IJCAT - International Journal of Computing and Technology, Volume 3, Issue 7, July 2016
ISSN : 2348 - 6090
www.IJCAT.org

Comparative Analysis of Data Cleaning Tools Using SQL Server and Winpure Tool

[1] Dr. Abdelrahman Elsharif Karrar, [2] Moez Mutasim Ali

[1] Taibah University
Saudi Arabia

[2] University of Science and Technology
Sudan

Abstract - Data cleaning based on similarities involves identification of "close" tuples, where closeness is evaluated using a variety of similarity functions chosen to suit the domain and application. Current approaches for efficiently implementing such similarity joins are tightly tied to the chosen similarity function. In this paper, we compare between two cleaning tools. The two cleaning tools considered are Microsoft SQL Server2012 Data Quality Services and Winpure clean and match software. Data Quality Services is a knowledge-based system that performs both computer-assisted and interactive cleansing and matching processes using the created knowledge base. WinPure Clean & Match 2009 is the latest edition, following on from the award-winning Clean & Match 2007. It builds upon its data duplication module and now features advanced fuzzy matching logic to identify and remove more duplications. The comparison between the above two tools is carried out using academic datasets, with the Weather dataset as its input.

Keywords **- Data Cleaning, Data Cleansing, Data Quality Services, Winpure, Datasets.**

1. Introduction

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. [1] Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary. [2]

Databases and data warehouses require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain "dirty data" is high. Furthermore, data warehouses are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. 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. [3]

2. Data Cleaning

Data cleaning also called data cleansing or data scrubbing, is a process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. [4]

An organization in a data-intensive field like banking, insurance, retailing, telecommunications, or transportation or Universities might use a data cleaning tools to systematically examine data for flaws by using rules, algorithms, and look-up tables.

Typically, a database cleaning tools includes programs that are capable of correcting a number of specific types of mistakes, such as adding missing zip codes or finding duplicate records. Using a data cleaning tools can save a database administrator a significant amount of time and can be less costly than fixing errors manually.

2.1 Data Cleaning Approaches

In general, data cleaning involves several phases

- **Data analysis**: In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required. In addition to a manual inspection of the data or data samples, analysis programs should be used to gain metadata about the data properties and detect data quality problems. [5]

- **Definition of transformation workflow and mapping rules**: Depending on the number of data sources, their degree of heterogeneity and the "dirtiness" of the data, a large number of data transformation and cleaning steps may have to be executed. Early data cleaning steps can correct single-source instance problems and prepare the data for integration. Later steps deal with schema/data integration and cleaning multi-source instance problems, e.g., duplicates. The schema-related data transformations as well as the cleaning steps should be specified by a declarative query and mapping language as far as possible, to enable automatic generation of the transformation code. In addition, it should be possible to invoke user-written cleaning code and special purpose tools during a data transformation workflow. The transformation steps may request user feedback on data instances for which they have no built-in cleaning logic. [5]

- **Verification**: The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated, e.g., on a sample or copy of the source data, to improve the definitions if necessary. Multiple iterations of the analysis, design and verification steps may be needed, e.g., since some errors only become apparent after applying some transformations. [5]

- **Transformation**: Execution of the transformation steps either by running the ETL workflow for loading and refreshing a data warehouse or during answering queries on multiple sources. [5]

- **Backflow of cleaned data**: After (single-source) errors are removed, the cleaned data should also replace the dirty data in the original sources in order to give legacy applications the improved data too and to avoid redoing the cleaning work for future data extractions. For data warehousing, the cleaned data is available from the data staging area. The transformation process obviously requires a large amount of metadata, such as schemas, instance-level data characteristics, transformation mappings, workflow definitions, etc. For consistency, flexibility and ease of reuse, this metadata should be maintained in a DBMS-based repository. To support data quality, detailed information about the transformation process is to be recorded, both in the repository and in the transformed instances, in particular information about the completeness and freshness of source data and lineage information about the origin of transformed objects and the changes applied to them. [5]

3. Data Cleaning Tools

Data cleaning tools are software's applications that help users to clean data by identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing modifying or deleting this dirty data.

Most of the cleaning tools go throw the five phases that mentioned in 2.1. Focusing more specifically on data cleaning, there are many techniques in the research literature, and many products in the marketplace. The space of techniques and products can be categorized fairly neatly by the types of data that they target. Here we provide a brief overview for two data cleaning tools, using academic data to clean it with these two tools, and at last we compare the final results of these tools.

One of the most popular tools is Microsoft SQL Server 2012. Organizations can use SQL Server 2012 to efficiently protect, unlock, and scale the power of their data across the desktop, mobile device, datacenter, and either a private or public cloud. SQL Server 2012 has made a strong impact on organizations worldwide with its significant capabilities. It provides organizations with mission-critical performance and availability, as well as the potential to unlock breakthrough insights with pervasive data discovery across the organization. Finally, SQL Server 2012 delivers a variety of hybrid solutions you can choose from.

Microsoft SQL Server 2012 provides some services to deal with the big data. One of these services is Data Quality services which used to build a knowledge base and use it to perform a data cleaning in different tasks, including correction, enrichment, standardization, and de-duplication.

There are some other data cleaning tools, the most known is WinPure Clean & Match 2013. This is award-winning list cleaning, data cleansing and data deduplication software, because it offers a collection of affordable data cleaning tools.

3.1 Microsoft SQL Server 2012 Data Quality services

Data cleansing in Data Quality Services (DQS) includes a computer-assisted process that analyzes how data conforms to the knowledge in a knowledge base, and an interactive process that enables the data steward to review and modify computer-assisted process results to ensure that the data cleansing is exactly as they want to be done. [6]

The data cleansing feature in DQS has the following benefits:

Identifies incomplete or incorrect data in your data source (Excel file or SQL Server database), and then corrects or alerts you about the invalid data.

Provides two-step process to cleanse the data: computer-assisted and interactive.

- **The computer-assisted** process uses the knowledge in a DQS knowledge base to automatically process the data, and suggest replacements/corrections.
- **Interactive** allows the data steward to approve, reject, or modify the changes proposed by the DQS during the computer-assisted cleansing.

- Standardizes and enriches customer data by using domain values, domain rules, and reference data. For example, standardize term usage by changing "St." to "Street", enrich data by filling in missing elements by changing "1 Microsoft way Redmond 98006" to "1 Microsoft Way, Redmond, WA 98006".
- Provides a simple, intuitive, and consistent wizard-like interface to the user to navigate data and inspect errors amongst a very large set of data.

DQS categorizes the data under the following five tabs:

- Invalid: values that failed a domain rule or reference data.
- New: Valid values for which DQS does not have enough information.
- Suggested: Values for which DQS found suggestions.
- Corrected: Values that are corrected by DQS.
- Correct: Values that were found correct.

The following figure displays how data cleansing is done in DQS:

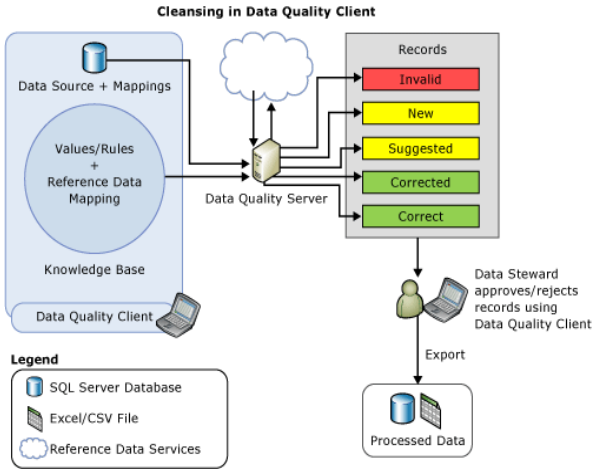

Figure (1) Cleaning in Data Quality Client [6]

- A data quality project tends to have two elements to it. One is an initial fix to clean up bad data. This is known as a data cleansing project. As the name implies, the end goal is to have a set of clean data. The tools look through the data, transforming values to match a standard, flagging outlying values that might be anomalies and suggesting changes that could be made. It also hunts for possible duplicates through data matching, applying policies to look for entries in the database that might refer to the same thing.
- The second part of a data quality project is what happens next to keep the data clean. As with Master Data Management, this isn't a fix-once act. It's very easy for data quality issues to creep back in after the cleansing has taken place so an implementation of Data Quality Services needs to bear in mind what should happen next. The processes and policies need to be defined to ensure that the data quality knowledgebase is used in future to maintain the quality of the data. It's also important to identify the data stewards who will be responsible for fixing any problems the knowledgebase flags.
- It's also important to think of the knowledgebase as an on-going project. The set of knowledge and rules within the knowledgebase can grow over time, bringing more control and accuracy to data. As more data passing through the knowledgebase, it becomes more tuned to picking out anomalies and better at identifying what the correct value should be.

- A Data Quality Services project should include both the plan for how to clean the data initially and how to maintain quality moving forward.
- Before either of these can start, however, we need to define what we want our data to look like. A key part of a data quality project is working out what we want to be the correct value. Once we've done that, we can start applying the rules to change the other values so you end up with consistency across our whole data set.
- So when we're starting to work with Data Quality Services, first take a look at our existing data and decide what we'd like it to look like. Then we can do data cleansing and data matching to give ourselves a clean and accurate set of data to start with. Then we need to hook our knowledgebase into our processes to ensure data quality moving forward.
- Before creating data cleansing project must have a relevant knowledge base to use in the data quality project for the cleansing, after creating a knowledge base project the cleansing project can be created to use that knowledge base project on it.

3.2 WinPure Clean & Match 2013 Software

- WinPure are a worldwide leading provider of list and data quality solutions that are powerful, simple to use, inexpensive and most importantly can be used by anyone rather than just IT specialists or data cleaning experts. [7]
- WinPure Clean & Match 2013 is the latest edition. It builds upon its data deduplication module and new features advanced fuzzy matching logic to identify and remove more duplications. Already acclaimed for its easy-to-use interface and powerful data cleansing functions, WinPure have now added fuzzy matching logic onto its data deduplication module. It provides a range of data services that are aimed at further improving data quality and providing a clean starting point for ongoing database management.
- Businesses around the world are now using WinPure software to help improve the quality of their information, helping them to increase profitability through more accurate data, and reducing costs by eliminating duplications, spelling errors and mistakes. [7]

3.3 Comparing Data Quality Services with WinPure Clean and Match

Microsoft Data Quality Services and WinPure Clean and Match 2013 Software are two of the leading and powerful tools in the data cleaning world, the Data Quality Services is provided by Microsoft SQL Server 2012, and the WinPure Clean and Match is provided by WinPure company.

Comparing between these two tools is so hard because every tool have its own features.

The final result of Data Quality Services goes throw the following four stages:

- **A mapping stage** where users can identify the data source to be cleansed, and map it to required domains in a knowledge base.
- **A computer-assisted cleansing stage** where DQS applies the knowledge base to the data to be cleansed, and proposes/makes changes to the source data.
- **An interactive cleansing stage** where data stewards can analyze the data changes, and accept/reject the data changes.
- **The export stage** that lets users export the cleansed data.

In the third stage there is a confidence level value for the suggested correct answers, this value is based on the knowledge base that has been built in DQS against a high-quality data set that was created before we start the cleaning process.

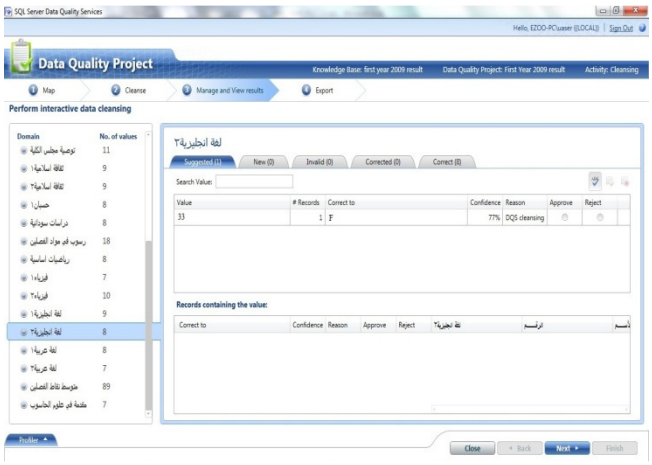

Figure (2) Correct Confidence Level Value

After performing all the stages The DQS provides statistics, these statistics about the source data and the

cleaning results that enable users to make informed decisions about data cleansing, in the last step we export the data and the final result was:

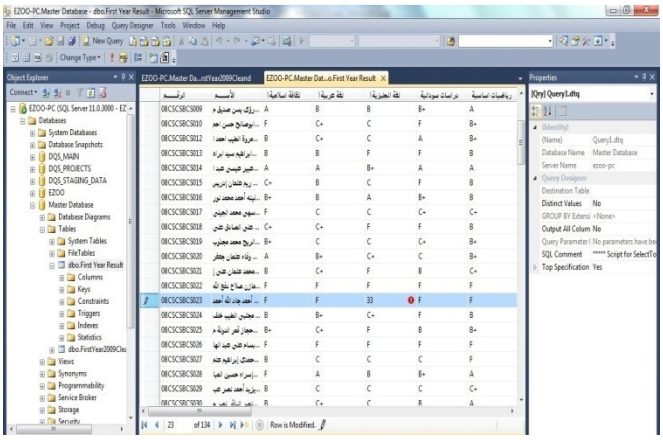

Figure (3) Data in SQL Server before Cleaning

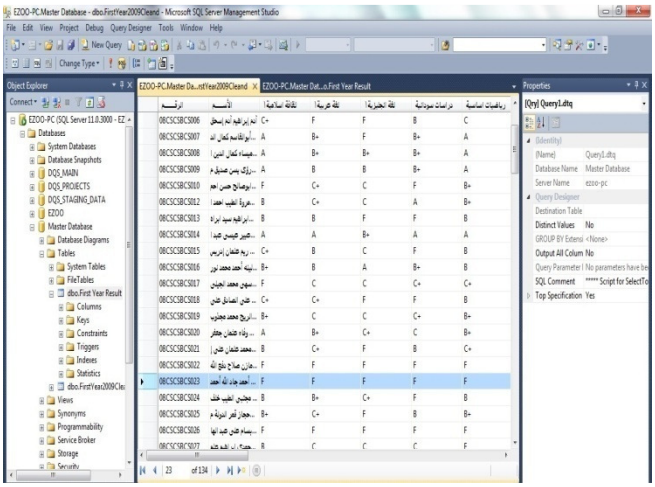

Figure (4) Data in SQL Server after Cleaning

The final result came by importing the academic data into the WinPure Clean and Match 2013 software using , and using two of the seven powerful cleaning modules that provided by WinPure Clean and Match 2013 software, to easily clean, correct, standardize and remove duplications from these academic data in a matter of minutes, rather than hours.

The first module used here is Statistics module, the main idea of this module is to identify which filled/columns have missing values, how much of data is fully populated (eg. How many missing names or postcodes exist on the data, how many contacts have missing email addresses, etc). [8]

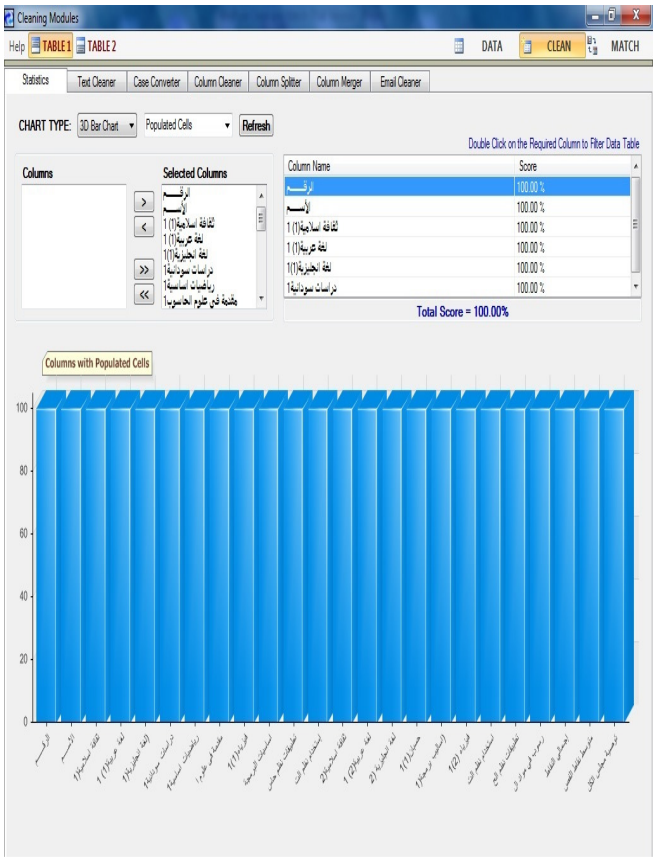

Figure (5) Statistics Module

The second module is Text Cleaner module, a very powerful module that will quickly and effectively remove unwanted characters from data. At a press of a button, the text cleaner automatically remove non-printable characters, leading or trailing spaces, and even repetition of certain characters. [8]

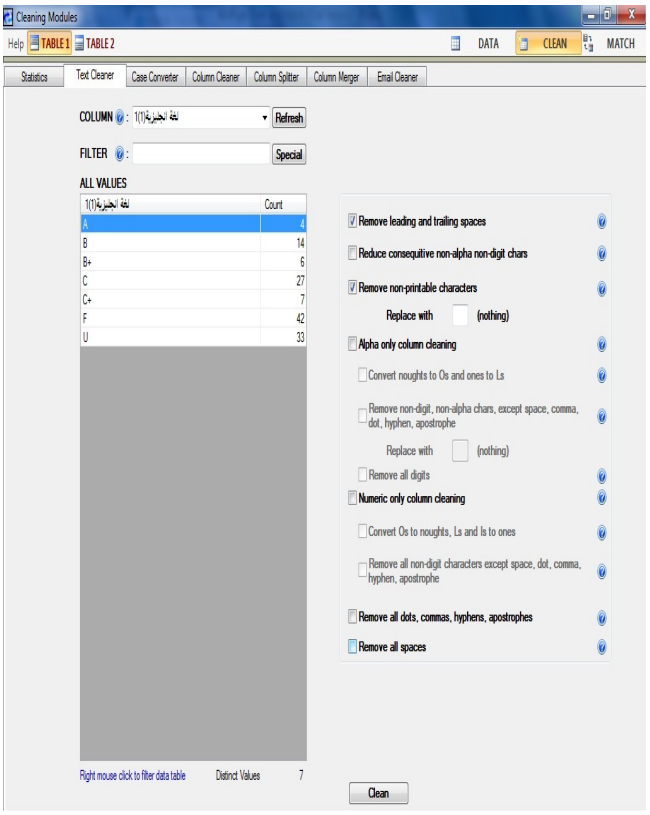

Figure (6) Text Cleaner Module

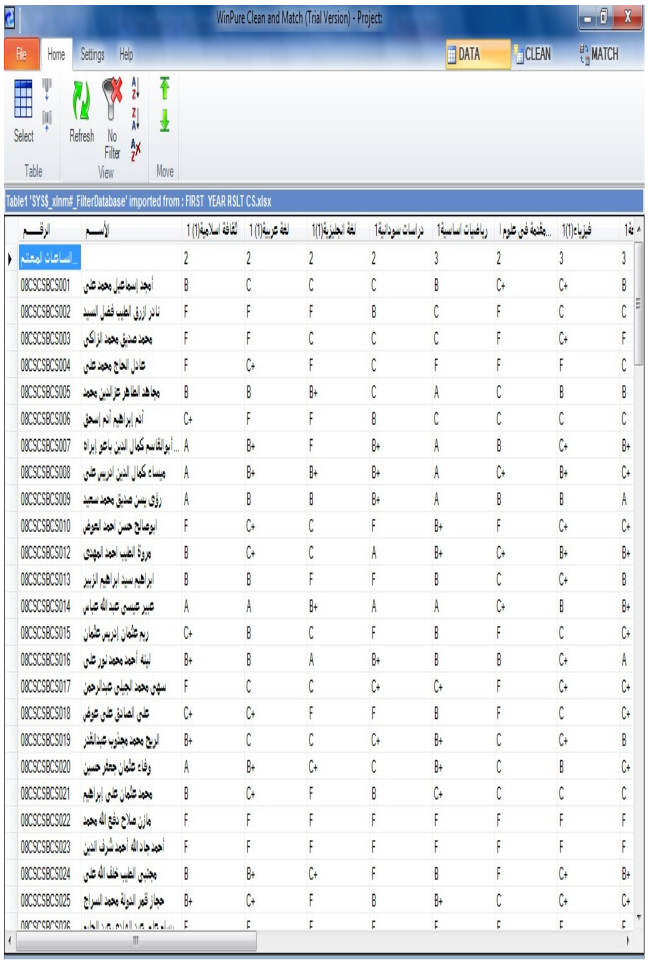

Figure (7) Data in WinPure after Cleaning

Table (1): Comparison between Data Quality Services and WinPure Clean and Match 2013 Software

Cleaning Tool Name	Import and Export	Cleaning Process	Accuracy	Performance time
Microsoft SQL Server 2012 Data Quality Services	**complex and go throw several steps**	**go throw several steps**	**depend on the created knowledgebase richness**	**depend on the hardware consideration and data size**
WinPure Clean and Match 2013 Software	**easy and it's just a click of a button**	**done with a click of a button**	**depend on the used cleaning module**	**depend on data size**

4. Discussion

In the Comparison between Data Quality Services (DQS) and WinPure Clean and Match 2013 Software we depends on four elements:

1. How each tool import and export the files.
2. How is the cleaning process flow in each tool.
3. The cleaning accuracy.
4. The cleaning performance time.

Referring to Table (1) we can infer that WinPure Clean and Match 2013 Software takes less time in cleaning process than Data Quality Services, because it didn't give consideration to the hardware.

Data Quality Services Provide several steps to begin the cleaning process, in the other way WinPure Clean and Match 2013 Software provide easy

Data Quality Services Provide options for an automated process to clean the source data or manually go over the cleansing results and fix issues that are found, in the other way WinPure Clean and Match 2013 Software don't provide the manual option.

References

[1] E. Rahm, "Data Cleaning: Problems and current approaches", 2004.

[2] V. G. Surajit Chaudhuri, Raghav Kaushik, "A Primitive Operator for Similarity Joins in Data Cleaning," 2006.

[3] M. Li Lee, T. Wang Ling, Y. Teng Ko, "Cleansing data for mining and warehousing", August 2003.

[4] A. Chapman, "Principles and Methods of Data Cleaning," July 2005.

[5] M. Hellerstein, "Quantitative Data Cleaning for Large Databases", February 27, 2008.

[6] Microsoft®, "Data Quality Services", 2012.

[7] D. Leivesley, "WinPure Clean & Match 2013, Powerful Data Quality Software Featuring Advanced Fuzzy Matching Data Deduplication," 2013.

[8] WinPure®, "WinPure Clean & Match 2013," 2013.