International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Preserve Quality Medical Drug Data toward Meaningful Data Lake by Cluster

Areen Al-Hgaish, Wael Alzyadat, Mohammad Al-Fayoumi, Aysh Alhroob, Ahmad Thunibat

Abstract*: Big data is facing many challenges in different aspects, which appear in characteristics such as: Velocity, Volume, Value and Veracity. Processing and analysis of big data are challenging issues to acquire quality information in order to support accurate medical drug practice. The quality of data taxonomy is indicated by three basic elements: are meaningful, predication and decision-making. These elements have been encouraged in previous work that focused on the same challenges of big data. Consequently, the proposed approach preserves the quality of medical drug data toward meaningful data lake by clustering. It consists of four components. Data collection and pre-processing represent the first component in the data lake. Profile data is treated with semi-structured data to clean it up. The second component is extracting data through enforcing rules on whole data to produce different groups and generate weight based on constraints within groups. In component three, data is organized and clustering. This component complies with schema profiling referring to component two in the data lake. Weight outputs of component three are inputs for component four, where K-Mean clustering is applied to obtain different clusters. Each cluster presents an alternative drug to achieve meaningful drug data that is consistent with component three in the data lake.This paper addressed two main challenges; the first challenge is extracting meaningful data from big data; whereas the second challenge is using big data technique with K-Mean clustering algorithm. An experimental approach was followed through using Food and Drug Administration (FDA) data and symptoms in R framework. ANOVA statistical test was carried out to calculate sum of square error, P- Value and F-Valuefor the evaluation of variances between clusters and variances within clusters. The results showed the efficiency of the proposed approach.

Keywords:Data Lake, K-Mean Clustering, Big Data, Semi-structured Data.

Manuscript published on 30 September 2019
* Correspondence Author
AreenAL-Hgaish*,Department of Software Engineering, Faculty of Information Technology, Isra University, Amman, Jordan. Email: areen.metib@gmail.com
Wael ALzyadat,department of Software Engineering at Al-Zaytoonah University of Jordan, Amman, Jordan. Email: wael.alzyadat@zuj.edu.jo
Mohammad Al-Fayoumi, Department of Software Engineering, Faculty of Information Technology, Isra University, Amman, Jordan. Email: mafayoumi@iu.edu.jo
AyshAlhroob,Department of Software Engineering, Faculty of Information Technology, Isra University, Amman, Jordan. Email: aysh@iu.edu.jo
Ahmad Thunibat, department of Software Engineering at Al-Zaytoonah University of Jordan, Amman, Jordan Email:a.thunibat@zuj.edu.jo

I. INTRODUCTION

The development of infrastructure systems in information technology at recent time has led to the explosion of huge amounts of data. This data was characterized by high velocity, huge volume, and high variety called big data. It includes multiple forms of data, such as: structured, unstructured and semi- structured data, from multiple data sources (e.g. social media, call phones, GPS, sensors, healthcare and IOT). `

The massive amounts of data require several tools to be analyzed. Methods used must be effective and capable of storage data in repositories that realize these quantities of data at lowest possible cost, taking into consideration the cost of storage. Data lake concept was introduced to find solutions to challenges of big data in a centralized repository that captures and stores raw data in the same original formats described bymetadata[1].

Data lake is characterized by the ability to adapt with new data formats with high flexibility and store different datasets from various data sources in the same data repository, such as structured data from traditional Database Management System, as well as different types of structured data, such as: images, videos, voice, texts, e-mails and IOT. This data contains attributes which are undefined previously. Data lake is characterized by flexibility compared to traditional decision support systems that need predefinition of attributes for a specific type of data. It also works to identity data structures when using data in real time[2, 3].

The challenge of data value is not clear or useless. Data lake works on discovering relationships among datasets whose contents are interlinked, analyzing and indexing them to extract high-value data with lower cost, as well as considering risks that data may face in the data lake in terms of dataquality [4].Raw data need description and interpretation, especially with the increasing number of data sources, which motivates to organize and describe data correctly and accurately to achieve high quality [5]. Lack of metadata services is attributed to the inability of data analysts to discover all datasets in the data lake. The challenge is how to deal with massive amounts of data from different data sources to extract meaningful and high-quality data[1].

There is a significant growth in the process of data generation, where some studies indicated that more than 44 Zettabyte will be generated in 2020, more than 80% of which is unstructured data.

This massive growth limits data movement. Also, low cost of data lake allows for data replication and consumption of storage space. However, it constitutes a risk in terms of data governance, including data quality and data security, all of which can be constraints on data, so that data may be unable to move. This is called data gravity [2, 6].

In thispaper, we will study semi-structured data (drug data) to produce meaningful data through obtaining alternative drug names. Each drug will have a brand name, a scientific name, a dose and strength. Our data source is: Food and Drug Administration (FDA).

This paper is structured as follows: Section 2, illustrates related previous work. In Section 3, our approach consisting of four main components is presented. The experimental approach using R framework is shown in Section 4. Finally, the results and conclusion of our work are presented in Section 5.

II. RELATED WORK

The data lake is a low-cost storage physical environment based on Hadoop technology. Data lake is a concept which embraces all enterprise data and moves them into one physical place. The concept addresses the veracity and the volume of big data characteristics. As the concept is linked to the big data wave, data lake methodology is applied with Hadoop. Data warehouse is associated with traditional relational database. So for structured data, it is viewed as more expensive but also as more mature. The data lake is associated with mixed types of data and datasets[7]. It is less expensive as it is based on Hadoop storage, but it is less mature. The tall array was used to handle out-of-memory data. Meaningful and accurate data was obtained from big data which is characterized by volume and veracity, where a large dataset was used and divided into small chunks commensurate with the memory (distributed filesystem)[8].The data lake is viewed as a methodology. However, it's not only a methodology, but rather an actual new data architecture solution composed by hardware, software and conceptual design[2].

According to ref[9], the data lake faces the following challenges:

1. Lack of ability to determine data quality, because the data lake focuses on data storage.
2. Because data structures are not predefined and less mature, the data lake requires metadata management.
3. Data nature requires that data is analyzed from scratch.

The workflow process is different between traditional warehouses and data lakes according to Khine and Wang[9], where processing in a warehouse is done by following ETL process. (E) Data is extracted from the databases first, (T) then come processing, cleaning data and converting them before the loading process. Last process is loading data into the warehouse (L). This approach is called schema-on-write approach. This means that data needs to be predefined for the schema before loading them. Data lakes do not need to predefine schemas and therefore the approach used in them is called the schema-on-read approach. Data is collected from sources to the data lake, which is the first process. The second process is adding metadata when loading data relying upondata analysis process, because a data lake does not contain predefined schemas data is then transformed into appropriate models by using metadata added in pre-processing for use in applications and services.

KAYAK works on data management to accelerate the data preparation process and data is processed in the data lake. Analysis tools work in KAYAK using schema-on-read through direct access to datasets which are store in the distributed file system, which contains a metadata-catalogue that includes the profile for each dataset and associates datasets with each other, contributing to handle metadata and place data in the data lake in a manner suitable to facilitate its use. It also features tolerance which works on production preview of results thereby helping to continue the analysis process and calculations rapidly through increased execution. KAYAK provides a series of primitives that are predefined for the processes of preparation, analysis and data management in the data lake, collects metadata explicitly through specialized tools and then stores metadata attributes in a centralized catalogue to facilitate access to it through any task[10].

A flexible virtual data lake was created to retrieve large geospatial datasets that are integrated and specific. The klimatic architecture was executed in three stages, aiming to include geospatial data covering all regions as well as many variables and years. The first stage is crawling and scraping publicly accessible data. The largest number of public geospatial files is collected from data sources and downloaded. The second stage in klimatic architecture is extracting metadata and indexing. Each dataset is added to an extraction queue and processed. The third stage is loading data into virtual data lake storage. The content of the datasets is converted to relational format to accelerate retrieval and integration. After data is loaded into the data lake, the query response is provided through a query interface using Flask and Python[11]. Both[10, 11]pointed out the importance of metadata to links and content of the datasets is converted to relational format.

Data is collected in the data lake and managed in NoSQL databases where schema is free and flexible, which allows heterogeneous data to be stored easily. A series of derived schema versions was used instead of a global scheme or a single scheme, taking into account inclusion dependencies and discovery of unnecessary changes such as moved or copied data by using algorithms to derive dependencies. In this work, the process of development of schema versions was understood through knowledge of structures, data semantics and changes that may occur over time to make the analysis process effective, useful and yield the required results[12].

Quality and meaningful data were obtained in the healthcare sector where Westra et al [13],electronic healthcare records (EHRs) have been used effectively and clinical and managerial data has been reused to improve quality and facilitate healthcare processes. Administrative data describes the context of healthcare. Data in repositories is missing and needs to be standardized and contained in data repositories side by side with clinical data.

Network-based model was used to extracting knowledge from the data stored in a data lake to uniformly represent the structured, semi-structured and unstructured sources of a lake, which is one of successful solutions for managing big data[14]. Web log files have been analyzed to improve websites and create meaningful data by identifying link connections that occur on websites. Three tasks are performed: pre-processing, pattern discovery and pattern analysis. In the pre- processing phase, data is collected to create a dataset before the appropriate process to provide structured, reliable and integrated data sources through data cleansing, transaction identification, coordination and aggregation in meaningful sessions. The web log structures are then determined and every access to the web page is recorded in the web server's log file. Transaction identification creates meaningful clusters of references per user. After pre-processing, log entries are broken into logical clusters. The path analysis method is applied to analyze pre-processed web log data files and perform link analysis betweenweblinkstoextractimportantlinkrules, patterns and statistics and remove irrelevant rules and statistics between large groups of web links[15].

CLAMS system works to enforce complex quality constraints on large datasets in the data lake, store heterogeneous data and record data sources. First, data is transformed in its various formats (semi-structured and unstructured) to RDF triples, then loaded into HDFS at the stage of data ingestion. Then, constraints are built in the next stage by providing interactive user interfaces to examine detected quality rules. A set of specific constraints is effectively enforced through activating an algorithm that works to detect errors and data that does not match with specific constraints[16].CLAMS system is consistent with our approach, where the first step in the approach is to collect data obtained from its sources and transform it into semi-structured format. Rules will be enforced on all data obtained based on the scientific name of the drug to be converted into a structured form. This step produces a different set for each scientific name of drug; i.e.we make a cluster based on scientific names of drugs (Value). After obtaining organized sets of data for scientific names in the previous step, we enforce constraints on the resulting sets to generate weights for items in each group. In the final stage, we apply K-Mean clustering to group data based on weight to collect the minimum points, producing low variation within clusters. This variation increases across clusters. Through these procedures in our approach, we reach the acquisition of data that has similarity attributes to obtain alternative drugs. So, we obtain high quality and meaningful data in the data lakeby utilizing and exploiting the same data obtained from data sources.

In another work consistent with our approach, the framework for content metadata management (CM4DL) is considered. It represents all types of profiles and supports analytical discovery of content relationships in the data lake. This work refers to the importance of metadata in possible formats to support a process information profiling, where data profile describes the value of dataset and schema profile describes the schema of dataset to arrive at information profiling, where different attributes from different datasets are matched[17].

We indicated the importance of metadata in the data lake, where metadata was extracted and indexed. Schema versions have been used instead of global schema to understand data structure, change over time and date history. This effectively contributes to data analysis yielding meaningful results. Metadata, schema, ontology and RDF are all operations to find relationships between different datasets before conducting data analysis.

Many keys were mentioned to create a successful data lake. Machine learning algorithms were applied to reach business value[9]. This corresponds to our approach to achieve meaningful and quality data. In our approach, we applied clustering algorithm.

III. PROPOSED APPROACH

The approach consists of four components. The first one is data collection and pre-processing component to obtain clear data content and present datasets to be the input to component two that involves the clustering method to find groups that are similar in attributes and start the first round of clustering. The third component organizes data clusters, where it generates weights of the groups that resulted from the previous step. The last component is analyzing data, where K-Mean clustering is applied which represents the second round of clustering to acquire meaningful drug data. Each component involves treating the characteristics of big data as follows; the first component achieves volume, while the second and third components achieve variety and component four achieves value. This ensures the main objective of the paper to be fulfilled. We will demonstrate the design of our approach via each component and illustrate the levels of processes. Figure 1 illustrates the data follow of theapproach.

A. Data Collection and Pre-process

The first level of the first component is to input big datasets as a first step to apply this approach.

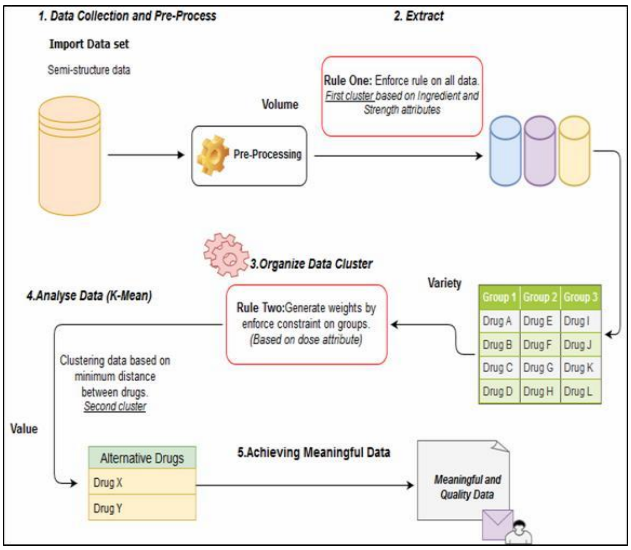

Fig. 1.Approach to Preserve Quality of Medical Drug Data toward Meaningful Data Lake by Clustering.

Datasets are applied from Food and drug Administration (FDA), identifying drug products approved on the basis of safety and effectiveness and containing a number of attributes that support our research requirements to arrive at alternative drugs. Fourteen attribute and 35840 records are considered.

Step 1: Determining the dataset address: (https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files). Downloading Orange Book dataset, which contains the kind ofthe drug file content in semi- structured from. The file format is: Comma Separated Values (CSV), which means not fully standardized, where attributes are clear and facile to handle.

Table I : Description of Dataset Attributes[1].

Attribute Name	Data Type	Description
Ingredient	Character	The active ingredient of the drug (Scientific Name)
Dosage Form; Route of Administration	Character	The drug dosage form and route
Trade Name	Character	The brand name of the product
Applicant	Character	The firm name holding legal responsibility for the new drug application
Strength	Character	The potency of the active ingredient
New Drug Application Type	Character	The type of new drug application approval
New Drug Application (NDA) Number	Integer	The FDA number assigned to the application
Product Number	Integer	The FDA number assigned to identify the application products
Therapeutic Equivalence (TE) Code	Character	The TE code which indicates the therapeutic
Approval Date	Character	The date the product was approved
Reference Listed Drug (RLD)	Character	Drug product approved under section 505(c)
Reference Standard (RS)	Character	Drug product selected by FDA that an applicant is seeking approval of an ANDA
Type	Character	The group or category of approved drugs
Applicant Full Name	Character	The full name of the firm holding legal responsibility for the new drug application

Step2: Presenting and extracting attribute contents. TableI explains the attributes based on three elements, which are: attribute name, data type and description.

After finishing the level of data collection, view and awareness attributes, the second step in component one is pre-processing.

After presenting data, we need to iterate processes and methods to make both attributes and contents clear and ensure that they exist in a manner which is consistent with our requirements. To define all value effects in the final result, such as missing values, spaces and duplicated rows, we detect and deal with them at the pre-processing level which is preparing data. In this paper, we adopted the R framework data analysis, which provides a comprehensive set of pre-processing tools and unsupervised tools, including the clustering algorithm that we will use. At this level of processing, the data isnoisy.

We first apply pre-processing techniques to facilitate the data analysis process and data adaptation with the algorithm. After applying a number of pre-processing techniques, the data obtained is considered a high-quality and reliable data which is ready to be applied to data analysis, and clustering algorithm and techniques of pre- processing that we will use.

- **Selected Attributes**

The technique of selected attributes is the first technique we will use to filter data based on attributes and contents. We identify our important attributes and get rid of irrelevant attributes or duplicate. We define the attributes required in our field to complete the rest of the data cleaning operations as the table IIshows the attributes selected.

Table II. Selected Attributes in Our Research Scope.

Attribute Name	Description
Trade Name	The brand name of the product
Ingredient	The active ingredient of the drug (Scientific Name)
Strength	The potency of the active ingredient
Dosage Form; Route of Administration	The drug dosage form and route

After applying the selected process and identifying the important attributes, the contents of each attribute must be checked in terms of proper existence. Missing values that appear in the class attributes mean that there is a content that should be ignored and solved by deleting outliers.

- **Remove marginal spaces**

We remove the spaces between data contents, so that they do not affect the results in a negative way or take wrong values or ignored values from the content.

- **Eliminate duplicated rows**

We apply the last technique of pre- processing which is to remove the duplicates, where we delete the duplicated rows. This is important to make lateranalysis processes more sufficient and of better quality.

In the data lake, the nature of data must be consolidated to arrive at meaningful data that we obtain at the pre-processing level. At this level, data volume is also reduced by using 'Preprocess' Package in the R framework (https://cran.r-project.org/web/packages/PreProcess/index.html). At this level, it is focused on attributes and content in order to input clean data to the second component.

Represents preprocessing technique is consisted with the work of first component in the data lake which is profiling data component which descript the content value data.

[1]https://www.fda.gov/drugs/drug-approvals-and-databases/orange-book-data-files

B. Extraction Component

Results of the pre-processing procedure will be the input to the extraction component. First, we enforce retrieval rules**(Rule one)** that match obtaining divided data (groups) by investigating the matching between the content attributes to reach different trade names with the same attributes (ingredient and strength).

Rule One:

If Drug X (Ingredient&& Strength) = = Drug Y

(Ingredient&& Strength) Then put in same group

Metadata has been added to this component by enforcing rule one as a basis for establishing relationships between the contents of drug names. At this level of the approach, different data sets are created that have relationships connecting each group (Rule One).Metadata contributes to our data maturity, because the data structure is not predefined (schema-on-read). This level presents the first round of clustering of the approach.

C. Organizes Data Clusters.

The produced groups are entered to the third component for generating weights of these groups.

We use groups of properties in the previous component that include category and numeric data (mixed data). There are correlation relationships between the active ingredient attribute and its strength (matching) with reference to rule one. Also, the dose attribute can be identical between drugs within each group and may be different. We used these properties for creating new features todiscover unobserved relationships in our data through generating a feature weight for each drug within groups, via enforcing constraints on dose attribute**(Rule two)**. This makes the K-Mean clusteringalgorithm work better in the next component, as it deals with numerical data.

Finding the dose-shape attribute: Dose; Route; i.e. (X;Y). If the attribute is identical in the group, it takes the same weight for drugs. If there is a partial match in the attribute; i.e. (Z;Y), it takes a weight which is close to the previous weight. Whenever there is any degree of similarity for a close dose attribute, it takes a closer weight and vice versa.

Rule Two:

If (Dose==Dose && Rout==Rout)

Weight =N

Else if (Dose==Dose&& Rout! =Rout)
Weight = N+1

Else if (Dose! =Dose&& Rout==Rout)
Weight = N+1

Else if (Dose! =Dose&& Rout! =Rout)
Weight = N+2

Extraction component and organizing data clusters component comply with the work of the second component in the data lake, which is schema profiling component which describes the schema of datasets and deals with variety data, because the natural data is found in schema-on-read format, through two components dealing with discovering relationships.

D. Analyzing Data,

Clustering for the second time in our approach by using K-Mean clustering in groupingweights based on the minimum distance.

This is done by identifying the centroid of each group, then the distance between objects and centroid is calculated. The last step gathers results based on minimum distance and puts them in the same cluster number. This is complementary to the first round of clustering in component two, but with obtaining more accurate relationships. K-Mean algorithm classifies elements based on features which are similar when the distance between elements is equal to zero.We get meaningful data, where drugs having minimum distance proximity or a distance equal to zero are alternative drugs.

IF (Drug X. Weight &&Drug Y. Weight)

= = distance zero between drugs

THEN return to alternative drug

The work of component four complies with that of the third component in the data lake which is the information profiling component where metadata is profiled to match different attributes from different datasets.

Enforced two rules; the first rule is for dividing data into different groups which is the first round in the clustering process, while the second rule is used to generate feature weights to extract relationships in our dataset to input the results to the second round of clustering (K-Mean).

Accessing meaningful and quality data at the end of our approach helps support the medical sector to obtain alternative drugs. In the next section, we present and discuss the experiments carried out to realize and evaluate our approach.

IV. EXPERIMENT

We present the experimental approach to preserve the quality of medical drug data toward meaningful data in the data lake by clustering. We take each component separately and describe all experimental steps using R framework, explain steps of work and analyze the results of each component. The FDA considers that medicines are equivalent and alternative if three conditions are met: if the drug is found with the same active ingredient; corresponding strength; same dose and route administration. Depending on previous equivalent conditions of the drugs, we applied the first technique in pre-processing, which is selecting attributes manually in our scope of research. Table IIshows name attributes selected from the dataset and the rest of attributes have been deleted, because they are out of our research scope. Some attributes contain missing values that cannot be used.

The four most important attributes in our research are asfollows: we selected the first attribute which is the trade name which refers to the brand names of the drugs.

The second attribute is the ingredient which represents the composition of chemical drugs or the so-called active ingredient of a drug. This ingredient can be replicated in more than one trade name for a drug. The third attribute is strength, which represents the potency of ingredient in each drug. We can find a trade name of a drug containing an ingredient that has one or more strength. The last attribute selected was dose form and route of administration.

Through the last three attributes, we look at similarities and differences between drug names. In the data collection component, we create object which includes these attributes for the procedure of data analysis in R repository and any change which occurs in data within this object in the component is taken into account. By ingredient, strength and dose for each drug, medications can be obtained which are considered alternatives, provided that similarity exists in the active ingredient, its strength and dose. In case of ingredient similarity and strength difference, it is not considered an alternative drug. Selection of attributes process reduces volume of data to become 35840 rows and four variables instead of fourteen variables, we had. The four most important attributes have been selected in ourresearch.

In the pre-processing operation, we remove the duplicates, where we delete the duplicated rows based on matching in the trade name, ingredient, strength and dose. 20873 duplicate rows have been deleted, where the data was 35840 rows before the deletion process. After deleting the duplicates, the number of rows became 14977.

Figure 2 illustrates the results of data collection and pre-processing component and the change observed in data volume. Pre-processing techniques have significantly reduced volume of data and data became clearer after elimination. Data has become of quality after applying pre-processing techniques. Pre-processing techniques which we used in our approach in the first component of the data lake represent profiling data, which works to describe the value content of the datasets.

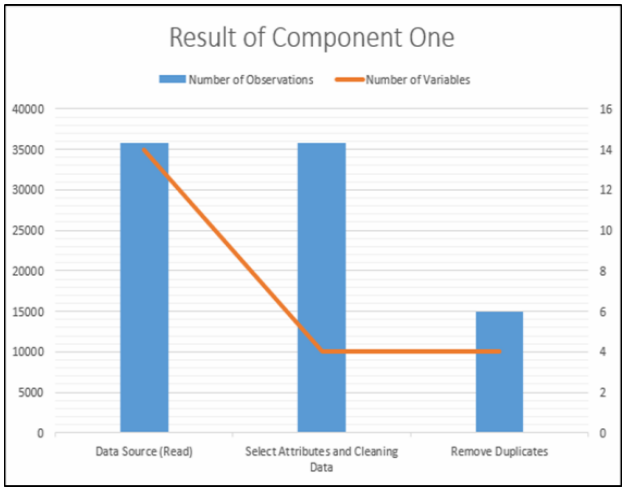

Fig. 1. Data Collection and preprocess.

Depending on equivalence conditions of the drugs which were previously mentioned in the first component based on FDA, we have built a rule one to divide data into different groups.

Each group presents different drug names with content matching in terms of ingredient and strength. Figure 3 illustrates the results of each group produced, where the ingredient value is OH and the strength value is five for five different drug names. This means that resulting groups are equal in the active ingredient and its strength. The number of groups produced at the extraction component is 8568 groups indicating the first round of clustering where data has become more valuable than previously.

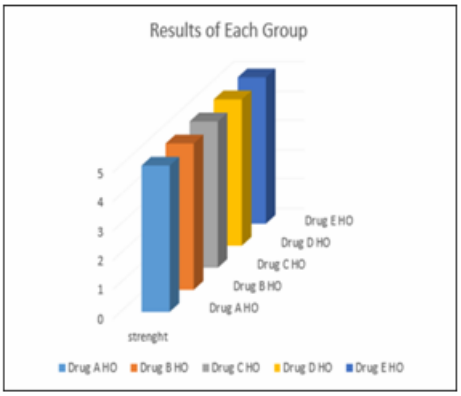

Fig. 3.Results of Each Group.

Through the resulting groups, this component represents the first round of clustering in our approach via our previous processes. Here, we used the filter function in R framework.

In this component, we handle the property of data Varity, because the nature of this data is schema-on-read, where we work to find relationships between different attributes.

The last process in our approach is to measure the clustering rate using ANOVA statistical analysis, which determines whether the variability between groups is larger than the variability of the observations within the groups. Sum of Square Error (SSE), F-Value and P-Value are used. The evaluation results of our approach. The value of SSE represents the variation in the group. The greater the error value, the greater the signal of variation in the group. The value is positive when approaching zero or being equal to zero. The value of our SSE equals 0.08170503, which indicates the lack of variation in the groups and yields very positive results.

F-value is used to determine whether the variability between groups is larger than the variability of the observations within the groups. We obtained an F-Value of 14.823 indicating that the probability is low enough. Where the P-Value is less than the F-value, we can conclude that not all the groups are equal.

The P-Value indicates effects among groups. We obtained a P-Value equal to 0.0001186, which is less than the value of alpha α (0.05).

Applying our approach preserves quality medical drug data (semi-structured) toward meaningful data in the data lake by clustering. Using R-framework allows access to meaningful data of high-quality and arrival at alternative drug names that have the same effectiveness. This contributes to supporting medical opinion to pharmacists, the health sector and investors in this field.

V. RESULT AND CONCLUSIONS

The experimental results to provide meaningful alternative medical drugs by applying an approach that preserves the quality of medical drug data toward meaningful data in the data lake by clustering. It also provides a comparison with relevant studies on clustering algorithms and shows how suitable the results we achieved are by ANOVA statistical test using R framework for the evaluation of variances between clusters and variances within clusters.

The biodiversity of humans and wildlife as well as the conservation of wild areas have been studied [18], where K-Mean clustering was used to obtain homogeneous groups. To determine the differences among groups and relationships among them through using AVOVA test. Results are shown in TableIII.

Table III: Comparison of Results with Sponarski[18]

Factor	F-Value	P-Value
Attitudes toward wolf	369.43	0.001
Wolf Management	1184.47	0.001
Our approach	14.823	0.0001186

The resulting P-Value equal to 0.001 and in our approach the P-Value is equal to 0.000186. Our result appears better than that of the previous work, as the P-Value is closer to zero, indicating greater effect and variability among the groups. The P-Value is less than 0.05, which emphasizes the existence of significant differences between groups and substantial relationships relating data each group.

The sum of square error was used between clusters obtained according to Li, Gao, & Jiao[19].

In the mentioned previous work, a Sum of Square Error (SSE) ratio of 0.0125 was obtained using Feature Weighted Fuzzy K-Mean (FWFKMe), while a ratio of 0.1554 was obtained using Fuzzy K-Mean (FKMe) indicating that data has multi-dimensional features and is divided into three categories. Our experimental results showed a Sum of Square Error ratio of 0.0817 using K-Mean clustering algorithm. Two rules were built using semi-structured data format and our approach has achieved more accurate clusters compared with feature weighted clustering algorithm (FKMe).

Table IV: Comparison of Results with Li, , & Jiao[19].

Resulted	SSE
Our Approach	**0.0817**
Fuzzy K-Mean(FKMe)	**0.1554**
Feature Weight K-Mean (FWFKMe)	**0.0125**

Figure 4 shows a comparison of results between Fuzzy K-Mean (FKMe), Feature Weight K-Mean (FWFKe) in the previous work and our approach which preserves the quality of medical drug data (semi-structured) toward meaningful data in the data lake by clustering.

This paper addressed two main challenges; the first challenge is extracting meaningful data from big data; whereas the second challenge is using big data technique with K-Mean clustering algorithm.

The approach consisted of four components to handle the characteristics of big data component one handles volume, component two and three handle variety and component four handles value. The approach is suitable for data that has big data characteristic, such as Volume, Variety and Value.

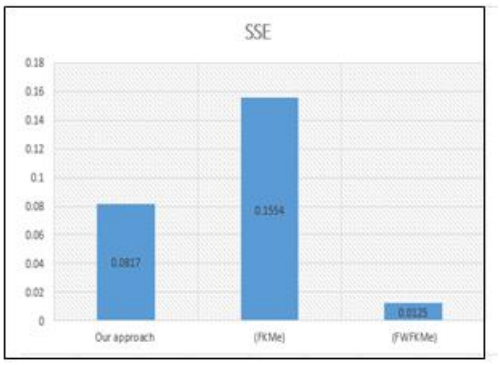

Fig. 4.Comparison of Our Approach with (FKMe) and (FWFKMe) in Terms of Sum of Square Error (SEE).

The paper summarized as follows:

- The approach was utilized by R framework to provide clusters to support medical opinion by obtaining alternative drug names that correspond in practice to a medical drug.
- The approach treats the 3Vs of big data with an appropriate mechanism.
- Our approach used K-Mean clustering with categorical and numerical data with big data characteristics.
- It proposes a new approach to extract meaningful and quality data in order to represent the semi-structured data of a data lake.
- It characterized data in the form of characters with no specific categorization, where weights are generated for data in a systematic manner based on similarity of attributes through the enforcement of constraints (extract metadata).

The effectiveness of the approach has been demonstrated through FDA dataset. According tothe obtained results, data is divided and weights are generated by applying K-Mean algorithm which is considered powerful tool for obtaining alternative drug names that correspond in practice to a drug. In the future, multi-rules, multi-clustering algorithms and many feature selections can be used for more accurate results, as well as to obtain dataset content with more attributes to combine more than one drug and give them an alternative drug in case of a compound active ingredient of the drug. The approach used in this paper preserves the quality of medical drug data toward meaningful data in the data lake by clustering in big data scope. It has a positive effect that supports pharmacists and doctors to diagnose patients and give the appropriate alternative drug if the requested drug is not available, provided that the alternative drug has the same efficiency. It also helps health care domain, where big data is utilized to take the right decision. Last but not least the approach supports increasing investments in the health sector as well as in pharmaceutical companies.

REFERENCES

1. Halevy, A., et al. Goods: Organizing google's datasets. in Proceedings of the 2016 International Conference on Management of Data. 2016. ACM.
2. Madera, C. and A. Laurent. The next information architecture evolution: the data lake wave. in Proceedings of the 8th International Conference on Management of Digital EcoSystems. 2016. ACM.
3. Alzyadat, W.J., et al., FUZZY MAP APPROACH FOR ACCRUING VELOCITY OF BIG DATA. Compusoft, 2019. 8(4): p. 3112-3116.
4. Hai, R., S. Geisler, and C. Quix. Constance: An intelligent data lake system. in Proceedings of the 2016 International Conference on Management of Data. 2016. ACM.
5. Terrizzano, I.G., et al. Data Wrangling: The Challenging Yourney from the Wild to the Lake. in CIDR. 2015.
6. ALzyadat, W. and A. Alhroob, Development Planning in the Big Data Era: Design References Architecture.
7. Fang, H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. in Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on. 2015. IEEE.
8. Al_Zyadat, W.J.a., et al., Securitizing big data characteristics used tall array and mapreduce. International Journal of Engineering & Technology, 2018. 7(4): p. 5633-5639.
9. Khine, P.P. and Z.S. Wang. Data lake: a new ideology in big data era. in ITM Web of Conferences. 2018. EDP Sciences.
10. Maccioni, A. and R. Torlone, Crossing the finish line faster when paddling the data lake with kayak. Proceedings of the VLDB Endowment, 2017. 10(12): p. 1853-1856.
11. Skluzacek, T.J., K. Chard, and I. Foster. Klimatic: a virtual data lake for harvesting and distribution of geospatial data. in Parallel Data Storage and data Intensive Scalable Computing Systems (PDSW-DISCS), 2016 1st Joint International Workshop on. 2016. IEEE.
12. Klettke, M., et al. Uncovering the evolution history of data lakes. in Big Data (Big Data), 2017 IEEE International Conference on. 2017. IEEE.
13. Westra, B.L., et al., Achieving" meaningful use" of electronic health records through the integration of the nursing management minimum data set. Journal of Nursing Administration, 2010. 40(7/8): p. 336-343.
14. Giudice, P.L., et al., An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. Information Sciences, 2019. 478: p. 606-626.
15. Das, R. and I. Turkoglu, Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Systems with Applications, 2009. 36(3): p. 6635-6644.
16. Farid, M., et al. CLAMS: bringing quality to Data Lakes. in Proceedings of the 2016 International Conference on Management of Data. 2016. ACM.
17. Alserafi, A., et al. Towards information profiling: data lake content metadata management. in Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. 2016. IEEE.
18. Sponarski, C.C., et al., Heterogeneity among rural resident attitudes toward wolves. Human Dimensions of Wildlife, 2013. 18(4): p. 239-248.
19. Li, J., X.-B. Gao, and L.-C. Jiao, A new feature weighted fuzzy clustering algorithm. Acta Electronica Sinica, 2006. 34(1): p. 89.

AUTHORS PROFILE

AreenMetib AL-Hgaish , Master degree of Software Engineering , interest in Big Data, Software Engineering processes, and Artificial Intelligence.

Wael JumahAlzyadat,Assistant Professor of Software Engineering**,** currently works atAl-Zaytoonah University of Jordan. His research area encompasses the area of Software Analysis, Intelligence System, Streaming Data, and Big Data. Moreover, established more than 20 published articles and achieved two copyrights.

Mohammad AhmadAl-Fayoumi, Professor of Software Engineering and Security at Isra University of Jordan.

Aysh M. Alhroobis an associate professor of software engineering in Isra University, Jordan. PhD (2010) from University of Bradford, UK. 2010. Aysh joined to Isra University in Jordan as Assistant Professor in Faculty of Information Technology, Software Engineering Department. Aysh has published 20 research papers in international journals and conferences. In addition to, Aysh has published his first book in software testing, 2010.

Ahmad Thunibat,now works as Head of Software Engineering department at Al-Zaytoonah University of Jordan, Amman, Jordan.He obtained his PhD degree in software engineering from the National University of Malaysia in 2012. His Research Interest include software testing, information systems acceptance, requirement engineering, mobile technology