

**DATA CLEANSING IN DATABASES**

**Dissertation submitted in partial fulfilment of the requirements for the**

**Degree of**

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

**RONY MATHEW**

**11301526**

**Supervisor**

**Mr. NIKHIL SHARMA**



**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**LOVELY PROFESSIONAL UNIVERSITY**

**PHAGWARA, PUNJAB (INDIA)**

**NOVEMBER 2017**

## **DECLARATION STATEMENT**

I hereby declare that the research work reported in the Dissertation II in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr.Nikhil Sharma. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

**RONY MATHEW**

**11301526**

## **SUPERVISOR'S CERTIFICATE**

This is to certify that the work reported in the M.Tech Dissertation on Data Cleansing in databases, submitted by Rony Mathew at Lovely Professional University, Phagwara, India is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Name: Nikhil Sharma

Date: 29/11/2017

## **ACKNOWLEDGEMENT**

I would like to express our deepest appreciation to all those who provided me the possibility to complete this dissertation proposal. I would like to express my special gratitude to my dissertation mentor, **Mr Nikhil Sharma**, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my dissertation and especially in writing this proposal.

Without great support of LOVELY PROFESSIONAL UNIVERSITY, it would be not possible to complete this proposal. The University has always been very responsive in providing necessary information, and without their generous support plan would lack in accurate information on current developments.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and will continue to work on their improvement, to attain desired career objective.

Hope to continue cooperation with all of you in the future.

**RONY MATHEW**

## **ABSTRACT**

Data cleansing is a way to maximize the accuracy of data in a system. It's different to data purging, which usually focuses on clearing space for new data, concentrating instead on ridding syntax errors, typographical errors or fragments of records. The goal of data cleansing is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases. In a perfect world, your business data would always be spot on. Every name, every number, every address would be 100% reliable, so that you can get the most out of your marketing efforts. Unfortunately, the reality of accurate information is somewhat lacking, regardless of your niche. Even more unfortunate is that inaccurate data costs you money.

Data cleansing allows you to improve the quality of your data, which in turn leads to smarter decisions within your business. And by increasing the accuracy and consistency of your data, you can improve your response rates and increase revenue. With accurate information, you can quickly and conveniently reach customers directly, notifying them of time sensitive promotions or information fast.

The bottom line is that bad data can affect many facets of a business. This does depend on how much your company relies on its data and how you organize and manage it, but even if data isn't a major priority, any data you do have should be accurate. It's the only way you can use it effectively.

# CHAPTER-1

## INTRODUCTION

---

Data cleansing is the procedure of finding and correcting imprecise or noisy data from a dataset. The procedure is used in databases where wrong, incomplete, imprecise or inappropriate piece of the data are observed and subsequently corrected, replaced or deleted. Industry organizations widely depend on knowledge whether it is the reliability of users' details or make sure correct invoices are sent and given to the customers. To make sure that the user data is utilized in the most useful and helpful manner that can increment the essential nature of the brand, marketing companies should give value to data excellence. Organization and deciding that the data is dirt free can produce considerable industry data. Marketing companies can halt more hassles like large cost come in processing mistakes, physical problem shooting, wrong invoice knowledge and shipments to imprecise address by cleaning the data. The detail of the user is every time altering because of replacement or further facts which have to be altered and the new details should imitate in the dataset. Marketing organizations can gain a huge class of advantages by cleaning data that may direct to reducing outfitted economy and incrementing takings.

These are some of the benefits of data cleansing:

- **Increases the effectiveness of client acquisition behavior:**  
Marketing organizations can significantly increment their customer acquisition doings by cleaning their data as a more capable manner list having correct data can be produced. During the industry process, company organizations should make sure that the data is correct, up-to-date and clean by periodically following data goodness routines. Fresh data may also make sure the biggest profits on email and postal campaigns as new manners of encountering outdated details are less. Various-channel user knowledge can also be maintained seamlessly which gives the organization with a chance to carry out unbeaten advertising campaigns in the coming times as they would be vigilant of the procedures to successfully reach out to the objective customers.
- **Increases Quality Decision Making Process:**  
The milestone of quality result production in a industry organization is client data. According to Sirius Decisions, dataset in an standard B2B company doubles each 12-18 months and while the dataset may be dirt free at the start, inconsistencies can happen at any time. Though various organizations falls to increase data quality management. However, more of them don't have details of the previous time value organize was conducted on the user's data. Correct knowledge and increased data are important to result production. High quality data can maintain good analytics and all-round marketing intellect which may assist increased decision production as well as

implementation. At the last, having correct data can help marketing organizations make increased decisions which may add to the rise of the industry in the good run.

- Streamlines Industry Practices:

Eradicating replicate data from the dataset can help marketing organizations to streamline industry procedures and keep a lot of economy. Data cleaning may also help in deciding if a specific job metaphors within the company can be altered and if those stages can be included somehow else. If trustworthy and correct sales knowledge is there, the act of a product or a service in the sell can be easily utilized. Data cleansing along with the correct analytics can also add the project to recognize a chance to start a good product or service in the sell which the users may like, or it can show different advertising avenues that the industry can seek. For example, if a industry campaign is a failure, the marketing project can see the different other industry channels that have the better user reply data and realize them.

- Increases efficiency:

Having a quality and appropriately maintained dataset can help marketing organizations to make sure that the customers are making the good use of their occupation hours. It can also stop the staff of the company from connecting clients with out-of-date knowledge or produce void vendor records in the system by easily serving them to make job with quality records thereby improving the staff's effectiveness and production. This is the reason for clean data decreasing the danger of attack as the staff has right to use to correct seller or user data when expenses or refunds are started.

- Increases profits;

Marketing organizations that work on increasing the regularity and improving the precision of the data can significantly increase their reply rates which falls in improved profits. Quality data can assist marketing organizations to extensively low the number of incoming mails. If there are any time responsive knowledge or promotions that the company wants to convey to their users directly, correct knowledge can help in connecting the users conveniently and quickly. Replicate data is another factor which can be effectively removed by data cleansing. According to Sirius Decisions, the economical force of replicate data is straight comparative to the time that it residue in the dataset. Replicate data can considerably ditch the company's capital as they will have to provide twice as much on a single user. For example, if many mails are sent to the same user, they might get displeased and might totally lose awareness in the company's goods and services.

#### Data Goodness

Top quality data desires to gain a set of value criterion. They are:

- Legitimacy: Extend to which the dealings match to described industry facts or constraints (see also legitimacy (figures)). While latest catalog method is used to create data-capture processes, legitimacy is nearly simple to make sure: incorrect data

rises mainly in heritage contexts (wherever constraints were not made in software) and where incorrect data-detecting technology was utilized (e.g., spreadsheets, wherever it is very tough to limit what a customer picks to put into a group, if group justification is not used). Data facts go down into the subsequent groups:

- Data Constraints – e.g., datae in a exacting column should be of a exacting datatype, e.g., Boolean, numbers (real or integer), date, etc.
- Range features: naturally, numbers or dates be supposed to go down inside a assured series. That is, they have smallest quantity and/or greatest acceptable datae.
- compulsory Constraints: Certain columns cannot be empty.
- Unique features: A field, or a mixture of fields, should be only one of its kind crossways a database. For example, no two people be able to contain the equal common security number.
- Set-Membership features: The datas for a column arrive from a set of separate datas or codes. For example, a person's sexual category may be Female, Male or unidentified
- Foreign-key features: This is the additional common container of set association. The set of datas in a column is distinct in a column of a different table that contains exceptional datas. For example, in a US taxpayer dataset, the "state" column is necessary to belong to one of the US's distinct states or regions: the place of allowable states/regions is kept in a disconnect States table. The word foreign key is rented from relational database expressions.
- Regular expression patterns: Frequently, text fields will contain to be validated this way. For example, phone numbers could be necessary to have the model (999) 999-9999.
- Cross-field justification: Assured situation that make use of various fields have to hold. For example, in laboratory drug, the total of the components of the disparity white blood cell tally have to be the same to 100 (because they are each and every one percentages). In a hospital database, a patient's date of release from hospital cannot be previous than the date of admittance.
- Precision : The amount of consistency of an assess to an average or a correct data - notice also correctness and accuracy. Accuracy is too firm to gain during data-cleaning in the normal case, since it needs using an outside basis of data which contain the correct data: that "gold normal" data is frequently not available. precision have been gained in some cleaning context, especially client phone data, by by means of outer dataset that equal up zip codes to various location (country and state), and also assist confirm that addresses inside these zip codes truely survive.
- Fullness: The measure to which all essential needs are recognized. Incompleteness is roughly not possible to make with data cleansing methods: one cannot suppose details that were not collected while the data in query was originally in record. (In a number of contexts, e.g., discussion data, it might be to fix incorrect by departing back to the initial foundation of data, i.e., re-questioning the matter, but even this do not ensure hit for the reason that of issues of recall - e.g., in an interview to collect data on food expenditure, no one is expected to identify closely what people ate six months before.



In the matter of these systems that contains assured columns must not be void. One might effort about the trouble by appointing a data that indicate "not known" and "lost", but supply of defaulting datas does not entail that the data has been completed inclusive.

- Integrity : The level where a group of events are similar in crosswise systems (see also integrity). Nonintegrity happens where two data values in the dataset oppose each another: e.g., a user is record in two various system as have two dissimilar present details, and just one of it may be precise. Correcting integrity issue is not all the time capable: it needs a diversity of policies - e.g., decide which data were kept more newly, which data set is expected to be more consistent (the other knowledge might be exact to a particular association), or just demanding to detect the fact by checking every data sets (e.g., call up the user).
- Similarity : The level to which a dataset events are mentioned by the similar units of assess in every system ( observe component of measure). In datasets united from various locales, load may be kept either in pounds or kilos, or have to be changed to a same assess using an arithmetic conversion.

It is essential for organization to have an efficient dataset, both for ensure capable contact with their users and maintain fulfillment standards. Data Cleansing or data scrubbing is the procedure of detecting and correct incorrect data from a dataset. With orientation to user data, data cleansing is the procedure of maintain steady and correct (clean) customer dataset through recognition & deleting of inprecise (unclean) data. Here, incorrect data stand for any data that is inprecise, not complete, or incorrectly formatted.

The final target of data cleansing and maintain a correct client dataset is to produce a "single client view" which means that here is only single evidence for each user that has all their valid data. This procedure of data cleansing and maintain clean user database provides variousple profit to organizations, counting:

- Data cleansing is a serious measure that help in maintain conformity with the Data Protection Act
- Data cleansing help to decrease expenditure in the form of spam emails and save on mail finance.
- Maintain a clean dataset allow for swift position of valid client data and decreases service reply time.
- It also increases the service value as all valid data is situated at similar place and outcomes in improved user practice.
- Quality client data can give more precise vision knowledge important to good sales directing and organization.

Though, keeping a fresh customer dataset is a hard task. client data is increasingly vibrant and tend to go out-of-date easily. Additional, many companies, based on diverse policies have various datasets. This lead to the single client being there on many datasets with bits of valid knowledge under each policy. A little steps that can increase in consolidate the client data preserve quality dataset are as follow:

### Process to Clean client Data:

- **Data audit:** The primary step towards data cleansing, is the total audit of all client datasets. The audit should be completed using statistical and dataset procedures to sense duplicancies and inaccuracy. The knowledge ought to be used to suppose features and location of anomaly, which may lead to origin reason of the issue.
- **Use Variousple method:** The procedure of auditing of a dataset must not be imperfect to analysis during statistical or dataset procedures and further steps like collecting exterior data and compare it with inner data can be use. in addition, if an company has constraint of time and effort, it can utilize the facilities of outer telemarketing company. Though, in this approach, the company needs to be careful with respect to the brand representation and the way of operational of exterior company.
- **Merge Data:** The procedure of cleansing the dataset must not be partial to just the recognition and elimination of dirty (inprecise) data from client database. It must be used as an occasion to merge customer data and extra knowledge like email addresses, phone numbers or extra contacts should be included whenever feasible.
- **Feedback:** The association should set up a direct mechanism where any inprecise knowledge gets report and gets efficient into dataset. For example, there must be a control and criticism method for emails as well as any email which is not delivered owing to an wrong address, should be report and the unfounded email address cleanse from the client data.
- **Repeat:** People's live are gradually more flattering dynamic and so connected details like addresses, telephone number, company email-id, alter regularly. Thus, the procedure of data cleansing be supposed to not be consideration of as a one-time procedure; as an alternative, it must made a part of the normal workflow. normal weed of inprecise knowledge and update customer dataset is the only way towards ensure clean client dataset.

#### 2.1 A Systematic Approach on Data Pre-processing In Data Mining

In this paper, they described Knowledge Discovery in Databases (KDD) procedure which help to reduce the complexity in dataset and provides good analysis and ANN training. For this, they used gathered data from field and soil testing labs. Data preprocessing is a procedure which needs manual work and time. Sampling, transformation, denoising, normalization and feature extraction are different processes of data preprocessing. They found not complete, noisy, not consistent, poor quality data in the database. They use binning, clustering, data combination, data transformation, data reduction, Data aggregation and attribute subset collection, Dimensionality decrease, Numerosity drop techniques to use noisy data.

Data cube aggregation, Wavelet Transforms, Principal Component Analysis, Sampling, Data Discretization and Data summarization, Top-down discretization or splitting, characteristic Subset collection, Dimensionality decrease, Bottom-up discretization are the data reduction methods used in this paper.

#### 2.2 Classifier Based Duplicate Record Elimination for Query Outcomes from Web Databases

In this paper, supervised data matching methods are used. A new record matching process is introduced known as Unsupervised Duplicate Elimination (UDE) to identify and delete data from records. Three dividers named weight factor resemblance summing classifier, hold up vector machine classifier and single class hold up vector machine classifier are iteratively working with UDE. Problem Definition, Element Identification, Ontology matching are the processes in duplicate detection and identification. Ontologies should be presented in textual and graphical format. Normally, graphical format are proposed for simple understandability. Entity-Relationship (ER) models are famous for depict the main relations of the record features.

#### UDE ALGORITHM :

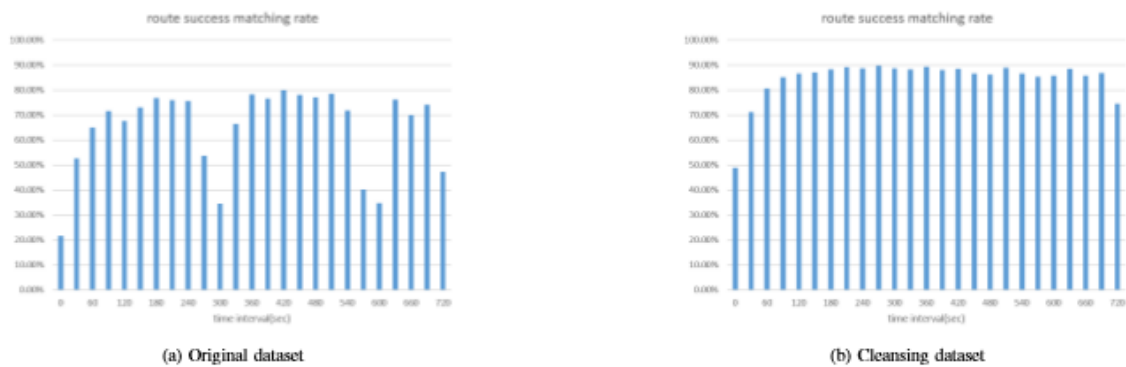
- 1) Weight factor Similarity Summing Classifier ( $p_i = \sum_{v \in D} v_i$ )
- 2) Support vector Machine Classifier
- 3) single-Class Support vector Classifier
- 4) Evaluation matrix

$$\text{Precision} = \frac{\text{Number of correctly identified duplicate pairs}}{\text{Number of all identified duplicate pairs}}$$

Replication detection and replicate elimination is the main subject of this paper. To overcome this problem, Unsupervised Duplicate Elimination (UDE) is used.

## 2.3 Trajectory Data Cleansing Using HMM

Vehicle trajectory database contains GPS errors and low sampling rate. It is used to be cleansed before taken for training of data. In this paper, they prefer a HMM (Hidden Markov Model) based system to reconstruct the dataset. They have used database from OpenStreetMap of Beijing, taxi trajectory database from microsoft research lab, Asia. The produced system is of three modules, noise filter module, HMM map match module, and Aroute searching module. They prefer a HMM based trajectory data cleaning system to clean and rebuild the missed traveling routes of vehicles from the given trajectory database. They extend the hidden Markov process to match the trajectory data on the digital road map and prefer a group of formulae for the transition probability or observation probability. They increase the quality of the low sampling rate trajectory database .They carry out wide experiments to authenticate the performance of the method.

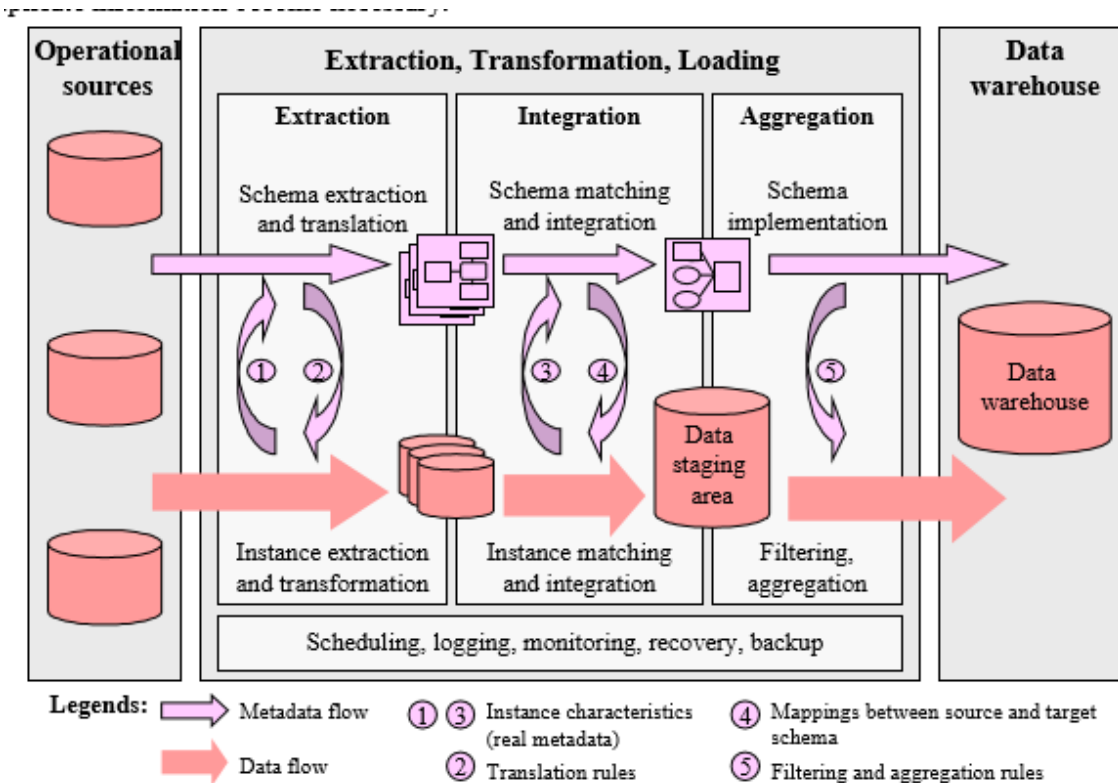


## 2.4 An efficient domain-independent algorithm for detecting approximately duplicate database records

In this paper, they presents an algorithm for identifying clusters of copy data. There are three facts behind this algorithm. First is, a process of smith-waterman algorithm is used to find out minimum edit-distance. Second, union algorithm is used to find out duplicate relationships. Third, the algorithm use a precedence queue of cluster subsets to react accordingly.

## 2.5 Data Cleaning: Issues and Current Approaches

In this paper, they give an overview of important data cleansing solution approaches. They have described ETL process in detail. The ETL process can be shown as,



They also shown various data cleansing issues in schema level and instance level. They discussed single source and various source issues. The various phase of data cleansing are Data analysis, Definition of transformation workflow and mapping rules, Verification, Transformation, Backflow of clean data.

## 2.6 An Efficient Duplication Record Detection Algorithm for Data Cleansing

This paper was proposed to review and differentiate different algorithms of empirical techniques to find out most efficient algorithms in terms of effectiveness and correctness. They have differentiate all the empirical techniques and found out the best one is DCS++. After successfully implemented the DCS++ algorithm, they produced a better variation in terms of efficiency and accuracy. The difference is implemented in C#. They also discussed single source and various source issues under schema and instance level. The described empirical techniques are blocking and windowing.

With the access of precise string matching algorithm, they use that the correctness of proposed algorithm is similar as correctness of DCS++ algorithm and there is no increase in number of comparison. The output are not terrible as 100% accuracy and 70.63% remember is achieved. planned Algorithm is implement with the customized Recursive Algorithm which is performing more precisely and effectively than of DCS++ by Recursive Algorithm with low threshold datas but with increased threshold datas which have same output.

## 2.7 Different Similarity procedures to Identify Duplicate Records in Relational Databases

In this paper, four different similarity measure tests are verified. Besides the use of method quality, the option of same events is very significant for data contrast used in those method. The Standard Blocking Method used in SIMR gave the most excellent outcomes with Jaro resemblance measure by all trial of copy records identification excellence. In the research that go after and in the version of the accessible or in the expansion of the new copy records recognition methods, it would be helpful to examine the element which would, on an exacting number of records, carry out the choice of the most appropriate parallel procedure to evaluate the date on some characteristic records stage.

## 2.8 Study of Data Cleaning & Comparison of Data Cleaning Tools

This research paper uses an general idea of data cleaning issues, data goodness, cleansing procedures and assessment of data cleansing tool. Set of criteria of data quality,



They described about data auditing, workflow specification, workflow execution, post processing/control. They made comparison among different data cleansing tools like MS EXCEL DATA CLEANER, RAPIDMINOR, WINPURE CLEAN & MATCH. every tool have its own exact features which are depending upon the facts we can make use of the tool to dirt free data. In future work we can verify other functionality of these tools and propose own.

## 2.9 BigDancing: A System for Big Data Cleansing

This paper describes an important obstruction since data cleaning often comprise expensive calculations like enumerate pairs of tuples, maintaining dissimilarity joins, and selling with user-defined function. In this paper, they present BigDancing, a large Data Cleaning system to begin efficiency, scalability, and accessibility issue in data cleaning. The structure can perform on top of the majority similar general point data dealing out platforms, range from DBMSs to MapReduce-like works.

BigDancing take these policies into a sequence of changes that helps dispersed computation and different optimization, such as common scans and particular joins operator. New outcomes together copied and genuine database shows that BigDancing performs present baseline system up to extra than two instructions of size lacking dedicatng the value given by the fix algorithms.

## 2.10 A Review of Data Cleansing Concepts – Achievable Goals and Limitations

The paper gone into investigate a number of research works performed in the region of data cleaning. A careful analysis into these accessible works was learnt to decide the realizable goals and the issues that arise based on the approach performed by the researches. The recognition of issues by the majority of these researches has led into the growth of many framework and system to be performed in the region of data warehousing.

In this paper, they grouped data quality issues into syntactic anomaly which worry data format and datas for data presentation. Lexical errors named discrepancy between the arrangement of data items or the particular format. Domain error identify wherever the known data for a characteristic A do not be conventional to the probable field arrangement. irregularity deal with non uniformed exercise of data and other abbreviation which usually are visible while dissimilar currency arrangement is use to identify member of staff wages.

### 2.11 A Comparative Analysis of Data Cleansing Tools

This paper represent a comparison and analysis of data cleansing tools and features and benefits of each tool. furthermore it show relative study of data cleansing tools and decide the top one. There are a great diversity of tools obtainable that can use to hold data cleaning. in addition, a lot of arithmetic programs has data corroboration build in, which can choose some error mechanically.

### 2.12 A Study over Issues and Approaches of Data Cleansing/Cleaning

The paper identify the data value issues that are adorned by data cleansing and provide an summary of the central approach as a resolution. moreover in data warehousing, the data cleansing is an significant piece of the commonly-calle ETL procedure and is BI Tool can be included with ETL Tool or vice-versa. The present technique that support the data cleansing are to discuss. Elements of data cleansing are:



### **2.13 E-Clean: A Data Cleaning Framework for Patient Data**

This scheme uses the take out, change and weight model as the scheme main procedure model to serve up as a rule for the execution of the scheme. In addition that, parsing technique is also use for the recognition of unclean data. The way that they decide for similar attributes is regular expression. Along with those data cleansing algorithm, k-Nearest Neighbors algorithms is chosen for the data cleansing element of the program since it is easy to recognize and simple to apply.

They also discussed the advantages and disadvantages of categorization and Regression Trees (CART), k-Nearest Neighbor, Genetic Algorithms.

### **2.14 A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse**

In this paper,the frameworks accessible for data cleansing offer the basic service for data cleansing such as quality collection, structure of token, collection of cluster algorithms, collection of similar functions, collection of removal functions and combine functions. The research work deal concerning the novel frameworks for data cleansing. They also present a result to hold data cleansing procedure by by means of a novel frameworks intend in a chronological orders.

The steps are as follows:

- A. Selection of attributes
- B. Formation of tokens
- C. Selection of clustering algorithm
- D. Similarity computation for selected attributes
- E. Selection of elimination function
- F. Merge

The fresh frameworks consist of six basics: collection of attribute, arrangement of token, collection of clustering algorithms, resemblance computation for chosen attribute, collection of removal functions and Merging. The frameworks will be practical to build up a great data cleansing tool by means of the accessible data cleansing technique in a chronological order.

### **2.15 DUPLICATERECORD DETECTION FORDATABASECLEANSING**

As a product of this research work, contrast amongst standard duplicate elimination algorithms (SDE), sorted neighborhood algorithms (SNA), duplicate elimination sorted neighborhood algorithms (DE-SNA), and adaptive duplicate detection algorithms (ADD) is provide. An example is also develop which correspond to that adapt duplicancy detection algorithms is the best resolution for the crisis of duplicancy record detections. For estimated similar of data record, string matching algorithm has been implement and it is completed that the outcome are much improved with recursive algorithms with word bases.

Adaptive replica discovery algorithms outperform the additional algorithm in terms of correctness but not in competence as these algorithms take more time for implementation



since of both behaviors. The algorithms is field self-determining and the dissimilarity planned for estimated similar increase the accurateness of the outcome.

#### 3.1 Problem Formulation

From the current data cleansing algorithms and techniques, we defined a problem in the following areas.

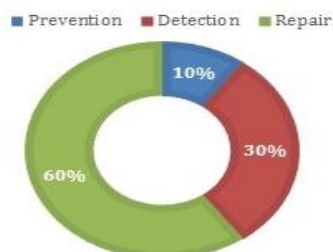
- A) ACCURACY
- B) CONSISTENCY
- C) TIMELINESS
- D) DATA ADDED
- E) ACCESSIBILITY

An optimized algorithm need to be developed which consists of all these properties and should be easy to handle and maintain. We formulated the research work on the basis of various existing algorithms and techniques like wave filtration which is used by google, mean,median and mode methods which are found in research methodologies inorder to develop an optimized data cleansing algorithm that trains all the noises in a dataset and correct it with the most appropriate method.

Data have value if they assure the supplies of the proposed use. There are several factor comprise data value, includes precision, totality, reliability, appropriateness, trust worthy, and interpretability. Timeliness also affects data value. Two other factor affects data quality are believability and interpretability. Believability reflects how much the data are to trust by user, while interpretability reflects how simple the data are unstated.

- Increases the Efficiency of Customer Acquisition Activities:
- Increases Decision Making Process
- Streamlines Industry Practises
- Increases Productivity

**TYPICAL CLEANSING %**



## CHAPTER-4

### SCOPE OF THE STUDY

---

Each industry and business require the fresh and noise free data. Data warehouse fill and always restore huge quantity of data from diversity of sources so the likelihood that some of source contain dirty data is far above the ground. Data cleansing is used so that the accuracy of their data is vital to keep away from incorrect termination. Data cleansing is essential step in any data- associated project. The require of data cleansing is for the increment of the data mining effect. Today's real-world datasets are highly vulnerable to noisy, misplaced, and incoherent data due to their naturally giant size (often numerous gigabytes or more) and their expected beginning from various, diverse sources. Low excellence data will lead to inferior mining outcomes. There is huge scope of data cleansing, while each association is by means of data and data can be from more than one sources , so to extract value and efficient result data cleansing is too essential.

The field is still in the nascent stages and will slowly find its applications in many new areas apart from industry where it is currently booming. IoT, Healthcare, HR, Education, Governance, Agriculture are some of the relatively untouched areas for analytics till now and the ones where you will see more proliferation in the coming years. Data cleansing is a procedure to convert a raw data which is distinctly collected into the useful knowledge. It is widely used for implementation of data into a useful knowledge from effective data collection. However, it totally depends on nature of data that how to do its interpretation in an appropriate way.

- Data cleansing process the work in such a manner that it allows industry to more proactive to grow substantially.
- It optimizes large database within the short time and works industry intelligence which is more vital to organizational growth.
- It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.
- It includes tree-shaped structure to understand the hierarchy of data and representation of the set of knowledge described in the database.
- Brings a genetic way of classification of different sets of data items to view the data in the quick glance.

Data cleansing tools clean through datasets and identify earlier hidden patterns in one step. A model of pattern discovery is the study of retail sale data to distinguish seems not related product that are regularly purchase jointly. Other model discovery issues include detect false credit card communication and identifying abnormal data that could correspond to data entry keying error.

## CHAPTER-5

### OBJECTIVE OF THE STUDY

---

Objectives of the study are:

- To increase existing data preprocessing algorithms and techniques which are used by organization and industry.
- To calculate efficiency and run time taken by each techniques.
- To measure the degree of belief on each algorithm and the level of noises each dataset contains.
- To study large datasets and analyze various errors and noises that occurs.
- To create a strong data cleansing technique than an existing one, which is easy to use by any user.
- To have the universal perceptive of data cleaning.
- To learn the data cleaning tool.
- To cleanse the data using different data cleansing tools like Ms Excel Data cleaner, RapidMinor and Winpure Clean & Match.
- To understand the various methods used by companies.

### PROPOSED RESEARCH METHODOLOGY

---

The hypothetical study has been done from different sources like journals, research papers, books and internet. Data Cleaning tools has been used for the clean-up of different kind of excel data . On the source of outcomes obtained evaluation of tools has been done to find the finest tool for data cleansing. In my work, a bank dataset is taken and trained it using wave filter algorithm and implemented using python on anaconda platform.

Filter is a procedure that removes some unnecessary workings or feature from dataset. Filtering is a set of signal processing, the major quality of filter being the whole or fractional control of some feature of the signal explanation needed. Most frequently, this means remove some frequency or regularity band. though, filter do not completely act in the occurrence domain; mainly in the field of data cleansing many other targets for filtering exist. Correlation can be detached for assured frequency component and not for other without have to act in the occurrence field. Filter is broadly used in electronics and telecommunication, in radio, television, audio recording, radar, control systems, music synthesis, data mining, image processing, and computer graphics.

The technique which is going to use is wave filtering which is used by Google and also some common techniques like mean,median and mode inorder to handle the outliers in dataset. This is going to implemented in Python on anaconda platform and the currently training dataset is of a bank which contains about 1,00,000 data. Python is a language that is remarkably easy to learn, and it can be used as a stepping stone into other programming languages and frameworks. Most automation, data mining, and big data platforms rely on Python. Traditional Python gives you just a basic platform where you have to install your desired packages manually ( this even does not have NumPy and Pandas installed), Anaconda gives you just everything.

Anaconda contains porting for all the popular python libraries that can be used in data science. The most important being scikit-learn, numpy,pandas, scipy etc. Plus it also comes with the jupyter notebook and Ipython distribution. So, it saves you from importing numerous libraries separately. It has a large package manager. It comes with a lot of helpful things preinstalled and all set to run for a classic machine learning project.

## **CHAPTER-7**

### **EXPECTED OUTCOMES**

---

The expected outcome is a dataset which contains no incomplete, incorrect, imprecise or inappropriate parts of the data and then replacing, modifying, or deleting the dirty data. The outcome comes from a noisy dataset. The cleansed data can be used for data mining and data analytics. An algorithm is needed to be proposed which is more optimized than existing ones and is easy to use and maintain. The efficiency will be considered by means of running time and how precise the data is cleaned.

## CHAPTER-8

### SUMMARY & CONCLUSIONS

---

---

Data cleansing is very essential component of data mining. From the study we can observe there are diverse types of issues in data cleansing. Data cleansing method and approach depends on the kind of data which we wish for to fresh and according to that we relate exacting method. The study also presents a assessment of data cleansing tool and determine the better tool. Each tool has its own precise feature and depends upon the data we can utilize the tool to fresh data. In future works we can test other functionality of the tools.

A fresh database that is used for email advertising campaigns can considerably decrease bounce backs and efficient knowledge can be altered into a key industry benefit. There is no suspicion that data cleansing can aid marketing organizations to realize industry goals with ease. Marketing organizations can also utilize the services of an expert data cleansing service supplier for improved arrival on asset on their advertising behavior. Data cleansing can not only keep time and money for marketing organizations, but it can also make sure that the company achieves overall outfitted effectiveness.

## REFERENCES

---

- [1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth “Knowledge Discovery and Data Mining: Towards a Unifying Framework” in KDD-96 Proceedings 1996, AAAI.
- [2] G.Kalpana, R.Prasanna Kumar, T.Ravi “Classifier Based Duplicate Record Elimination for Query Outcomes from Web Databases” in 2010, IEEE.
- [3] Alvaro E. Monge, Charles P. Elkan “An efficient domain-independent algorithm for detecting approximately duplicate database records” in University of California.
- [4] Arfa Skandar, Mariam Rehman, Maria Anjum “An efficient Duplication Record Detection Algorithm for Data Cleansing” in International Journal of Computer Applications (0975 – 8887) Volume 127 – No.6, October 2015.
- [5] Qin Wang, Min-Te Sun, Kazuya Sakai “Trajectory Data Cleansing Using HMM” in 2017 46th International Conference on Parallel Processing Workshops.
- [6] Jasdeep Singh Malik, Prachi Goyal, Akhilesh K Sharma ” A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules”
- [7] S.S.Baskar, Dr.L.Arockiam, S.Charles “A Systematic Approach on Data Pre-processing In Data Mining” An international journal of advanced computer technology, 2 (11), November-2013
- [8] Dulaga Hadzic, Nermin Sarajlic, and Jasmin Malkic “Different Similarity Measures to Identify Duplicate Records in Relational Databases” in IEEE, 2016.
- [9] Erhard Rahm, Hong Hai Do “Data Cleaning: Issues and Current Approaches” in Microsoft Research, Redmond, WA.
- [10] Sapna Devi, Dr. Arvind Kalia "Study of Data Cleaning & Comparison of Data Cleaning Tools” in IJCSMC, Vol. 4, Issue. 3, March 2015, pg.360 – 370.
- [11] Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, Si Yin, “BigDancing: A System for Big Data Cleansing.”in University of Waterloo.
- [12] Benilda Eleonor V. Comendador, Lorena W. Rabago and Bartolome T. Tanguilig III “An Educational Model Based on Knowledge Discovery in Databases (KDD) to Predict Learner’s Behavior Using Classification Techniques” in Technological Institute of the Philippines Quezon City, Philippines.
- [13] Nidhi Choudhary “A Study over Issues and Approaches of Data Cleansing/Cleaning” in International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 2, February 2014.



- [14] Shivangi Rana, Er.Gagan Prakesh Negi, Kapil Kapoor “A Comparative Analysis of Data Cleansing Tools” in Volume 6, Issue 4, April 2016, International Journal of Advanced Research in Computer Science and Software Engineering.
- [15] Mariam Rahman, Vatcharapon Esichaikul “DUPLICATE RECORD DETECTION FOR DATABASE CLEANSING” in 2009 Second International Conference on Machine Vision.
- [16] Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin, Ong Siong Lee “E-Clean: A Data Cleaning Framework for Patient Data” in 2011 First International Conference on Informatics and Computational Intelligence.
- [17] Boye A. Høverstad, Axel Tidemann and Helge Langseth “Effects of Data Cleansing on Load Prediction Algorithms” in IEEE, 2013.
- [18] Ricardo Almeida, Paulo Oliveira and Luís Braga, João Barroso “Ontologies for Reusing Data Cleaning Knowledge” in 2012 IEEE Sixth International Conference on Semantic Computing.
- [19] J. Jebamalar Tamilselvi and Dr. V. Saravanan “A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse” in IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5, May 2008.
- [20] Kofi Adu-Manu Sarpong, John Kingsley Arthur “A Review of Data Cleansing Concepts – Achievable Goals and Limitations” in International Journal of Computer Applications (0975 – 8887) Volume 76– No.7, August 2013.
- [21] Ray Y. Zhong and George Q. Huang, Qingyun Dai “A Big Data Cleansing Approach for n-dimensional RFID-Cuboids” in Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.
- [22] Ioannis D. Schizas “Distributed Data Cleansing via a Low-Rank Decomposition” in 2013 IEEE.
- [23] Liz Aranguren Pachano, Taghi M. Khoshgoftaar, and Randall Wald “Survey of Data Cleansing and Monitoring for Large-Scale Battery Backup Installations” in 2013 12th International Conference on Machine Learning and Applications.

**KDD** : Knowledge Discovery in Databases

**UDE** : Unsupervised Duplicate Elimination

**ER** : Entity Relationship

**HMM** : Hidden Markov Model

**ETL** : Extraction, Transformation, Loading

**CART** : Classification And Regression Trees

**SDE** : Standard Duplicate Elimination

**ADD** : Adaptive Duplicate Detection

**DE-SNA** : Duplicate Elimination Sorted Neighborhood Algorithm