# 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.**

## I. INTRODUCTION

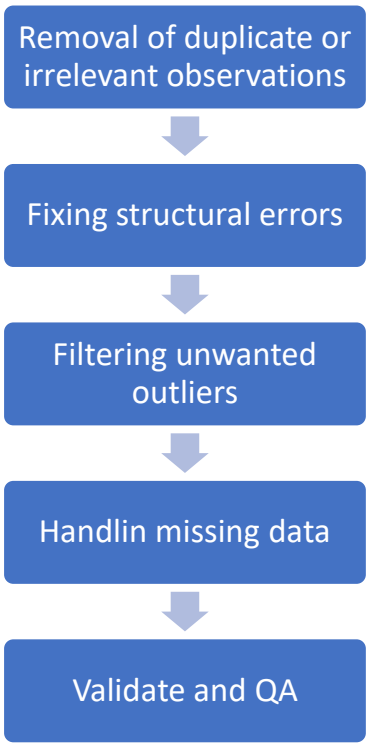
Data cleaning is the process that prepares data for analysis by removing duplicate or unnecessary data, modifying or updating data for analysis that is incorrect, missing, irrelevant, duplicated or improperly formatted.

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:




- 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

  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?

## 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.

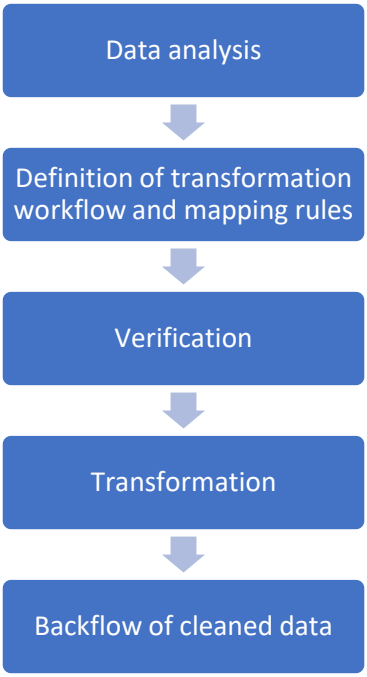


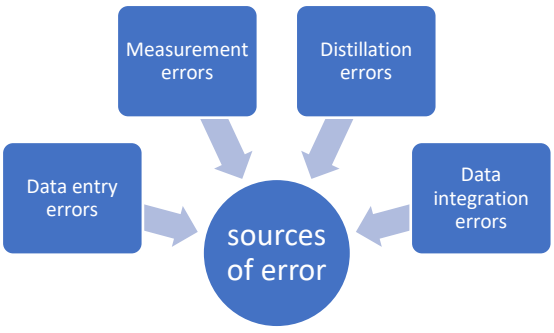

Figure: Data cleaning processes

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.

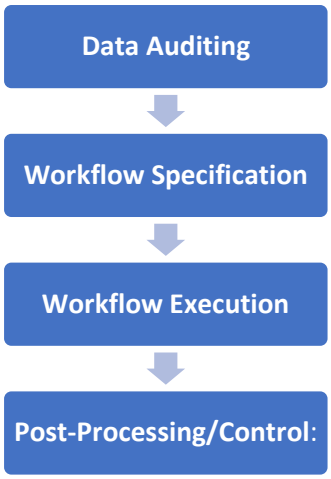



**Figure: DATA CLEANING PROCESS**

**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.

**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<br>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: comparative analysis of string data cleaning algorithms**

## ***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<br>Rapid Minor<br>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<br>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.

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


**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.
- 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.