A Conceptual Framework for Data Cleansing – A Novel Approach to Support the Cleansing Process

Kofi Adu-Manu Sarpong
Valley View University,
Accra-Ghana
Faculty of Science
(Department of Computer Science)
P.O. Box VV 44, Oyibi-Accra

Joseph George Davis
Kwame Nkrumah University of Science and Technology
Kumasi-Ghana
College of Science
(Department of Computer Science)

Joseph Kobina Panford
Kwame Nkrumah University of Science and Technology
Kumasi-Ghana
College of Science
(Department of Computer Science)

ABSTRACT
Data errors occur in various ways when data is transferred from one point to the other. These data errors occur not necessarily from the formation/insertion of data but are developed and transformed when transferred from one process to another along the information chain within the data warehouse infrastructure. The main focus for this study is to conceptualize the data cleansing process from data acquisition to data maintenance. Data Cleansing is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse. Poor data or "dirty data" requires cleansing before it can be useful to organizations. Data cleansing therefore deals with identification of corrupt and duplicate data inherent in the data sets of a data warehouse to enhance the quality of data.

The research was directed at investigating some existing approaches and frameworks to data cleansing. The research attempted to solve the gaps identified in some data cleansing approaches and came up with a conceptual framework to overcome the weaknesses which were identified in those frameworks and approaches. This novel conceptual framework considered the data cleansing process from the point of data is obtained to the point of maintaining the data using a periodic automatic cleansing approach.

General Terms
Data warehousing, data cleansing, data sets

Keywords
Conceptual Framework, data cleansing process, gap analysis, dirty data

1. INTRODUCTION
The issue regarding data cleansing has been tackled by most authors. Several frameworks have been proposed over the years [1, 2, 3, 4, and 5]. Muller et al., discusses several approaches/frameworks that deal with the cleansing of "dirty data" including Potter's Wheel, Intelliclean, AJAX, ARKTOS among others. Marcus and Maletic in their studies identify methods of error detection and discussed three methods for data cleansing – define error type, search error instances and correct errors in order to ensure the quality of data [6]. The quality of data is often evaluated to determine usability and to establish the processes necessary for improving data quality [8].

In the previous work of the author on data cleansing concepts, a comparative study was conducted on the different approaches to the data cleansing and review of these concepts on data cleansing was achieved. This extensive work conducted by the author on data cleansing has been the background to which this research work is been conducted. From the study, several weaknesses were identified with the existing approaches which have informed the development of a conceptual framework to overcome the weakness of these approaches [10, 11]. The conceptual framework discussed in this paper incorporates factors that enhanced the data cleansing process and the entire data cleansing structure within a data warehouse. The process is categorized into three parts: extract, load/clean and validate/export. Listed below are the steps involved in the entire cleansing process.

1. Data is extracted from different sources (internal and external sources)
2. Data type is obtained (files/multimedia databases)
3. Data is then loaded and a copy is stored in a central repository (cache) and shows loading details
4. An interactive graphical user interface serves as the ease of use interface
5. The data is obtained from the cache and we search and remove the error instance being worked on.
6. The cleaning process is applied based on an algorithm specified and shows the cleaning details
7. Data is validated depending on the user's perception and data warehouse is updated with clean data
8. The clean data is exported to a periodic automatic cleaning for data maintenance
9. The unclean data is exported to a knowledge base engine for routine checking before forwarding it to the periodic automatic cleaning and vice versa for data maintenance.

2. THE PRINCIPLE OF DATA CLEANSING
Quality data is essential to the sustenance of organizations. Despite how hard organizations try some of the data they collect turn to be "dirty". This is why data cleansing has proven to be a major research concern to discover ways to detect and clean "dirty data" within the database. Data cleansing is also referred to scrubbing. In simple terms it means to clean the errors which are not in accordance with business rules. In order to detect and clean the wrong data,

incomplete and duplicated data, several algorithms or cleansing rules/algorithms which have been predefined to deal with "dirty data" by means of statistics, data mining and other predefined cleaning rules to improve data quality have been discussed by several authors [2, 3, 4]. The principle of data cleansing therefore is illustrated by the figure 1 below

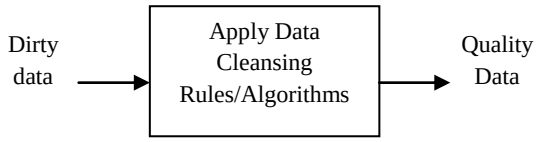

Figure 1: Basic data cleansing principle

3. DATA CLEANSING PROCESS PERSPECTIVE

Data cleansing is viewed as a process [6]. The cleansing process in this research begins when the data has been imported from different sources (internal/external) data locations. Here, there is an option to select the type of error the user wants to detect and correct from the database. When the data has been loaded into the application, the searching process is implemented and then the list of clean data and duplicated or missing data items are displayed in separate windows.

The clean data is then exported to another file for maintenance. According to [5, 9], data is audited with the use of statistical methods to detect anomalies and contradictions. This eventually gives an indication of the characteristics of the anomalies and their locations. The detection and removal of anomalies is performed by a sequence of operations on the data known as the workflow. It is specified after the process of auditing the data and is crucial in achieving the end product of high quality data [4, 5]. In order to achieve a proper workflow, the causes of the anomalies and errors in the data have to be closely considered. If for instance we find that an anomaly is a result of typing errors in data input stages, the layout of the keyboard can help in manifesting possible solutions [9]. The workflow is executed after its specification is complete and its correctness is verified. The implementation of the workflow should be efficient even on large sets of data which inevitably poses a trade-off because the execution of a data cleansing operation can be computationally expensive [5]. After executing the cleansing workflow, the results are inspected to verify correctness. Data that could not be corrected during execution of the workflow are manually corrected if possible. The result is a new cycle in the data cleansing process where the data is audited again to allow the specification of an additional workflow to further cleanse the data by automatic processing [5].

Table 1: Critical factors aiding in data cleansing

Key Parameters	Indicators
File format	Files and multimedia databases
Interface	Graphical user interface for interactivity and ease of use
Detailed cleaning process	Loading time, cleansing time and data statistics
Maintenance	Update clean data periodically

Data cleansing has proven to be essential in making "dirty data" reach its quality state. In order to achieve data quality, there are four critical factors supporting the data cleansing process which are considered in this paper. These factors along with its key indicators are listed in table 1 above. Detailed explanations of these factors are given in section 4 below.

4. THE CONCEPTUAL FRAMEWORK

The conceptual framework as shown in figure 2 below describes the adapted framework to enhance data cleansing in a data warehouse employable by companies in this research work. The framework combines propositions made by other authors in the area of data cleansing. Much attention has not given to the critical factors supporting the data cleansing process as indicated in table 1 and must be considered to enhance the data cleansing process. The parameters and its associated indicators are discussed below. Our extension to these frameworks necessitates the idea that several researches ongoing in other parts of the world are to facilitate the use of data cleansing systems by companies for efficient reporting and data analysis. First of all, the key parameters in the conceptual framework are discussed and then compared to the existing frameworks. The framework performs three basic operations – extract, load and clean and validate and export. Other activities under the basic operations must be performed in order to achieve a clean data. In the extraction phase, the data is obtained from either internal or external sources. This data could be a text file or any of the multimedia databases (Oracle, SQL, MS Access, and MSSQL Server). At the next phase – load/clean, the data extracted is loaded in order to prepare it for the cleansing operation. A cache copy of the data is loaded into memory. Outliers such as duplicates, missing data, incomplete data or any other error is detected for onward cleansing. The final phase takes care of validation as well as data export. Here the user perception regarding the data is considered. The user then validates clean data when satisfied with the quality of the data cleansing performed on the "dirty" data. The clean data is then exported to the Periodic Automatic Cleansing Area (PACA) for storage and maintenance. The data perceived by the user as unclean is sent to the knowledge base engine (KBE) for onward cleansing before finally sending it to the PACA for storage and maintenance.

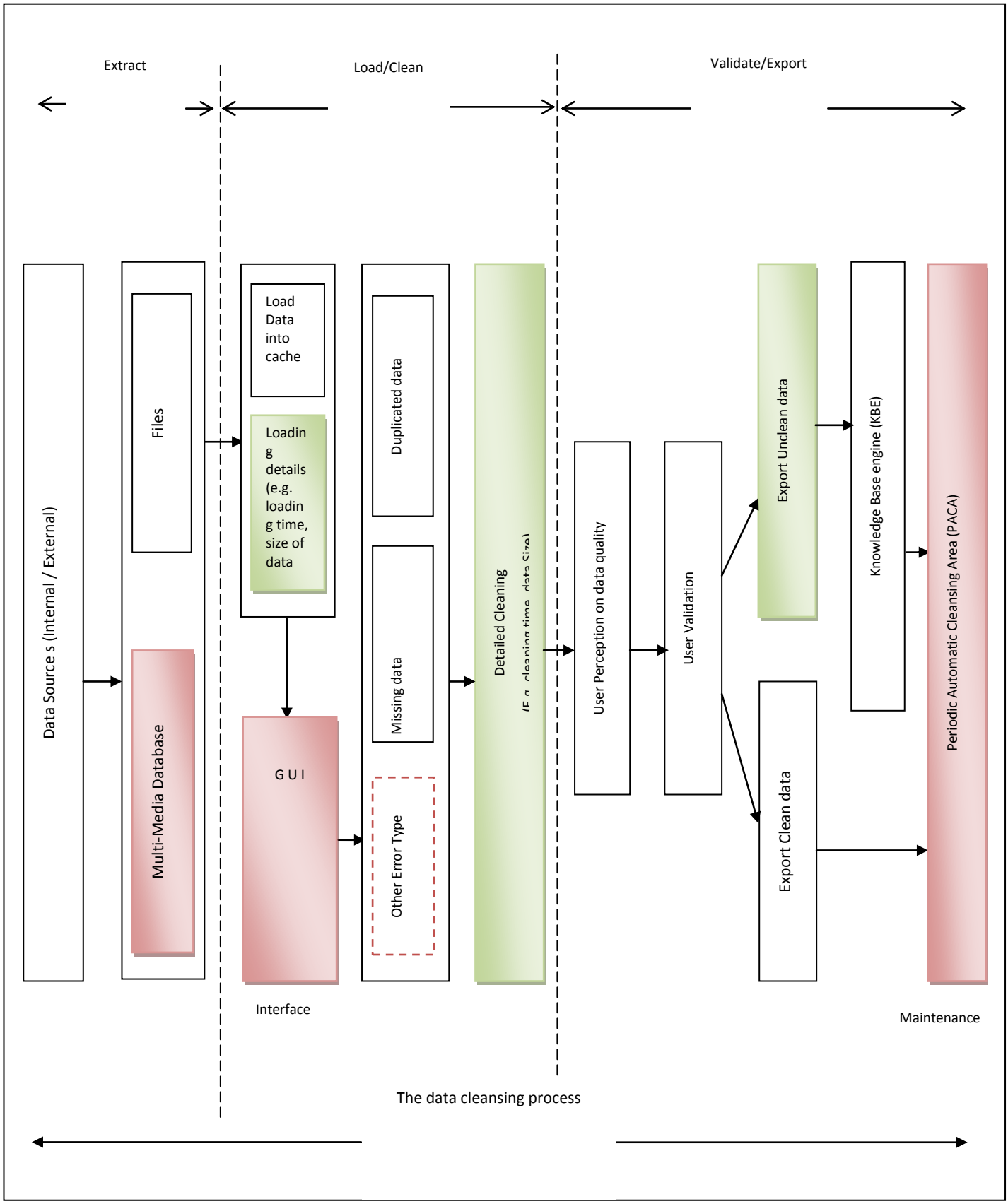

Fig 2: Conceptual framework (iCleaner)

4.1 FILE FORMAT

One of the most challenging tasks is writing a program to consider all file formats. According to [7], in their paper, an analysis of data storage and retrieval of file format system, explained that in the modern GUI world the software and hardware developers are facing a challenging trouble to use different types of file formats and storage devices.

This is because there are several file formats which include image file formats (examples GIF, JPEG, PNG, TIFF), Audio File formats (examples WAV, WMA, MP3), Video File format (examples AVI, DAT, MP3), Text File format (examples PDF, HTML, RTF) and Source File format and Database file format and many more [7]. That is why we realized from literature that most of the data cleansing tools supported only text files. It is therefore necessary to research into how other file formats could be included in the data cleansing process. The frameworks discussed in [1 – 4] all focused and dealt with one file format that is text file format. This was a limitation in these frameworks which is a strength the concept built upon in this framework.

4.2 USER INTERFACE

The user interface employed by most of the tools reviewed is not interactive. Once the user must get involved, an interactive interface is required when building a data cleansing tool. With regards to interactivity, Potter's Wheel, Intelliclean, ARKTOS and AJAX have an interactive interface which allows the user to use the system with ease. The study revealed that the framework used by [1] is very interactive and has a spreadsheet-like interface. The framework discussed by [2] has an interactive user interface; however, the framework requires little input from the end user. According to the work done [3], ARKTOS is discussed as been highly interactive and has a graphical user interface from loading and executing validations on loaded files. With regards to the work done by [4], which is AJAX, it was discovered that the user interface was complex and hence was not user friendly to non-technical persons/users. Comparing the works discussed with the framework examined in this paper, the interface has proven to highly interactive and has graphical user interface for loading, cleansing, validating and exporting data files, making it easy to use by all users.

4.3 DETAILED CLEANSING PROCESS

Evaluating the work done by [1-4], it revealed that almost all the data cleansing tools were silent on the data cleansing process. Details such as, the time it took for the cleansing to take place, number of records found with errors and the number of cleaned records, etc. had not been given in the works reviewed in literature. This is an important aspect of the cleansing process and as such must be given attention.

4.4 MAINTENANCE

One critical question that remained unanswered in literature is how the cleaned data would be maintained? The maintenance of data is important since a clean data might be corrupted when it is kept in the same repository with other files. Hence, it is an essential part of the data cleansing process. In maintenance, particular attention is given to the existing cleaned data that might become erroneous after some time when dirty data is introduced to it. To maintain the clean data, database administrators are required to set the period for the maintenance schedule. This will enable the automatic maintenance feature to perform a routine check within the specified time frame to maintain the clean data within the central repository. Depending on the period for data analysis and period set for reporting, the periodic automatic update could be done every forth night, monthly, quarterly, or yearly. Data maintenance begins when the clean data is exported to the central repository for management to use for reporting and analyzing purposes. In this framework, the periodic automatic cleansing area (PACA) is responsible for performing automatic cleansing on data exported to the area when the user has duly validated the clean data. The unclean data which goes through the Knowledge Based Engine (KBE) is also exported to the PACA when the KBE has thoroughly worked on the data in its engine.

5. COMPARATIVE ANALYSIS OF EXISTING TECHNIQUES

This novel framework is compared to other existing techniques to substantiate its efficiency. The comparison is going to be done with reference to the file format, the user interface, the cleansing process and maintainability. From the table 2 below, the parts indicated N/A simply means the factor under consideration is not applicable by the framework in question.

Table 2: Comparative analysis of existing techniques

Key Factor	Potter's Wheel	Ajax	Arktos	iCleaner (new framework)
File format	Text	Text	Text	Text/multimedia databases
User interface	Each of these techniques comes with an interactive interface but the difference lies in the level of complexity.			
Cleansing process	N/A	N/A	N/A	Considers: loading tine, cleansing time and data statistics
Maintain-ability	N/A	N/A	N/A	Considered maintenance using the PACA

6. CONCLUSION

The framework discussed in this research has considered the weaknesses in other existing approaches. Since data is an essential commodity to every organization, it is paramount to find solution to dirty data in order to obtain the right data for data analysis and reporting. The development of this framework is an attempt to bring to light a new approach to the data cleansing process. This is to help future research directions in the area of algorithm development and the creation of a software tool by software developers to accomplish the data cleansing with little effort. Though many researchers have proposed several algorithms, a novel algorithm supporting the framework will be proposed in future work.

7. ACKNOWLEDGMENTS

I am grateful to God for his directions. I express my thanks to my wife Adwoa Agyeiwaa Adu-Manu and our lovely son Nana Yaw Adu-Manu Sarpong for their encouragement and support throughout the writing process and their contribution towards the development of the paper.

8. REFERENCES

[1] Raman V and Hellerstein J.M, Potter's Wheel: An Interactive Data Cleaning System, Proceedings of the 27th VLDB Conference, Roma, Italy, 2001, pp. 1-10.

[2] Mong L.L, Tok W.L and Wai L.L.(2000). IntelliClean : A Knowledge-Based Intelligent Data Cleaner, ACM, pp. 290-294

[3] Panos V., Zografoula V, Spiros S., and Nikos K.(2000). ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp.1-6

[4] H. Galhards, D. Florescu, D. Shasha, E. Simon. (May 2000). AJAX: An extensible data cleaning tool. Proceedings of the ACM SIGMOD on Management of data, Dallas, TX USA, pp. 21-22.

[5] Heiko Muller, Johann-Christoph Freytag. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing, pp. 21.

[6] Jonathan I. Maletic, Andrian Marcus. (2000). Data Cleansing: Beyond Integrity Analysis, pp. 8.

[7] R. Mariappan and B. Parthasarathy. (2009). an analysis of data storage and retrieval of file format system, Indian Journal of Science and Technology, vol.2 No. 9, pp. 38-40.

[8] Dongre Kuldeep (2004). Data cleansing strategies, pp. 10

[9] Rahm, E., Do, H.H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bull, vol. 23 No. 4, pp. 3-13

[10] Adu-Manu, K.S and Arthur J.K. (2013). A Review of Data Cleansing Concepts – Achievable Goals and Limitations, International Journal of Computer Applications (0975 –8887), vol. no 76, pp. 19-22.

[11] Adu-Manu, K.S and Arthur J.K. (2013). Analysis of Data Cleansing Approaches regarding Dirty data – a Comparative Study, International Journal of Computer Applications (0975 –8887), vol. no 76, pp. 14-18.