A Comparison Study of Data Scrubbing Algorithms and Frameworks in Data Warehousing

Hamed Ibrahim Housien*
School of Information science & Engineering
Central South Univesity
Changsha, 410083, China
* Research & Development
Ministry of Higher Education& Scientific Research
Baghdad, Iraq

Zhang Zuping
School of Information science & Engineering
Central South Univesity
Changsha, 410083, China

Zainab Qays Abdulhadi**
School of Information science & Engineering
Central South Univesity
Changsha, 410083, China
** Ministry of Higher Education& Scientific Research
Baghdad, Iraq

ABSTRACT

In these days, many organizations tend to use a Data Warehouse to meet the requirements to develop decision-making processes and achieve their goals better and satisfy their customers. It enables Executives to access the information they need in a timely manner for making the right decision for any work. Decision Support System (DSS) is one of the means that applied in data mining . Its robust and better decision depends on an important and conclusive factor called Data Quality (DQ), to obtain a high data quality using Data Scrubbing (DS) which is one of data Extraction Transformation and Loading (ETL) tools. Data Scrubbing is very important and necessary in the Data Warehouse (DW). There are growing relationships to get high DQ and effective DS. The use of DS algorithms is a solution to the constraints that limit the DQ which leads to weak decisions and the burden of the high financial costs. These constraints are: dirty data, noise data, missing values, inconsistency, uncertain data, ambiguous, conflicting, duplicated records and similar columns. The Sources and causes of these constraints are many, including: input error, merge data from different sources, difference in representing the same information, etc. In addition there are more than 35 sources and causes of the poor-quality data that arise at the stage of the ETL process. This paper present comparison and analysis for DS algorithms and the pros and cons of each algorithm, accuracy and time complexity. Additionally, it present a comparative and analysis of the Data Scrubbing Frameworks and determine the best framework.

Keywords

Data scrubbing, Data warehousing, Data Quality, Extract-Transform-Load (ETL).

1. INTRODUCTION

Data Mining is a knowledge discovery process used to extract hidden pattern from data to support decision making, one of the major steps of it is the data warehousing. Figure 1 illustrates the performance improvement of data mining enhanced Integration Decision Support System (IDSS).

Decision Support System (DSS) is a conceptual framework for a process of supporting managerial decision making, usually by modeling problems and employing quantitative models for solution and analysis. It refers to the information system supporting the company's decision-making activities due to the rapidly changing business situations [1].

Data warehousings (DWs) are increasingly being used by many organizations in many sectors to improve their operations and to better achieve their objectives. DW enables executives to access the information they need to make informed business decisions [2]. Therefore DW has to deliver highly aggregated, high quality data from heterogeneous sources to decision makers [3], Data warehouse typically provides a simple and concise view around particular subjects by excluding data that are not useful in the decision support process.

Data Quality(DQ) in a DW are of great importance, and critical factor. Many firms have problems to ensure DQ. Ensuring high-level DQ is one of the most expensive and time-consuming tasks to perform in data warehousing projects [4]. Many DW projects have failed halfway through due to low DQ. Therefore, data has to be transformed and cleaned before it is loaded into the warehouse so that downstream data analysis is reliable and accurate. This is usually accomplished through an Extract-Transform-Load (ETL) process[5]. Use data quality and extraction, transformation, and load (ETL) tools to automate the continuous detection, cleansing, and monitoring of key files and data flows.Implement data quality checks or audits at reception points or within ETL processes [6]. Stringent checks should be done at source systems and a data integration hub. There are four stages to classify Data Quality in the data warehouse, This paper focus on Data Quality in ETL stage, one of the major steps of ETL stage is Data Scrubbing.

Data scrubbing(DS) is the first important pre-process step and most critical in a Business Intelligence (BI) or Data warehousing project [5]. To have High quality data, all inconsistencies should be identified and dissolved from data before creating a data warehouse. Data scrubbing is the major starting step to have a high quality data warehouse to make valid decisions in decision support systems in a reasonable time. Data scrubbing some time called Data Cleaning or Data Cleansing.Some difference between Data Cleaning and Data Cleansing describe in [7], Figure 2 illustrated Data Scrubbing and ETL tools.

The remainder of the papers is as follows. Section 2 describes the literature review in the field of Data Scrubbing. Section 3 showing the background of Data Quality in Data warehouse. Section 4 explains comparisons and analyses of Data Scrubbing Algorithms. Section 5 & 6 describes comparisons of Data Scrubbing Algorithms and analysis testing respectively. Section 7 gives conclusions and recommendations for future work.

2. Literature Review

Several researchers in data processing developed DS techniques to optimize DQ. Dictionary based on Data Cleaning is very commonly used.

Hasimah H. Mohamed et al. has a table of advantages and disadvantages of three data mining algorithms such as Classification and Regression Trees(CART), Genetic Algorithm and k-Nearest Neighbor, Among those of data cleaning algorithms, They select k-NN (k-Nearest Neighbor) algorithm for the data cleaning part of the E-Clean system. One of the reasons they select k-NN algorithm is it is very simple to understand and easy to implement. Another reason is K-NN algorithm is very effective for large scale of data, which has fulfilled the most important project requirement to clean the clinical databases that have accumulated large quantities of patient's information [8].

Xuhui Chen and Xinghua Zhang have proposed a method which based on the genetic neural network to handle missing values. This method uses two algorithms respectively. First , Genetic Algorithm due the feature is easy to find the optimal solution region, and find the optimal initial values of network parameters. Second, use the searching capability of BP (backpropagation) algorithm to search for the best model parameters optimal solution space. The specific steps of the proposed method or algorithm it's four steps: Initialize, the design of fitness of function, The genetic treatment, and Generate new population. This method uses the global search ability of genetic algorithm and the nonlinear mapping ability of neural networks, and the algorithm greatly improves the prediction accuracy of the data. So it is an effective way to improve DQ in the forecast data such as power data of high precision numerical, and the algorithm can be fully achieved in the stored procedures of SQL (Structured Query Language) [9].

Vaishali Rajeev Patel and Rupa G. Mehta have proposed modified K-means algorithm (MK-means) which provides a solution for automatic initialization of centroids and analyzes the performance of MK-means algorithm with integration of cleaning method and normalization techniques which shows the improvement in the performance of MK-means algorithm.This algorithm works on basis of minimizing squared error function. K-means uses convergence criteria to minimize the sum of squared error (SSE). Reduced sum of squared error generates better results for K-means algorithm. This method proposed a solution to initialize the centroids automatically. One of the limitations of Naive K-means algorithm is focused and MK-means Algorithm is proposed which provides a solution for automatic initialization of centroids and performs preprocessing task like managing the missing values using various techniques and normalization of data. The outperformed results the improvement in MK-means with integration of cleaning and normalization techniques. Selection of the most appropriate attribute to play major role in the centroid initialization process is a major challenge for the algorithm. Major shortcomings of MK-means are to pass number of clusters initially and Outliers are present in the process. In future, shortcomings of MK-means will be taken care of by further research [10].

Mauricio A. Hernandez and Salvatore J. Stolfo has detail the SNM (Sorted Neighborhood Method) that is used by some to solve merge/purge and present experimental results that demonstrates this approach may work well in practice but at great expense.The sorted neighborhood method for solving the merge/purge problem can be summarized in three phases, First: Create Keys, Second: Sort Data, and the last: Merge. They describe the sorted-neighborhood method as a generalization of band jons and provide an alternative algorithm for the sorted-neighborhood method based on the duplicate elimination algorithm, This duplicate elimation algorithms taked advantage of the fact that "matching" records will come together during different phases of the Sort phase. Due to space limitations.They show a means of improving the accuracy of the results based upon a multi_pass approach that succeeds by computing the Transitive Closure over the results of independent runs considering alternative primary key attributes in each pass [11].

ZHONG Jia Qing et al. has improve SNM (Sorted Neighborhood Methods) , the improved algorithm of the basic idea is to use a relatively small sliding window, select a keyword of database to execute the SNM algorithm, store the similar serial number of the record after the sort, and then successively select other keywords in the database independently to perform SNM algorithm, and add new similar record number of to similar record storage after each execution, then they can get the serial number of all possible duplicate records, and then possible duplicate records intuitive method to clean up [12].

Luyi Mo et al. have develop efficient algorithms to compute the quality of this query under the possible world semantics. They address the challenging problem of computing the PWS-quality score of a probabilistic top-k query. They have developed efficient algorithms to evaluate the quality of U-kRanks, PT-k, and Global-topk queries. They also investigate the problem of cleaning a probabilistic database to achieve an optimal gain of quality under a limited budget. After this , They applied it on Efficient Data Cleaning Algorithms, these algorithms, (Dynamic Programming (DP), RandU, RandP and Greedy) and find better Effectiveness and Efficiency algorithm of data cleaning among them. That Greedy always have a close-to-optimal performance, and it is the best among all heuristics [13].

Mortadha M. Hamad and Alaa Abdulkhar Jihad present an enhanced technique to clean data in the DW by using a new algorithm that detects and corrects most of the error types and expected problems, such as lexical errors, domain format errors, irregularities, integrity constraint violation, and duplicates, to improve the quality of the data. The system gives a detailed report on the types of errors detected and reported according to the sources of data. This is important to improve the quality of data in the future and to increase the processing speed of the data that come from those sources. Their main focus has been on achieving a good quality of the data. In spite of that the pace of implementation of this algorithm is adequate.They proposed enhanced algorithm has been designed and implemented such that it well scales to large amounts of data processing without a significant degradation of the most of relative performance issues. Dealing with a large number of data items affect the time spent only, as when the amount of data increases the time required to address them will increase also [14].

Arindam Paul et al. attempt has been made to provide a hybrid approach HADCLEAN for cleaning data which combines modified versions of Personal Name Recognizing Strategy (PNRS) and Transitive closure algorithms.The proposed approach is explained using a sample data. Each algorithm is applied one after the other to obtain the cleaned data. The modified version of PNRS which work on fields (Attributes).The modified version of PNRS is to use an organization specific dictionary, along with a standard

dictionary, for checking the spelling mistakes. This is important because most of the verbal data present in DWs are official data and contain organizational jargons, sometimes even limited to a particular organization. but not to the fields like "Name" and "Address" as they are not found in any dictionaries with clarity. So here comes the role of modified version of transitive closure Algorithm. The modified version of TC which work on records (Tuples). The modified version of TC in a way to fully automate it without any manual intervention. This is primarily based on using more than one key to match the records into one group or rather saying that these records are the same. The modified version of TC divide the key at two levels, each level is applied one by one. First level, they order the keys based on decreasing priority of uniqueness / importance. (Topmost priority, middle priority, low priority). Now after categorizing the attributes, they apply following rules on the records to find out the related records [5].

Kazi Shah Nawaz Ripon et al. have proposed a novel domain-independent technique for better reconciling the similar-duplicate records. They also introduce new ideas for making similar-duplicate detection algorithms faster and more efficient. Additionally , they also propose a significant modification of the transitivity rule. Finally, they propose an algorithm that incorporates all these techniques for similar-duplicate detection into a domain-independent environment [15].

3. Background

DQ criteria hierarchy as it has been described in [16], it is the degree to which data meet the specific needs of specific customers, which contains several dimensions. Low DQ costs businesses vast amounts of money every year. Defective data lead to breakdowns in the supply chain, poor business decisions, and inferior customer relationships management [14].

DQ dimension of DW is the first step to DQ improvement. Below are DQ dimensions in DW development process illustrated in [3].

DQ and ETL tools are used to automate the continuous detection, cleansing, and monitoring of key files and data flows. They are also used to implement DQ checks or audits at reception points or within ETL processes. Stringent checks should be done at source systems and a data integration hub [6].

The low DQ arises from four phases at DW: at Data Sources, at Integration and Data Profiling, at ETL and at Data Base Modeling (Schema). This paper focus on the ETL phase .

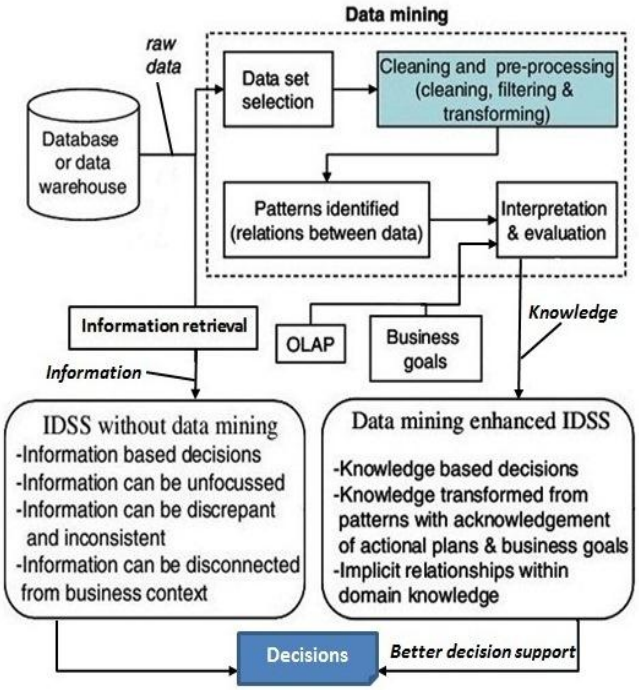

Fig: 1 Data mining enhanced (IDSS) for better DSS [17]

One consideration is whether DS is more appropriate at the source system, during the ETL process, at the staging database, or within the DW. A data cleaning process is executed in the data staging area in order to improve the accuracy of the DW. The data staging area is the place where all 'grooming' is done on data after it is culled from the source systems. Staging and ETL phase is considered to be the most crucial stage of data warehousing where the maximum responsibility of DQ efforts resides. It is a prime location for validating DQ from source or auditing and tracking down data issues. There may be several reasons of DQ problems at this phase. There are reasons of low DQ : Dirty Data ,Noise Data, Missing value, Missing Tuple, inconsistency, uncertain data, ambiguous, conflicting, duplicate records, lexical errors, domain format errors, irregularities, integrity constraint violation and similarity columns. In addition, There are more than 35 sources and causes of the poor-quality data that arise at this phase, some of the identified reasons are shown in [18].

The efficiency of DW is mainly dependent on ETL component on its architecture. Design and implementation of ETL is considered as a supporting task for DW. In a typical DW project ETL consumes a large fraction of time, money and effort to do the correct functionality and adequate performance. Due to different applications objectives and characteristic of their data, to maintain ETL will be a main concern [7].

4. Comparison & Analysis of Data Scrubbing Algorithms

This section presents two tables of comparison and analysis for DS Algorithms. Table 1 shows Time Complexity and Accuracy with the methods used whereasTable 2 shows the pros and cons of each algorithm.

Table 1. Time Complexity & Accuracy with methods used

Method	Time Complexity	Accuracy
1.K-Nearest Neighbor	Standard KNN $O(m * n + n * \log n)$ Improved KNN $O(\log n + m * n_1 + n_2 * \log n_2)$	96%
2.Genetic	$o(\log(n))$	Not Mentioned
3.Genetic Neural Network	$O(\log(n)) + O(Pn_h^2)$	97%
4.Naïve K-Means Clustering	$O(KNT)$	Not Mentioned
5.The sorted neighborhood method with multi-pass approach	$O(N \log N)$ If $w < [\log N]$ Otherwise $O(wN)$	90% for w > 4
6.Dynamic Programming	$O(C^2 \mid Z \mid)$	Not Mentioned
7.Greedy	$O(C \mid Z \mid \log \mid Z \mid)$	Not Mentioned
8.Rule based	Not Mentioned	Not Mentioned
9.Combines modified versions of PNRS and Transitive closure algorithms	Not Mentioned	89%

Table 2. The pros and cons of each algorithm

Advantages	Disadvantages
*simple and easy to learn. *robust to noisy training data. *effective for large training data.	*high computation complexity. *memory limitation. *poor run-time performance if the training set is large.
*Can solve problems with multiple solutions. *Its execution technique is not dependent on the error surface. * can locate the solution in the whole domain, it does not solve complex constraints problems easily, especially for exact constraints. * huge evaluations for generation and population are sometimes time-consuming. *it working in a random population.	*variant problems cannot be solved. *cannot assure constant optimization response times. * premature and local convergence.
* most suitable for large datasets having continuous values.	*Suffers from the shortcomings of passing a number of clusters and initial centroids preliminary. * It does not work effectively on categorical data. * with the same input parameter; the clustering results are completely different.
*Sorting requires a few machine instructions to compare two records, and thus has the potential for the largest constant factor. *Semantic identification method, using dual-threshold position weight increase user's workload, reducing the efficiency of the algorithm, but at the same time improves the accuracy and robustness of the algorithm.	*The effectiveness of this approach is based on the quality of the chosen keys used in the sort. *The complexity of this mapping is, at worst, log B.
*It solves each subsubproblem just once and then saves its answers in a table, thereby avoiding the work of recomputing the answer every time it solves each subsubproblem.	*high computation complexity.
*It is an efficient solution and gives a close-to-optimal quality improvement. *it has a close-to-optimal performance, and it is the best among all heuristics.	*very high computation complexity for reasonable quality.
*It is able to clean the data completely, addressing all the mistakes and inconsistencies in the data or numerical values specified. *The most important of which are Pragmatic, Interactive, Easy to Implement, Extensible, Performance, and Universality. *proposed comprehensive algorithm for data cleaning for DW.	*Time taken to process huge data is not as important as obtaining high quality data, since a huge amount of data can be treated one-time and thus the processing time will not be needed or consumed again. Hence, our main focus has been on achieving a good quality of the data. In spite of that the pace of implementation of this algorithm is adequate.
Enhance the PNRS algorithm by 1.5 fold times	*The modified version of it has prioritization of the attribute keys which is data specific. This cannot be automated and needs manual intervention.

Table 2 shows that:

1. There are more disadvantages in K-Nearest Neighbor. During the training period, It just simply stores the training instances and postpones most computations to classifying period, which leads to tremendous computational cost. It does not take into account of the contributions of various attributes, which give impacts on the accuracy of classifying. It describes in [8].
2. There are more advantages in Genetic Algorithm. It does not have much mathematical requirements about the

optimization problems. Due to their evolutionary nature, It will search for solutions without regard to the specific inner workings of the problem. It describes in [8].

3. There are more disadvantages in Combines modified versions of PNRS and Transitive closure algorithms. For example the errors in the field “date of birth” can be corrected. For e.g. year of birth 1850 maybe not be correct for an employee of a particular firm. It can be auto corrected to 1950. There is always a room for such data specific corrections. It describes in [5].

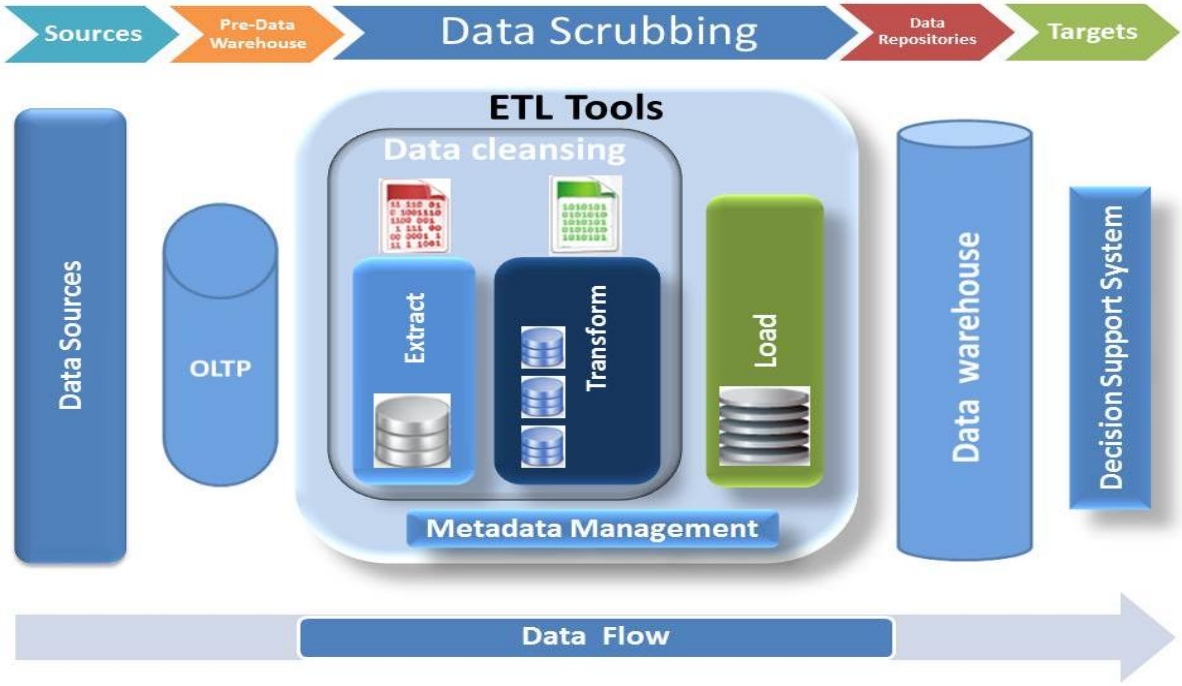

Fig 2: Data Scrubbing and ETL Tools (Adapted from [19],[20],[21]**)**

5. Comparative and analyses of Data Scrubbing Frameworks

The next subsections will describe three frameworks: two published frameworks "Wrangler" and "An Enhanced Technique for Data Cleaning" ,and one commercial framework "WinPure Clean & Match 2012".

5.1 Data Wrangler

It is an interactive system for creating data transformations. Wrangler combines direct manipulation of visualized data with automatic inference of relevant transforms, enabling analysts to iteratively explore the space of applicable operations and preview their effects. Wrangler leverages semantic data types (e.g., geographic locations, dates, classification codes) to aid validation and type conversion. Interactive histories support review, refinement, and annotation of transformation scripts. User study results show that Wrangler significantly reduces specification time and promotes the use of robust, auditable transforms instead of manual editing [22]. The following are some properties of a "Wrangler" framework:

1. Wrangler couples a mixed-initiative user interface with an underlying declarative transformation language.
2. It provides short natural language descriptions—which users can refine via interactive parameters—and visual previews of transform results, to convey the effects of data transforms. These techniques enable analysts to rapidly navigate and assess the space of viable transforms.
3. It’s interactive history viewer supports review, refinement, and annotation of these scripts. Wrangler’s high-level language supports a variety of runtime platforms: Wrangler scripts can be run in a web browser using JavaScript or translated into MapReduce or Python code.
4. It builds on this prior work to contribute novel techniques for specifying data transforms.
5. To help analysts understand the effects of an operation before they commit to it, Wrangler’s natural language transform descriptions are augmented by novel transform previews that visualize transform results. In concert, these techniques help analysts iteratively hone in on a desired transformation.

5.2 An Enhanced Technique for Data Cleaning

It gives a solution to handle data cleaning process by using a new algorithm that detects and corrects most of the error types and expected problems. This solution works on the quantitative data and any data that have limited values [14]. The following are some properties for "An Enhanced Technique for Data Cleaning" framework:

1. This solution offers the user interaction by selecting the rules and any sources and the desired targets. Each step from the algorithm is well suited for different purposes.
2. It has attempted to solve all the errors and problems that are expected, such as Lexical Error, Domain Format Error, Irregularities, Integrity Constraint Violation, Duplicates, Missing Value, and Missing Tuple.
3. The user selects any rules needed in the data cleaning system from some rules that are used in verification and process of data.
4. The system gives a detailed report on the types of errors detected and reported according to the sources of data. This is important to improve the quality of data in the future and to increase the processing speed of the data that come from those sources.

5.3 WinPure Clean & Match 2012

Its a comprehensive data cleansing, data deduplication software and a data listing software solution. It helps to clean mailing lists, spreadsheets, marketing databases and electronic mails. The software carries out data deduplication and offers the option of basic as well as advanced search. It helps to merge duplicate records on one or two lists which help to make the merging process effective and easier. "WinPure Clean & Match 2012" offers some interesting features such as 'Safe Merge' options that make sure that no data is lost while merging the records [23]. The following are some properties for "WinPure Clean & Match 2012" Software:

1. It determines a data duplication on the list to achieve the results set.
2. It helps in selecting the master record from each duplicate group, Each group of duplicates provides the ability to select a Master record. This Master record will be the record from each duplicate that will be kept. However this master record has missing values in columns.

   So, for each of these duplicate groups we want to remove the duplicate AND also want to populate all the missing values in the master record, to give us a more accurate and populate record for master record.
3. After Execute free trial version , the software will automatically remove all the duplicated records and populate the missing values.

6. Analysis Testing

This section describes the analysis and compared three frameworks as described in the previous section, after tested these frameworks. Exporter from the data used in this test, first: from "Data Wrangler", used migration file to describe unemployment rates as a percentage of total labour force, and test it on Data Wrangler framework. It consists of 23 records, and it is divided into two main categories of man and woman. These categories were further divided into native and foreign born. These fields have been divided into a span of 4 years. The missing values in this file is 17.66 %, All records are unique. The second: from "An Enhanced Technique to Clean Data", used Customer file and tested it on "An Enhanced Technique to Clean Data" framework, it consists of 9000 records, each record having 12 fields. The Missing Values for this file is 6.91%, 900 unique records and 10 exact duplicates per records. The third: Sales file, tested it on "WinPure Clean & Match 2012" framework. It consists of 10000 records, each record having 6 fields. The Missing Values for this file is 5.56% , 400 unique records and 25 exact duplicates records.

Table 3 clarifies the files used in the test, with a number of records, fields, missing values, duplicates records before testing.

Table 3. Files used and properties

File name	No. of Records	No. of Fields	Missing values	Duplicates Record
Migration	23	4	17.66 %	non-existent
Customer	9000	12	6.91 %	90%
Sales	10000	6	5.56 %	96%

Table 4 shows the results of comparison between the three above frameworks from several aspects, the most significant (Delete/Fold/unfold, Missing values, Availability, Duplication, Illegal values elimination, Misspellings, Varying value representations, File formats, Ease of usage).

For example on Delete/Fold/unfold, delete all rows when a certain column equals to a certain value.

"Data Wrangler" deals with missing values by using the their above or below values.

"An Enhanced Technique for Data Cleaning" deals with missing values by Average, most frequently, and constant values.

"WinPure Clean & Match 2012" deals with missing values by usning another table as a master to fill the missing values. Missing values of a single table handling is not supported. Additionally, "WinPure Clean & Match 2012" deals with Illegal values elimination by Range Constraints, Regular expression patterns and Unique Constraints.

Table 4.The results of comparison between three frameworks

Frameworks / Problems	Data Wrangler	An Enhanced Technique	WinPure Clean & Match 2012
Delete /Fold /unfold	Yes, Support conditional split or merging.	No	Yes, limited to a set of options, it is not flexible.
Missing values	Yes	Yes	Yes
Availability	Website /Desktop	Request from the author	Desktop
Duplication	No	Yes	Yes, Uses Fuzzy Matching
Illegal values elimination	No	No	Yes
Misspellings	No	No	No
Varying value	No	No	Yes

representations			
File formats	Text Files, Excel	Excel	Text Files, Excel, Commercial DBMS
Ease of usage	Moderate	Moderate	High

7. Conclusions and recommendations for future work

This paper presents two tables of comparison and analysis of DS Algorithms (purification data) with the methods used, the pros and cons of each algorithm, Accuracy and Time complexity. Additionally, It presents a comparison and analysis of the DS Frameworks and determines the best framework.

The future work will design and implement a comprehensive algorithm for the task of DS for DW. In addition it will develop and improve the effective and efficient DS framework.

8. REFERENCES

[1] Efraim Turban, Ramesh Sharda and Dursun Delen, "Decision Support and Business Intelligence Systems", 9th edition, 2011.

[2] S. Sumathi and S.N Sivanandam, "Data Marts and Data Warehouse: Information Architecture for the Millennium", Studies in Computational Intelligence (SCI),Springer-Verlag Berlin Heidelberg 2006.

[3] Munawar, Naomie Salim and Roliana Ibrahim, "Towards Data Quality into the Data Warehouse Development", IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, 2011.

[4] R.R Nemoni and R Konda, "A Framework for Data Quality in Datawarehouse", In J. Yang et. Al (Eds): UNISCON 2009, Springer-Verlag Berlin Heidelberg, 2009.

[5] Arindam Paul, Varuni Ganesan, Jagat Sesh Challa and Yashvardhan Sharma, "HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses", IEEE, 2012.

[6] Wayne W. Eckerson, "Data Quality and the Botton Line, Achieving Business Success through a Commitment to High Quality Data", The Data Warehousing Institute,Available at: www.dw-institute.com , Accessed on Jan 2013.

[7] Negin Daneshpour and Ahmad Abdollahzadeh, "Data Engineering Approach to Efficient Data Warehouse: life cycle development revisited", IEEE, 2011.

[8] Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin, and Ong Siong Lee, " E-Clean: A Data Cleaing Framework for Patient Data", First International Conference on Informatics and Computational Intelligence, IEEE, 2011.

[9] Xuhui Chen and Xinghua Zhang, "Extract-Transform-Load of Data Cleaning Method in Electric Company", International Conference on Artificial Intelligence and Computational Intelligence", IEEE, 2010.

[10] Vaishali Rajeev Patel and Rupa G. Mehta, "Performance Analysis of MK-Means Clustering Algorithm with Normalization Approach", World Congress on Inforation and Communication Technologies, IEEE, 2011.

[11] Mauricom A. Hernandez and Salvatore J. Stolfo, "The Merge / Purge Problem for Large Databases", Department of Computer Science, Columbia University, New York, 1995.

[12] Zhong Jia Qing, Zhang Yi Fang and Lu Zhi Gang, "Research of Data Cleaning Algorithm in Data Warehouse", China Academic Journal Electronic Publishing House, 2009.

[13] Luyi Mo, Reynold Cheng, Xiang Li, David Cheung and Xuan Yang, "Cleaning Uncertain Data for Top-k Queries, Department of Computer Science University of Hong Kong, Hong Kong, 2012.

[14] Mortadha M. Hamed and Alaa Abdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse", Development in E-systems Engineering, IEEE, 2011.

[15] Kazi shah Nawaz Ripon, Ashiqur Rahman and Atiqur Rahaman, "A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates", Journal of Computers, Vol.5, No.12, Decem, 2012.

[16] Israr Ahmed and Abdul Aziz, "Dynamic Approach for Data Scrubbing Process", International Journal on Computer Science and Engineerign, Vol.02, No.02, 2010.

[17] Shaofeng Liu, Alex H.B Duffy, Robert Ian Whitifield, Iain M. Boyle, "Integration of decision support systems to improve decision support performance", Springer-Verlag London Limited, February 2009.

[18] Ranjit Singh, and Kawaljeet Singh, "A Decriptive Classification of Causes of Data Quality Problems in Data Warehousing", IJCSI International Journal of Computer Science Issues, Vol.7, Issue 3, No2, May 2010.

[19] Carlo Vercellis, "Business Intelligence:Data Mining and Optimizationfor Decision Making", John Wiley & Sons, Ltd., 2009.

[20] "Figure 1: Basic ETL Functionality", Available at: http://gerardnico.com/wiki/dit/etl_become_di , Accessed on Jan 2013.

[21] "Figure: Overview of Data Warehousing Infrastructure", Available at http://174.37.163.146-static.reverse.softlayer.com/data-warehousing/data-warehousing-overview.asp , Accessed on Jan 2013.

[22] Sean Kandel, Andreas Paepcke, Joseph Jellerstein and Jeffery Heer, "Wrangler: Interactive Visual Specification of Data Transformation Scripts", ACM Human Factors in Computing Systems (CHI), May 2011.

[23] WinPure Ltd, "Merging Duplicate Records –The Easy Way",Available at: http://www.winpure.com/blog/merging-duplicate-records/ , Accessed on Jan 2013.