Data Cleaning: A Framework for Robust Data Quality In Enterprise Data Warehouse

[1]Chinta Someswara Rao, [2]J Rajanikanth, [3]V Chandra Sekhar, [4]Dr. Bhadri Raju MSVS
[1,2,3,4]Dept. of CSE, SRKR Engineering College, China Amiram, Bhimavaram, AP, India

Abstract
Now a day's every second trillion of bytes of data is being generated by enterprises especially in internet. To achieve level best decision for business profits, access to that data in a well-situated and interactive way is always a dream of business executives and managers. Data warehouse is the only viable solution that can bring that dream into reality. The enhancement of future endeavors to make decisions depends on the availability of correct information that is based on quality of data underlying. The quality data can only be produced by cleaning data prior to loading into data warehouse since the data collected from different sources will be dirty. Once the data have been cleaned it will produce accurate results when the data mining query is applied. So correctness of data is essential for well-formed and reliable decision making. In this paper we propose a framework which implements robust data quality to ensure consistent and correct loading of data into data warehouses that is necessary to disciplined, accurate and reliable data analysis, data mining and knowledge discovery.

Keywords
Data Cleaning, Data warehouse, Data Mining, KDD

I. Introduction

Data warehouse of an enterprise consolidates the data from multiple sources of the organization/enterprise in hoping to provide a unified view of the data that can be used for enterprise wide decision making, reporting, analyzing and planning and a large number of other data analysis tasks. The processes performed on data warehouse for above mentioned activities are highly sensitive to quality of data. They depend upon accuracy and consistency of data. Data cleaning is to deal with the dirty data in data warehouse so as to keep high data quality. The principle of data cleaning is to find and rectify errors and inconsistencies for the data [1-4]. Data Cleaning is a sub task of data preprocessing and is necessary to make Quality and Productive strategies and to take correct decisions by business decision makers. After Preprocessing, data is loaded onto Data warehouse, and then Data Mining Queries will be applied on the cleaned data.

Data received at the data warehouse from external sources usually contains errors, e.g. Spelling mistakes, inconsistent conventions across data sources, and/or Missing fields, Contradicting data, Cryptic data, Noisy values, Data Integration problems, Reused primary keys, Non unique identifiers, inappropriate use of address lines, Violation of business rules etc [5-7]. However data cleaning is not generalized i.e the attributes will be different from one domain to other domain. Once the attributes differs the Rules that we apply on the database or flat files will be different.

In this framework we collected data from Notepad, Microsoft Word, Microsoft Excel, Microsoft Access and Oracle 10g. We have used Decision Tree Induction Algorithm for filling out the Missing Values in different data sources mentioned above. We provided Solutions to clean Dummy Values, Cryptic Values, and Contradicting data. We assumed a database named "Indiasoft" with Nine Attributes including int, char, String, Date data types, which contains dirty data. After applying algorithms the data from different sources is cleaned and inserted into Oracle 10g database which can be used as a Data Warehouse.

II. Related Work

Data cleansing is a relatively new research field. The process is computationally expensive on very large data sets and thus it was almost impossible to do with old technology. The new faster computers allow performing the data cleansing process in acceptable time on large amounts of data. There are many issues in the data cleansing area that researchers are attempting to tackle. They consist of dealing with missing data, determining record usability, erroneous data, etc. Different approaches address different issues. Here we discuss some related research addresses these issues of data quality [8-11].

Maheswara Rao [12] introduced a new framework to separate human user and search engine access intelligently with less time span. And also Data Cleaning, User Identification, Sessionization and Path Completion are designed correctly. The framework reduces the error rate and improves significant learning performance of the algorithm.

Aye et al [13] introduced a new data preprocessing technique to prune noisy data, irrelevant data, reduce data size and to apply pattern discovery techniques. This paper mainly focuses on data extraction and data cleaning algorithms. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data.

Suneetha et al. [14] discussed a preprocessing algorithm for data cleaning, user identification and session identification. The Web usage patterns are presented using snowflake schema method.

Marguart et al. [15] described that the quality of data is an important issue in data mining, and 80% of mining efforts spend to improve the quality of data. The data quality depends on accuracy, completeness, consistency, timelines, believability, interpretability and accessibility. Tasawar et al[16] presented a complete preprocessing technique such as data cleaning algorithm, filtering algorithm, user and session identification is performed. They proposed a new hierarchical sessionization algorithm that generates the hierarchy of sessions

E. Rahm and H. Hai Do [17] classify data quality problems that can be addressed by data cleaning routines and provides an overview of the main solution approaches. Data cleaning is especially required when integrating heterogeneous sources of data. Data from divergent sources should be addressed together with schema-related data transformations. In data warehouses, data cleaning is a major part of the so-called ETL process. The article also presents contemporary tool support for data cleaning process.

V. Raman and J. Hellerstein [18] consider the cleansing of data errors in structure and content as an important aspect for data warehouse integration. Current solutions for data cleaning involve many iterations of data "auditing" to find errors, and long-running transformations to fix them. Users need to endure long waits, and often write complex transformation scripts. Authors presented Potter's Wheel, an interactive data cleaning system that tightly integrates transformation and discrepancy detection. Users gradually build transformations to clean the data by adding or undoing transforms on a spreadsheet-like interface; the effect of a transform is shown at once on records visible on screen. These transforms are specified either through simple graphical operations, or by showing the desired effects on example data values. In the background, Potter's Wheel automatically infers structures for data values in terms of user-defined domains, and accordingly checks for constraint violations. Thus users can gradually build a transformation as discrepancies are found, and clean the data without writing complex programs or enduring long delays.

Chiara Francalanci and Barbara Pernici [19] in this paper suggested that the quality of data is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements. The assessment of data quality dimensions should consider the degree to which data satisfy user's needs. User expectations are clearly related to the selected services and at the same time a service can have different characteristics depending on the type of user that accesses it. The data quality assessment process has to consider both aspects and, consequently, select a suitable evaluation function to obtain a correct interpretation of results. This paper proposes a model that ties the assessment phase to user requirements. Multichannel information systems are considered as an example to show the applicability of the proposed model.

Alkis Simitsis [20] in this paper confirmed Extraction-Transformation-Loading (ETL) tools as pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse. In previous line of research, the author has presented a conceptual and a logical model for ETL processes. This paper describes the mapping of the conceptual to the logical model. First, it is identified that how a conceptual entity is mapped to a logical entity. Next, the execution order in the logical workflow using information adapted from the conceptual model has been determined. Finally, the article provides a methodology for the transition from the conceptual to the logical model.

In this section, the differing views of data cleansing are surveyed and reviewed. A general framework of the data cleansing process is presented and a set of general methods that can be used to address the problem is presented.

III. System Architecture

A. Existing System

In the existing system ,the TDQM framework [21-22] is done which advocates continuous data quality improvement by following cycles of define, measure, analyze and improve. Though it is easy to implement and manageable in enterprise environment for data cleansing it has many drawbacks.

1. Most of the existing systems are only limited to duplicate records elimination.
2. Less interactive from user point of view.
3. Problems in identifying dirty data.
4. Problems in selecting appropriate algorithm i.e one algorithm can't clean all types of dirty data.

B. System Architecture

Considering the data cleansing as a vital part of ETL in enterprise level data warehouse development, to achieve data quality, we are going to propose a pragmatic, interactive and easy to implementable data cleansing framework.

Rules configuration data base: Rules configuration data base is a central repository that comprises three types of rules, which are data extraction rules, transformation rules and business rules. These rules are the driving force throughout the data cleaning process for data cleaning. These rules enter into rules configuration database based on data profiling results, user experience and data warehouse model requirements to execute data cleansing process. The data extraction rules needed to extract required data from larger data set that needs to clean up. These rules mostly based on data profiling input after data profiling of source systems. The transformation rules define what parameters, functions and approaches are required to clean it. The data cleansing process use that transformation rules to clean data according to inputs as shown in fig. 1. The transformation rules can be data formatting type, removal of duplication records, default values of missing values and other related inconsistencies etc.

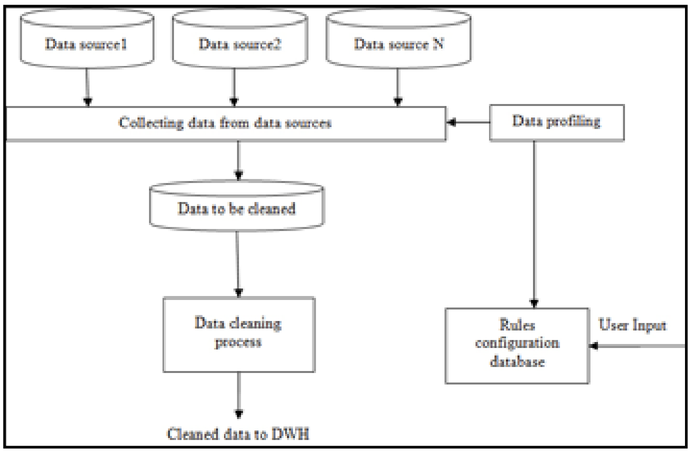


Fig. 1: System Architecture

C. Data Cleaning Stages

Data quality improvement is achieved through data cleansing which has four stages namely, investigate, standardize reduplication, and survivorship. Each of these stages is explained in the following subsections [23-24].

1. Investigation Stage

This is the auditing phase in which client data is analyzed to identify different errors and patterns in the data. This stage requires all or a sample of data to discover the data quality. Investigation results include frequency reports on various tokens, labels and record patterns. These reports provide the basis for tuning standardization rule sets for a given customer data.

2. Standardization Stage

In this stage data is transformed to a standard uniform format agreed by the customer. This involves segmenting the data, canonicalization, correcting spelling errors, enrichment and other cleansing tasks using rule sets. This stage usually requires iterative tuning of rule sets. The tuning can be done on the whole data or a sample data-set.

3. De-Duplication/Matching Stage

This is an optional phase where standardized data from previous stages is used to identify similar or duplicate records within or across datasets. This stage is configured by providing parameters for the blocking and the matching steps. Blocking is used to reduce the search scope and matching is used to find the similarity between the records by using edit distance or other known matching methods. It generally takes little iteration to decide on these thresholds in practice.

4. Survivorship Stage

In this stage the customer decides what data has to be retained after the match stage. If data is being merged from different sources then how the overall merging should take place are defined using rules. This is done by incorporating the inputs from the customer who decides the credibility of data sources using his business knowledge.

D. Data Cleansing Process

The data cleansing process takes two inputs

1. Data required to be cleaned
2. Rules of cleansing from rules configuration database

This is the area where actual data cleansing processing done based on rules from rules configuration repository and output of this process provides error-free and consistent data that is ready to load into data warehouse. This output data is standardized, uniform, accurate and complete with accordance to business. The cleaned data not only provides data quality but expedite the processing speed and performance of overall ETL process.

E. Problem Statement

Develop a framework with a list of data cleaning techniques and heterogeneous data sources in a easy way to understand and select the options. The user selects the data source and supplies the data cleaning technique on the data source. The cleaned data should be produced on single data source and multiple data sources at a time and integrated into a Relational database management system such as 10g.

F. Algorithm

In this paper we user Decision Tree Induction algorithm has been used to fill the missing values. For each attribute one decision tree will be constructed. Decision tree consists of a set of nodes. Each node is a test on an attribute. Two branches will come out from each node except for the leaf nodes. One branch is the "Yes" and the other is "No". The Missed value is filled with the Leaf Nodes.

```
Algorithm: Generate _decision _tree.
Method:
(a). Create a node N
(b). If SAMPLES are all of the same class c then
(c). Return N as a leaf node labeled with the class c
(d). If attribute_list is empty then
(e). Return N as a leaf node labeled with the most common
      class in SAMPLES;//majority visting
(f). Select test-attributes, the attribute among attribute_list
      with the highest information gain.
(g). Label node N with test attribute
(h). For each known value ai of test-attribute
(i). Grow a branch from node N for the condition_test_
      attribute=$a_i$
(j). Let $s_i$ be the set of samples in samples for which test_
      attribute=$a_i$
(k). If $s_i$ is empty then
(l). Attach a leaf labeled with the most common class in
      samples
(m). Else attach the node returned by Generate_decision_
      tree(si,attribute_list-test_attribute)
```

IV. Implementation

A. Missing Values

Missing Values are replaced by Values in the leaf nodes of Decision Tree. Initially Data in the form of rows is collected from different Data Sources. Each row is a combination of several attributes. Each attribute is separated by comma ",". The Nine Attributes we considered in this project for cleaning are,

```
Empid, Empname, DOB, Age, Gender, Highqual,
Income,Experience,CompanyName
```

The following are the examples of missing values in the rows:

```
ct1568,,,50,,b.tech,5,7,wipro,
,ramesh,,60,eee,,,30,ibm,
,,,,m,b.e,,,infosys,
,aaa,,50,,b.tech,25000,7,Dell,
,ramesh,,65,ccc,,3,,ibm,
,,,,m,b.e,,15,infosys,
,bbb,,50,,b.tech,5,7,Satyam,
,ramesh,,65,f,,4,23,ibm,
,,,,m,b.e,3,9,Convergys,
```

B. Dummy Values

Internet is a place where Dummy Values generates in a huge range. For example, Email account registrations, download registrations.. etc.. There is no Algorithm exists at present to identify the dummy Values. Each domain follows it's own method to identify and solve dummy Values. The following are the examples of Dummy values.

```
abc  xyz  pqr  efg  hfg  yur tur wws hbn
aaaa  bbbb  cccc  dddd  eeee  ffff  gggg  hhhh  iiii  jjjj
abcdefg  higksls  bvvbv  ddkd  lslsls  slss  qqq  ppppp
```

After identifying these kinds of values, we replace them with a global value "Indiasoft" .Later when a Data Mining query is applied, whenever mining process finds value "Indiasoft" it immediately knows that it is a dummy value. The mining will not be done based on this kind of values. So accurate results will be obtained by cleaning dummy values in Preprocessing.

C. Cryptic Values

A simple Example for Cryptic Values is CGPA. If User enters his CGPA as 9.0, the cleaning process should convert it to the percentage before inserting into the database.

Since most of the universities prefer Percentages than CGPA.

D. Contradicting Data

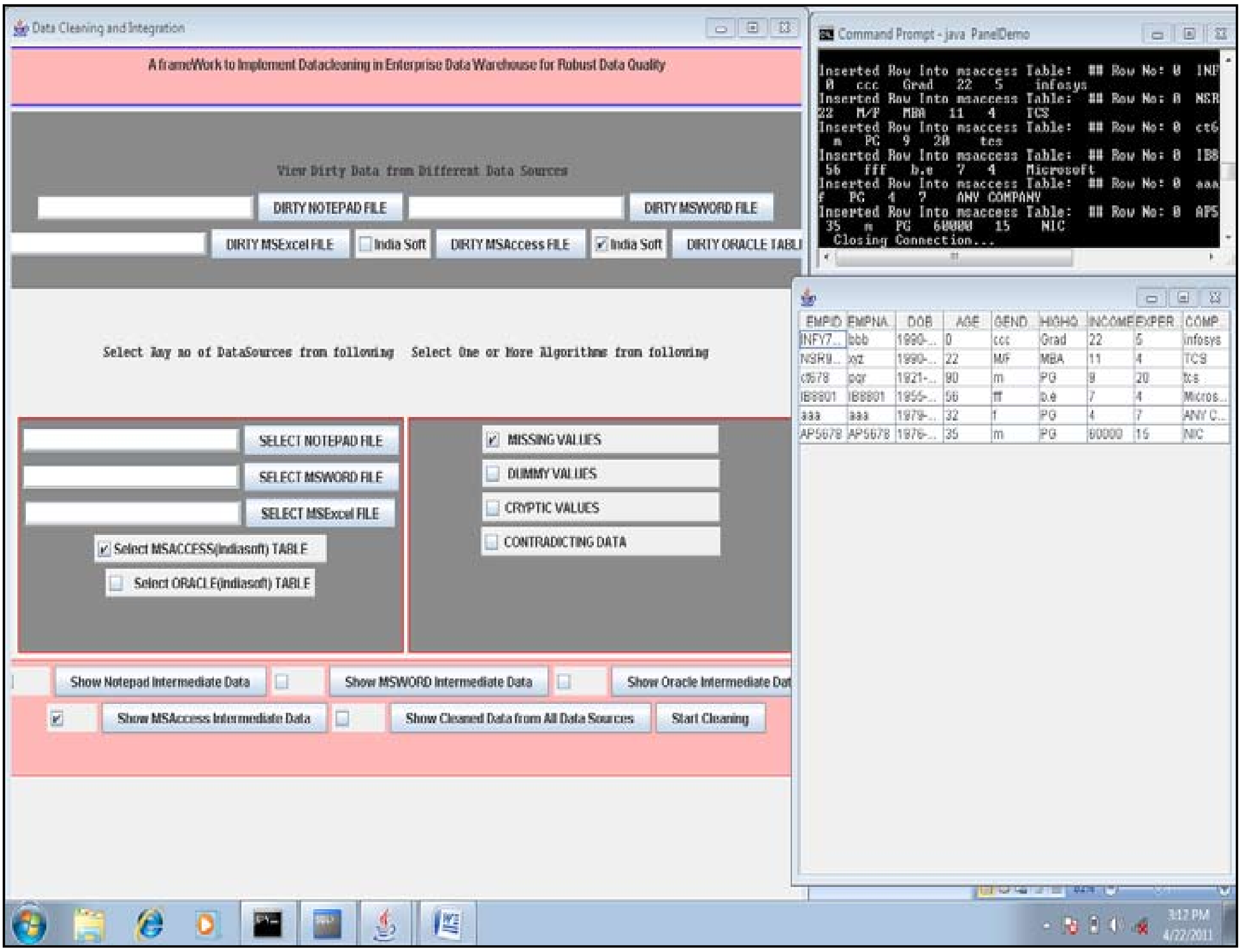
An Example for Contradicting data is Date of Birth and Age. If Age and Year of DOB mismatches the cleaning process identifies these kinds of Contradicting data. Then the Year of DOB is changed to 2011+Age.

V. Test Cases

In general a test case is a set of test data and test programs and their expected results. A test case in software engineering normally consists of a unique identifier, requirement references from a design specification, preconditions, events, a series of steps (also known as actions) to follow, input, output and it validates one or more system requirements and generates a pass or fail.

Test Case: 1(shown in fig. 2)

Input Specification	**Output Specification**:	**Result**
User Selects the Microsoft Access Data Source & Missing Values Algorithm. Clicks Start Cleaning button Clicks show Intermediate data button.	Displays filled Missing Values in a Table format.	Pass


Data Cleaning and Integration
A frameWork to Implement Datacleaning in Enterprise Data Warehouse for Robust Data Quality
View Dirty Data from Different Data Sources
DIRTY NOTEPAD FILE
DIRTY MSWORD FILE
DIRTY MSExcel FILE
□ India Soft
DIRTY MSAccess FILE
☑ India Soft
DIRTY ORACLE TABLE
Select Any no of DataSources from following
Select One or More Algorithms from following
SELECT NOTEPAD FILE
SELECT MSWORD FILE
SELECT MSExcel FILE
☑ Select MSACCESS(indiasoft) TABLE
□ Select ORACLE(indiasoft) TABLE
☑ MISSING VALUES
□ DUMMY VALUES
□ CRYPTIC VALUES
□ CONTRADICTING DATA
□ Show Notepad Intermediate Data
□ Show MSWORD Intermediate Data
□ Show Oracle Intermediate Data
☑ Show MSAccess Intermediate Data
□ Show Cleaned Data from All Data Sources
Start Cleaning
Command Prompt - java PanelDemo
3:12 PM
4/22/2011

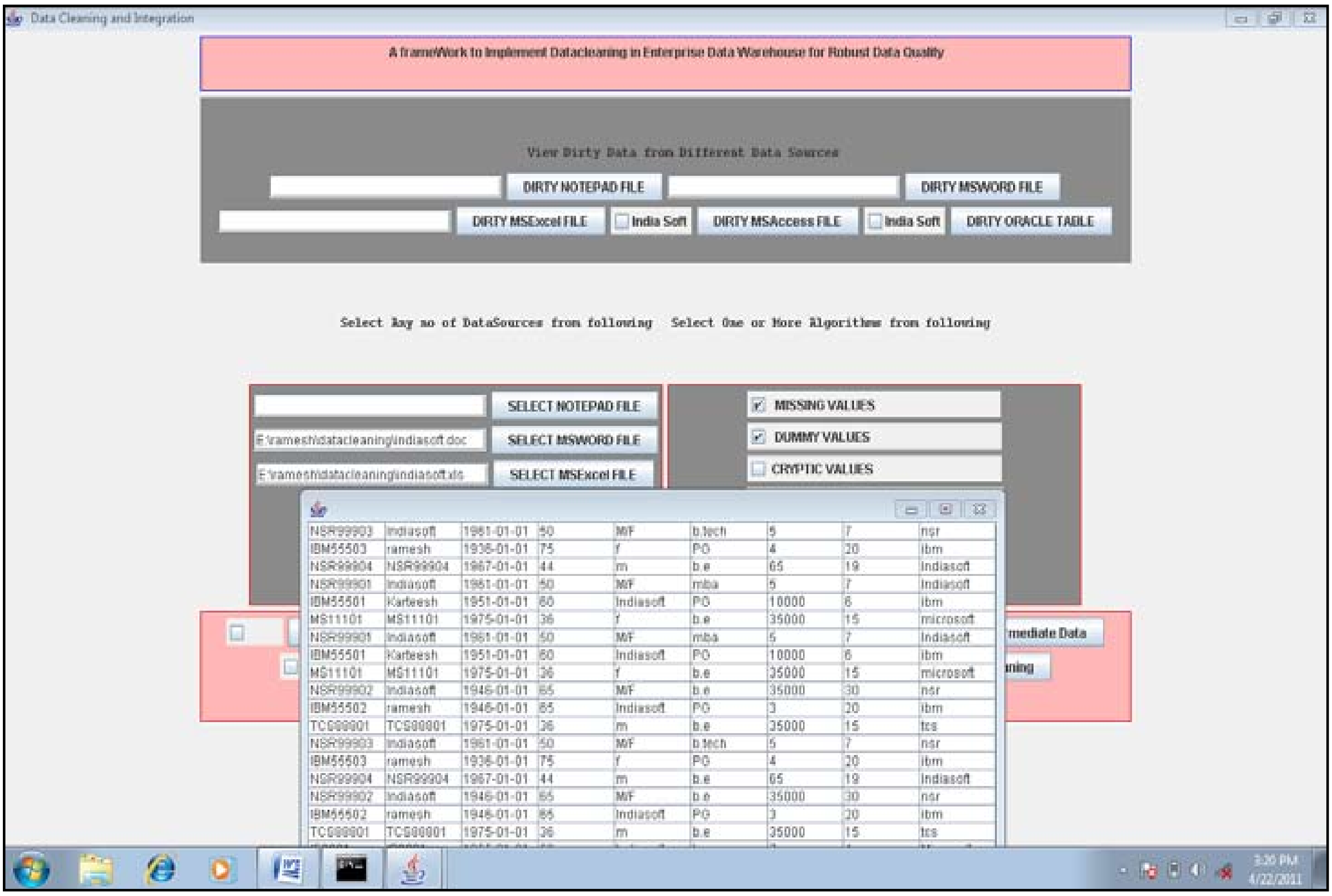
```
Inserted Row Into msaccess Table:  ## Row No: 0  INF
 0   ccc   Grad   22   5    infosys
Inserted Row Into msaccess Table:  ## Row No: 0  NSR
22   M/F   MBA   11   4    TCS
Inserted Row Into msaccess Table:  ## Row No: 0  ct6
  m   PG   9   20    tcs
Inserted Row Into msaccess Table:  ## Row No: 0  IB8
 56   fff   b.e   7   4    Microsoft
Inserted Row Into msaccess Table:  ## Row No: 0  aaa
f   PG   4   7    ANY COMPANY
Inserted Row Into msaccess Table:  ## Row No: 0  APS
 35   m   PG   60000   15    NIC
  Closing Connection...
```

EMPID	EMPNA	DOB	AGE	GEND	HIGHQ	INCOME	EXPER	COMP
INFY7...	bbb	1990-...	0	ccc	Grad	22	5	infosys
NSR9...	xyz	1990-...	22	M/F	MBA	11	4	TCS
ct678	pqr	1921-...	90	m	PG	9	20	tcs
IB8801	IB8801	1955-...	56	fff	b.e	7	4	Micros...
aaa	aaa	1979-...	32	f	PG	4	7	ANY C...
AP5678	AP5678	1976-...	35	m	PG	60000	15	NIC

Fig. 2: Test case 1

Test Case: 2 (shown in fig. 3)

Input Specification	Output Specification:	Result
User Selects the Microsoft Word, Microsoft Excel Data Sources & Missing Values, Dummy Values Algorithm. Clicks Start Cleaning button Clicks show Intermediate data button.	Displays filled Missing Values, Replaced dummy values in a Table format.	Pass


Data Cleaning and Integration
A frameWork to Implement Datacleaning in Enterprise Data Warehouse for Robust Data Quality
View Dirty Data from Different Data Sources
DIRTY NOTEPAD FILE
DIRTY MSWORD FILE
DIRTY MSExcel FILE
□ India Soft
DIRTY MSAccess FILE
□ India Soft
DIRTY ORACLE TABLE
Select Any no of DataSources from following
Select One or More Algorithms from following
SELECT NOTEPAD FILE
SELECT MSWORD FILE
SELECT MSExcel FILE
☑ MISSING VALUES
☑ DUMMY VALUES
□ CRYPTIC VALUES
mediate Data
ning

E:\ramesh\datacleaning\indiasoft.doc

E:\ramesh\datacleaning\indiasoft.xls

NSR99903	Indiasoft	1961-01-01	50	M/F	b.tech	5	7	nsr
IBM55503	ramesh	1936-01-01	75	f	PG	4	20	ibm
NSR99904	NSR99904	1967-01-01	44	m	b.e	65	19	Indiasoft
NSR99901	Indiasoft	1961-01-01	50	M/F	mba	5	7	Indiasoft
IBM55501	Karteesh	1951-01-01	60	Indiasoft	PG	10000	6	ibm
MS11101	MS11101	1975-01-01	36	f	b.e	35000	15	microsoft
NSR99901	Indiasoft	1961-01-01	50	M/F	mba	5	7	Indiasoft
IBM55501	Karteesh	1951-01-01	60	Indiasoft	PG	10000	6	ibm
MS11101	MS11101	1975-01-01	36	f	b.e	35000	15	microsoft
NSR99902	Indiasoft	1946-01-01	65	M/F	b.e	35000	30	nsr
IBM55502	ramesh	1946-01-01	65	Indiasoft	PG	3	20	ibm
TCS00001	TCS00001	1975-01-01	36	m	b.e	35000	15	tcs
NSR99903	Indiasoft	1961-01-01	50	M/F	b.tech	5	7	nsr
IBM55503	ramesh	1936-01-01	75	f	PG	4	20	ibm
NSR99904	NSR99904	1967-01-01	44	m	b.e	65	19	Indiasoft
NSR99902	Indiasoft	1946-01-01	65	M/F	b.e	35000	30	nsr
IBM55502	ramesh	1946-01-01	65	Indiasoft	PG	3	20	ibm
TCS00001	TCS00001	1975-01-01	36	m	b.e	35000	15	tcs

Fig 3: Test case 2

VI. Conclusions and Futurework

We designed sample database of Employees with attributes such as Empid, Empname, Income,…etc and named this database as "Indiasoft". We have taken some sample instances for each attribute however data cleaning is more effective by considering more instances for an attribute. Then most probable values can be filled in Missing Values, Contradicting values. We assumed some Dummy Values before comparing with the database values so that we replaced this with "indiasoft". Whenever Data mining queries finds this indiasoft it immediately knows that it is a dummy value. So mining will not go through that path. Identifying & cleaning of Noisy Values is left as future exercise. In the future a data-cleansing product or application designed and developed using this framework that can be building in-house or for commercial use. Warnings are issued and stored for any records that do not meet cleansing standards, so that they may be recycled through a modified cleansing process in the future.

References

[1] S.Jeffery, G.Alonso, M.Franklin, W.Hong, J.Widom "Declarative support for sensor data cleaning", Pervasive, May 2006.

[2] Ali Mirza Mahmood, Mrithyumjaya Rao Kuppa,"A novel pruning approach using expert knowledge for data-specific pruning", Engineering with Computers, Springer, pp. 21–30, 2011.

[3] E.Rahm, H. Do. Data cleaning: Problems and current approaches.IEEE Data Eng. Bull., 23(4):3-13, 2000.

[4] Suneetha K.R, R. Krishnamoorthi,"Data Preprocessing and Easy Access Retrieval of Data through Data Ware House", Proceedings of the World Congress on Engineering and Computer Science, Vol.1, pp.978-988, 2009.

[5] Xianjun Ni, "Design and Implementation of WEB Log Mining System", International Conference on Computer Engineering and Technology, IEEE, pp. 425-427, 2009.

[6] Mosavi,"Multiple Criteria Decision-Making Preprocessing Using Data Mining Tools", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, pp. 26-34, 2010.

[7] J.Rao, S.Doraiswamy, H.Thakkar, L.S.Colby,"A Deferred Cleansing Method For RFID Data Analytics", Int. Conf. on Very Large Data Bases (VLDB06), pp. 175–186. September 2006.

[8] Ballou, D., Tayi, K.,"Methodology for Allocating Resources for Data Quality Enhancement", CACM, pp. 320-329, 1989.

[9] Redman, T.,"Data Quality for the Information Age", Artech House, 1996.

[10] Redman, T.,"The Impact of Poor Data Quality on the Typical Enterprise", CACM, Vol. 41, pp. 79-82, 1998.

[11] Wang, R., Storey, V., Firth, C.,"A Framework for Analysis of Data Quality Research", IEEE Transactions on Knowledge and Data Engineering, Vol. 7, No. 4, pp. 623-639,1995.

[12] Maheswara Rao.V.V.R, Valli Kumari.V,"An Enhanced Pre-Processing Research Framework for Web Log Data Using a Learning Algorithm", Computer Science and Information Technology, pp. 01–15, 2011.

[13] Aye.T.T,"Web Log Cleaning for Mining of Web Usage Patterns", International Conference on computer Research and Development, IEEE, pp. 490-494, 2011.

[14] Suneetha K.R, R. Krishnamoorthi,"Data Preprocessing and Easy Access Retrieval of Data through Data Ware House", Proceedings of the World Congress on Engineering and Computer Science, Vol. 1, pp. 978-988, 2009.

[15] Marquardt.C, K. Becker, D. Ruiz,"A Preprocessing Tool for Web usage mining in the Distance Education Domain", In Proceedings of the International Database Engineering and Application Symposium (IDEAS' 04), pp.78-87, 2004.

[16] Tasawar Hussain, Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence", IEEE Conference on Emerging Technologies, pp.21-26, 2010.

[17] E. Rahm, H. Hai Do,"Data Cleaning: Problems and Current Approaches", University of Leipzig, Germany, [Online] Available: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf

[18] Raman, V.; Hellerstein, J.M.: Potter's Wheel: An Interactive Framework for Data Cleaning. Working Paper, 1999. http://www.cs.berkeley.edu/~rshankar/papers/pwheel.pdf.

[19] Cinzia Cappiello, Chiara Francalanci, Barbara Pernici. "A Self-monitoring System to Satisfy Data Quality Requirements", OTM Conferences, pp. 1535-1552, 2005.

[20] Alkis Simitsis, Kevin Wilkinson, Umeshwar Dayal, Malú Castellanos: Optimizing ETL workflows for fault-tolerance. ICDE, pp. 385-396, 2010.

[21] Yang W. Lee, Diane M. Strong,"Knowing-Why About Data Processes and Data Quality", Journal of Management Information & Systems, Vol. 20, pp. 13-39,2004.

[22] Ganesan Shankaranarayanan, Mostapha Ziad, Richard Wang,"Managing Data Quality in Dynamic Decision Environments: An Information Product Approach", Journal of Data Management, 2003.

[23] Yang Lee, Diane Strong, Beverly Kahn, Richard Wang", AIMQ: A Methodology for Information Quality Assessment", Information & Management, Vol. 40, Issue 2, pp. 133-146, 2002.

[24] Beverly Kahn, Diane Strong, Richard Wang,"Information Quality Benchmarks: Product and Service Performance", Communications of the ACM, pp. 184-192, 2002.