An Algorithm For Rule Base Design In Data Cleaning

Al-Amin, Ovijit Roy, Md. Alomgir Hossain
Student, Department of Computer, Science and Engineering
IUBAT – International University of Business Agriculture and Technology Dhaka, Bangladesh
aminal23023@gmail.com

Student, Department of Computer Science and Engineering
IUBAT – International University of Business Agriculture and Technology Dhaka, Bangladesh
Ovijit786@gmail.com

Senior Lecturer, Department of Computer Science and Engineering
IUBAT – International University of Business Agriculture and Technology Dhaka, Bangladesh
alomgir.hossain@iubat.edu

Abstract: Data quality is a critical factor for Data Warehouse and Data Integration System. Improving the quality of data is important because it is used in the process of decision support, which requires accurate data. Data cleaning is the process of identifying and removing or correcting error in the data. Data cleaning is a key precondition for analysis decision and data integration. Rule base design is the new face for data cleaning. So we decide that, we choose this new technique and design its model with algorithm and justify the existing model for the data cleaning.

Keywords: Data Quality, Data Warehouse, Data Integration, Types of Error, Source of Error.

1. INTRODUCTION
Data cleaning is the process to clean data from different errors that makes data dirty. Dirty data hampers to take major statistical and strategic decision in business and other purposes. So data needs to be clean. In data warehouses, Data are actually extracted from local operational databases and materialized as a global database. In the case of data integration systems, Data cleaning is a process that needs to be performed during query processing when data are returned from the source databases. But both contexts are lead to the point of quality data that helps in the process of decision support. Data Warehouse is collection of information from many databases. It supports business analysis and decision making. Data integration involves combining data residing in different sources and providing users with a unified view of them. Quality Data refer that error free data after using data cleaning. Clean data is Quality Data. The main objective of the data cleaning is to reduce the time and complexity of process and increase the quality and error free data. [1][2]

II. LITERATURE REVIEW
According to the (Venkatesh Ganti and Anish Das Sarma), Data cleaning is the process of starting with raw data from one or more sources and maintaining reliable quality for applications. Data cleaning, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data (Erhard Rahm, Hong Hai Do). Now most of cases are using algorithm base data cleaning. Where different algorithm for different errors. For this reason it becomes more complexity. The Rule base design is developed in a Paper named “Design and implementation of data cleaning: Rule base” by “YAN hao[3] and DIAO Xing-chun[4]”.

III. METHODOLOGY
According to the analysis, model design and algorithm and some other different criteria we think that our analysis goes to qualitative and quantitative method both. For making this paper, we collect data from secondary source and we also see some research paper according to our topic which is already published.

IV. ANALYSIS OF DATA QUALITY PROBLEMS
At the start of the informational method, each MIS system is isolated from others. Therefore info is difficult to exchange. Obviously, the mixing of the MIS systems is an inexorable trend. To support call well, effective data sorting and cleaning square measure justifiably required to induce data reliable and proper. On the opposite hand, as for brand new developed data system.

V. DATA QUALITY
Data quality is the degree to that data meets the precise needs of specific customers that contain many dimensions. Poor data quality prices businesses big amounts of money each year. The following square measure characteristics and measures of information quality [6]:
- Definition Conformance
- Completeness (of values)
- Validity (Business rule conformance)
- Accuracy (to the Source)
- Precision
- Non-duplication (of occurrences)
- Derivation Integrity
- Accessibility
- Timeliness
A. Types of Error in Data
Lexical errors: Lexical error is a sequence of characters that does not match the pattern of any token. Format errors: Format errors are the errors caused due to missing system files. Integrity Constraint Violation: when a duplicate row is inserted into a table that has a unique constraint or index then it occurs. Duplications: If you store two copies of the same data in your database, it takes twice as much space. Missing Value: Missing values, occur when no data value is stored for the variable in an observation. Missing Tuple: Database instances from tuples but if the tuples are showed null values in a table that is must be filled then it occurs.

B. Source of Error in Data
An understanding of the sources of data errors is helpful each in planning data assortment and duration techniques that mitigate the introduction of errors, and in developing applicable post-hoc data cleansing techniques to detect and ameliorate errors. several of the sources of error in databases fall under one or a lot of the following categories:

• Data Entry Errors
• Measurement Errors
• Distillation Errors
• Data Integration Errors

VI. COMPARISON AND ANALYSIS OF DATA CLEANING TOOLS
Many business tools for data cleanup square measure developed to support data detection and data cleanup and conversion. They are divided into 2 types: abnormal data detection tools and conversion tools (for abnormal data cleaning)[7]. there are 2 forms of data analyzing tools: prototype extraction and data processing. MIGRATIONARCHITECT Cevoke software package may be a quite data model analyzing tool. Data processing tools, for example, WIZRULE defines and complete data cleanup and changing per some patterns and rules, and analyzes the synchronic linguistics, model and look frequency of data kind. Some specific cleanup tools and ETL tools, such as SAS [8], produce a storage zone supported database management system and manage data in a very uniform method.

VII. DATA CLEANING MODEL DESIGN
In order to satisfy the strain for dynamic growth of data quality issues, we tend to hints an information cleansing data modeling supported rules, as shown in Figure one. In the modeling, totally different information quality downside has different cleansing rules and cleansing rules ar hold on within the rule base. By rules management module, the modeling implements dynamic definition and modification of cleaning rules and so enhances the information modeling’s ability to wear down this issue. Before information cleaning method is finished, applicable rules are dynamically chosen from rule base and these elite rules kind the cleansing task. Then the information cleansing process is finished in step with the cleansing task. This style improves the pliability of the information modeling.

Fig 1: Data Cleaning Model Design

Fig 2: Rule Base Design

VIII. RULE BASE DESIGN
Cleaning rule chiefly involves 2 forms of information. On the one hand, improvement rule involves the information data) improvement needs, such as information concerning the trigger-cleaning events. On the other hand, improvement rule involves the knowledge of corresponding improvement algorithms[9], as a result of one cleaning demand perhaps have variety of various cleaning algorithmic programs and every algorithm has its own characteristics applying to specific conditions, thus improvement requirements and improvement algorithms have to be compelled to be outlined independently. At identical time, it has to offer interface within the modeling to take care of improvement algorithms because improvement algorithms themselves square measure forever want to be changed and dilated. In the style of the rule base, we have a tendency to adopt the planning that rule base includes of 2 forms of base: strategy base and algorithmic program base. Within the strategy base, we have a tendency to store knowledge cleaning needs. At identical time, cleaning algorithms square measure hold on within the algorithmic program base. When it comes to knowledge improvement method, relationship between strategy and algorithmic program is quickly established according to the particular wants by the "strategy algorithm associated definition ", as shown in Figure 2.
IX. CLEANING STRATEGY DESIGN
Cleaning Strategy consists of the following steps [10]:
1. Develop a Data Quality Plan
2. Standardize Contact Data at the Point of Entry
3. Validate the Accuracy of Your Data
4. Identify Duplicates of Your Data
5. Append Data

X. PROPOSED ALGORITHM
1. Start
2. Fetch all data from the Database
3. Strategy 1:
   If data are duplicated in 3 fields into 2 tuples Go to step 4
   4. Clean the duplicated data
   5. Repeat step 3 and 4
   6. Show the unique or cleaned data
   7. End

XI. FLOWCHART OF SYSTEM

A. Module Selection and Design
Strategy 2:
If data are duplicated in 2 fields into 2 tuples Go to step 4
Select * From TABLE;

Fetch all data from database and the amount of data is 10
Strategy 1:
Select * From TABLE WHERE 1 GROUP BY FIELD1, FIELD2;

Clean duplicate data and fetch unique data and the amount of data is 7

Strategy 2:
Select * From TABLE WHERE 1 GROUP BY FIELD1, FIELD2, FIELD3;

Clean duplicate data and fetch unique data and the amount of data is 5

XII. IMPLEMENTATION AND FINDINGS
Here we are just create a system that will takes the input and save into the database. The system will accepted the Name, Mobile No, Address, and Email when it goes to inserted into the database. It will not see that the data that I inserted is it duplicated or not. Database will store the data with duplicated fields. But the system will check that database fields that without Name; Mobile No, Address and Email are match with other fields or not. If it is match then it will not show into the system but that duplicated data will preserve into the database. So that following this system we de-duplicate the data and get the unique data from the database.

A. Module Selection and Design
Strategy 2:
If data are duplicated in 2 fields into 2 tuples Go to step 4
Select * From TABLE;

Fetch all data from database and the amount of data is 10
Strategy 1:
Select * From TABLE WHERE 1 GROUP BY FIELD1, FIELD2;

Clean duplicate data and fetch unique data and the amount of data is 7

Strategy 2:
Select * From TABLE WHERE 1 GROUP BY FIELD1, FIELD2, FIELD3;

Clean duplicate data and fetch unique data and the amount of data is 5

Fig 3: System Flowchart

Fig 4: Database of our System

Fig 5: Fetch All Data

Fig 6: Clean Data by Strategy 1
B. Findings
According to the strategy the amount of data is decreasing and the purified data are coming to face. According the strategy the whole code remain same only the change on selection of data. Thus the complexity becomes less.

Fig 8: Bar chart of data amount

XIII. CONCLUSION
High data quality is a general requirement in any information system right now. Data cleaning is a key precondition for analysis decision and data integration. More and more researchers focus on data cleaning work and try to make high data quality. In this paper, we design a data cleaning algorithm against the rule based design. We found that almost everywhere in every condition there is using algorithm base data cleaning process and it is very complexity because of the algorithm that is change according to the condition due to the raw information. In this design, the algorithm will remain same only our rules and queries will be change according to the user strategy. This will decrease the complexity of data cleaning processes and improve data quality.

References


