Analysis and recommendations of data cleaning process for various types of data

A project report submitted in partial fulfilment of the requirements of the award of the degree of

Bachelor of Technology

In

Computer engineering

By

Sejal Jain, Registration number: PCE20CS171

Rahul Khandelwal, Registration number: PCE20CS152

Yash Tripathi, Registration number: PCE20CS200

Sanskar Sharma, Registration number: PCE20CS166

Sachin Yadav, Registration number: PCE20CS161

Under the guidance of

Dr. Neelam Chaplot, Associate Professor,

Department of computer science and engineering

(Session 2021-2022)

Department of Computer Engineering

Poornima College of Engineering

ISI-6, RIICO Institutional Area, Sitapura, Jaipur – 302022

January, 2022

Department Certificate

This is to certify that Ms Sejal Jain, registration no. PCE20CS171, of the Department of Computer Engineering, has submitted this project report entitled "Analysis and recommendations of data cleaning process for various types of data" under the supervision of Dr. Neelam Chaplot, working as Associate Professor in department of Computer Engineering as per the requirements of the Bachelor of Technology program of Poornima College of Engineering, Jaipur.

Dr. Surendra Kumar Yadav	Mrs. Sonam Gour
Head of Department	Coordinator- Project
Computer Engineering	

CANDIDATE'S DECLARATION

I hereby declare that the work which is being presented in this project report entitled "Analysis and recommendations of data cleaning process for various types of data" in the partial fulfilment for the award of the Degree of Bachelor of Technology in (Computer Engineering), submitted in the Department of Computer Engineering, Poornima College of Engineering, Jaipur, is an authentic record of my own work done during the period from July 2021 to December 2021 under the supervision and guidance of **Dr Neelam Chaplot, Associate Professor**.

We have not submitted the matter embodied in this project report for the award of any other degree.

Dated:
Place: Jaipur

SUPERVISOR'S CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Signature)

Dated: Dr. Neelam Chaplot

Place: Jaipur Associate Professor,

Computer Science Engineering

ACKNOWLEDGEMENT

I would like to convey my profound sense of reverence and admiration to my supervisor Dr. Neelam Chaplot, Associate Professor **Department of Computer Engineering, Poornima College of Engineering,** for his intense concern, attention, priceless direction, guidance and encouragement throughout this research work.

I am grateful to **Dr. Mahesh Bundele**, Director of Poornima College of Engineering for his helping attitude with a keen interest in completing this dissertation in time.

I extend my heartiest gratitude to all the teachers, who extended their cooperation to steer the topic towards its successful completion. I am also thankful to non-teaching staff of the department to support in preparation of this dissertation work.

My special heartfelt gratitude goes to **Dr. Surendra Kumar Yadav, HOD, Department of Computer Engineering, Ms. Sonam Gour, Project Coordinator, Department of Computer Engineering, Poornima College of Engineering**, for unvarying support, guidance and motivation during the course of this research.

I would like to express my deep sense of gratitude towards management of Poornima College of Engineering including **Dr. S. M. Seth**, Chairman Emeritus, Poornima Group and former Director NIH, Roorkee, **Shri Shashikant Singhi**, Chairman, Poornima Group, **Mr. M. K. M. Shah**, Director Admin & Finance, Poornima Group and **Ar. Rahul Singhi**, Director Poornima Group for establishment of institute and providing facilities my studies.

I would like to take the opportunity of expressing my thanks to all faculty members of the Department, for their kind support, technical guidance, and inspiration throughout the course.

I am deeply thankful to my parents and all other family members for their blessings and inspiration. At last, but not least I would like to give special thanks to God who enabled me to complete my dissertation on time.

Sejal Jain, Department of Computer Engineering, PCE20CS171
Rahul Khandelwal, Department of Computer Engineering, PCE20CS152
Yash Tripathi, Department of Computer Engineering, PCE20CS200
Sanskar Sharma, Department of Computer Engineering, PCE20CS166
Sachin Yadav, Department of Computer Engineering, PCE20CS161

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENT

1.ABSTRACT

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

2. INTRODUCTION

Data cleaning is the process that prepares data for analysis by removing 1duplicate or unnecessary data, modifying or updating data for analysis that is incorrect, missing, irrelevant, duplicated or improperly formatted.

Data cleaning can be used for analysis of the data quality in a data source, manually approving/ rejecting suggestions.

Data cleaning not only means removing unnecessary or irrelevant data but also means to perform actions such as fixing spelling and syntax errors, standardizing data sets and correcting the mistakes like filling the missing values, empty fields, and identifying duplicate records.

Data cleaning ensures the ease of finding recent and important documents when required. Data cleaning also makes us aware of the fact that if we have our important information like (bank account number) then it can be very risky.

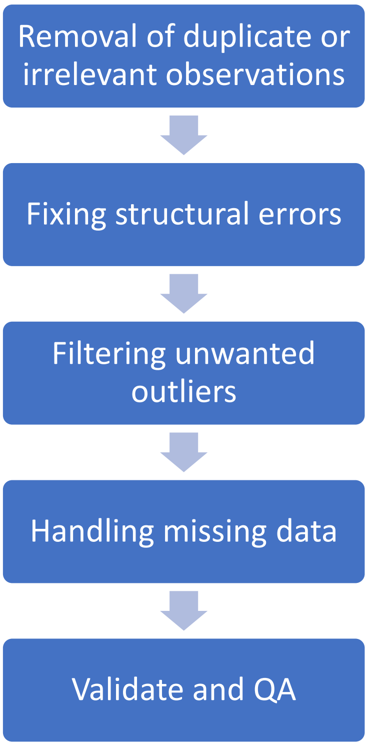

Removal of duplicate or irrelevant observations
Fixing structural errors
Filtering unwanted outliers
Handling missing data
Validate and QA

Figure 1: Basic steps for data cleaning

Step 1: Removal of delicacy from the data or irrelevant observations

During data collection when we combine data from multiple sources there are opportunities that we have created duplicate data that means having similar entries more than once. Also, there are chances that we have some irrelevant observations also, all these create problems when we analyse the data.

Step 2: Fixing up of the structural errors

Structural errors are found where there is strange or incorrect naming conventions. For example, we find "-" and "NULL" value but they should be analysed as same category.

Step 3: Filtering out the unwanted outliers

Outliers are the improper data entry in a data set.

For example: age column containing -18 as age but we know that age of a person will not be negative.

Another example: Gender column containing beautiful, we know that gender of a person can either be male, female or transgender but beautiful isn't the gender of a person, its characteristic or feature of a person.

Step 4: Efficient handling of missing data

Missing entries can't be ignored as many algorithms won't give correct results if the values from a data set are missing.

Ways to deal with missing values of the data set:

Drop the observation having missing value.

But dropping observations will be equal to dropping information so be careful of this before removing observations.

Other than this we can also provide values to the missing entries.

Step 5: Validate and QA

This includes answering basic questions like

1. Does this data make sense?
2. Does data follow appropriate rules for its field?
3. Does this data support the facts of our working theory?
4. Can we find a particular trend in the data?

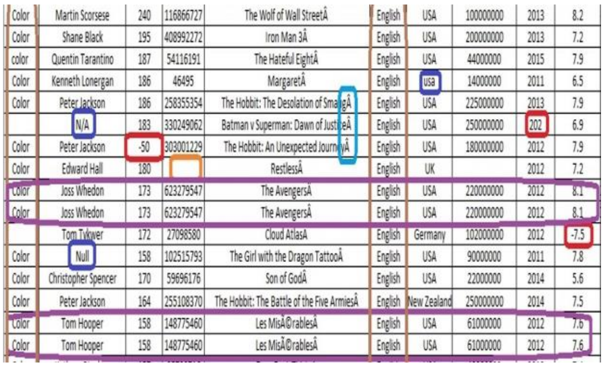
3. Problem Statement and Objective

3.1 Problem Statement

Analyse a dataset and then apply different types of data cleaning methods on it according to the requirement of the data set. Analysis the data set cleaned through various methods then comparing them statistically and then finding the best results out of the various applied methods.

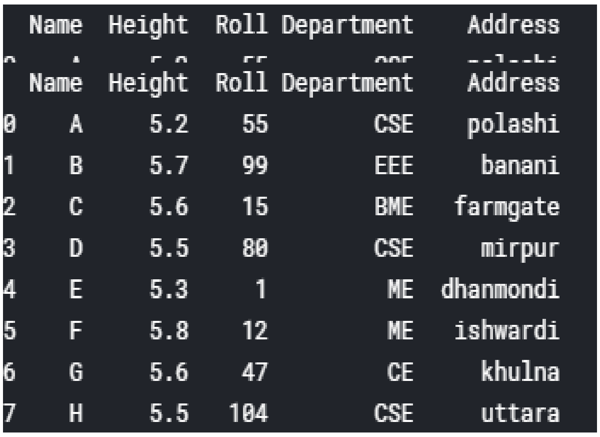
3.2 Objective

- To apply the general understanding of data cleaning.
- To analyze a dataset.
- To use the data cleaning tools.
- To apply various data cleaning methods on a data set



Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasÂ	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

Figure 2: Example of Unclean Data Set



```
  Name  Height  Roll Department    Address
  Name  Height  Roll Department    Address
0    A     5.2    55        CSE    polashi
1    B     5.7    99        EEE     banani
2    C     5.6    15        BME   farmgate
3    D     5.5    80        CSE     mirpur
4    E     5.3     1         ME  dhanmondi
5    F     5.8    12         ME   ishwardi
6    G     5.6    47         CE     khulna
7    H     5.5   104        CSE     uttara
```

Figure 3: Example of Clean Data Set

4. Literature Review

4.1 Paper 01:

A Review on Data Cleansing Methods for Big Data

By: Fakhitah Ridzuan and Wan Mohd Nazmee Zainon (2019)

Summary:

Fakhitah Ridzuan in **"A review on data cleaning methods for Big Data [2019]"** discussed about the various data cleaning processes that include

- Data analysis

  A process for identification of errors, inconsistent, incorrect or missing entries in a data set.
- Definition of transformation workflow and mapping rules

  It defines the detection and elimination of anomalies performed by the sequence of operations on the data. It is specified after data analysis to gain knowledge about the existing anomalies.
- Verification

  In this phase the correctness and effectiveness of the transformation workflow are evaluated.it consist of multiple iterations to ensure that errors are being corrected.
- Transformation

  It's executed to refresh the data in the data warehouse. Detail information of transformation process must be recorded to support data quality.
- Backflow of cleaned data

  Finally, after all the errors are removed the uncleaned data is replaced with cleaned data.

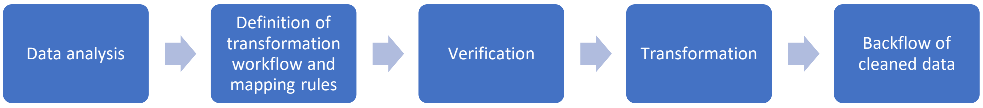
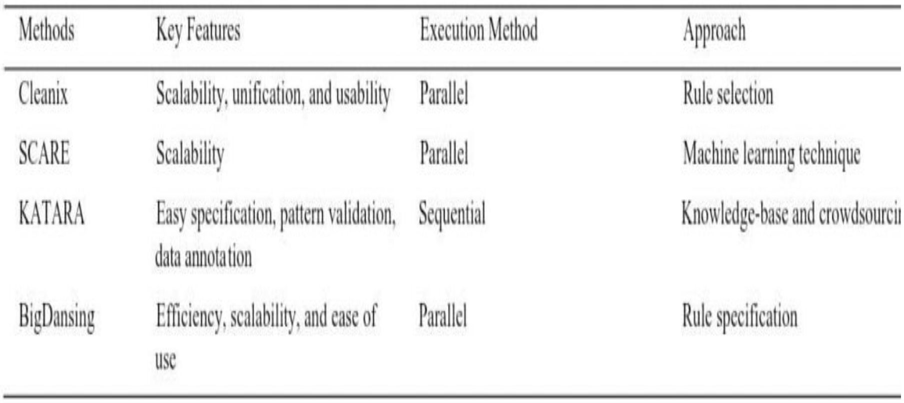


Figure 4: Data cleaning processes

The discussion further continues to the data cleaning for big data.

Various meth Most of the organization depend on the data-driven decision making, thus information system is closely related to the business process management to leverage their processes for competitive advantage. Nowadays, the amount of data keeps increasing, but the quality of the data is decreasing as many of the data collected is dirty.

Various data cleansing approaches are available to solve this issue but data cleansing remains as a challenge in order to cope with the criteria of big data. Some of the approaches are not suitable for big data as it has a significant amount of data that need to be processed at a time. Despite the availability of existing frameworks to address data cleansing for big data, but the value and veracity of the data often left out when designing the approaches. Besides, the need for domain expert in undeniable as an expert is needed to verify and validate the data before it can undergo an analysis processed for big data cleaning were proposed which are as follows:



Methods	Key Features	Execution Method	Approach
Cleanix	Scalability, unification, and usability	Parallel	Rule selection
SCARE	Scalability	Parallel	Machine learning technique
KATARA	Easy specification, pattern validation, data annotation	Sequential	Knowledge-base and crowdsourci
BigDansing	Efficiency, scalability, and ease of use	Parallel	Rule specification

Figure 5: Analysis processed for big data cleaning

4.2 Paper 02:

An Overview Study on Data Cleaning, Its Types and Its Methods for Data Mining (2018)

By: S. Lakshmi

Summary:

S. Lakshmi and etl **(2018)** discussed about the sources of errors in data. These are discussed below:

Data entry errors

Data is often corrupted at entry time by typing errors or misinterpretation or misunderstanding of the data source.

Measurement errors

There are certain cases in which data is intended to be measured by some physical process in the world like: population size, speed of vehicle, growth of an economy etc.

Data integration errors

In most of the cases the data is collected from multiple sources and while merging this data there are chances that the data may get duplicated, merging task requires attempts to resolve inconsistencies across the databases

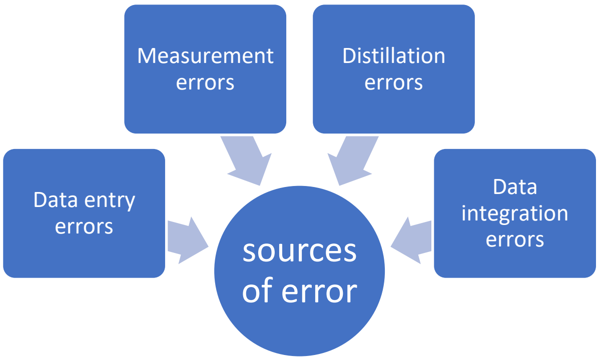


Figure 6: Sources of errors in data

The discussion further continued towards the approaches to improve data quality.

Data entry interface design

For human data entry, most of the errors in data can be avoided through judicious design of data entry interfaces, one key aspect of this was the specification database integrity constraints, including data type checks, bounds on numeric values, the prevention of references to non-existent data.

Organizational management

Archiving and analysis to minimize opportunities for error; automating data capture; capturing metadata and using it to improve data interpretation; and incentives for multiple parties to participate in the process of maintaining data quality.

Data preparation is an important issue for both data warehousing and data mining, as real-world data tends to be incomplete, noisy, and inconsistent. Data preparation includes data cleaning, data integration, data transformation, and data reduction. Data cleaning routines can be used to fill in missing values, smooth noisy data, identify outliers, and correct data inconsistencies. Data integration combines data from multiples sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration. Data transformation routines conform the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0 to 1.0. Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content. Concept hierarchies organize the values of attributes or dimensions into gradual levels of abstraction. They are a form a discretization that is particularly useful in multilevel mining. Automatic generation of concept hierarchies for categoric data may be based on the number of distinct values of the attributes defining the hierarchy. For numeric data, techniques such as data segmentation by partition rules, histogram analysis, and clustering analysis can be used. Although several methods of data preparation have been developed, data preparation remains an active and important area of research.

4.3 Paper 03:

Study of Data Cleaning & Comparison of Data Cleaning Tools (2015)

By: Sapna Devi and Dr. Arvind Kalia

Summary:

The data quality problems are further classified as single source problems and multisource problems and the process of data cleaning.

DATA CLEANING PROCESS

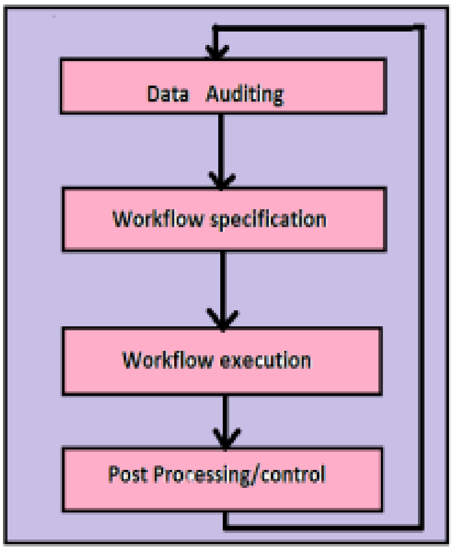


Figure 7: Data cleaning processes

Data Auditing: Auditing in simple words means conducting an official inspection. So, data auditing implies process to conduct data audit to access data quality or utility for a specific purpose.

Workflow Specification: We can perform multiple operations to detect and eliminate common order problems. This is called the data cleansing workflow. It is specified to gain information about the inconsistencies present in data.

Workflow Execution: The data cleaning workflow is executed after specification and verification of its correctness.

Post-Processing/Control: After performing the cleansing workflow the results are checked to again verify the correctness of operations performed. Within the controlling step the tuples that could not be corrected initially are inspected intending to correct them manually.

Data cleaning is very necessary part of data mining. From the above study we can see that there are different types of problems in data cleaning. 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 presents a comparison of data cleaning tools and determines the best tool. Each tool has its own specific features and depending upon the data we can use the tool to clean data. In future work we can check other functionality of these tools and suggest own. Data cleaning tools discussed in this paper were:

A. **MS EXCEL DATA CLEANER** Excel Files Data Cleaning Utility is a useful adding for Excel to Clean and Organize Data. It is fast & Reliable and you can save your precious time & Money.

Text cleaner: A Collection of Tools to Clean Text in Selected Cells.

Duplicate Cleaner: A Collection of Tools to eliminate duplicate entries.

Data Organizer: A Collection of Tools to Organize Data.

Text Organizer: A collection of tools to organize the text.

B. RAPIDMINOR: Rapid minor is a software platform that is used for data mining. We can also clean our data using Rapid Minor. This tool contains various operators for data cleaning or data cleansing. It released on 2006, latest version available is Rapid Minor. 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.

C. WINPURE CLEAN & MATCH: It is a comprehensive data cleansing, data deduplication software and a data listing software solution. It helps to clean mailing lists, spreadsheets, marketing databases and electronic mails. The software carries out data deduplication and offers the option of basic as well as advanced search. It helps to merge duplicate records on one or two lists which help to make the merging process effective and easier. "Win Pure Clean & Match " offers some interesting features such as “Safe Merge” options that make sure that no data is lost while merging the records.

4.4 Paper 04:

A Review of Data Cleansing Concepts – Achievable Goals and Limitations (2013)

By: Kofi Adu-Manu Sarpong and John Kingsley Arthur

Summary-

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. In view of this a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of data cleansing. The papers reviewed in this report looked at critical aspects of data cleansing and the various types of data that could be cleansed. Several algorithms have been proposed in the various works discussed.

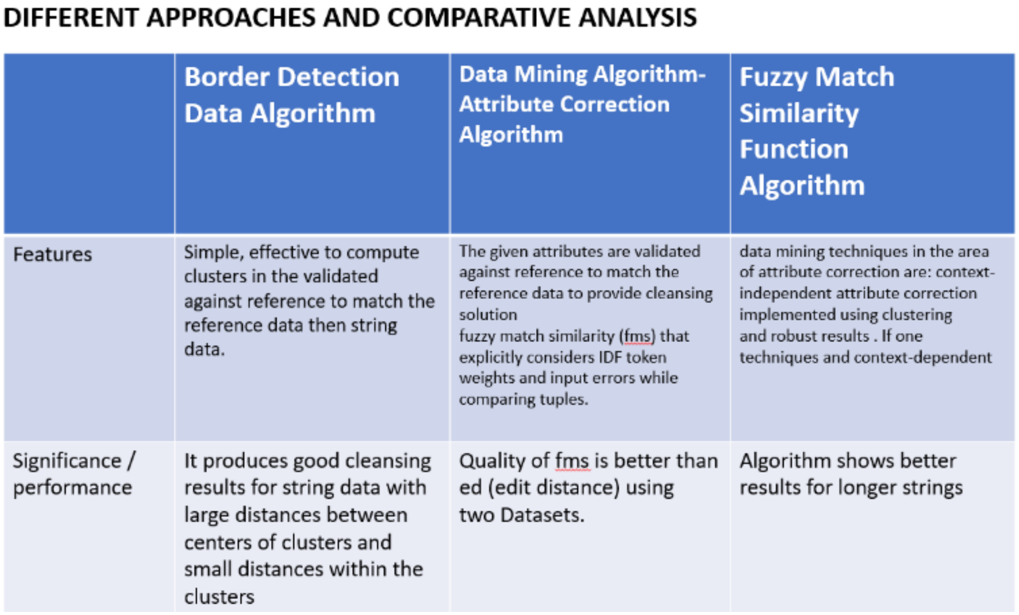
According to, the classification of data quality problems can be divided into two main categories: single-source and multiple-source problems. At the single-source, Rahm and Do divide these into schema level and instance level related problems without considering the occurrence in a single relation. The single-source problems deal with attribute, record, record type and source whereas the multiple-source problems deal with naming conflicts, schema-level conflicts and the identification of overlapping data which refers to same real-world entity.

4.5 Paper 05-

DATA CLEANING: CURRENT APPROACHES AND ISSUES

By: Vaishali Chandrakant Wangikar and Ratnadeep R. Deshmukh (2011)

Summary- Various data cleaning algorithms and techniques are presented in the paper but each method can be used to identify a particular type of error in the data. The technique suitable for one type of data cleaning may not be suitable for the other. data cleaning has a wide variety of situations that need to cater efficiently by some comprehensive data cleaning framework. Future research directions include the review and investigation of various methods to address wide area of data cleaning. A better integration of data cleaning approach in the frameworks and data decision processes should be achieved.



DIFFERENT APPROACHES AND COMPARATIVE ANALYSIS

	Border Detection Data Algorithm	Data Mining Algorithm- Attribute Correction Algorithm	Fuzzy Match Similarity Function Algorithm
Features	Simple, effective to compute clusters in the validated against reference to match the reference data then string data.	The given attributes are validated against reference to match the reference data to provide cleansing solution fuzzy match similarity (fms) that explicitly considers IDF token weights and input errors while comparing tuples.	data mining techniques in the area of attribute correction are: context-independent attribute correction implemented using clustering and robust results . If one techniques and context-dependent
Significance / performance	It produces good cleansing results for string data with large distances between centers of clusters and small distances within the clusters	Quality of fms is better than ed (edit distance) using two Datasets.	Algorithm shows better results for longer strings

Table 1: Different approaches and comparative analysis of data cleaning algorithms

Table 2: Comparison Table

S. No	Paper title	Author's Name	Year	Approach used	Finding	S/w and H/w Required
1	A Review on Data Cleansing Methods for Big Data	Fakhitah Ridzuan and Wan Mohd Nazmee Zainon	(2019)	Investigating data cleansing	Data cleaning process	Laptop / computer with good amount of storage
2	Study of Data Cleaning & Comparison of Data Cleaning Tools	By: Sapna Devi and Dr. Arvind Kalia	(2015)	Using basic software to do data cleaning	Finding best choice between the software available	MS Excel Rapid Minor Win Pure Clean and Match
3	An Overview Study on Data Cleaning, Its Types and Its Methods for Data Mining	By: S. Lakshmi	(2018)	Approach for improving data quality	Sources of error in data	Laptop / computer with good amount of storage
4	A Review of Data Cleansing Concepts– Achievable Goals and Limitations	Kofi Adu-Manu Sarpong and John Kingsley Arthur	(2013)	Data cleaning problems and current approaches	Steps of data cleaning	System/desktop System/desktop
5	Data Cleaning: Current Approaches and Issues	Vaishali Chandrakant Wangikr and Ratnadeep Deshmukh	(2011)	Border Detection Data Algorithm Fuzzy Match Similarity Function Algorithm	Data cleaning algorithm for checking spellings of the string in the data set	System/desktop

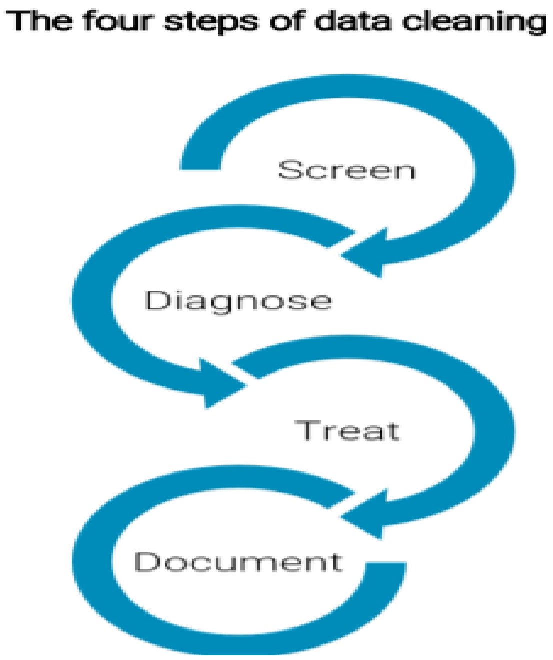


Figure 8: Steps of data cleaning

Screening: In this step we check if we have sufficient amount of data or not. Then we check outliers, inconsistencies and strange patterns present in the data. Finally, after susception, analysis results are given.

Diagnosing: In this step we check if there is any data missing from the data set, errors present in the data set, validity of the entered data. If suspects are present till, then we didn’t perform diagnose.

Treatment: This stage involves entering correct data and removing the undesired and not useful data from the data set.

Document: In this step of data cleaning, we maintain the change log and archive raw data and old values.

5. Conclusion

In this report we mentioned the meaning, use and importance of data cleaning and data cleaning methods. We also discussed basic methos for performing data cleaning. Important steps include, Removal of duplicate or irrelevant data, fixing of structural errors, filtering out unwanted outliers, handling the missing data, validation and question and answer. The discussion continues by stating different sources of errors in the data, data pre-processing, algorithms for string data cleaning. In the overview of data cleaning, we discussed the definition of data cleaning, benefits of data cleaning, characteristics of quality data and screening, diagnosing, treatment and documentation processes involved in data cleaning.

Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration. Data transformation routines conform the data into appropriate forms for mining. Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content. Concept hierarchies organize the values of attributes or dimensions into gradual levels of abstraction. Automatic generation of concept hierarchies for categoric data may be based on the number of distinct values of the attributes defining the hierarchy. For numeric data, techniques such as data segmentation by partition rules, histogram analysis, and clustering analysis can be used. Although several methods of data preparation have been developed, data preparation remains an active and important area of research.

6.Future Scope

In future we would like to do deep study of more data cleaning algorithms and analyse the data sets with different cleaning approaches. Based on the comparison we would find out the best suitable technique for a particular type of data set.

Analysis and recommendations of data cleaning process for various types of data

Ms. Sejal Jain
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

Mrs. Neelam Chaplot
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

Mr. Yash Tripathi
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

Mr. Rahul Khandelwal
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

Mr. Sanskar Sharma
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

Mr. Sachin Yadav
Computer Science And Engineering
Poornima College of Engineering
Jaipur, India

***Abstract*— In this research paper we will discuss about the meaning, need, process and importance of data cleaning. We will also discuss about data pre-processing and the methodology such as screening, diagnosing, treating and documenting involved in the process of data cleaning and comparison of string data cleaning algorithms like border detection algorithm, data mining algorithm and fuzzy match algorithm and compare them according to task performed by them. By the end of this research paper, you will get an overview of data cleaning, benefits of data cleaning and characteristics of quality data.**

***Keywords* Data cleaning, data mining, data pre-processing, algorithms, machine learning, artificial intelligence.**

1.Introduction

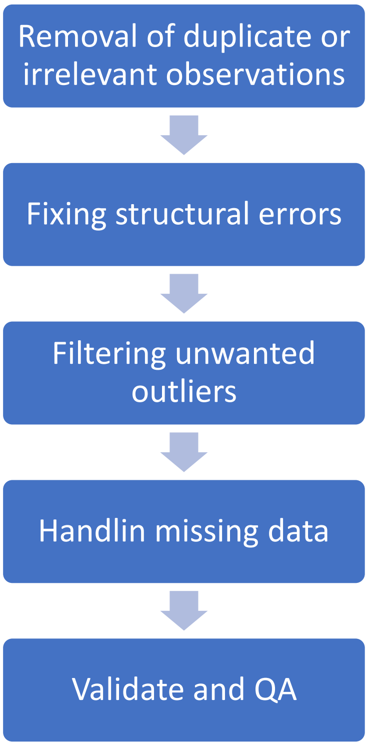
Data cleaning is the process that prepares data for analysis by removing duplicate or unnecessary data, modifying or updating data for analysis that is incorrect, missing, irrelevant, duplicated or improperly formatted.

Data cleaning can be used for analysis of the data quality in a data source, manually approving/ rejecting suggestions.

Data cleaning not only means removing unnecessary or irrelevant data but also means to perform actions such as fixing spelling and syntax errors, standardizing data sets and correcting the mistakes like filling the missing values, empty fields, and identifying duplicate records.

Data cleaning ensures the ease of finding recent and important documents when required. Data cleaning also makes us aware of the fact that if we have our important information like (bank account number) then it can be very risky.

figure: Basic steps for data cleaning:


Removal of duplicate or irrelevant observations
Fixing structural errors
Filtering unwanted outliers
Handlin missing data
Validate and QA

Step 1: Removal of delicacy from the data or irrelevant observations

During data collection when we combine data from multiple sources there are opportunities that we have created duplicate data that means having similar entries more than once. Also, there are chances that we have some irrelevant observations also, all these create problems when we analyze the data.

Step 2: Fixing up of the structural errors

Structural errors are found where there is strange or incorrect naming conventions. For example, we find "-" and "NULL" value but they should be analysed as same category.

Step 3: Filtering out the unwanted outliers

Outliers are the improper data entry in a data set.

For example: age column containing -18 as age but we know that age of a person will not be negative.

Another example: Gender column containing beautiful, we know that gender of a person can either be male, female or transgender but beautiful isn't the gender of a person, its characteristic or feature of a person.

Step 4: Efficient handling of missing data

Missing entries can’t be ignored as many algorithms won’t give correct results if the values from a data set are missing.Ways to deal with missing values of the data set:

1. Drop the observation having missing value.
   But dropping observations will be equal to dropping information so be careful of this before removing observations
2. Other than this we can also provide values to the missing entries.

Step 5: Validate and QA

This includes answering basic questions like

5. Does this data make sense?
6. Does data follow appropriate rules for its field?
7. Does this data support the facts of our working theory?
8. Can we find a particular trend in the data?

II LITERATURE REVIEW

Fakhitah Ridzuan and etl in **[2019]** discussed about the various data cleaning processes that include

- Data analysis
  A process for identification of errors, inconsistent, incorrect or missing entries in a data set.
- Definition of transformation workflow and mapping rules
  It defines the detection and elimination of anomalies performed by the sequence of operations on the data. It is specified after data analysis to gain knowledge about the existing anomalies.
- Verification
  In this phase the correctness and effectiveness of the transformation workflow are evaluated. It consists of multiple iterations to ensure that errors are being corrected.
- Transformation
  It’s executed to refresh the data in the data warehouse. Detail information of transformation process must be recorded to support data quality.
- Backflow of cleaned data
  Finally, after all the errors are removed the uncleaned data is replaced with cleaned data.

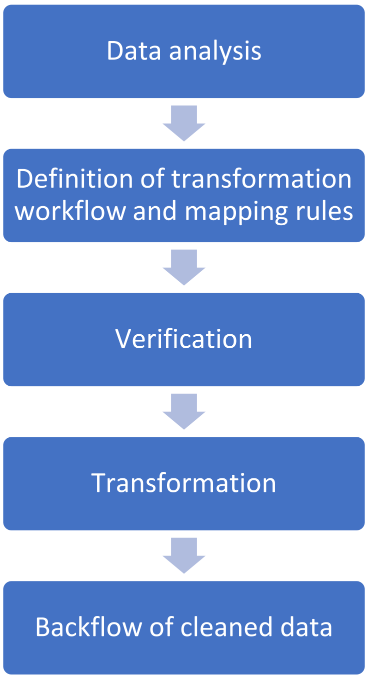
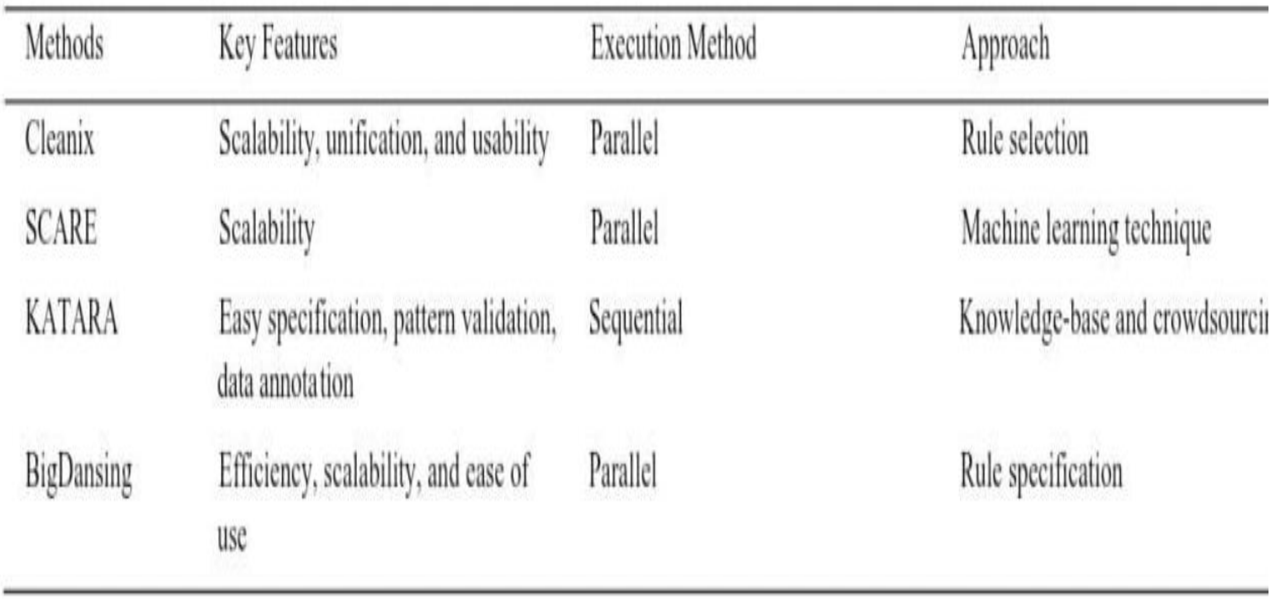


Figure: Data cleaning processes

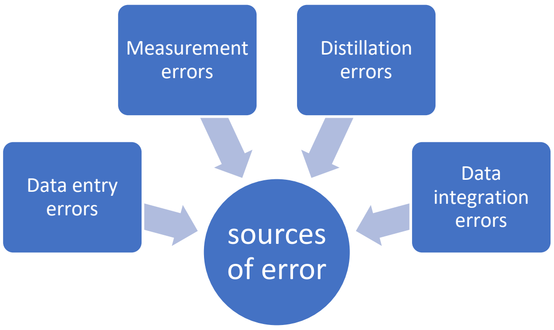
The discussion further continues to the data cleaning for big data

Various methods for big data cleaning were proposed which are as follows:



Methods	Key Features	Execution Method	Approach
Cleanix	Scalability, unification, and usability	Parallel	Rule selection
SCARE	Scalability	Parallel	Machine learning technique
KATARA	Easy specification, pattern validation, data annotation	Sequential	Knowledge-base and crowdsourci
BigDansing	Efficiency, scalability, and ease of use	Parallel	Rule specification

S. Lakshmi and etl **(2018)** discussed about the sources of errors in data. These are discussed below:



Data entry errors

Data is often corrupted at entry time by typing errors or misinterpretation or misunderstanding of the data source.

Measurement errors

There are certain cases in which data is intended to be measured by some physical process in the world like: population size, speed of vehicle, growth of an economy etc.

Data integration errors

In most of the cases the data is collected from multiple sources and while merging this data there are chances that the data may get duplicated, merging task requires attempts to resolve inconsistencies across the databases.

The discussion further continued towards the approaches to improve data quality.

Data entry interface design

For human data entry, most of the errors in data can be avoided through judicious design of data entry interfaces, one key aspect of this was the specification database integrity constraints, including data type checks, bounds on numeric values, the prevention of references to non-existent data.

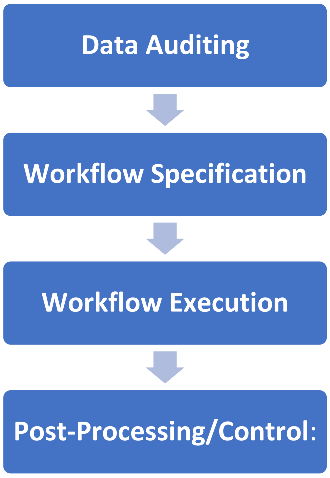
Organizational management

Archiving and analysis to minimize opportunities for error; automating data capture; capturing metadata and using it to improve data interpretation; and incentives for multiple parties to participate in the process of maintaining data quality.

Sapna Devi and etl (2015)

Discussed that data quality problems are further classified as single source problems and multisource problems and the process of data cleaning.

Figure: DATA CLEANING PROCESS



Data Auditing: Auditing in simple words means conducting an official inspection. So data auditing implies process to conduct data audit to access data quality or utility for a specific purpose.

Workflow Specification: We can perform multiple operations to detect and eliminate common order problems. This is called the data cleansing workflow. It is specified to gain information about the inconsistencies present in data.

Workflow Execution: The data cleaning workflow is executed after specification and verification of its correctness.

Post-Processing/Control: After performing the cleansing workflow the results are checked to again verify the correctness of operations performed. Within the controlling step the tuples that could not be corrected initially are inspected intending to correct them manually.

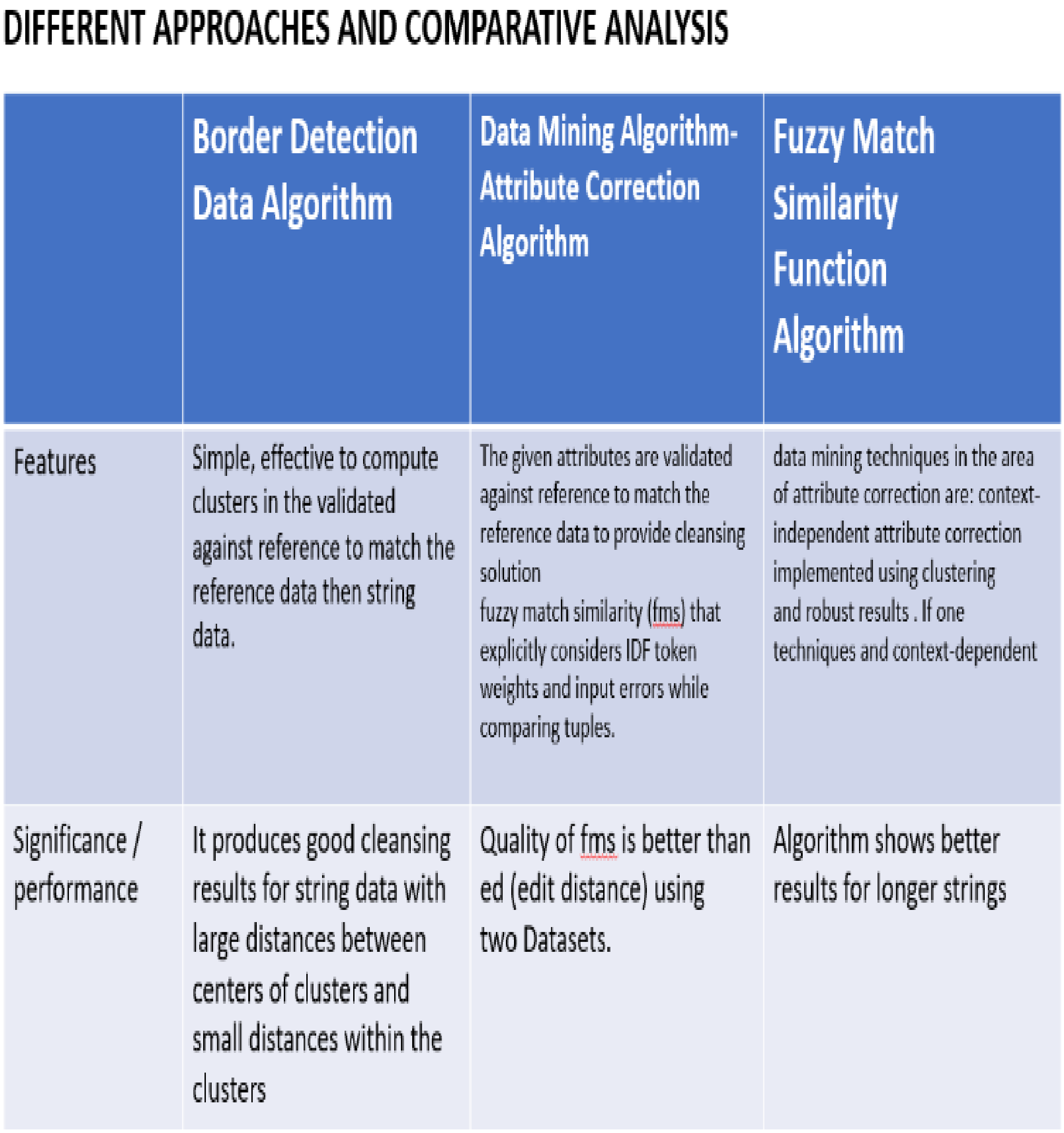
Kofi Adu-Manu Sarpong and etl [2013] talked about data warehousing, importance of data cleaning, problems, methods and challenges in the data cleaning, data cleaning problems and current approaches, matching algorithms with a duplicate detection system, open user involvement in data cleaning for data warehousing quality. To conclude, data cleaning has become an important step

for organizations as they need quality data to improve the services they give to their customers.

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. In view of this a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of data cleansing. The papers reviewed in this report looked at critical aspects of data cleansing and the various types of data that could be cleansed. Several algorithms have been proposed in the various works discussed.

According to, the classification of data quality problems can be divided into two main categories: single-source and multiple-source problems. At the single-source, Rahm and Do divide these into schema level and instance level related problems without considering the occurrence in a single relation. The single-source problems deal with attribute, record, record type and source whereas the multiple-source problems deal with naming conflicts, schema-level conflicts and the identification of overlapping data which refers to same real-world entity

Vaishali Chandrakant Wangikar and etl (2011)” talked about the various algorithms for string data cleaning which are discussed below.



DIFFERENT APPROACHES AND COMPARATIVE ANALYSIS

	Border Detection Data Algorithm	Data Mining Algorithm- Attribute Correction Algorithm	Fuzzy Match Similarity Function Algorithm
Features	Simple, effective to compute clusters in the validated against reference to match the reference data then string data.	The given attributes are validated against reference to match the reference data to provide cleansing solution fuzzy match similarity (fms) that explicitly considers IDF token weights and input errors while comparing tuples.	data mining techniques in the area of attribute correction are: context-independent attribute correction implemented using clustering and robust results . If one techniques and context-dependent
Significance / performance	It produces good cleansing results for string data with large distances between centers of clusters and small distances within the clusters	Quality of fms is better than ed (edit distance) using two Datasets.	Algorithm shows better results for longer strings

<u>**Table: comparative analysis of string data cleaning algorithms**</u>

Table: Comparison Table

S. No	Paper title	Author's Name	Year	Approach used	Finding	S/w and H/w Required
1	A Review On Data Cleansing Methods For Big Data	Fakhitah Ridzuan And Wan Mohd Nazmee Zainon	(2019)	Investigating Data Cleansing	Data cleaning process	Laptop / computer with good amount of storage
2	Study Of Data Cleaning & Comparison of Data Cleaning Tools	By: Sapna Devi and Dr. Arvind Kalia	(2015)	Using Basic Software to Do Data Cleaning	Finding best choice between the software available	MS EXCEL Rapid Minor Win Pure Clean and Match
3	An Overview Study on Data Cleaning, Its Types and Its Methods for Data Mining	By: S. Lakshmi	(2018)	Approach For Improving Data Quality	Sources of error in data	Laptop / computer with good amount of storage
4	A Review of Data Cleansing Concepts– Achievable Goals and Limitations	Kofi Adu-Manu Sarpong and John Kingsley Arthur	(2013)	Data Cleaning Problems and Current Approaches	Steps of data cleaning	System/desktop
5	Data Cleaning: Current Approaches and Issues	Vaishali Chandrakant Wangikr and Ratnadeep Deshmukh	(2011)	Border Detection Data Algorithm Fuzzy Match Similarity Function Algorithm	Data cleaning algorithm for checking spellings of the string in the data set	System/desktop

III OVERVIEW OF DATA CLEANING

Screening: In this step we check if we have sufficient amount of data or not. Then we check outliers, inconsistencies and strange patterns present in the data. Finally, after susception, analysis results are given.

Diagnosing: In this step we check if there is any data missing from the data set, errors present in the data set, validity of the entered data. If suspects are present till, then we didn't perform diagnose.

Treatment: This stage involves entering correct data and removing the undesired and not useful data from the data set.

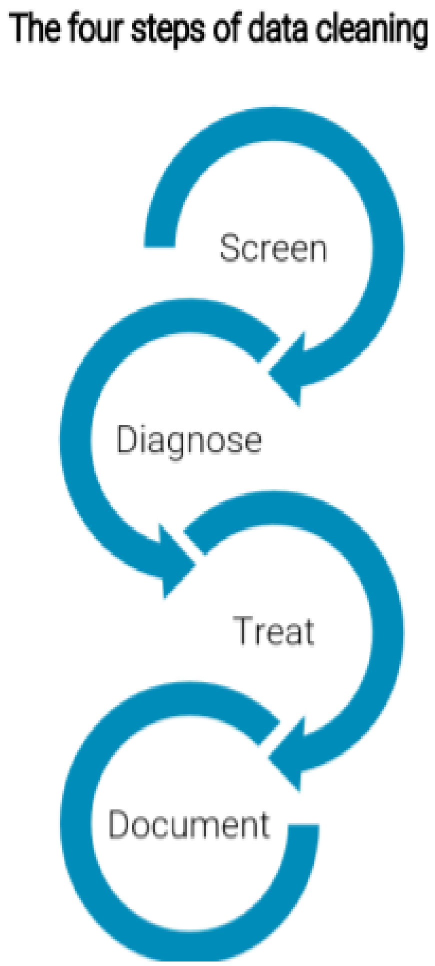
Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Benefits of data cleaning include:

- Minimal risk of error generation when multiple sources of data are at play.

Document: In this step of data cleaning, we maintain the change log and archive raw data and old values.



- Fewer errors make clients happier so the employees are less-frustrated.
- Identification of errors, missing values and incorrect values at early stage makes data cleaning easy.
- **Characteristics of quality data**
- **Validity.** The degree to which our data satisfies pre-defined constraints.

- **Accuracy.** Closeness of the data to the true values.
- **Completeness.** Most important required data is known to us.
- **Consistency.** Ensuring data consistency within the same dataset and/or across multiple data sets.
- **Uniformity.** Data specification using the same unit of measure.

Conclusion

In this paper we mentioned the meaning, use and importance of data cleaning and data cleaning methods. We also discussed basic methos for performing data cleaning. Important steps include, Removal of duplicate or irrelevant data, fixing of structural errors, filtering out unwanted outliers, handling the missing data, validation and question and answer. The discussion continues by stating different sources of errors in the data, data pre-processing, algorithms for string data cleaning. In the overview of data cleaning, we discussed the definition of data cleaning, benefits of data cleaning, characteristics of quality data and screening, diagnosing, treatment and documentation processes involved in data cleaning.

References

[1] Fakhitah Ridzuan and Wan Mohd Nazmee Zainon “A review on data cleaning methods for Big Data” 2019

[2] S. Lakshmi and Dr SV “An Overview Study on Data Cleaning, Its Types and Its Methods for Data Mining” 2018

[3] Sapna Devi and Dr. Arvind Kalia “Study of Data Cleaning & Comparison of Data Cleaning Tools” 2015

[4] Kofi Adu-Manu Sarpong and John Kingsley Arthur “A Review of Data Cleansing Concepts–Achievable Goals and Limitations” 2013

[5] Vaishali Chandrakant Wangikr and Ratnadeep Deshmukh “DATA CLEANING: CURRENT APPROACHES AND ISSUES” 2011

[6] Sonka, Steven. (2016) “Big Data Characteristics.” International Food and Agribusiness Management Review 19 (A): 7-12.

[7] Arturas Mazeika Michael H. B¨ohlen: Cleansing Databases of Misspelled Proper Nouns, Clean DB, Seoul, Korea, 2006

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

[9] Data Cleansing – A Novel Approach to Support the Cleansing Process" International Journal of Computer Applications, Volume 77– No.12, September 2013.

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