ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Principals and Methods of Data Cleansing for Removing Erroneous Data From Database.

Jay Kumar M. Purohit[1], Dr S. B. Kishor[2].

1. ResearchScholar, Gondwana University, Gadhchiroli.
2. HOD,Dept of Computer Science, S.P.Collage, Chandrapur.

Manuscript Info

Manuscript History:

Received: 14 December 2015
Final Accepted: 19 January 2016
Published Online: February 2016

Key words:
Data Harmonization, Legacy Data, Data Cleansing, Outliers, Geocode, etc.

Corresponding Author

Jay Kumar M. Purohit.

Abstract

Cleansing data from impurities is an integral part of data processing and maintenance. In order to fulfill the requirement it demands certain kinds of technological methods to improve the quality of data, to provide the Healthy data sources which further gives rise pure and nourished data which can be more efficient then the preliminary data . This paper includes a research work as well as field work from various location and studies of data cleansing problems, approaches and methods.

Copy Right, IJAR, 2016,. All rights reserved.

Introduction:-

Data Cleansing or Cleaning is the process of determining the faulty data or dirty data, inaccurate data as well as incomplete and unreasonable data from the database. The inconsistencies are mainly caused because of mistakes made by user while entering the data; mainly due to typographical errors (typing errors).

Data Cleansing like data validation performs validation & it performs very strong and strict validation as compare to the data validation for ex:-

- Rejecting an home address without pin code
- Rejecting such entries which have most of the similar entries that matches the existing known records

Data cleansing also improves the data quality by making short codes into certain meaningful form such as:

St->street,rd->road,Ny->newYork,Nu-Deli->NewDelhi,c-pur-Chandrapur etc.These process of standardizing data is also known as Harmonization of data [(1)(2)]

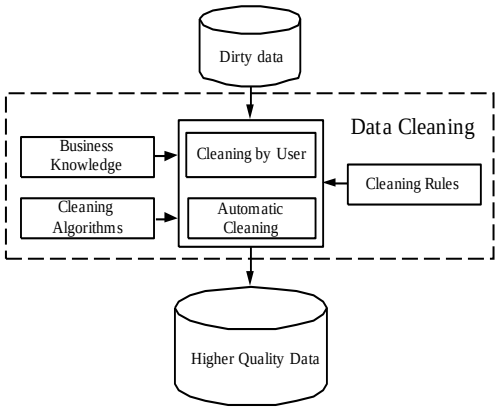
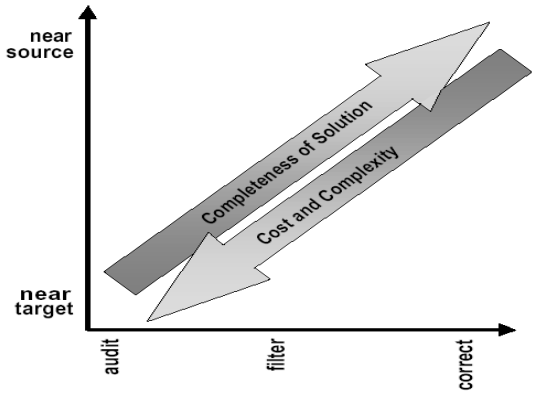


Fig.1 The Principle of Data Cleaning

Characteristics Of Data Cleaning:-

- **Error Reduction**:-The major objective of Data Cleaning is to improve quality of data so that it becomes fit for use.
- **Error Understanding**:-One of the views of Data Cleaning is to develop good understanding about the errors so that error propagation can be controlled & Data quality can be further improved.
- **Error Documentation**:-Error documentation is one of the important parts of the Data Cleansing process. It involves correction or changes made to the errors are documented (noted).
- **Error Preventition:**-prevention of error is always recommended other than later detecting and correcting the errors. In primary species database.
- **Merge & Purge**:-Eventually it is required to merge the two or more Databases. it gives rise to new kind of errors that is error of duplicate records. Such process of merging the different list of information from different sources and removing the duplicate records is called as purging. [3]



Data cleansing economics:-

Clean and Accurate CRM (Customer Relationship Management) data is essential for effective sales, Marketing and customer management strategies. It's an economical method of making CRM Data useable again and delivers range of benefits all of which improve your company profitability & productivity. Every Business depends on finding the best tool that will well equip their business towards success & Data Cleaning tool is one of it.

The major objective of this paper is to discuss about the importance of Neat & Clean Data in Database or in Data Warehouses as well as various challenging issues related with Data Cleansing or Data Scrubbing. Data cleansing or Data Scrubbing is the act of detecting and correcting, removing corrupt or inaccurate records from record set, table, or Database.

Hackers try to submit harmful and bad data directly to server in order to conquer the security. programs such as **Code Injections, SQL Injections, Buffer Overflows** are responsible for breaking server security and exploit data hence Data Cleansing are essential methods to protect server data. Data Cleansing prevents submitted faulty and dirty data from execution as well as makes server attacks impact less. It afterwards adds or includes data to log (database) which sends to administrator for further investigations.

Data cleansing vs data validation:-

The term Data Validation refers to rendering data or inspecting the data into predefined format during the time of submission. For ex- In Column of E-mail address the correct E-mail address is expected so it should be examined at the time of entry; where valid E-mail address is expected other than that of invalid e-mail address. Invalid e-mail address means error or an attempt to injection attacks.

By using suitable Data Validation techniques injections attacks and error can distinguished and properly handled. Errors can be overcome by using reasonable messaging were as attacks can be rejected and minimized by using accurate data cleaning techniques. Data Cleaning and Data Validation when combined together it helps to develop robust programs which deals with server security, protections and leads to strong benefits to business success.

Pros & cons:-
The pros and cons of Data Cleansing techniques are as follows:-
Pros:-

1. It sieges injections Attacks.
2. It prevents Buffer Overflows.
3. Protects server and Networks.
4. Secures Database Information.

Cons:-
1. Additional programming cost.
2. Additional programming time.

Role of data cleansing in i.t & business administration:-
Now a day's role of Computer is only limited up to computerized typewriter as well as computational device. But in the era of A.I it is expected that computer must supply more reasonable, scientific, technical answers to the query's & problems. They must use neural networks to Data Mining in order to produce more actionable reports.

In order to provide more reasonable and scientific answer to provide computer needs to provide more subjective data rather than objective. To store such type of information strong data storing techniques with relational data storage is required; to make this information more functional and actionable each of these facts must be recognizable and distinguishable.

Since we are using more complex data storing techniques and it is required that data stored in these techniques are easily recognizable and distinguishable at the time of collection (acquisition) so we need neural network with neural reorganization process to properly identify and integrate acquired data. They must use sophisticated matching algorithms that take into account phonetics, abbreviations and subject specific terminology along with statistical probabilities.

Data cleansing is an important part of data acquisition. it consist of following 4 steps,

1. Establish table like structure to represent basic validation rules.
2. Data acquired or collected as per the rules of validation table.
3. Seeking out entire acquired database which fails to satisfy validation rules.
4. Locating improper data element from the acquired dataset should be removed or updated as per the validation rules.

Scope and limitations:-
The scope of this Research is to explore the erroneous data and correct it. So that quality of data is improved and chances of re-occurring of error is less. for cleaning the data some software tools are developed such as Biogeomencer, Biota which are specifically reserved for primary species data or species occurrences of data.[(4, 2005)]

Data Cleansing is time consuming and expensive process. So after having performed data cleansing achieving data collection free of errors, one would want to avoid re-cleansing of data in its entirety after some values in data collection is changed.

Conclusion:-

Our data cleaning approach satisfies several requirements. First of all it detects and removes all major error and inconsistencies in data. Algorithm for data cleaning and data transformation specified and useful for other data sources. Data cleaning is an essential task in order to get correct and qualitative data.[4]

In this paper, we have found that how data cleansing is essential and important for business organization and server security. Data cleansing play vital role for business improvements, it focuses on various tools & techniques for data validation as well as find erroerneous data from the database. it is also found that most of the Datawarehouses,databases,business organization are using data cleansing software such as **Win pure, Data Ladder** etc,that produce correct and precise data for decision making in business.

The famous philosopher Ben Franklin says "An ounce (small quantity) of prevention is worth a pound of cure".

References:-

1. Wikipedia Free Encylopedia.
2. (n.d.). jan zing. Research on Data Acquisition and cleansing .
3. (n.d.). Hernandez M.A.; Stolfo S.J. Real world data is Dirty .
4. (2005). principal and methods of data cleansing.(primary species database).