TELKOMNIKA, Vol. 16, No. 2, April 2018, pp. 834 ∼ 842
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/telkomnika.v16.i2.7669

Data Cleaning Service for Data Warehouse: An Experimental Comparative Study on Local Data

Arif Bramantoro
Faculty of Computing and Information Technology in Rabigh
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: asoegihad@kau.edu.sa

Abstract

Data warehouse is a collective entity of data from various data sources. Data are prone to several complications and irregularities in data warehouse. Data cleaning service is non trivial activity to ensure data quality. Data cleaning service involves identification of errors, removing them and improve the quality of data. One of the common methods is duplicate elimination. This research focuses on the service of duplicate elimination on local data. It initially surveys data quality focusing on quality problems, cleaning methodology, involved stages and services within data warehouse environment. It also provides a comparison through some experiments on local data with different cases, such as different spelling on different pronunciation, misspellings, name abbreviation, honorific prefixes, common nicknames, splitted name and exact match. All services are evaluated based on the proposed quality of service metrics such as performance, capability to process the number of records, platform support, data heterogeneity, and price; so that in the future these services are reliable to handle big data in data warehouse.

Keyword: *Data Cleaning Service, Data Warehouse, Data Quality, Local Data*

1. Introduction

Data warehouse is a relational database for questioning and analyzing by further processing. It is obtained from several transactions from other sources. Integrated data warehouse is an integration of files, sources and other records. Several services are used to ensure good data, such as data cleaning and data integration service within enterprise. Subject oriented data warehouse is a subject centric model involving several subjects, such as vendor, product, sales and customer [1].

A good data warehouse must focus on proper analysis and cleaning of data rather than daily service transaction and operations. This kind of model is required by most enterprises. The model must be simple and related to the data cleaning objective. It should also avoid data which are not required for transaction and decision making operations. Nonvolatile nature of data is important for these operations. It should be physically separated away from the application. This separation helps data for recovery and other time consuming mechanisms, such as loading and retrieval of data [2]. Time variant is the period of time which is involved in the data storage in data warehouse. This is the element of time. The decisions taken during the process is very important, therefore the generated trend reports are significant elements in data warehousing. A proper decision support system works successfully with this trend report. There are several commercial applications such as customer relationship and business applications using utilizing data cleaning service.

A proper scheme is involved during the development of data warehouse. The questions and analysis are completed in the designing stage. Meaningful access of relevant data is required together with the generated values. The extraction of the source is important, therefore it must be very clean without any unrelated sources. The provided service is data cleaning in any big enterprise. The input data input are rechecked and tested before they are allocated to a specific data warehouse. The loaded data are separated from technical specification and process [3].

Received October 25, 2017; Revised February 13, 2018; Accepted March 14, 2018

Several automatic executions are also made to eliminate error in the data. Identifying the incomplete data is a hindrance for processing of data. It makes the corrections more complicated. A service called back flushing is used to recheck the data cleaning frequently. Installation of data occurs in the first instance for the model from other sources. Monitoring service is used for the recovery of data at different levels from huge to small quantity. The amount of load is also related to the process in warehouse. Hence, caution steps are noteworthy to process the data smoothly.

ETL is the process of Extract, Transform, and Load of data. It means that the extraction of relevant data is followed by the transformation and consequently the loading of data in the warehouse. Extract is a method of data extraction which occurs in data warehouse from the allocated resources. The consolidation of these resources also takes places in the separated system that is allocated for each level of processing. This step of extraction takes the data into another level called transforming [4].

Transform is a mechanism to involve the data extraction which is converted from previous form and placed in the data warehouse without any errors in the data. The source of data needs a proper manipulation of all methods. It follows a set of functions to extract data into the warehouse without any modification to the existing data. The technical and other requirements are validated to meet the requirements. There are several transformations involved in the process, such as selecting only particular information and assigning them with specific functions. Coding the data with values is a concrete service of data cleaning which occurs automatically. Another form of data cleaning service is to encode the result into a new value and to combine data values of two different methods. The form of data can be simple or complex method. The path of data may be failed or successful, which both methods involve in handling data in a specific program. For example, the model can be a translated code in an extracted data.

Load is a process of data handling which is important together with the targeting range of information. Few data can be overwritten with other non-updated or updated data. The selection of the design is also important together with a proper understanding of the available choices related to time and business requirements. The complex model of system is to allow several changes to be updated and uploaded. The overall quality of data in the data warehouse environment is validated by utilizing ETL mechanism [5].

The objective of this paper is to identify the causes for data quality problem. Particular methodology and experiments are adopted to address the problem. It is expected that it contributes to better data quality in data warehouse. Data cleaning and duplicate elimination services are the appropriate methods to improve the data quality. Experiments are conducted to provide the results and comparisons between de-duplication services. The research approach as it is presented here is novel due to the level of implementation by utilizing service oriented approach as an evident support to the conducted survey.

In this research, only data cleansing service is considered as the main task. The rest of the data quality services such as completeness and historical reputation services remain as a future work to compose more services in a service-oriented system.

2. Data Quality in Data Warehouse

Data warehousing is a promising industry for several government organization and private institutions, which involves several confidential data storage with regards to internal security. With the enormous amount of data, the responsibility of organization becomes critical when it comes to security concerns [6]. The assurance of data quality is the primary objective of any management levels. There is an increased potential data quality and its irregularities. Data warehouse is adopted by the organization to improve the relationship between customers, client and management. Thus, improving the efficiency for the entire organization is required.

Data quality is defined as the measured performance or the loss of data in an organization [2]. The purpose of data quality measurement is to identify the missing data from the system. The quality of data attained for the data warehouse model assures the inputs on the client side. However, one user is different from another user. The data must be simple, consistent and full of understanding. The abundance of data increase the burden on the system side. The quality

of data is critical together with the identification of irregularities. The key quality of data and its dimension metrics are important to understand the effective quality improvement.

Data quality has the importance due to the use of data warehouse system. Data quality is measured in each phase of operations. Metrics are selected to ensure measurement of data quality and analysis. The selection of metrics is critical to the final result which directly affects to the customer relationship. Quantifying data is important to save the cost and improve market standards in a competitive economy [7]. In this paper, data quality and quality of service metrics are combined to improve the confidence of data quality process.

With an increasing technology and enormous data inputs in industry, the authorities need to improve the quality of data in enterprise. There are several problems faced by the enterprises in order to maintain and sustain their quality of service in delivering the project. The types of data are classified into intrinsic, contextual, representative and accessible. Performance is the factor of quality standard in any enterprise trademarks. The addition of data must be static rather than dynamic in order to efficiently avoid irregularities during monitoring the process of quality standard improvements. The consumer must be carefully considered when there are data sent by the client to the enterprise [8]. A common data quality framework includes a loop of activity in weighting its cost and benefit as illustrated in Figure 1.

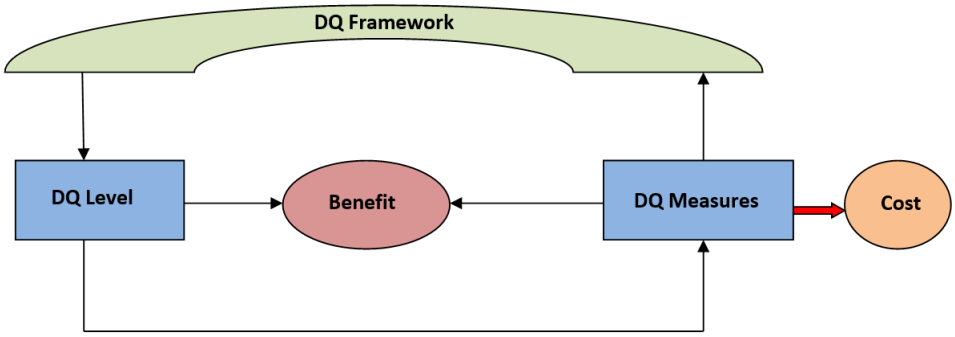


Figure 1. Data Quality Process Loop

3. Classification of Data Quality

Modern data quality improvement approach requires a real time scenario with the preference to avoid operational and analytical models [9]. The correctness and assurance of data quality is measured after during the improvement. The data quality issues requires to be handled for designing services for data warehouse without any quality problems. The identification of problems caused by poor data is examined to derive a proper procedure. Inaccurate information by the customer is another cause for a decline in quality. Unlike conventional approach, there are several other proximity and time variant issues that must be given a consideration in modern approach. The source of data in modern data warehouse is related to the data quality improvement in data warehouses. The fields are filled by the ones in unstructured forms. These issues are improvements and advancement for modern research in data warehouse compared to the conventional method of Inmon [10] and Kimball research [11].

According to Data Warehouse Institute [12], data quality improvement includes the correction on defective data to ensure the achievement of minimum level of data quality standard. It is also mentioned that data are required to be flawless without any irregularities. It has to meet the standard requirement of the compatible application. The quality of data required by user is different from the one required by the organization. Strict rules are used to avoid an improper data processing. Validation is made at particular level where data are equipped with pin numbers or passwords. The frequent data errors are considered as a common phenomenon, however the model developed for data quality in data warehouse is regularly adaptive to all changes. Hence,

data in high quality can be used in operations, decision making process and modeling. In addition, the quality information indicates which data model needed by data warehouse.

The probability of errors that can lead to a decline in data quality is required for records and protocol distribution in a network. The calculation of technical information and the requirement protocol proposed by the enterprise have to be fulfilled to achieve data quality. The assured mechanism for development of these protocols can benefit the enterprise by providing data quality management in a large scale. The goal is to meet market standards rather than to adopt low cost protocols that may lead to a failure of the suggested model. Many organization and government agencies are involved with huge database collection [13]. The importance of data quality becomes a big concern to achieve results and experiments required by a client. If this is not taken seriously, several complications may arise due to a failure in data quality which affects the customer relationship model at any level of processes in enterprise.

An effective risk management is needed for a system to learn from its deficiencies. The designed protocols must be in such a way to cope up with the risk and deliver the required standard results. The policy makers must decide the risk strategies to comprehend the desired data quality standards. Further management of the risk mitigation protocols for data quality improvement and the desired policy formulation play a major role based on data quality requirements. The agent for risk mitigation approach is assigned after several testing levels, since they are going to play a major role in the enterprise working level. The decisions are taken from the policy of the risk mitigation for data quality approaches. The management of any enterprise should pay an attention to Llyods approach [14] of data quality model and risk mitigation standards.

4. Duplicate Elimination Test Bed

Duplicate elimination is one of the important concrete services in data cleaning service composition. The main objective of data cleaning service is to maintain data quantity. It is a service-oriented method to remove duplicated data which may be represented by the user more than one time. The general idea is a matching process that enables to identify duplicated data.

One important aspect during the search of the duplication of the same records is the ambiguity of data. There are several experiments conducted to convince the duplicate elimination. Several services are used during the matching process on those experiments. However, only few of them give the desired results.

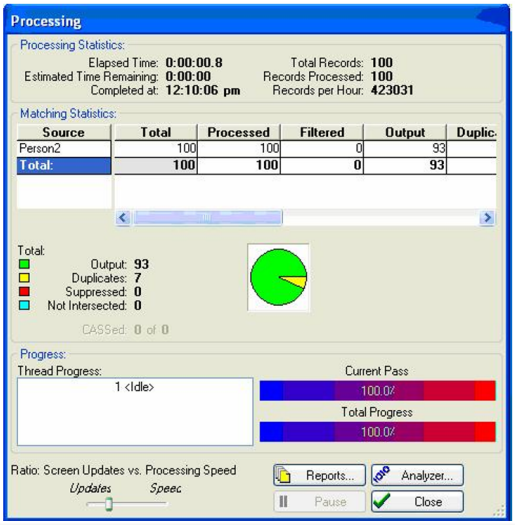
The duplicate information are displayed and recorded in the form of a table together with the indication of its percentage. The effective services are generally chosen to get successful results of the improvement on data quality standards. The aim of this research is to compare the duplicate elimination services and find out which ones perform better. The comparison is generally based on two parameters. The first parameter is the time to detect the errors in the data that alter the system and environment. Additional time is required to improve the quality of data in the process system of any functionality. The second parameter is the memory that determines on the effectiveness of the data quality.

The services required for the experiments are available from the following service providers: WinPure Clean and Match (referred as WinPure), DoubleTake3 Dedupe & Merge (referred as DoubleTake), WizSame (referred as the same name), and Dedupe Express (referred as DQ-Global).

Before the comparison between services are conducted, the experimental test bed needs to be developed on real data in local Saudi Arabia. During the first experiment, there are eight duplicates from the data set manually selected from data warehouse and further examined by the duplicate detection services as shown in Figure 2. It is important to note that the data with high privacy are preserved. Due to the limitation of the page, only the result of the duplicate data detected by DoubleTake service is presented in this paper as illustrated in Figure 3 which has seven duplicate data. In this figure, DoubleTake service provides some information that might be different from other services, such as the number of suppressed records and the rate of records per hour. Hence, this research standardizes the service output as the number of duplicate records to ease the comparison. The quality of service is included for the performance analysis as well.

PersonID	Nation.	NAME	Sex	Status	BirthDate	PassportNo
416301	351	MOHAMMED YOUNES GHULAM ALI	1001	1001		
416300	351	MOHAMMED YOUNES GHULAM ALI	1001	1001	01/07/1942	375917
507100	351	MOHAMMAD BHUTTA	1001	1002	01/07/1940	245467
507101	351	MOHAMMAD SAEED BHUTTA	1001	1002	01/07/1940	245467
515900	511	ROSSI A. ROGER	1001	1002	07/10/1928	4131
515901	511	ROSSI ROGER	1001	1002	07/10/1928	4131
525900	351	MUHAMMAD YAQUB	1001	1002	01/07/1927	448411
525901	351	Mohammad Yaqub	1001	1002	01/07/1927	448411
677701	514	OLIVIER ARDINO	1001	1002	01/07/1908	4869210
677700	514	OLIVIER ARDUINO	1001	1002	01/07/1908	4869210
591000	351	MOHAMMAD ZAKIR MALIK	1001	1001	05/06/1947	841901
591001	351	MOHAMAD ZAKR MALIK	1001	1001	05/06/1947	841901
2235	2	Robert Smith	1001	1001	20/04/1923	5567777
2234	2	Bob Smith	1001	1001	20/04/1923	5567777
466401	1	MR. MAJED ALKHAMEES	1001	1002	01/03/1983	221554
466402	1	MAJED ALKHAMEES	1001	1002	01/03/1983	221554

Figure 2. First Experiment Dataset



416301	351	MOHAMMED YOUNES GHULAM ALI	1001	1001		
416300	351	MOHAMMED YOUNES GHULAM ALI	1001	1001	01/07/1942	375917
515900	511	ROSSI A. ROGER	1001	1002	07/10/1928	4131
515901	511	ROSSI ROGER	1001	1002	07/10/1928	4131
525901	351	Mohammad Yaqub	1001	1002	01/07/1927	448411
525900	351	MUHAMMAD YAQUB	1001	1002	01/07/1927	448411
466401	1	MR. MAJED ALKHAMEES	1001	1002	01/03/1983	221554
466402	1	MAJED ALKHAMEES	1001	1002	01/03/1983	221554
677701	514	OLIVIER ARDINO	1001	1002	01/07/1908	4869210
677700	514	OLIVIER ARDUINO	1001	1002	01/07/1908	4869210
507100	351	MOHAMMAD BHUTTA	1001	1002	01/07/1940	245467
507101	351	MOHAMMAD SAEED BHUTTA	1001	1002	01/07/1940	245467
591001	351	MOHAMAD ZAKR MALIK	1001	1001	5/6/1947	841901
591000	351	MOHAMMAD ZAKIR MALIK	1001	1001	5/6/1947	841901

Figure 3. Duplicate Data Detected By DoubleTake

Due to the limitation of the page, only one experimental test bed is presented in this paper. A summary of total five experiments is presented in Table 1.

Table 1. Summary of Five Experiments

	WinPure	DoubleTake	Wizsame	DQGlobal
Experiment 1	50%	88%	75%	88%
Experiment 2	25%	75%	67%	33%
Experiment 3	50%	90%	90%	80%
Experiment 4	88%	50%	75%	63%
Experiment 5	17%	100%	92%	83%

5. Comparisons Between Services in Duplicate Detection

In this comparison, there is a finer granularity based on the previous experiment test bed. Each service processes the same set of records so that the detection capability of all services can be justified. All the records for the comparison are based on the duplicate types. The comparisons are made based on the predefined duplication types. There are seven duplication types as follows:

1. Different spelling and pronunciation comparison.
   The duplicated records examined in this comparison and the examination result by running four services are illustrated in Figure 4. Due to the existence of different languages in Saudi Arabia, inconsistent name transliterated from another language is not uncommon. It is interesting to note that the service provided by WinPure is unable to detect any records with different spelling and pronunciation.

Group No	PersonID	Nation.	NAME	Sex	Status	BirthDate	PassNo
1	2021200	351	WALI MOHAMMAD MOHAMMAD ABDULLAH	1001	1002	01/07/1937	392904
	4903234	351	WALI MOHAMMAD ABDULLAH	1001	1002	01/07/1937	392904
	4489509	351	WALI MUHAMMAD MOHAMMAD	1001	1002	01/07/1937	392904
2	2008600	610	TACK O. RIOLLE HOVER	1001	1002	01/07/1916	2559861
	2008603	610	TACK RIOLLE HOVER	1001	1002	01/07/1916	2559861
3	1723400	610	JOHN ROUMELIOTIS	1001	1001	01/09/1931	2376821
	1723401	610	JACK ROUMELIOTIS	1001	1001	01/09/1931	2376821
4	1720100	112	AMAL MOHAMAD KALIL ABOSAOUAN	1002	1002	01/07/1949	27534
	1720102	112	Ms. AMAL MOHAMAD KALIL ABOSAOUAN	1002	1002	01/07/1949	27534
5	1521200	510	MR.DONALD DARGE ANGUS	1001	1002	06/07/1927	674205
	2829300	510	DONALD DARGE ANGUS	1001	1002	06/07/1927	674205
6	1522100	510	LESLIE RONALD FRANCIS	1001	1002	02/12/1930	148121
	2279348	510	LESLIE RONALD FRNCIS	1001	1002	02/12/1930	148121

Duplicate ID No.	WinPure	DoubleTake	Wizsame	DQGlobal
1		Detected		
2		Detected		Detected
3		Detected	Detected	Detected
4		Detected	Detected	Detected
5		Detected	Detected	
6		Detected	Detected	Detected
7		Detected	Detected	
8		Detected		
9				Detected
10			Detected	
No. of detected Duplicates	0	8	6	5
percentage of Detection	0%	80%	60%	50%

Figure 4. Different Spelling and Pronunciation Duplicated Records and Examination Results

2. Comparison based on misspellings.
   The duplicated records examined in this comparison and the examination result by running four services are illustrated in Figure 5. This comparison provides less percentage of the detected records than the previous comparison. It can be inferred that the misspelling cases have more variants in the records. The next comparisons are not shown as a figure due to the limitation of the page.

3. Comparison based on name abbreviation.
   The duplicated records are examined in this comparison and the examination results by running four services. In this comparison, the records with name abbreviation are handled more accurately by the services, except for WinPure service.

4. Comparison based on honorific prefixes.
   The duplicated records are examined in this comparison and the examination result by running four services. It is interesting to note that DQGlobal service underperforms in this experiment.

5. Comparison based on common nicknames.
   The duplicated records are examined in this comparison and the examination results by running four services. In this comparison, DoubleTake service is unable to detect.

Group No	PersonID	Nation.	NAME	Sex	Status	BirthDate	PassNo
1	250900	511	MARC RAYMOND SOULIER	1001	1002	21/08/1943	24177
	98776	511	MARC RAYMOND SOULIER	1001	1002	21/08/1943	24177
2	67200	111	MOHAMAD KHALED HUSSNI	1001	1002	14/07/1942	374518
	67203	111	Dr. MOHAMAD KHALED HUSSNI	1001	1002	14/07/1942	374518
3	144200	110	George Robert	1001	1002	01/07/1917	11 53040
	253282	110	George Bob	1001	1002	01/07/1917	11 53040
4	131500	610	Christine LIONEL WILLIAMS	1001	1002	21/01/1933	1034761
	725393	610	Chris Lionel Williams	1001	1002	21/01/1933	1034761
5	218300	112	SAMI NAJIB ABOU CHAHLA	1001	1002	04/05/1945	23606
	263838	112	SAMI NAJB ABOU CHAHLA	1001	1002	04/05/1945	23606

Duplicate Group No.	WinPure	DoubleTake	Wizsame	DQGlobal
1			Detected	
2			Detected	
3		Detected	Detected	Detected
4			Detected	
5		Detected	Detected	Detected
6		Detected	Detected	Detected
7			Detected	
8		Detected		Detected
No. of detected Duplicates	0	4	7	4
percentage of Detection	0%	40%	70%	40%

Figure 5. Misspellings Duplicated Records and Examination Results

6. Comparison based on splitted name.
   The duplicated records are examined in this comparison and the examination result by running four services. In this comparison, WinPure service underperforms again.

7. Comparison based on exact match.
   The duplicated records are examined in this comparison and the examination result by running four services. Exact match feature is important for some cases which need a specific handling, such as to investigate the internal mistake of data warehousing.

A complete comparison summary for seven duplicate types is illustrated in Figure 6. It can be inferred that WizSame service has the highest reliability in any comparison criteria amongst other services, although it has no peak performance in term of the number of detected records.

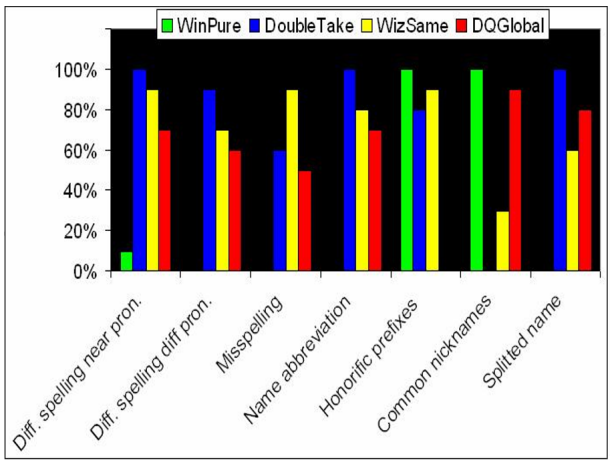


Figure 6. Seven Type Duplicates Examination Results

6. Quality of Data Cleaning Services

In addition to the evaluation for quality of data, there is a requirement to assess the data cleaning service based on the quality of service. There are several quality of service metrics are taken into account in this paper, such as performance, capability to process the number of

records, data heterogeneity, and price. for The performance of the service is broken down into two metrics: processing time and memory. Time is an important factor which is mostly taken into account in most algorithm comparisons which is calculated based on the processed records. 1000 records are considerably enough to be taken into account for this comparison. The time spent by each service on the processing of 1000 records is being calculated. The results of these record manipulation depends on the system environment. Therefore, the comparison between all of these services is conducted in the same environment. The environment related to experiments are kept consistent on all four services. Modifying the environment may affect the overall performance of these services.

The result of the processing time evaluation is presented in Figure 7 (a). It shows that WizSame and WinPure utilized less CPU time for processing 1000 records. Accordingly, DQGlobal took the maximum time for processing 1000 records. Figure 7 (b) presents the comparison between the memory utilization of examination services. In this evaluation, both DoubleTake and WizSame had an optimal performance, while DQGlobal and WinPure had more memory consumption for the processing of 1000 records.

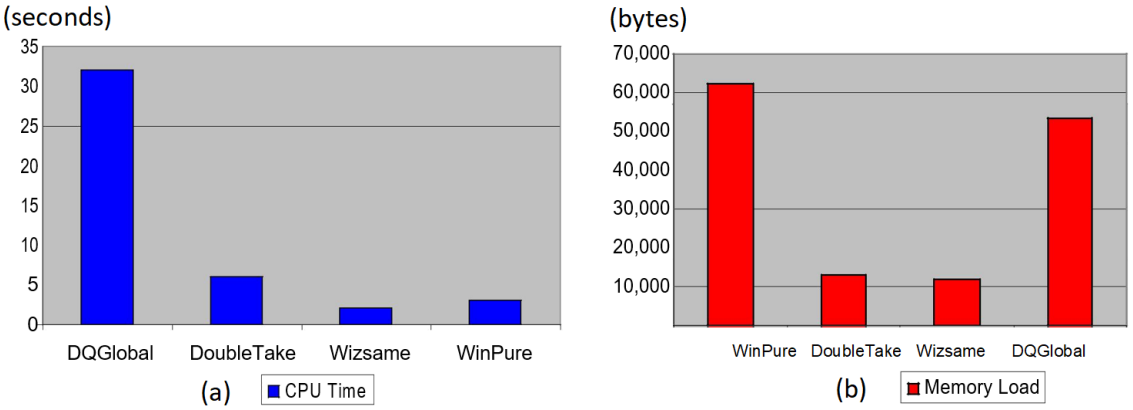


Figure 7. Time Spent and Memory Utilization on The Processing of 1000 Records

The capability of each service to process the records is an important metric for data cleaning service. The comparison describes how many records that each service can process the removal of duplication. WinPure was able to process 250,000 records at maximum. DoubleTake was able to process twenty million records at maximum. WizSame was able to process one million record at maximum. DQGlobal was able to process one million records at maximum.

Different service has different capability to process particular data format. It is considered as a data heterogeneity metric. WinPure is able to process Text File, MS Excel, MS Access, Dbase, MS SQL Server. DoubleTake is able to process MS Excel, MS Access, Dbase, Plain Text File, ODBC, FoxPro, MS SQL Server, DB2 and Oracle. WizSame is able to process dBase, MS SQL, MS Access and Oracle, Plain text file, Dbase, ODBC, OLE DB. DQGlobal is able to process MS Access, Paradox, MS Excel, DBF, Lotus, FoxPro, and Plain text file. This comparison shows that DoubleTake runs more data formats compared to other services. WizSame scores second for running more data formats in removing duplication.

Price is another quality of service metrics considered in this paper. The service that has high price is not feasible for particular users. The price for purchasing the license of the applications is in a wide range by the time this paper is written. WinPure costs $949.00, DoubleTake costs $5,900.00, WizSame costs $2,495.00 and DQGlobal costs $3,850.00. This comparison implies that DoubleTake has the highest price compared to the rest of the services. However, since we wrap all these applications as services, the cost is minimized by paying only for the executed services.

7. Conclusion

In data warehousing, data cleaning service plays an important roles in many domains. If the data is not clean and full of anomalies, the resultant data have a lot of issues, such as data integration and query errors. In order to get the best form of the extracted data, it is important to clean the data as an initial step. Data redundancy should be removed to maintain the data integrity. This research provides an overview about the quality of data to be used in data warehousing and to analyze, practice and experiment the concept of data quality by utilizing real local data. Hence, this research has two contributions. First, it surveyed of data quality in the environment of data warehouse and the data integrity analysis. Second, it compared the services that can remove the duplication of data through some real experiments. The experiments were conducted based on the performance measures so that it could be determined which service is more effective for the removal of data duplication. The comparison is considered as an aid for users to select the best services depending on their needs, especially in the scope of Saudi Arabia.

Acknowledgement

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia. The author, therefore, gratefully acknowledges the DSR technical and financial support. The author also thanks Mshari AlTuraifi for conducting the experiments in Saudi Arabia.

References

[1] B. Moustaid and M. Fakir, "Implementation of business intelligence for sales management," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 5, no. 1, pp. 22–34, 2016.
[2] I. Khliad, "Data warehouse design and implementation based on quality requirements," *International Journal of Advances in Engineering and Technology*, pp. 642–651, 2014.
[3] L. Robert, "Data quality in healthcare data warehouse enviornments," *34th Hawaii International Conference System on System Sciences*, pp. 9–1, 2001.
[4] G. Shankaranarayanan, "Towards implementing total data quality management in a data warehouse," *Journal of Information Technology Management*, vol. 16, no. 1, pp. 21–30, 2005.
[5] A. Amine, R. A. Daoud, and B. Bouikhalene, "Efficiency comparaison and evaluation between two etl extraction tools," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 3, no. 1, pp. 174–181, 2016.
[6] R. Archana, R. S. Hegadi, and T. Manjunath, "A big data security using data masking methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 2, pp. 449–456, 2017.
[7] H. Marcus, K. Mathias, and K. Bernard, "How to measure data quality; a metric approach," *Twenty Eighth International conference on Information System, Montreal*, pp. 1–15, 2007.
[8] H. Frederik, Z. Dennis, and L. Anders, "The cost of poor quality," *Journal of industrial Engneering and Management*, pp. 163–193, 2011.
[9] K. Rahul, "Data quality in data warehouse problems and solution," *Journal of Computer Engineering (ISOR-JCE) ISSN-2278-0661, Volume 16, Issue1*, pp. 18–24, 2014.
[10] B. Inmon, "Data warehousing 2.0 architecture for next generation of data warehousing," Tech. Rep., 2010.
[11] R. Kimball, M. Ross, J. Mundy, and W. Thornthwaite, *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence Remastered Collection*. John Wiley & Sons, 2015.
[12] United States Department of Interior CIO, "Data quality management guide," Tech. Rep., 2008.
[13] Q. Sun and Q. Xu, "Research on collaborative mechanism of data warehouse in sharing platform," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 1100–1108, 2014.
[14] LLyods, "Solvency ii-section 4-statistical quality standards," Tech. Rep., 2010.