International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 10005-10014
© Research India Publications
http://www.ripublication.com

An ETL Based Framework For Data Cleaning In Multi Data Source

[#1]Agusthiyar.R and [#2] Dr. K. Narashiman
[#1] Asst.Prof (Sr.G), Department of Computer Applications, Easwari Engineering College, Chennai, ramagusthiyar@gmail.com
[#2]Professor& Director, AUTVS Center for Quality Management, Anna University, Chennai, knman@annauniv.edu

Abstract

In this age of information, there is a huge availability of data which creates a need for data cleaning. In this context, data cleaning solutions are very important for Tera data users. Normally, data cleaning deals with detecting, removing errors and inconsistencies in large data sets. For any real world data set, doing this task by hand is completely out of the question as it involves huge amount of human resource and time. Several organizations spend millions of dollars per year to detect data errors. This paper proposed a framework for data cleaning, which in turn provides an approach to managing data cleaning solutions using ETL based framework. By using this ETL concept the noisy data and inconsistencies can be removed with minimum effort and in a simplified manner.

Keywords: Noisy data**,** Data cleaning, Data quality, Data warehouse, ETL (Extract, Transform and Load)

Introduction

In a decade, many researchers have faced problems regarding the quality of data since data warehousing techniques have become important in decision support information systems. [3] The quality of data in the large real world data set depends on a number of issues, but the source of the data is the crucial factor. Data entry acquisition is inherently prone to errors, both simple and complex. For data cleaning, in a single data source can be dealt with attribute selection or feature selection method to detect and remove errors significantly. This method gives the quality data for the end users or business community for homogeneous data sources. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems data cleaning process becomes crucial. Data warehouses require and provide extensive support for data cleaning. They load and

continuously refresh huge amounts of data from a wide variety of sources using ETL process. Furthermore, data warehouses are used for decision making, so that the correctness of the data is vital to avoid wrong decisions. For instance, duplicated or missing information will produce incorrect or misleading statistics ("garbage in, garbage out"). Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing [1]. In this paper, an ETL based framework provides solution for data cleaning. In the next section, literature study has been made. In section 3, data quality problems are classified, section 4, a methodology and framework has been proposed and an ETL based framework has explained, section 5, sample data has tested by some ETL tools and in section 6, conclusion has been made.

Related Works

This paper investigates current data cleaning methods, approaches, data quality oriented data mining and data warehousing frameworks and designs in the previous years. From this literature study, there is lacking in the standard data cleaning procedure for the researchers, and there is a survey about the data (field) error rates in the data acquisition phase are typically around 5% or even more using sophisticated measures for error prevention available. Recent studies have shown that 40% of the collected data in dirty in one way or another. During data cleaning, multiple records representing the same real life object are identified, assigned only one unique database identification, and only one copy of exact duplicate records is retained. A token-based algorithm for cleaning a data warehouse is the notion of "token records" was introduced for record comparison and the smart tokens are more likely applicable to domain-independent data cleaning, and could be used as warehouse identifiers to enhance the process of incremental cleaning and refreshing of integrated data [4]. Data cleaning is a process of identifying or determining expected problem when integrating data from different sources or from a single source. Many problems occur in the data warehouse while loading or integrating data. The main problem in data warehouse is noisy data. An attribute selection algorithm and token formation algorithm is used for data cleaning to reduce a complexity of data cleaning process and to clean data flexibly and effortlessly without any confusion [9]. Every attribute value forms either a special token like birth date or an ordinary token, which can be alphabetic, numeric, or alphanumeric. These tokens are sorted and used for record match. The tokens also form very good warehouse identifiers for future faster incremental warehouse cleaning. The idea of smart tokens is to define from two most important fields by applying simple rules for defining numeric, alphabetic, and alphanumeric tokens. Database records now consist of smart token records, composed from field tokens of the records. These smart token records are sorted using two separate important field tokens. The result of this process is two sorted token tables, which are used to compare neighboring records for a match. Duplicates are easily detected from these tables, and warehouse identifiers are generated for each set of duplicates using the concatenation of its first record's tokens. These warehouse identifiers are later used for quick incremental record identification and refreshing [6].

Data Quality Mining

Data Quality Mining can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of Data Quality Mining is to detect, quantify, explain, and correct data quality deficiencies in very large databases [7]. The quality of the data in data mining is measured by some set of parameters.

Data quality parameters

- ✓ Accuracy - An aggregated value over the criteria of integrity, consistency and density
- ✓ Completeness – All values for a certain Variable are recorded;
- ✓ Consistency – It concerns contradictions and syntactical anomalies of data.
- ✓ Timeliness – It concerns the recorded value is not out of date.
- ✓ Believability – It ensures that the originality of the data
- ✓ Interpretability – It covers syntax and semantics
- ✓ Accessibility - It concerns availability, security and performance

The quality of data is often evaluated to determine usability and to establish the processes necessary for improving data quality. Data quality can be measured objectively and subjectively. Data quality is a state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. The hierarchy of data quality is shown in the Fig. 1:

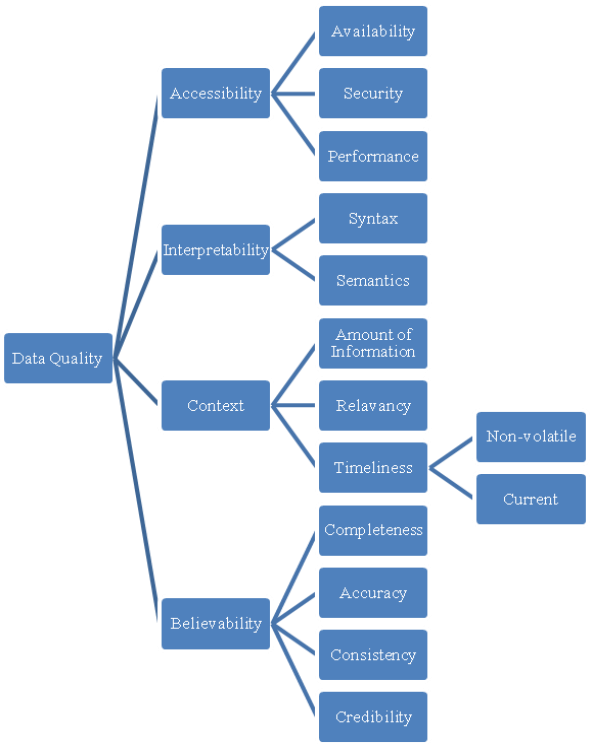
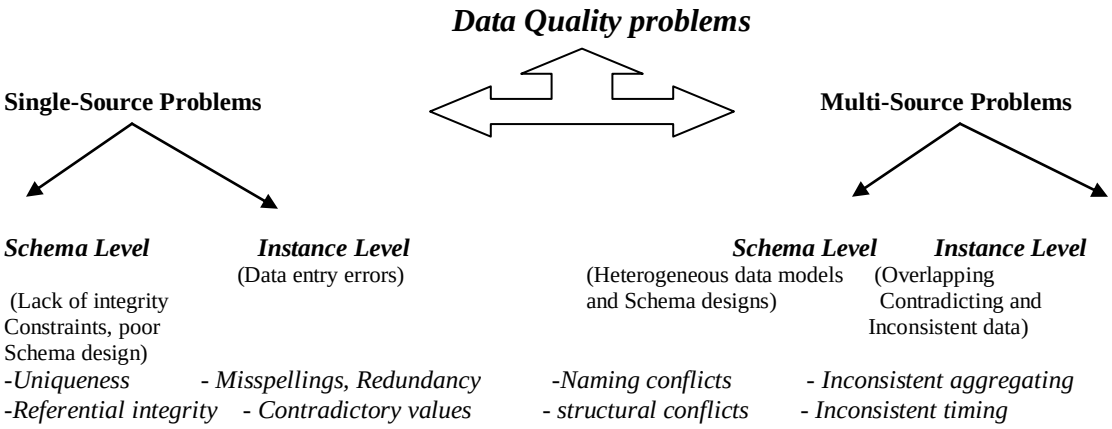


Figure 1: The Hierarchy of Data Quality

Data quality has two distinct aspects: one is the "correctness" of data (such as accuracy and Consistency), and the other one involves the appropriateness of data for some intended purposes.

Data producers and users generally assume that the purpose of data quality assurance is to provide the best data possibly[5].



Data Quality Problem

The data quality problem is roughly distinguished between single-source and multi-source problems and between schema- and instance-related problems. Schema-level problems are also reflected in the instances; they can be addressed at the schema level by an improved schema design (schema evolution), schema translation and schema integration. Instance-level problems, on the other hand, refer to errors and inconsistencies in the actual data contents which are not visible at the schema level. They are the primary focus of data cleaning. Fig. 2 also indicates some typical problems for various cases. While not shown in Fig. 2, the single-source problems occur (with increased likelihood) in the multi-source case, too, besides specific multi-source problems [1].

Frame Work and Methodology

Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing. During the ETL process it is required [1]. This paper proposes a framework based on ETL process. In the first phase data would be selected from the multiple data source for data cleaning (Extraction), then in the second phase, detect and remove the noisy and inconsistencies in the data source were done using appropriate methods (Transformation). In the third phase cleaned data will be loaded into the final data warehouse (Loading).

Methodology

A. Select the two tables of noisy data from different source
B. Combining them for noise removal and data cleaning
C. Choose the attribute value, which is unique in the table
D. Compare the attributes values, which is same in the table
E. Find noisy and duplicates attribute values based on the unique attribute
F. Filtering the attributes, which is found noisy and duplicates
G. Apply the appropriate data mining algorithm or technique for segregating the data
H. Get the quality data and Load into the data warehouse

Select the two tables of noisy data from different source

Data cleaning in single data source is explained by number of researchers in last decade. All these findings and implementations are in and around single data source problems and methods. So we mainly concentrate on multi data source for implementation and testing of this framework. For that, we have to select the two tables of noisy data from different source. After that the attribute selection is very important when comparing two attributes of two tables. Select the attributes of noisy data such as duplicates, inconsistencies etc.

Combining them for noise removal and data cleaning

When we select two tables of attributes, if the attribute values are same in the two tables then the tool is required to merge the two tables and find the duplicates and inconsistencies of both. For that combining process is required.

Choose the attribute value, which is unique in the table

For comparing two tables, there is a unique attribute in the table is required. Both the tables, two or more attribute values are same and others are differ, the unique attribute of the table is used to find the originality of data and assign as there are duplicate values in the attributes of the table. For example, Date of Birth (DOB) is the unique attribute in the following example of Table1 & Table2.

Compare the attributes values, which is same in the table

There are two tables are selected for noise detection and noise removal. Before merging both tables of the attributes the values of the attributes are compared. If there is a duplicate exists, then the attributes of the tables were merged by the tool and then applying for further process.

Find noisy and duplicate attribute values based on the unique attribute

The unique attribute is the one of the important factor to find the duplicate attribute values for both the tables. If there are ‘n’ number of attributes in the tables, out of that ‘n-3’ attributes values have duplicate entry and data inconsistencies. The unique attribute of both the tables are used to check and find the duplicates and noisy data existence.

Filtering the attributes, which is found noisy and duplicates
Filtering is the process done by the ETL tool, for which the attributes have duplicate and noisy values. In this step, the data inconsistencies have been replaced by replace missing values operator and the tool giving to the correct values for the missing value attribute and duplicate values are removed by the remove duplicate operator of the ETL tool.

Apply the appropriate data mining algorithm or technique for segregating the data
After removing data inconsistencies and duplicates values, the appropriate data mining algorithm has to apply to the data set for segregating the data for decision making. Based on the dataset size and complexity, the clustering, association rule and classification methods are to be applied. The specific data mining algorithm has to apply in the dataset for refine the data.

Get the quality data and Load into the data warehouse
All the above steps have been repeated until the dataset get refinement. If the data is refined, it will load into the data warehouse for further process. Now the quality data could have the capability to apply any business process and decision making for company development.

Algorithm

Input: Table of noisy attributes (T1, T2…Tn) & (A1, A2…An)

Output: Cleaned data (Quality data)

Begin

For i =1 to n times

Select table from different source

Merge them for processing

Choose the same attribute (A1 …An from T1 and A1…An from T2)

Choose the unique attribute ($A_{uni\ q}$ from T1 and T2)

Find noisy and duplicate attributes

Remove noisy attribute values

Apply data mining algorithm for noise & duplicate removal

Get the cleaned data (quality data)

End

Where T1, T2…Tn are tables and A1, A2…An are attributes of the tables & Auniq is the unique attribute.

Framework

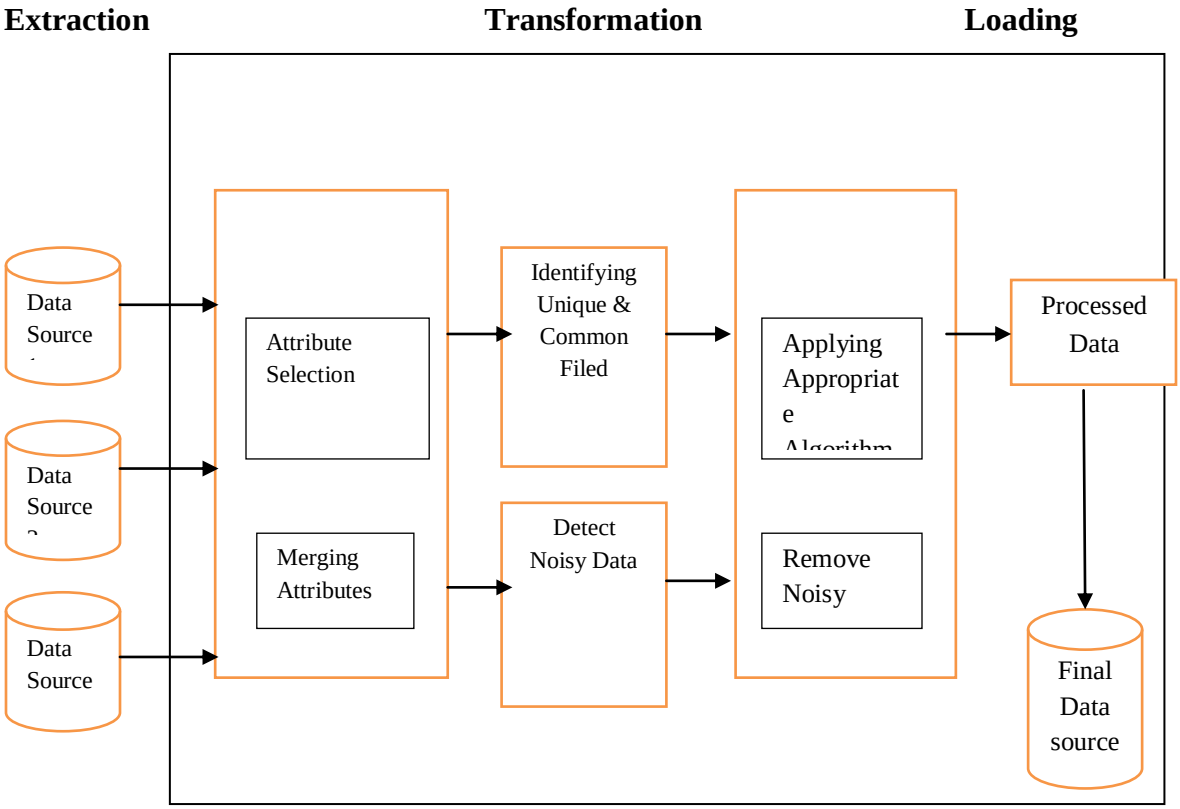


Figure 3: An ETL based Framework for Data Cleaning in Multi Data Source

An Example

To experiment this framework and idea by some artificially designed databases, which have some data quality problems as mentioned above. This example gives that the working experience of this framework mechanism towards data quality issues and measures.

In India, during elections voters play an important role to elect the right candidate. In this regard, the Election Commission of India maintains the voter's information database district-wise and state-wise. Information about the voters is collected with the aid of Government staff members. For instance, students residing in rural areas get their voter id from their respective assembly constituency. After their graduation or post graduation studies students migrate to metropolitan cities owing to varied job opportunities. If the student decides to get a new Voter's ID from the constituency where he/she resides, in this scenario, there is a duplication of Voter's ID exist. In this proposed work, this issue will be resolved using ETL based data cleaning procedure.

This framework takes two tables as examples. The following Table1 shows the information of Voter's ID samples from different districts of Tamil Nadu and the Table 2 shows the information of Chennai metropolitan City Voter's ID samples including Tiruvallur and Kanchipuram districts. The Unique Attribute (Auniq) is required to find the duplicates from two tables. Here DOB is the Unique Attribute to find duplicates in the tables. In Table1 & Table2 the 1st and 5th row contains the information of a Voter from multi sources of data. The 1st and 5th row contains the duplicate values of the following attributes Elector name, Relation name, DOB and Sex. By the Unique Attribute DOB, we confirm that from the two rows of data, there is a duplicate data entry in the Voter's Information exists. Then we have to apply this data set to this proposed framework, it will detect the duplicates, remove the noisy attribute values and gives the cleaned (quality) data set to the data repository.

Conclusion and Future Work

This paper has investigated current data cleaning problems and approaches, based on these investigations, a simplified ETL based framework is proposed for data cleaning towards quality oriented data warehousing, as it provides an approach to managing data cleaning solutions by ETL based framework, and it will detect and remove the noisy data and inconsistencies in a simplified manner. This framework adopts the existing data cleaning methods implementations and also improves the efficiency of data cleaning methods in data warehouse applications. In this paper, the first phase is used to select the data and it would be selected from the multiple data source for data cleaning (Extraction), and in the second phase, detecting and removing the noisy and inconsistencies in the data source were done by the appropriate methods (Transformation). In the third phase cleaned data will be load in to the final data warehouse (Loading) for further process.

This research work developed for data cleaning in different data sources and different dimensions mentioned in the Figure. To demonstrate this proposed framework for schema level issues and instance level issues regularly. Finally this research paves way for researchers to create a common data cleaning tool for all the issues of single and multi data source based on the proposed ETL based framework.

Table 1: Voter's information of different districts of Tamil Nadu

S. No	ID Card No	Assembly Constituency No & Name	Part No	Serial No	Elector Name	Relation Name	DOB	Sex	Voter's Address
1	FXJ4094161	211 - Ramanathapuram	107	945	Sathiq basha	Ibrahim	29/05/1980	Male	66-1 /109 Ramanathapuram Kansahib street, Ward 15
2	BXH0177945	352-Sivagangai	250	1859	Vijaya sree	Laksmi	12/12/1970	Female	11 / Maruthupandiar Nagar, Sivagangai
3	ZXA1897826	118- Madurai East	110	599	Vigneshwaran	Ramanathan	9/6/1974	Male	106A/ Annai theresa Nagar, Samayanallur

4	CPQ74 15296	598-Tirchirappali	77	354	Nataraj an	Chokk alinga m	20/03 /1978	Male	2-10/ Nakkeerar Street, Puduvayal
5	KHH33 19613	200-Paramakudi	16	754	Balu	Rajan	10/11 /1975	Male	#266/Thiruvalluv ar Nagar, Main road, Paramakudi

Table 2: Voter's information of Chennai Metropolitan City

S. No	ID Card No	Assembly Constituency No & Name	Part No	Serial No	Elector Name	Relation Name	DOB	Sex	Voter's Address
1	XBG0798481	7-Maduravoyal	152	1173	Sathiq basha	Ibrahim	29/05/1980	Male	3-14 / Krishna Street, Senthamil Nagar, Ramapuram, Chennai - 600089
2	AXH0187945	11-Villivakkam	141	1281	David rajan	Susai-yabbar	15/12/1977	Male	11 / Ramasamy street, Villivakkam, Chennai
3	YFX1689276	108- Avadi	112	1009	Rahini	Iyampe-rumal	3/9/1981	Female	106A/ Bharathidasan colony, Avadi
4	WSE2879642	10-Nesapakkam	50	1751	Pancha-varnam	Ramu	25/03/1978	Female	66/ Vivekanandar Street, Nesapakkam
5	XBH1895674	113-Sriperumbatur	178	500	Balu	Rajan	10/11/1975	Male	#188/Ram Nagar, Sriperumbatur

References

[1] Erhard Rahm Hong Hai Do, Data Cleaning: Problems and Current Approaches, pp: 1-10

[2] Hernandez, M.A.; Stolfo, S.J.: *Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem*. Data Mining and Knowledge Discovery 2(1):9-37, 1998.

[3] Taoxin Peng, A Framework for Data Cleaning in Data Warehouses, Napier university, Edinburgh, UK, 1-6.

[4] Lukasz Ciszak, Application of Clustering and Association Methods in Data Cleaning, Proceedings of the International Multiconference on Computer Science and Information Technology, pp : 97-103, 2008 IEEE.

[5] R. Kavitha kumar, *D*r. Rm. Chadrasekaran, Attribute Crrection-Data Cleaning using Association Rule and Custering Methods, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.2, pp: 22-32, March 2011.

[6] Paul Jermyn, Maurice Dixon, Brain J Read, Preparing Clean Views of Data for Data Mining, London Guildhall University, UK, pp: 1- 15.

[7] Jochen Hipp and Et .all , Data Quality Mining, DaimlerChrysler AG, Research & Technology, Ulm, Germany, Wilhelm-Schickard-Institute, University of T¨ubingen, Germany , 2000, pp: 1 – 6.

[8] Helena Galhardas - Daniela Florescu - Dennis Shasha - Eric Simon, An Extensible Framework for Data Cleaning.

[9] Timothy E. Ohanekwu, C.I. Ezeife, A Token-Based Data Cleaning Technique for Data Warehouse Systems, PP:1-6

[10] Christie I. Ezeife, Timothy E. Ohanekwu, Use of Smart Tokens in Cleaning Integrated Warehouse Data, International Journal of Data Warehousing & Mining (IJDW), Vol.1 No.2,PP: 1-22, Ideas Group Publishers, April-June 2005.

[11] J. Jebamalar Tamilselvi , Dr. V. Saravanan, J. A Unified Framework and Sequential Data Cleaning Approach for a Datawarehouse, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5,PP: 117-121 May 2008.

[12] Jiawei Han, Micheline Kamber, *D*ata Mining: Concepts and Techniques, Publisher: Elsevier Science & Technology Books, March 2006, ISBN-13: 9781558609013