Lean Database: An Interdisciplinary Perspective Combining Lean Thinking and Technology

Jamil Razmak[1], Samir Al-Janabi[2], Faten Kharbat[3], and Charles Bélanger[4]
[1]College of Business, Al Ain University, UAE
[2]Department of Computing and Software, McMaster University, Hamilton, Canada
[3]College of Engineering, Al Ain University, UAE
[4]Faculty of Management, Laurentian University, Canada

Abstract: *The continuous improvement approach is key to achieve a sustainable competitive advantage for organizations in their business processes. Nowadays, organizational business processes are seen through an automated function under the umbrella of organizational information systems. The huge amount of automated business processes produces data embedded with a part of messy data that could provide corrupt data. This study uses a lean thinking concept integrated with the data cleaning approach to reduce the waste of data according to business requirements and to enhance continuous improvement as part of a data defect reduction strategy. A new approach of improving and cleaning data waste is proposed by combining data cleaning algorithm and lean thinking concepts. After testing the quality and scalability of the algorithm, along with the evaluation of a corrupt dataset, the results showed improvement in the corrupt dataset reduction, leading to higher organizational performance in business processes. This integration can help researchers and technologists to fully understand and benefit from interdisciplinary capabilities while building bridges between different fields.*

Keywords: *Lean database, interdisciplinary, lean thinking, data quality, data cleaning.*

Received March 3, 2020; accepted July 14, 2020
https://doi.org/10.34028/iajit/18/1/4

1. Introduction

Lean thinking is a business approach that uses "unconventional methods" [37], aiming at delivering superior value for customers by removing "non-value-adding" activities [11]. Lean is the concept firstly adapted by Toyota production system by focusing on the operational excellence on the one hand and eliminating the waste on the other hand [44]. The main goal for such a system is to maximize outputs while using minimum inputs. Outputs are the products with minimum faults and maximum quality whereas inputs are the human effort, inventory space, or the investment amount. However, the lean philosophy can be integrated in any field if the processes are mapped, the goals are measured, and the resources are managed [51]. Lean management has been successfully adopted and implemented in different fields and disciplines such as healthcare [42], sustainable business [11], environment [3], and construction [20].

Lean thinking has been applied into software engineering and software development. Software engineers applied the concepts of lean management to software engineering [37] to produce what is called "Agile Manifesto" [9]. In fact, Janes and Succi [37] stressed the importance of communication and collaboration between the "lean" as a concept and the "agile" as an approach. Different perspectives were suggested to implement and adapt the lean thinking concept into the software development. As examples, [47] have tackled the "lean" concept from several business side principles (i.e., eliminate waste, build quality, create knowledge, defer commitment, deliver fast, and respect people). Whereas [33] addressed the coding issues (such as scripted builds, automated testing, continuous integration, less code, short iterations, and user participation) as a main source of "Lean" development, [37] developed Lean software development process that avoided three issues:

a) Skepticism.
b) The guru approach.
c) Agile extremists.

In the same vein, Agile Manifesto was focused on the IT functions and technology in adopting agile principles and lean thinking concepts to benefit many companies such as Facebook, Microsoft, IBM, Google, Adobe, Spotify, Netflix, etc., [9, 52]. For example, [34] utilized the lean concept to improve the information to support the overall information systems infrastructure. He developed five key principles of a lean approach for information management in which the traditional elements are included: value, value stream, flow, pull, and continuous improvement. He also identified the main sources of "waste" in information flow: failure demand, flow demand, flawed flow, and flow excess. Redeker *et al*. [51] highlighted lean communication flow as a result (do you mean "as a result of or for?

The meaning is different) for the lean information based on the classification of [35].

Nowadays, the emergence of digital information, communication and technologies generate more and more complex and heterogeneous data flowing from anywhere, anytime, and any device [7]. Many studies use practical and systematic approaches to apply the process improvement techniques, such as lean thinking that focuses on waste reduction through resource optimization [22, 23], and to sustain performance in organizations according to the available data used [7]; [21, 40]. The accumulated data in these organizations may include some corrupt data (wastes) that need to be cleaned. According to the saying "Garbage In, Garbage Out" (GIGO) in the computer science [45], these corrupt data will produce nonsense output or "garbage" in the organizational business processes that can negatively affect organizational performance.

However, there are no previous studies experimentally establishing the integration between lean thinking approach and technological algorithms to improve the quality of these data. Nevertheless, few researchers have tackled the use of lean thinking concept from the data side instead of the development side.

The core idea of lean thinking is doing more and more with less and less input to produce a value to the customer as an approach to foster innovation in software products [52]. Therefore, the main objective of this study is to integrate lean concepts into the database capabilities in order to enhance the value of data by eliminating the related wastes, along with understanding how the lean approach can be applied in software improvements. This study presents the concept of "*Lean Database*" based on the general lean thinking approach, which enhances the quality requirements in terms of data storage, access and retrieval. As we mentioned above, the critical integration point of technology and the lean thinking approach is a "*value*" from an interdisciplinary point of view, which is created by the provider to reach the customer and to guide technical decisions in the context of data cleaning that can be used as a means to gain business advantage in organizations.

In this setting, the present study contributes to the existing field of the interdisciplinary literature through establishing a linkage between lean thinking and data cleaning techniques. A conceptual framework was developed from the literature based on the integration between both of them to reduce the nonsense output (waste). With a similar purpose, relevant data cleaning algorithms were developed and tested, by considering lean thinking concepts to reduce the courted data to improve organizational business processes. As a result, this integration offers researchers and technologists an in-depth thinking out of the traditional box of creating a bridge between fields from both theoretical and practical sides. Furthermore, this integration paves the way for future studies exploring how lean thinking approach and technology capabilities can be used together to achieve higher performance in organizations within different settings.

Section 2 will explore the literature review in terms of data quality, building a bridge between lean thinking and data cleaning to reduce the waste sources in databases. Section 3 employs the technology capabilities to clean one or more of these wastes by applying an experimental algorithm as a measurement tool to evaluate the cleaning degree of these wastes. The final section will present some discussion and conclusion remarks.

2. Literature Review

2.1. Data Quality

Two quality dimensions have been shown in the literature: *data quality*, which refers to "dimensions, models, and techniques strictly related to structured data" and *information quality* when these models and techniques are related to "wider spectrum of information types" [4]. High-quality data refers to data that is "fit for use by data consumers" in terms of its usefulness and usability [56]. As a result, the value of data depends on its quality, which is considered as an intangible asset for modern organizations that reflects the reliability of their data [53, 54] as well as their trust worthiness [28, 36]. On the other hand, poor data quality will cost organizations and affect their profitability, such as problems in data quality may cause up to 40-60% of the revenue loss because of unclean data [1]. In addition, the quality of data affects the decisional and operational processes [4] and inter-organizational cooperation requirements in any organization [6]. Hence, statistical approaches have been linked to quality to reduce the defects and wastes in data and to provide a useful potential for its usage [28, 53, 54]. Some of these approaches, for example, have focused on detecting a "minimal number of updates to the data to correct the underlying inconsistencies [36].

The rule of data quality has been expanded to include numerous categories, including business entity rule, business attribute rule, data dependency rule, and data validity rule [48]. Poor data quality is a result of "dirty" data [16, 43]. The dirty data expression was coined in the 1998 literature to indicate poor data quality in large databases due to a range of problems from different sources [32, 43] classified three types of dirty data: missing data, noisy data, and inconsistent data. As well, a fourth category exists particularly when different data is integrated from different sources [1, 49], such associal media [36], multiple databases from the same [46] or different sources [15], inconsistent entries [46], changing data over time [15], real-time sensors (such as OCR or scanner readers)

[4],oras a result of the poor data entry of the web-forms [15].

2.2. Source of Waste in Data

Several issues affect the quality of data and make it unclean (dirty). The following are some examples increasing the waste in data according to the literature [16]. These waste sources may exist in information to form the next lean phase related to unclean database management, as mapped in Figure 1.

- *Lack of Integrity Constraints*: Integrity constraints means that all instances of a database schema should follow the same procedures at all-time [4]. Lack of input standardization and details will result in poor identified constraints, which will affect the quality of the database and its integrity. For example, in the case of a company, person or city name, the lack of standard entry through specific constraints will lead to incorrect, weak, or misleading information. This problem will be connected to waste in queries and reports as main services in the database management.
- *Out-of-Date Values*: One of the most important features for the data is timeliness [4]. When integrating data from multiple data sources, different data sources might enter data about the same entity at different points in time. Some data entries will lead to obsolete values. Changing the address of an employee over time is a good example of out-of-date values. Al-janabi and Janicki [1] stated that 2% of the data might become outdated in one month for customers and suppliers. This problem will affect the services for the database management in terms of queries, reports, and/or views.
- *Heterogeneous Schemas*: In multiple data sources, different schemas are used to represent entities. In such heterogeneous multi-database systems, values and entities may be inconsistent or suffer from "a loss of a clear identity" [4]. For example, one schema might contain a product dimension while another does not. When integrating those schemas, the missing attribute values could be padded with NULL values. This problem will be connected to waste in database design, which affects data models.
- *Different Data Entry Rules/ Format*: This issue occurs when integrating data from multiple data sources; hence, multiple sources might have different data entry requirements, rules or formats [34]. For instance, a person's full name might be entered starting with the first name in one data source, while starting with the last name in another one. This problem will be connected to both waste in data design and services since it will affect data models and reports respectively.
- *Duplicate Data:* duplicate entities will cause a variety of problems affecting the database services such as:
  - *Contradictory data*: Refers to multiple representations that may yield different information. For example, the telephone number of a customer may have the area code that reveals the territory of residence. If the city does not reflect the same area, then contradicting data occurs. In this case, the dependency of an attribute value is determined by another attribute [34].
  - *Overlapping data*: Occurs when integrating data from multiple data sources, the multiple representations may cover different or redundant data from different properties [46]. For example, a customer's telephone number may have the area code that reveals the geographical area. In reality, there is no need to add the area entity; however, if it exists, there is overlapping data.
  - *Semantic structure*: This may occur in the domains that have rich semantic structures (i.e., cultural systems, religious data, etc.,) in which multiple thesauri are used with kinds of relations between them [8].
- *Entry Errors*: This issue occurs when considering different modes of data entry (manual or automatic); errors may occur while typing, scanning, or Optical Character Recognition (OCR). Although the aim is to automate processes, software could be confused between the characters with similar outlook such as "1" and "I". This can also include the automated entry for a GPS, which can be inaccurate especially inside indoor environment. This problem will affect the services for the database management in terms of queries, reports, and/or views.

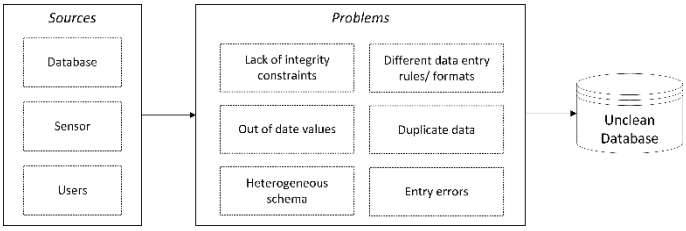

Figure 1. The leading causes of unclean database.

2.3. Lean Database

Several causes of waste have been shown in the literature that may affect the quality of database in any information system [34, 51] recently classified these waste under six categories regarding information systems. The first is the stock, which refers to a waste that could occur due to excessive information available in the database. The second is the motion; this waste requires manual intervention due to the lack of integration between systems. The third is the waiting,

waiting for intervention when the information does not flow in the information system. The fourth is the processes; this waste results from the delay in changing information, particularly when we have inaccurate information in the system. The fifth is the defect, which refers to flawed/inaccurate information that occurs in different information formats through lack of compatible standards in the information systems. The final one is the overproduction, which occurs through the excessive number of systems and multiple data sources. However, the role of lean thinking is reducing or eliminating the wastes not just in the database, but also in all kinds of information that cause these wastes [34]. Generally, lean thinking under its transformation concept employs several techniques "in order to remove waste and deliver improvements in specific areas" [34].

Therefore, lean thinking as a concept has been adopted in several sectors; it can be applied to data and information management to add value through a clean database. It "provides a way to do more and more with less and less human effort, less equipment, less time, and less space" [60], In other words, it provides a way to reduce the "loss" which points at managing the processes to eliminate any process that "does not lead to the ultimate goal" [51]. It is a fact that technology techniques through data cleaning algorithms also reduce the waste and increase the value in several industries. Databases are the heart of any technological system in the modern businesses [51]. However, providing customers with a messy data (waste data) in these databases will not meet the required expectations and the value from using these systems. There are three waste dimensions (services, design and hybrid) that cause unclean database based on lean information management [34, 51, 55]. Table 1 summarizes the waste dimensions for a database management based on the lean information proposed in the literature. These dimensions affect many parts in the database, which could be a result of dirty data and unclean database as mentioned above.

Table 1. Waste dimensions for a database management.

Waste dimension	Affected part	Waste example
Services	Queries Reports Forms	Lack of Integrity Constraints Out-of-Date Values Duplicate Data Entry Errors
Design	Data models	Heterogeneous Schemas
Hybrid dimension	Different parts	Different Data Entry Rules/ Format

Cleaning the dirty data and applying the lean concept will lead to a clean database. Therefore, the present study tried to achieve the quality to fit more than one application in terms of managing waste time and cost resulting from dirty data. It suggested a new concept called a "lean database management" as an approach to improve the organizational database by reducing waste and increasing value utilizing a lean standardized way guided by two of the main lean principles and empowered by the continuous improvement pillar.

2.4. Data Cleaning Techniques/ Lean Database

Data cleaning has been known as a technique to improve the quality of data [4, 48, 55], through detecting and removing errors, wastes and inconsistencies from the database. It has several terms with the same meaning in the literature such as data cleansing [32] or data scrubbing [53, 54]. On the other hand, several data repairing approaches exist such as tuple deletion [14] and value modification [10, 59]. Another approach is also based on the modification by using Conditional Functional Dependencies (CFDs) [50]. This approach employs users' interaction in which a user manually cleans a small set of unclean data only at the start of the algorithm. Constraints underlying those repairs would be inferred.

Another approach is related to data currency; it has been studied in various settings aiming to identify how promptly data are updated [5, 25]. In this domain for example, reliable timestamps may not be available in data [61], thus, data could be outdated or unclean. Another factor that may affect the quality of data is duplicate tuples. Data deduplication is the process grouping the tuples from one relation or more than one relation referring to the same real-world entity. Several approaches have been proposed for data deduplication such as probabilistic approach [26, 39], distance-based [29], rule-based [31], supervised learning [17], unsupervised learning [57], and active learning [18].

As a result, these technological approaches have been transforming its processes from/through following the basic lean concepts to complementing them with data cleaning principles to produce a value. As Figure 2 shows, lean thinking supports IT through reducing the waste (e.g., duplication), which can be done by utilizing the lean concepts through automated techniques that provide clean data. The transformation was driven by the need of IT companies to remain innovators in the highly competitive industry. As a result, this provides value for the data in the decision making process, which could benefit the customer in the end. All in all, this mixture makes technology act as the main enabler to improve performance and productivity, reduce costs and create value [12].

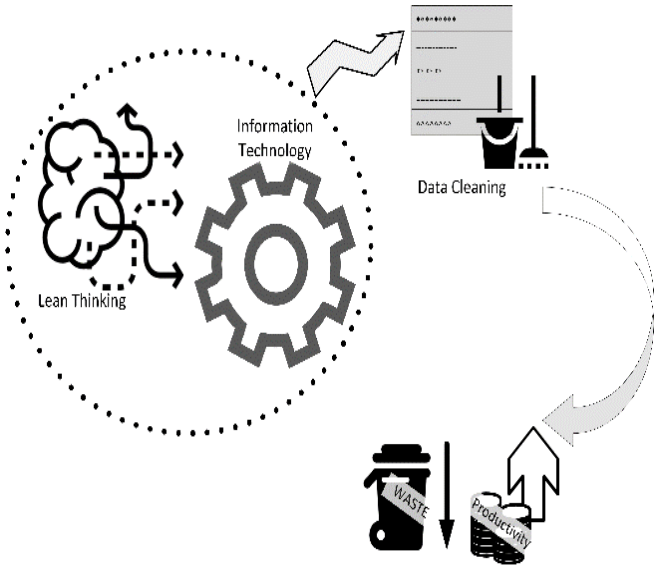

Figure 2. The integration between lean thinking and information technology capabilities to reduce waste.

3. Experimental Study: Integration of Data Cleaning Algorithm into Lean Database

This section presents a technique that may reduce the waste in the information system database in order to increase its efficiency and productivity and to enhance the value through a lean database concept. To determine the effectiveness of this technique in terms of the quality and performance, we conducted experiments to clean dirty data. The data-cleaning algorithm of our approach is based on the algorithm Corroborating Quality of Data Through Density Information (CURET) of [1] that cleans the dataset based on utilizing the density of the data. Algorithm 1 describes the cleaning steps. It starts by clustering the unclean dataset into sets of clusters, where each cluster contains related tuples that represent the same real-world entity. The representative tuple is found based on most current tuples and the weight of representatives of the attribute values that have no currency information in the clusters by utilizing density of data. The tuples are merged on the basis of the most trustworthy representatives and the most current attribute values. After that, the outliers that do not belong to any cluster are inserted into the cleaned data.

Algorithm 1: Data Cleaning Algorithm

Input: Unclean data D'
Output: Clean data D''
1: Cluster D' into sets of candidates
2:for each cluster do
3:Infer the most current tuples
4:Calculate the weight of the representatives
5:Calculate the trust scores of the representatives
6:Coalesce the tuples
7: insert outliers
8: return D''

3.1. Datasets

Using synthetic dataset, we evaluate our proposed techniques.

- **Synthetic Dataset**. We used the tool in [2] to generate a Workers dataset with 2000 records. We eliminated the duplicates using manual check and algorithm. For our experiments, we generated a dataset with twelve attributes: last name (LastName), first name (FirstName), company (Company), city (City), department (Dept), bank account number (BankAccount), social security number (SSN), phone (Phone), car (Car), salary (Salary), rank (Rank), and Date Of Birth (DOB). The dataset contains NULL values without timestamps available in it. It has been used to test the quality and performance of the techniques.
- **Hardware Setup**. For the experiments, we used Windows machine with 2.2GHz Intel Core 2 Duo CPU and 8GB of RAM.
- **Parameters**. Let dup_{rate} be the rate of duplicates. Let dup_{num} be the number of duplicates. Let N be the size of data D', including the duplicates. Let *error* be the rate of error (i.e. the fraction of the tuples in the duplicated tuples that are modified to have errors).

4. Experimental Results

Table 2 illustrates an instance of the corrupt dataset that contains data about workers. Tuples t_2-t_5 are grouped in one cluster as they refer to the same real-world entity. However, in order to merge duplicate tuples into one tuple, we need to find the correct attribute values in the cluster because the attribute values in these tuples are not all the same. For example, we need to decide what is the most accurate value of the last name. That is, whether it is “Jose”, “Jowse”, or “Joes”. Also, there is more than one value in the company, social security number, car, salary, rank, and date of birth. So we have more than data quality issue in this dataset including data deduplication of a dataset that contains outdated and inaccurate values in the duplicate tuples. In terms of determining which attribute values are the most current, the algorithm could decide that “2237” and “c” are the most current values of the salary and rank, respectively since salary and rank are generally increasing over time. In addition, assuming that tuples t_2 and t_5 are core tuples and t_3 and t_4 are border tuples, the algorithm could decide that “Jose” is the most accurate value of the last name by utilizing the density information.

Table 2. Instance of corrupted Workers table.

tid	Last Name	First Name	Company	City	Dept	BankAccount	SSN	Phone	Car	Salary	Rank	DOB
t_1	Shane	Gail	Vulputate LLC	Florence	dept4	3-443-5284895	514-731-694	+1(399)-678-7764	Dodge	807	e	25-10-91
t_2	Jose	Randall	Congue A Industries	Freeport	dept4	6-382-4853716	590-663-625	+1(646)-739-1342	Audi	2237	c	20-08-67
t_3	Jowse	Randall	Congue A Industries	Freeport	dept4	6-382-4853716	517-313-822	+1(646)-739-1342	Ford	1218	c	20-08-67
t_4	Joes	Randall	Congue A Industries	Freeport	dept4	6-382-4853716	590-663-625	+1(646)-739-1342	Ford	807	b	25-02-93
t_5	Jose	Randall	Industries A Congue	Freeport	dept4	6-382-4853716	590-663-625	+1(646)-739-1342	Audi	2237	c	20-08-67
t_6	Rami	Kim	Nibh Dolor LLP	Charleston	dept4	2-771-1477373	572-233-861	+1(631)-486-1074	Kia	1876	b	17-12-50
t_7	Lucas	Alexa	Id Associates	Ruby Valley	dept2	2-363-2855057	541-645-496	+1(523)-473-7521	Lotus	807	e	14-05-69
t_8	Kennedy	Brittany	Risus In Foundation	Mc Clelland	dept5	5-842-7912313	569-263-793	+1(462)-522-2794	Mazda	2331	a	08-02-66

Similarly, the algorithms could decide about the most accurate values of other attribute values. Finally, the algorithm could identify that ‘Jose”, “Randall”, ‘Congue A Industries”, ‘Freeport”, ‘dept4”, ‘6-382-4853716”, ‘590-663625”, ‘+1(646)-739-1342”, ‘Audi”, ‘2237”, ‘c”, and “20-08-67” are the most accurate and current values among the values of the tuples t_2, t_3, t_4, and t_5. The cleaned dataset is illustrated in Table 3.

Table 3. Instance of cleaned workers table.

tid	Last Name	First Name	Company	City	Dept	Bank Account	SSN	Phone	Car	Salary	Rank	DOB
t1	Shane	Gail	Vulputate LLC	Florence	dept4	3-443-5284895	514-731-694	+1(399)-678-7764	Dodge	807	e	25-10-91
t2	Jose	Randall	Congue A Industries	Freeport	dept4	6-382-4853716	590-663-625	+1(646)-739-1342	Audi	2237	c	20-08-67
t3	Rami	Kim	Nibh Dolor LLP	Charleston	dept4	2-771-1477373	572-233-861	+1(631)-486-1074	Kia	1876	b	17-12-80
t4	Lucas	Alexa	Id Associates	Ruby Valley	dept2	2-363-2855057	541-645-496	+1(523)-473-7521	Lotus	807	e	14-05-69
t5	Kennedy	Brittany	Risus In Foundation	Mc Clelland	dept5	5-842-7912313	569-263-793	+1(462)-522-2794	Mazda	2331	a	08-02-93

4.1. Quality Measuring

We adopted the F-measure that is commonly used in information retrieval, and it is used to evaluate the quality of the techniques. F-measure is defined as:
F-measure = 2. (Precision.Recall)/(Precision+Recall)
where *precision* is the ratio of true positives correctly retrieved to all the duplicates found, and recall is the ratio of true positives correctly retrieved to all the true duplicates in the ground truth.

Figure 3 illustrates the F-measure values for the *Workers* dataset at various values of the parameter *error*. We set dup_{rate}= 25%, dup_{num}= 4, and N = 3500. The *error* rates are 3%, 6%, 9%, 12%, and 15%, and the results for F-measure values are 0.89, 0.88, 0.86, 0.84, and 0.81, respectively. We observe that F-measure value decreases moderately as error increases. For *error* = 15%, the F-measure is 0.81. This indicates that cleaning quality is still good even with higher error rates.

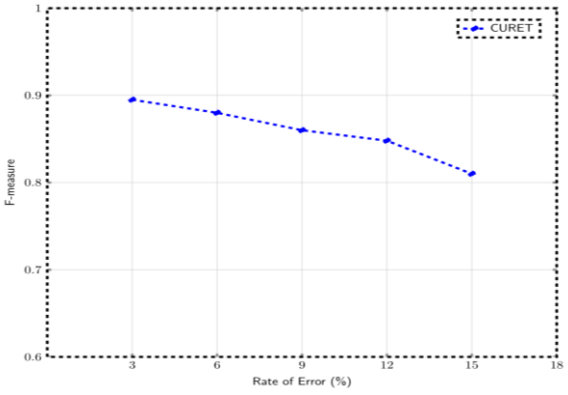

Figure 3. F-measures: Workers dataset.

4.2. Scalability Measuring

Figure 4 depicts the performance at various values of the parameter dup_{rate}. We set dup_{num}= 4, and N = 3500, and *error* =15%. As duplicates rate increases from 10% to 70%, the running time increases moderately.

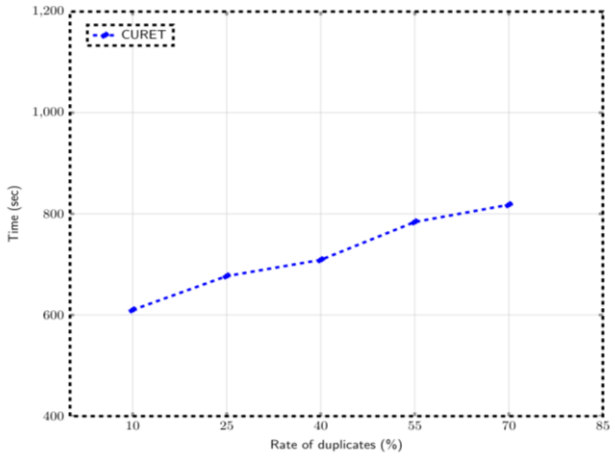

Figure 4. Scalability with the rate of duplications.

Figure 5 depicts the performance at various values of the parameter N. We set dup_{rate}= 25%, dup_{num}=4, and *error*=15%. The running time increases more rapidly with the increase of N from 1000 to 5000, as expected, due to increasing number of the duplicate tuples. As for run time complexity in terms of big O notation, the density-based clustering algorithm Density Based Spatial Clustering of Applications with Noise (DBSCAN) [24] is used in the clustering of step 1. DBSCAN is efficient as well as robust for outliers and arbitrary shaped clusters [38]. The worst case is when it

has to visit all the data records to decide whether they are core data points or not. In this case, the run time complexity is $O(n^2)$, where n is the total number of database records that need to be cleaned. The result of Figure 5 indicates practically this complexity.

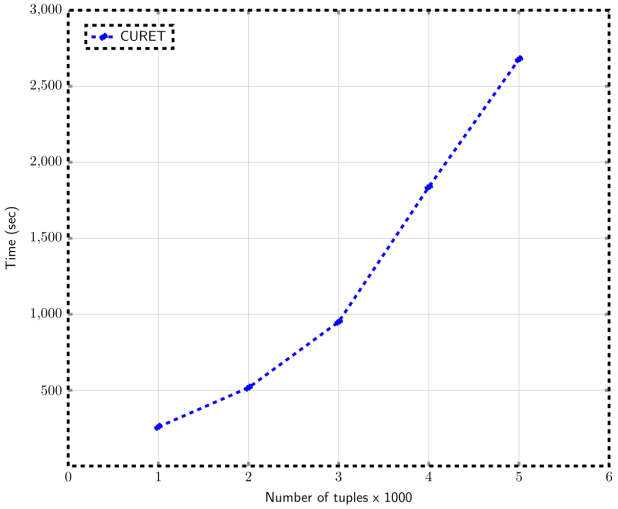

Figure 5. Scalability with the number of tuples.

5. Discussion

The integration of lean thinking in data cleaning and quality techniques requires advanced understanding of interdisciplinary studies [27] to build a bridge between fields and to solve the messy data problem as a part that hinders continuous improvement. In brief, data cleaning is an essential process to minimize inconsistency (e.g., duplications, errors, false outcomes), reduce waste and improve quality, which is the core of lean thinking that we need in technology. Nevertheless, we have to accept that quality aspects may differ from one application to another [49]. Therefore, several algorithms and methods were introduced in the literature to clean data, for example, "one-shot fix-ups and reconciliations" [43], which aims to fix the problems in a database when a *problem* occurs or if a fall-down was already there. However, some problems are temporarily solved -without solving the roots of the problem [15]. The second example is the automated rule-based taxonomy to select dirty data [34]; this taxonomy has led to a new classification of dirty data. In this direction, this approach is working on the idea of forming a relationship between the dimensions of data quality and types of dirty data according to the rule-based taxonomy of dirty data technique. The final example proposed by Cheng *et al.* [13] has several combinations for data cleaning strategies, which were built on four indicators: data volume, completeness, timeliness, and correctness. To that purpose, he put forth an equation for computing data quality between the indicators, assuming that each indicator has a specific weight.

However, these examples and others discussed data quality techniques in databases by focusing on integrity constraints, duplicate rows, missing values, and inconsistencies [13, 19]. While these techniques can be categorized from diverse viewpoints, "detection of similar duplicate records of structured data is an important issue in data cleaning research" [30]. Duplicate records will increase the storage of large amounts of data, hinder the traffic of data that is coming from several devices [41] and slow the aggregate query processing [58]. However, avoiding these issues will provide lean data in the database system [41], which is considered the core objective of the present study. In terms of data de-duplication, the present technique differentiates itself by utilizing the density of data to find the representative records between different data values in data sets that contain duplicated records with outdated and inaccurate data. It handles a variety of errors such as different data entry standards and transcriptions errors and uses an unsupervised learning algorithm that does not require a training set with pre-labeled duplicates. In addition, many data cleaning techniques need human interaction [58], while our technique handles cleaning issues efficiently without human interaction.

From an organizational perspective, the study implies that having clean data sources is essential for any organization to execute their strategic objectives successfully and to monitor the performance effectively. All the applications and its business processes (e.g., budgeting and forecasting, and modelling) that are used by different users such as executives, managers, supervisors, and operators, depend completely on the availability of clean data sources that accessible at the operational level. Clean data will improve these business processes, accelerate the decision making process, allocate organizational resources and improve the reporting process.

From a technical perspective, the study also implies that the suggested cleaning algorithm is essential in data warehouse as a repository of data that is intended to support decision-making. For example, ETL process (extraction, transformation, and load) is one of the most important steps in data warehousing. It starts by extracting or reading data from a variety of data sources such as Online Transaction Processing system (OLTP) databases and spreadsheets. In the transformation step, the data is transformed from its original format into another format that it is ready to be loaded into the data warehouse. Therefore, the algorithm will be valuable in the stage of data transformation so that the data warehouse can be loaded with cleaned data. Then, users such as business analysts can work with this data to perform tasks such as data mining and produce clean reports.

Finally, clean data algorithms are important for researchers, engineers, and data scientists alike because their work requires data with high quality to build on and to evaluate their algorithms and techniques. Errors in the data such as typographical errors, duplicates,

integrity constraints violations, and outdated data, would affect the effectiveness of the different algorithms they are using.

6. Conclusions

This study was an attempt to contribute to database management system by maintaining the best and most sustainable efficiency with the highest customer satisfaction at a minimum cost and time. The two main principles guiding the lean thinking are to act as a general concept to accelerate the flow of data value and information to the customer through reducing human interventions and allow the technology to accomplish the routine tasks in the database management system. However, database management system may look at data as input from different angles, which requires more attention from scholars and developers through working on cleaning processes of this data from theoretical and technical backgrounds. The cleaning process aims to reduce as much as possible all types of waste that cover delay in the processing of data through its duplication or changing information (out-of-date values), its loaded inaccurate information (entry errors), its lack of consistency and others.

Therefore, the contribution of the present paper suggested a new approach to clean partially these wastes through focusing on the lean concept in order to arrive at a more accurate database through decreasing human intervention and improving business process in organizations. Finally, our experiments led to an approach that cleaned the data through the duplication, which enhanced the value of information in the database. Finally, our experiments indicate that the waste represented by the unclean data can be reduced and thus turning the data in to more valuable asset to the decision maker that, for example, may take a managerial decision based on this cleaned data.

References

[1] Al-janabi S. and Janicki R., "Corroborating Quality of Data Through Density Information," *in Proceedings of SAI Intelligent Systems Conference*, London, pp. 1128-1146, 2016.

[2] Al-janabi S., Hamid A., and Janicki R., "DatumPIPE: Data Generator and Corrupter for Multiple Data Quality Aspects," *in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Sydney, pp. 589-592, 2017.

[3] Amerson A. and Parsons E., "Evaluating the Sustainability of The Gray-Whale-Watching Industry Along The Pacific Coast of North America," *Journal of Sustainable Tourism*, vol. 26, no. 8, pp. 1362-1380, 2018.

[4] Batini C. and Scannapieco M., *Data and Information Quality: Dimensions, Principles and Techniques*, Springer, 2018.

[5] Batini C. and Scannapieca M., *Data Quality Dimensions. Data Quality: Concepts, Methodologies and Techniques*, Springer, 2006.

[6] Batini C., Cappiello C., Francalanci C., and Maurino A., "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 16, 2009.

[7] Belhadi A., Kamble S., Zkik K., Cherrafi A., and Touriki E., "The integrated effect of Big Data Analytics, Lean Six Sigma and Green Manufacturing on the Environmental Performance of Manufacturing Companies: The Case of North Africa," *Journal of Cleaner Production*, vol. 252, 2020.

[8] Bellatreche L., Valduriez P., and Morzy T., "Advances in Databases and Information Systems," *Information Systems Frontiers*, vol. 20 no. 1, pp. 1-6, 2018.

[9] Birkinshaw J., "What to Expect from Agile," *MIT Sloan Management Review*, vol. 59, no. 2, pp. 39-42, 2018.

[10] Bohannon P., Fan W., Flaster M., and Rastogi R., "A Cost-Based Model And Effective Heuristic for Repairing Constraints By Value Modification," *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimor, pp. 143-154, 2005.

[11] Caldera S., Desha C., and Dawes L., "Exploring The Role of Lean Thinking in Sustainable Business Practice: A Systematic Literature Review," *Journal of Cleaner Production*, vol. 167, pp. 1546-1565, 2017.

[12] Cawley O., Wang X., and Richardson I., "Lean Software Development-What Exactly are We Talking About?," *in Proceedings of International Conference on Lean Enterprise Software and Systems*, Galway, pp. 16-31, 2013.

[13] Cheng H., Feng D., Shi X., and Chen C., "Data Quality Analysis and Cleaning Strategy for Wireless Sensor Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no.1 pp. 1-11, 2018.

[14] Chomicki J. and Marcinkowski J., "Minimal-Change Integrity Maintenance Using Tuple Deletions," *Information and Computation*, vol. 197, no. 1-2, pp. 90-121, 2005.

[15] Christen P., *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer Science and Business Media, 2012.

[16] Chu X., Ilyas I., Krishnan S., and Wang J., "Data cleaning: Overview and Emerging Challenges," *in Proceedings of the International Conference on Management of Data*, San Francisco, pp. 2201-2206, 2016.

[17] Cohen W. and Richman J., "Learning to Match

and Cluster Large High-Dimensional Data Sets for Data Integration," *in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Alberta, pp. 475-480, 2002.

[18] Cohn D., Atlas L., and Ladner R., "Improving Generalization with Active Learning," *Machine Learning*, vol. 15, no. 2, pp. 201-221, 1994.

[19] Corrales D., Ledezma A., and Corrales J., "From Theory to Practice: A Data Quality Framework for Classification Tasks," *Symmetry*, vol. 10, no. 7, pp. 248, 2018.

[20] Daniel E., Pasquire C., Dickens G., and Ballard G., "The Relationship between the Last Planner® System and Collaborative Planning Practice in UK Construction," *Engineering, Construction and Architectural Management*, vol. 24, no. 3, pp. 407-425, 2017.

[21] De Freitas J., Costa H., and Ferraz F., "Impacts of Lean Six Sigma over Organizational Sustainability: A Survey Study," *Journal of Cleaner Production*, vol. 156, pp. 262-275, 2017.

[22] De D., Chowdhury S., Dey P., and Ghosh S., "Impact of Lean and Sustainability Oriented Innovation on Sustainability Performance of Small and Medium Sized Enterprises: A Data Envelopment Analysis-Based Framework," *International Journal of Production Economics*, vol. 219, pp. 416-430, 2020.

[23] Dey P., Malesios C., De D., Chowdhury S., and Abdelaziz F., "The Impact of Lean Management Practices and Sustainably-Oriented Innovation on Sustainability Performance of Small and Medium-Sized Enterprises: Empirical Evidence from the UK," *British Journal of Management*, vol. 31, no. 1, pp. 141-161, 2020.

[24] Ester M., Kriegel H., Sander J., and Xu X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, pp. 226-231, 1996.

[25] Fan W., Geerts F., and Wijsen J., "Determining the Currency of Data," *ACM Transactions on Database Systems*, vol. 37, no. 4, pp. 1-46, 2012.

[26] Fellegi I. and Sunter A., "A Theory for Record Linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183-1210, 1969.

[27] Frodeman R., Klein J., and Pacheco R., *the Oxford Handbook of Interdisciplinarity*, Oxford University Press, 2017.

[28] Ghouzali S. and Larabi S., "Face Identification Based Bio-Inspired Algorithms," *The International Arab Journal of Information Technology*, vol. 17, no.1, pp. 118-127, 2020.

[29] Guha S., Koudas N., Marathe A., and Srivastava D., "Merging the Results of Approximate Match Operations," *in Proceedings of the 3th International Conference on very Large Data Bases-Volume 30*, Toronto, pp. 636-647, 2004.

[30] Guo A., Liu X., and Sun T., "Research on Key Problems of Data Quality in Large Industrial Data Environment," *in Proceedings of the 3rd International Conference on Robotics, Control and Automation*, Chengdu, pp. 245-248, 2018.

[31] Hernández M. and Stolfo S., "The Merge/Purge Problem for Large Databases," *ACM Sigmod Record*, vol. 24, no. 2, pp. 127-138, 1995.

[32] Hernández M. and Stolfo S., "Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 9-37, 1998.

[33] Hibbs C., Jewett S., and Sullivan M., *The Art of Lean Software Development: A Practical and Incremental Approach*, O'Reilly Media, 2009.

[34] Hicks B., "Lean Information Management: Understanding and Eliminating Waste," *International Journal of Information Management*, vol. 27, no. 4, pp. 233-249, 2007.

[35] Hölttä V., Mahlamäki K., Eisto T., and Ström M., "Lean Information Management Model for Engineering Changes," *World Academy of Science, Engineering and Technology*, vol. 42, no. 2010, pp. 1459-1466, 2010.

[36] Huang Y. and Chiang F., "Towards a Unified Framework for Data Cleaning and Data Privacy," *in Proceedings of International Conference on Web Information Systems Engineering*, Miami, pp. 359-365, 2015.

[37] Janes A. and Succi G., *Lean Software Development in Action*, Springer, 2014.

[38] Januzaj E., Kriegel H., and Pfeifle M., "Towards Effective and Efficient Distributed Clustering," *in Proceedings of Workshop on Clustering Large Data Sets*, Melbourne, 2003.

[39] Jaro M., "Advances in Record-Linkage Methodology As Applied to Matching The 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414-420, 1989.

[40] Kamble S., Gunasekaran A., and Gawankar S., "Achieving Sustainable Performance in A Data-Driven Agriculture Supply Chain: A Review for Research and Applications," *International Journal of Production Economics*, vol. 219, pp. 179-194, 2020.

[41] Küfner T., Uhlemann T., and Ziegler B., "Lean Data in Manufacturing Systems: Using Artificial Intelligence for Decentralized Data Reduction and Information Extraction," *Procedia CIRP*, vol. 72, pp. 219-224, 2018.

[42] Lee J., McFadden K., and Gowen C., "An Exploratory Analysis for Lean and Six Sigma Implementation in Hospitals: Together is Better?," *Health Care Management Review*, vol.

43, no. 3, pp. 182-192, 2018.

[43] Lee Y., Pipino L., and Wang R., and Funk J., *Journey to Data Quality*, The MIT Press, 2009.

[44] Liker J. and Morgan J., "The Toyota Way in Services: The Case of Lean Product Development," *Academy of Management Perspectives*, vol. 20, no. 2, pp. 5-20, 2006.

[45] Majiwala H., Parmar D., and Gandhi P., "Leeway of Lean Concept to Optimize Big Data in Manufacturing Industry: An Exploratory Review," *in Proceedings of Data Science and Big Data Analytics*, Singapore, pp. 189-199, 2019.

[46] Naumann F. and Herschel M., "An Introduction to Duplicate Detection," *Synthesis Lectures on Data Management*, vol. 2, no. 1, pp. 1-87, 2010.

[47] Poppendieck M. and Poppendieck T., *Implementing Lean Software Development: From Concept to Cash*, Pearson Education, 2007.

[48] Rahm E. and Do H., "Data Cleaning: Problems and Current Approaches," *IEEE Data Engineering Bull.*, vol. 23, no. 4, pp. 3-13, 2000.

[49] Ramadan B., "Indexing Techniques for Real-Time Entity Resolution," PhD Thesis, the Australian National University, 2016.

[50] Rammelaere J. and Geerts F., "Explaining Repaired Data with CFDs," *in Proceedings of the VLDB Endowment*, Los Angeles, pp. 1387-1399, 2018.

[51] Redeker G., Kessler G., and Kipper L., "Lean Information for Lean Communication: Analysis of Concepts, Tools, References, and Terms," *International Journal of Information Management*, vol. 47, pp. 31-43, 2019.

[52] Rodríguez P., Mäntylä M., Oivo M., Lwakatare L., Seppänen P., and Kuvaja P., "Advances in Using Agile And Lean Processes for Software Development," *Advances in Computers*, vol. 113, pp. 135-224, 2019.

[53] Salem R., "A Manifold Learning Framework for Reducing High-Dimensional Big Text Data," *in Proceedings of 12th International Conference on Computer Engineering and Systems*, Cairo, pp. 347-352, 2017.

[54] Salem R. and Abdo A., "Fixing Rules for Data Cleaning Based on Conditional Functional Dependency," *Future Computing and Informatics Journal*, vol. 1, no. 1-2, pp. 10-26, 2016.

[55] Salem M., Bouazizi E., Duvallet D., and Bouaziz R., "(m, k)-Firm Constraints and Derived Data Management for the Qos Enhancement in Distributed Real-Time DBMS," *The International Arab Journal of Information Technology*, vol. 16, no. 3, pp. 424-434, 2019.

[56] Strong D., Lee Y., and Wang R., "Data Quality in Context," *Communications of the ACM*, vol. 40, no. 5, pp. 103-110, 1997.

[57] Verykios V., Elmagarmid A., and Houstis E., "Automating the Approximate Record-Matching Process," *Information Sciences*, vol. 126, no. 1-4, pp. 83-98, 2000.

[58] Wang J., Krishnan S., Franklin M., Goldberg K., Kraska T., and Milo T., "A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data," *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Snowbird, pp. 469-480, 2014.

[59] Wijsen J., "Database Repairing Using Updates," *ACM Transactions on Database Systems*, vol. 30, no. 3, pp. 722-768, 2005.

[60] Womack J. and Jones D., "Lean Consumption," *Harvard Business Review*, vol. 83 no. 3, pp. 58-68, 2005.

[61] Zheng M., Tucek J., Qin F., and Lillibridge M., "Understanding The Robustness of SSDs Under Power Fault," *in Proceedings of the 11th USENIX Conference on File and Storage Technologies*, San Jose, pp. 271-284, 2013.

Jamil Razmak is an Assistant Professor in the Department of Management in the College of Business. He received his Master's in Business Administration and PhD in Interdisciplinary Studies of Business Technology Management from the Laurentian University in Canada. His primary research interests are in the field of business and technology management. Specifically, he is interested in e-health innovative technology and change management, DSS and business analytics.

Samir Al-Janabi received his Master and PhD degrees in Software Engineering from McMaster University, Hamilton, Canada. His research is motivated by the tremendous value in data. His primary research interests are broadly in the area of data management with a focus on data quality, databases, software engineering, and machine learning. He has extensive experience in software development in different aspects from analysis and design to implementation and testing.

Faten Kharbat is an Associate Professor in Computer Science, and received her PhD in Artificial Intelligence from the University of West of England, Bristol, UK. Her main research interest is learning classifier systems, cancer care, knowledge based systems, applying data mining techniques to marketing, information systems, enterprise social networking, and recently was involved in e-learning systems and quality of higher education.

Charles Bélanger is currently a Senior Business Professor with front line experience at the executive and middle management levels in complex organizations as well as in the private sector. He holds a PhD in Institutional Assessment and Quantitative Analysis from Florida State University. He has published extensively and received national and international awards for his achievements. He has consulted widely across the world in health management, vocational training and organizational audit.