



**Master Thesis**

**An Approach Preserve Quality Medical Drug Data  
(Semi-structure) Toward Meaningful Data Lake by Cluster**

**By**

Areen Metib Al-Hgaish

**Supervisor**

Prof. Dr. Mohammad Al-Fayoumi

**Co-supervisor**

Dr.Wael Al-Zyadat

**This Thesis was Submitted in Partial Fulfillment of the Requirements  
for the Master Degree of Software Engineering**

**Faculty of Graduate Studies**

**ISRA University**

**January 2019**

## **AUTHORIZATION STATEMENT**

I, Areen Matib Naser Akho Shaina, authorize Isra University to Provide Hard copies or soft copies of my thesis to libraries, institutions or institutions or individuals upon their request.

Name: Areen Metib

Signature:

Data:

## إقرار تفويض

انا الطالبة عرين متعب ناصر اخو صحبنة، أفوض جامعة الاسراء للدراسات العليا بتزويد نسخ ورقية من رسالتي ورقيا والكرونا للمكتبات او المنظمات او الهيئات والمؤسسات المعنية بالابحاث والدراسات العيا عند طلبها.

الاسم: عرين متعب ناصر اخو صحبنة

التوقيع:

التاريخ:

The undersigned have examined the thesis entitled "An Approach Preserve Quality Medical Drug Data (Semi-Structure) Toward Meaningful Data Lake by Cluster" presented by Areen Metib Akho Shaina, a candidate for the degree of Master of Science in Software Engineering and hereby certify that it is worthy of acceptance.

3/2/2019

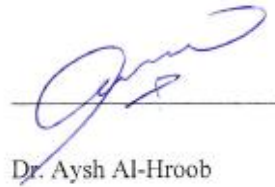
Date



Prof. Dr. Mohammad Al-Fayoumi

3.2.19

Date



Dr. Aysh Al-Hroob

3-19/c/14

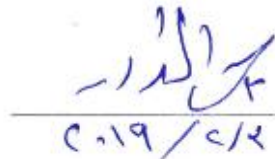
Date



Dr. Wael Al-Zyadat

3/19/c/14

Date



3-19/c/14

Prof. Dr. Ali Al-Dawood

## **DEDICATION**

**This thesis is dedicated**

**To my great father who couldn't wait to see his daughter submitting this thesis and to my affectionate mother who never stopped presenting me love and support being the constant source of motivation and encouragement.**

Areen Metib Al-Hgaish  
January 2019

## **ACKNOWLEDGMENT**

First and foremost, praise is due to God the Almighty for giving me strength, knowledge, ability and opportunity. Without his blessings, this achievement would not have been possible at all.

I would like to kindly present my thanks to my sincere supervisor Dr. Mohammad Al-Fayoumi, who has been so patient, helpful and cooperative in giving me support and advice.

Massive appreciation I'd like to express with very profound gratitude and personal thanks to my co-supervisor Dr. Wael Al-Zyadat who continuously encouraged and supported me.

I owe everything to my family for being patient and understandful while I set the normal flow of my life aside in order to focus on my research. This research would not have been possible to be accomplished without the support of a number of people at ISRA University, who helped me and provided me with knowledge and information, in addition to answering my questions. Without their valuable help, I would not have been able to achieve my goal.

They all deserve to be acknowledged as well.

# Table of Content

<b>DEDICATION</b>	<b>IV</b>
<b>ACKNOWLEDGMENT</b>	<b>VIII</b>
<b>LIST OF TABLES</b>	<b>XI</b>
<b>LIST OF FIGURES</b>	<b>XII</b>
<b>LIST OF ABBREVIATIONS</b>	<b>XIII</b>
<b>Abstract</b>	<b>XIV</b>
<b>1 CHAPTER ONE: INTRODUCTION</b>	<b>1</b>
1.1 Overview	1
1.2 Components of Data Lake	1
1.2.1 Data Profile Component	2
1.2.2 Schema Profile Component	2
1.2.3 Information Profile Component	2
1.3 Comparison between Data Lake and Traditional Data Warehouses	2
1.4 Motivation and Problem Statement	3
1.5 Research Questions.	3
1.6 Objectives	4
1.7 Scope	4
1.8 Technique	4
1.9 Conceptual Approach of Clustering Semi-structured Medical Data in Data Lake	4
1.10 General Research Methodology	5
1.11 Research Significance	6
<b>2 CHAPTER TWO: BACKGROUND AND RELATED WORK</b>	<b>8</b>
2.1 Overview	8
2.2 Keywords	8
2.2.1 Big Data	8
2.2.2 Clustering	8
2.2.3 Data Lake	8
2.3 Related Work	9
2.4 Similar Approaches	10
2.5 Tools and techniques	12
2.5.1 R Package Machine Learning	12
2.5.2 WEKA	12
2.5.3 OpenML	13
2.6 Framework	15
2.6.1 Hadoop Ecosystem	15
2.6.2 DS-Prox Approach Framework to Identify Similarities and Differentiates between Datasets	16
2.7 Discussion	17
<b>3 CHAPTER THREE: METHODOLOGY</b>	<b>19</b>
3.1 Overview	19
3.2 Design of an Approach to Preserve Quality of Medical Drug Data (Semi-Structured) Toward Meaningful Data Lake by Clustering	19
3.2.1 Component One: Volume (Data Collection and Pre-process)	19

3.2.1.1	Data Collection	20
3.2.1.2	Pre-processing	21
3.2.2	Component Two: Variety (Extracting Data)	23
3.2.3	Component Three: Variety (Organizing Data Clusters)	23
3.2.4	Component Four: Value (Analyzing Data)	24
3.2.5	Component Five: Value (Achieving Meaningful Data)	24
3.3	Measurement Results	25
3.3.1	Sum Square Error	25
3.3.2	F-Value	25
3.3.3	P-Value	25
3.4	Tools of Component Execution	25
3.5	Summary	26
<b>4</b>	<b>CHAPTER FOUR: EXPERIMENT AND DISCUSSIONS</b>	<b>27</b>
4.1	Overview	27
4.2	Experiment on an Approach to Preserve Quality of Medical Drug Data (Semi-Structured) Toward Meaningful Data Lake by Clustering	27
4.2.1	Component One: Data Collection and Pre-process	27
4.2.1.1	Data Collection	27
4.2.1.2	Pre-processing	28
4.2.2	Component Two: Extracting Data	32
4.2.3	Component Three: Organizing Data Clusters	33
4.2.4	Component Four: Analyzing Data (K-Mean)	35
4.2.5	Component Five: Achieving Meaningful Data	36
4.3	Measurement Results	37
4.3.1	Sum of Square Error (SSE)	37
4.3.2	F-Value	38
4.3.3	P-Value	38
4.4	Summary	38
<b>5</b>	<b>CHAPTER FIVE: RESULTS AND FUTURE WORKS</b>	<b>39</b>
5.1	Overview	39
5.2	Results	39
5.3	Answering the Research Questions	40
5.4	Contribution	41
5.5	Future Work	41
5.6	Conclusion	42
	<b>References</b>	<b>43</b>
	<b>Appendix A</b>	<b>46</b>
	<b>Appendix B</b>	<b>49</b>
	<b>Appendix C</b>	<b>52</b>



## LIST OF TABLES

<i>Table 1.1: Comparison between Data Lake and Traditional Data Warehouses</i>	3
<i>Table 2.1: Comparison between Similar Approaches</i>	11
<i>Table 2.2: Differences between Properties of R, WEKA and OpenML</i>	14
<i>Table 3.1: Description of Dataset Attributes</i>	21
<i>Table 3.2: Selected Attributes in Our Research Scope</i>	22
<i>Table 4.1: Data Collection Level</i>	28
<i>Table 4.2: Pre-processing Level</i>	29
<i>Table 4.3: Attributes Used and Not Used in Research Scope</i>	30
<i>Table 4.4: Component Two: Extracting Data (dividing data into Different groups)</i>	32
<i>Table 4.5: Component Three: Organizing Data Clusters (Generating Weight)</i>	34
<i>Table 4.6: Component Four Analyzing Data (K-Mean Clustering)</i>	35
<i>Table 4.7: Measurement Results</i>	37
<i>Table 5.1: Comparison of Results with (Sponarski et al., 2013).</i>	39
<i>Table 5.2: Comparison of Results with (Li, Gao, &amp; Jiao, 2006)</i>	40

## LIST OF FIGURES

<i>Figure 1.1: Clustering of Semi-structured Medical Data in Data Lake</i>	4
<i>Figure 1.2: General Methodology of Research Design (Kuechler &amp; Vaishnavi, 2008)</i>	5
<i>Figure 2.1: Layers and Components of the Hadoop Ecosystem</i>	16
<i>Figure 3.1: Design of an Approach to Preserve Quality of Medical Drug Data toward Meaningful Data Lake by Clustering</i>	20
<i>Figure 4.1: Results of data collection and pre-processing</i>	31
<i>Figure 4.2: Component Two (Results of Each Group)</i>	33
<i>Figure 4.3: Number of Groups Produced in the Approach</i>	36
<i>Figure 4.4: Plot of Clustering Result</i>	37
<i>Figure 5.1: Comparison of Our Approach with (FKMe) and (FWFKMe) in Terms of Sum of Square Error (SEE)</i>	40
<i>Figure 5.2: Implementation of Component one (Pre-processing)</i>	49
<i>Figure 5.3: Implementation of Component One (Data Collection)</i>	49
<i>Figure 5.4: Implementation of Component Two Extracting Data (Grouping Data)</i>	50
<i>Figure 5.5: Implementation of Component Two Extracting Data (Result of each group)</i>	50
<i>Figure 5.6: Implementation of Component Three Organizing Data Cluster (Generating Weight)</i>	51
<i>Figure 5.7: Implementation of Component Four Analyzing Data (K-Mean clustering)</i>	51

## LIST OF ABBREVIATIONS

	Abbreviation	Full Expression
1.	DL	Data Lake
2.	4V's	Volume, Velocity, Veracity and Value
3.	WEKA	Waikato Environment for Knowledge Analysis
4.	FDA	Food and Drug Administration
5.	HDFS	Hadoop Distributed File System
6.	SSE	Sum of Square Error

## **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.

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-Value. The results showed the efficiency of the proposed approach.

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