and manually make the changes to the data. Look at every detail in the data and make appropriate changes. This method can get expensive because a lot of people should be hired to clean the data in this process.
- *Automated Data Cleaning*: This process of data cleaning involves operation which are performed by the computer to clean the data by performing certain operations which are predefined by the users. In the automated data cleaning technique programs are written by the user which are used to solve a problem or a set of problems like spell check or deleting the files from a particular day.[6]

This method can be accurate and less costly. The automated data cleaning tools are often used by the organizations to improve badly formatted data from the marketing lists and CRM'S. It is not an easy task to manage big data, but it has become easier because of all the tools that are available for us now. Business these days should be properly handled, analyzed and all the raw data that is available should be stored properly and should be converted into valuable information than can be helpful in making the business more successful. All this can be done very easily using the automated data cleaning tools. Few tools that are available in the market right now are RapidMiner, Trifacta, Openrefine.

V. Importance of Data Cleaning Tools

Data cleaning is the most valuable process to the organization as the decisions are made analyzing the data. There are various kinds of data cleaning tools present in the market which are been used by various companies for the purpose of removing duplicate data, fixing the incorrect and incomplete values from the databases. These data cleaning tools save money by omitting bad records and deals with inaccuracies within seconds saving a lot of time and effort of the organization. According to studies organizations end up wasting upto 6 million dollars only because of no proper data cleaning tools.[7]

Data cleaning is very important to an organization which deals with data irrespective of the size of it. These organizations may include banks or other government enterprises, small to medium enterprises dealing with data and organizations in the field of marketing. In fact every organization dealing with data needs to have a data cleaning software. These tools should be used in regular basis as duplicate and incorrect data can grow quickly leading to decrease in business efficiency.[7]

The data cleaning software help to correct incorrect data or duplicate data such as the emails, phone numbers of the customers resulting in contacting them. One of the most preferred example in this scenario is due to duplicate data the marketing organizations may send multiple marketing advertisements to the same customer due to duplicate data resulting in losing the customer. These type of errors can be controlled by the organization within no time and save valuable resources of the organization as it would take much longer time to delete the records manually.

VI. Data Cleaning Tools

RapidMiner: This is a software platform mainly used for the phases of data science projects like data preparation, predictive analytics, machine learning, business analytics. There will be usage of two main data types in RapidMiner. They are- Nominal (Binomial, Polynomial) and Numeric (Integer, Data- time, Real). There four main tools to carry out each phase in the RapidMiner. They are-

- RapidMiner Studio
- RapidMiner Auto model
- RapidMiner Server
- RapidMiner Radoop

RapidMiner Studio: RapidMiner studio works on visual workflow designing process to improve the productivity of the data science compared to rapid prototyping of the ideas used for complex predictive models. RapidMiner studio improves the validation and prototyping of the predictive models rapidly. This happens due to the predefined connection, templates and reusable workflows. The database has more

than 1500 algorithms and predefined functions, this helps to construct very strong model for any complex use case.
RapidMiner Auto model: This makes use of automated machine learning that accelerates the job of data scientists during the design of machine learning models. The auto model analyzes data and identifies the similar issues such as missing values, correlations, and stability. This uses the best practices of data science as per the data selected.

RapidMiner Server: The RapidMiner server is a platform for collaboration, computation, and deployment. This also can increase the data science team's productivity. RapidMiner Server uses cases or projects and divides the resources to the teams. The RapidMiner Server has distributed architecture with high scalability. This can reduce the risk and downtime of the system.

RapidMiner Radoop: The complexity of the data prep and machine learning using hadoop and spark can be prevented by RapidMiner Radoop. This increases the in-memory functionality of the RapidMiner with the usage of sophisticated operators in Hadoop. Radoop has more than 60 operators used for data transformations and advanced predictive modeling.

System Requirements:
Minimum:

- Dual core
- Processor: 2 GHz
- RAM:4 GB
- Disk Space: >1 GB free space
- Resolution: 1280x1024

Recommended:

- Quad core
- Processor: 3 GHz or faster
- 16 GB RAM
- Disk Space: >100 GB free space

Features:

- *Data Access:* Can establish connection to any data source with any format and at any scale.
- *Data Exploration:* It can easily recognise the data quality issues and patterns
- *Data Blending:* It can create optimal data set for the predictive analysis.
- *Data Cleansing:* It can perform the data cleansing effectively that is used for advanced algorithms.
- *Modeling:* The models can be constructed efficiently with better and faster results.
- *Validation:* The model performance can be estimated confidently with accuracy.

Benefits: The tool when compared with the other tools like Weka, Spss, Sas due to it's features like scalability, stability, open-source and ability to solve complex problems. This provides services for model evaluation with the usage of cross validation and independent validation sets.
As per our survey, no other solution can offer these many procedures and define the optimal analysis process.

Major Clients: Domino's Pizza, McKinsey & Company, Acumen Solutions, Inc, SLALOM, LLC.

OpenRefine: OpenRefine is tool used for complex large datasets. The functions performed by openrefine are cleaning, transformation of the format, extension with web services and other external data. This is a very sensible data conditioning tool that is open source which is also known as Google refine. The operation of the OpenRefine is similar to the relational database i.e, it performs the operation in the form of rows and columns. Transformations, clustering and facets methods are used for cleaning process. Normalizing, denormalizing and format transformations are made by the tool.

System requirements:

- Java JRE
- Operating Systems: Windows, Linux, OSX
- RAM: 2 GB

Features:

- Importing, Filtering or faceting data.
- Editing: Clustering, Extending data (Column creation), rows and columns
- Understanding expressions and regular expressions
- Exporting and Recovery (undo or redo)

Benefits: The tool is open source and very easy to understand and can solve and analyse large data sets. This is useful for the programmers as it provides Java, Python and other programming languages. For non- programmers it provides basic operations without the necessity of code. Simple linking available for many other information sources like Wikipedia, IMDB, MusicBrainz, etc. For the clustering of the names it combines both machine and human review.
It takes a vast variety of data formats like files and online data repositories such as Spreadsheets and fusion tables.

Major Clients: Google

Trifacta Wrangler: Trifacta Wrangler is a tool used to clean and optimize the vast and clumsy data sets with accuracy. The data set is first imported to the Wrangler and the tool will structure and organize the data automatically. The machine learning algorithms provided by the wrangler will give the ability to perform common transformations and aggregations. After the completion of the process, it will be exported and used for data visualization and machine learning. Trifacta accepts vast variety of messy data like JSON, raw CSV, Excel files. It also accepts HDFS, S3, Redshift and other data sources.

System Requirements:

- RAM: 4 GB
- Disk Space: 2 GB hard disk
- Screen Resolution: >1280x720
- Internet connection: DSL or better
- Intel Pentium 4 or AMD Opteron Processor

Features:

- Interactive Exploration
- Predictive Transformation
- Intelligent Execution
- Collaborative Data Governance
- Flexible joins and Conditional CASE function

Advantages: This is the best data conditioning tool which has solution for all the types of users like customers, analysts, partners. This acts as a gap between raw data and analysis. This helps the users to understand even the complex data and automatically optimizes it. The biggest advantage of this tool is that it has good relationships with technology vendors at large scale, for data and analytics space. Their newly employed partner program can increase the data in wrangling ecosystem and can make trifacta as a leader in this competitive world.

Major Clients: Bell, CDC, CMS, Consensus, Google, GSK, Linkedin, etc.

VII. Comparative Study of Tools

We have studied the Data cleansing tools and have analyzed all the features and benefits from them for data cleaning. The three tools we have considered RapidMiner, OpenRefine and Trifacta have different modules to perform various functions as per the wants of the customer. Comparisons are made by taking specifications and features into consideration.

S.no	Tool Name	Release Year	Latest Version	License	OS
1	Rapid Miner	2006	V8.1 02-06-18	BSD	Cross-Platforms
2	Open Refine	2010	V2.7 06-18-17	AGPL	Windows, Linux, Mac OS
3	Trifacta Wrangler	2012	11-15-17	EULA	Windows 7 or later, OS X 10.10 or later

Tools / Problems	Rapid Miner	Open Refine	Trifacta Wrangler
Availability	Desktop, Mac & Linux	Desktop, Mac	Desktop & Mac
Missing values	Yes	Yes	Yes
Duplication	Yes	Yes	Yes
Illegal values Elimination	Yes	No	Yes
Varying Value Representation	No	No	Yes
Misspelling	Yes	No	Yes
Merge	Yes	Yes	Yes
File Format	TSV,CSV,XML,RDF TRIPLETS,JSON,GOOGLE SPREADSHEETS,GOOGLE FUSION TABLES	CSV,DATABASE,EXCEL, ACCESS,BINARY,XML	CSV,JSON,TXT,XLS,XLSX, TABLEAU DAT, EXTRACT
Ease of Use	Moderate	Moderate	High

VIII. Results

From the above comparisons and use of datasets into the following tools we can clearly state that open refine and trifecta are the best among the following tools. Openrefine is an open source tool which is free to use by the individuals as well as the organizations where Trifacta is not an open source but is capable of dealing with data cleaning, preparation in the best possible way with various operations involved and high processing speed. Rapid miner is limited with its operations such as duplication, missing values, merges and still can be used for data cleaning for these purposes.

IX. Conclusion

Data cleaning is one of the most important aspects in the organizations that deal with data. The important features and functionalities of the data conditioning, cleaning and readiness are discussed in the paper. The data quality issues, methods and phases of working are included. This paper also presents the comparison of certain data cleaning tools and suggests the best possible tool in terms of effectiveness, technical requirements, complexity and other important features. Maintaining good data quality will help us to analyse the data efficiently and it also helps to maintain good relationship with the clients. The comparison of tools and their ability to solve complex issues makes them beneficial than the traditional tools like R and SQL.

Acknowledgement

The authors would like to present their sincere gratitude to Prof. James Baldo Jr., Ph.d. for his valuable suggestions and comments to improve the standard of the paper.

References

[1] Neil D. Lawrence, "Data Readiness Levels", Amazon Research Cambridge and University of Sheffield. 6th April 2017, accessed on 28th March 2018.

[2] "The Six Primary Dimensions for Data Quality Assessment", Defining Data Quality Dimensions, accessed on 2nd April 2018.

[3] Erhard Rahm and Hong Hai Do "Data Cleaning: Problems and Current Approaches" University of Leipzig, Germany, IEEE Data Eng. Bull., 2000.

[4] "Where Does Dirty Data Originate & Why you need to know," *B2B Contact Data & Lead Generation by Social123*. [Online]. Accessed on 4th April 2018

[5] Heiko Muller and Johann-Christoph Freytag, "Problems, Methods, and Challenges in Comprehensive Data Cleansing" Humboldt-Universität zu Berlin zu Berlin; 2005.

[6] "Data Cleansing Strategy - Manual, Automated or Both," *Intelligent Data Group*, 24-Nov-2015, accessed on 10th April 2018.

[7] S. B, "How Businesses Can Benefit from Data Cleansing Software," The Data Cleansing Blog, 14-Mar-2018, accessed on 15th April 2018 [Online].

[8] www.rapidminer.com

[9] www.openrefine.org

[10] www.trifacta.co