Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/2125
DOI URL: http://dx.doi.org/10.21474/IJAR01/2125
ISSN NO. 2320-5407

RESEARCH ARTICLE

PRINCIPLES AND METHODS OF DATA CLEANSING FOR REMOVING ERRONEOUS DATA FROM DATABASE

Jay Kumar M. Purohit[1] and 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: 25 September 2016
Final Accepted: 27 October 2016
Published: November 2016

Key words:-
Extract Transform Load (ETL), ID Validation, Alphabetic Validation, Numeric Validation .

Abstract

Now days it's became universal truth that "To Error is human nature & to forgive, forget is God's nature". It means that making error is human right and there is no work done by the human being which is completely and 100% error free. For ex errors made by data entry operator while entering the data, error made at the time of data collections, error made by the researcher at the time sample selection as well as sample selection tools and techniques, some people thinks that errors in data files are acceptable at certain extent but there are some applications where clean data is essentially required where faulty data is never ever acceptable such as, In banking system is not acceptable to deposited or withdrawals money in or from wrong account. This paper is based on the concept about how to avoid dirty and faulty data to get populated in the databases as well as data files

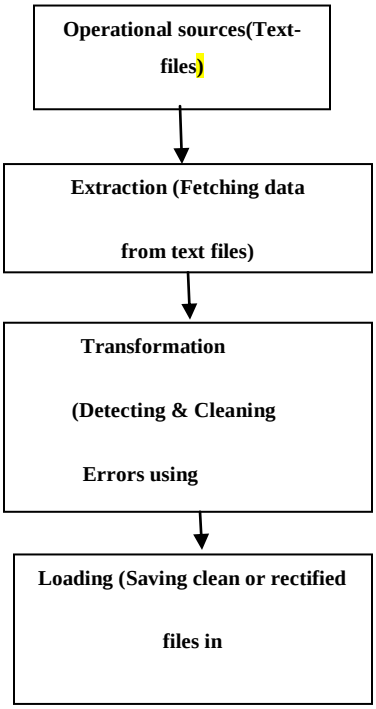
Introduction:-

Research work carried out by the researcher in this paper is about the designing an algorithm which eliminates the dirty,errorneous data from the database or data files,textfiles.Since incorrect data works as infectious virus which spreads from files to files and results in great economic losses, great expenses. Algorithm is designed such a way that it useful for data cleaning on any type of data sources, b'coz clean data is essential requirement for quality data.

In order to implement so designed algorithm in the form of real time software, Developer needs to pay attention, towards quality checking algorithm at the data entry point as well as correcting the corrupted data files which is full of faulty and corrupted data. Here we are trying to focus on data cleaning in text files by the process of ETL(Extract transform Load) process

Corresponding Author:- Jay Kumar M. Purohit.
Address:- ResearchScholar,Gondwana University, Gadhchiroli.

ETL functioning model:-



ETL system is so designed to work on any type of record set such simple textfiles, releated data files to correct the errors of type alphanumeric errors, invalid gender, invalid ID.Database stores data in tabular format and algorithm works on each field value depending on its type and nature.

In order to make transparent the functioning of ETL let us consider the example of college information system into which information about studentID, CourseID, FacultyID, DeptId is stored into database .The process initiates at the point of data entry in case of duplicate or redundant data it prompts the errors messages to the user and corrects the wrong entry ,the entry will be not submitted to the database until it get corrected.

Types of Errors:-

Here types of errors are considered in the college information system are as follows

1. Numeric values in place Non-numeric (Name, Gender, and City)
2. Non-numeric values in place of numeric (phone no, registration no, date)
3. Invalid or Redundant ID's
4. Invalid Gender.

ID Validation Algorithms:-

Step 1 start
Step 2.check for alphabet in input ID, Eliminate if occurs
Step 3 Concatenate id's.with preceding Zeros (0) as per following rules
3.1. If length of ID is equals to 1 the replace ID="00"+ID
3.2. If length of ID is equals to 2 then replace ID="0"+ID
3.3. If length of ID is greater than 3 then take only 3 characters eliminates rest
Step4 Change the ID as per the following rules
4.1. If student ID then ID="S"+ID
4.2. If Department ID then ID="D"+ID
4.3. If Course ID then ID="C"+ID
4.4. If Subject ID then ID="Sub"+ID
Step 5. Return clean ID

Sample Output:-

Before Cleaning:-

Sid	Sname	Gen	city	Co-no	Phno	Course id
S001	A.Rao	Mle	NAG	11OO23	9422907637	C1
S002	S.jyoti	FML	MUM	1122008	9420080013	C02

After Cleaning:-

Sid	Sname	Gen	city	Code no	Phn	Course id
S001	A.Rao	Male	Nagpur	110023	9422907637	C001
S002	S.jyoti	Female	Mumbai	1122008	9420080013	C002

Gender Validation Rules;-

Step 1 start

Step 2 Go through the alphabetic validation. check for alphabet in input Gender, Eliminate if occurs 'm','M','mle',Mle' with 'Male

Step 3 Go through the alphabetic validation. check for alphabet in input Gender, Eliminate if occurs 'f','F','fmle',fle' with 'Female'

Conclusion:-

In these paper researcher has proposed some of the data cleaning algorithms for databases and text files. It can detect errors, programmatically create valid values and refine the fields in the database.

The information age has meant that collections' institutions have become an integral part of the environmental decision making process and politicians are increasingly seeking relevance and value in return for the resources that they put into those institutions. It is thus in the best interests of collections' institutions that they produce a quality product if they are to continue to be seen as a value-adding resource by those supplying the funding.

Best practice for database information in museums and herbaria and institutions maintaining survey and observational information means making the data as accurate and possible, and using the most appropriate techniques and methodologies to ensure that the data are the best they can possibly be. To ensure that this is the case, it is essential that data entry errors are reduced to a minimum, and that on-going data cleaning and validation are integrated into day-to-day data and information management protocols.

References:-

1. Arup kumar Bhattacharjee,Atanu Mallick,Arnab Dey .Sadananda B.Data Cleaning in Text Files
2. Wikipedia Free Encylopedia
3. R. Cody, "Data cleaning 101," Proceedings for the Twenty-Seventh SAS User Group International Conference. Cary, NC: SAS Institute Inc
4. Dr. Mortadha M. Hamad and Alaa Abdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse". Computer
5. Science Department. University of Anbar, Ramadi, Iraq.
6. Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin and Ong Siong Lee, "E-Clean: A Data Cleaning Framework for Patient Data".
7. School of Computer Sciences. University Sains Malaysia Penang, Malaysia.