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:
قرار تفويض

انا الطالبة عرين متعب ناصر اخو صحينة، أعوض جامعة الابسرا للدراسات العليا بتزويد نسخ ورقية من رسالتي ورقياً والالكترونياً للمكتبات أو المنظمات أو الهيئات والمؤسسات المعنية بالابحاث والدراسات العليا عند طلبها.

الاسم: عرين متعب ناصر اخو صحينة
التوقع:
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 Arsen Motib 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

5/9/2019
Date
Dr. Wael Al-Zyadat

6/19/2019
Date
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-Hgaïsh
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

**DEDICATION**  
_________________________________________________________________________  IV

**ACKNOWLEDGMENT**  
_________________________________________________________________________  VIII

**LIST OF TABLES**  
_________________________________________________________________________  XI

**LIST OF FIGURES**  
_________________________________________________________________________  XII

**LIST OF ABBREVIATIONS**  
_________________________________________________________________________  XIII

**Abstract**  
_________________________________________________________________________  XIV

1  **CHAPTER ONE: INTRODUCTION**  
    1.1 Overview  
    1.2 Components of Data Lake  
      1.2.1 Data Profile Component  
      1.2.2 Schema Profile Component  
      1.2.3 Information Profile Component  
    1.3 Comparison between Data Lake and Traditional Data Warehouses  
    1.4 Motivation and Problem Statement  
    1.5 Research Questions.  
    1.6 Objectives  
    1.7 Scope  
    1.8 Technique  
    1.9 Conceptual Approach of Clustering Semi-structured Medical Data in Data Lake  
    1.10 General Research Methodology  
    1.11 Research Significance  

2  **CHAPTER TWO: BACKGROUND AND RELATED WORK**  
    2.1 Overview  
    2.2 Keywords  
      2.2.1 Big Data  
      2.2.2 Clustering  
      2.2.3 Data Lake  
    2.3 Related Work  
    2.4 Similar Approaches  
    2.5 Tools and techniques  
      2.5.1 R Package Machine Learning  
      2.5.2 WEKA  
      2.5.3 OpenML  
    2.6 Framework  
      2.6.1 Hadoop Ecosystem  
      2.6.2 DS-Prox Approach Framework to Identify Similarities and Differentiates between Datasets  
    2.7 Discussion  

3  **CHAPTER THREE: METHODOLOGY**  
    3.1 Overview  
    3.2 Design of an Approach to Preserve Quality of Medical Drug Data (Semi-Structured) Toward Meaningful Data Lake by Clustering  
      3.2.1 Component One: Volume (Data Collection and Pre-process)  

IX
**LIST OF TABLES**

Table 1.1: Comparison between Data Lake and Traditional Data Warehouses  __________3

Table 2.1: Comparison between Similar Approaches ______________________________11

Table 2.2: Differences between Properties of R, WEKA and OpenML ____________14

Table 3.1: Description of Dataset Attributes ________________________________21

Table 3.2: Selected Attributes in Our Research Scope ________________________22

Table 4.1: Data Collection Level __________________________________________28

Table 4.2: Pre-processing Level ____________________________________________29

Table 4.3: Attributes Used and Not Used in Research Scope __________________30

Table 4.4: Component Two: Extracting Data (dividing data into Different groups) _____32

Table 4.5: Component Three: Organizing Data Clusters (Generating Weight) ______34

Table 4.6: Component Four Analyzing Data (K-Mean Clustering) ______________35

Table 4.7: Measurement Results __________________________________________37

Table 5.1: Comparison of Results with (Sponarski et al., 2013). ______________39

Table 5.2: Comparison of Results with (Li, Gao, & Jiao, 2006)________________40
LIST OF FIGURES

Figure 1.1: Clustering of Semi-structured Medical Data in Data Lake ________________4
Figure 1.2: General Methodology of Research Design (Kuechler & Vaishnavi, 2008) ______5
Figure 2.1: Layers and Components of the Hadoop Ecosystem ______________________16
Figure 3.1: Design of an Approach to Preserve Quality of Medical Drug Data toward Meaningful Data Lake by Clustering ________________________________20
Figure 4.1: Results of data collection and pre-processing __________________________31
Figure 4.2: Component Two (Results of Each Group) ______________________________33
Figure 4.3: Number of Groups Produced in the Approach ___________________________36
Figure 4.4: Plot of Clustering Result ____________________________________________37
Figure 5.1: Comparison of Our Approach with (FKMe) and (FWFKMe) in Terms of Sum of Square Error (SEE) ________________________________40
Figure 5.2: Implementation of Component one (Pre-processing) ___________________49
Figure 5.3: Implementation of Component One (Data Collection) ________________49
Figure 5.4: Implementation of Component Two Extracting Data (Grouping Data) ____50
Figure 5.5: Implementation of Component Two Extracting Data (Result of each group) ___50
Figure 5.6: Implementation of Component Three Organizing Data Cluster (Generating Weight) ______________________________________________51
Figure 5.7: Implementation of Component Four Analyzing Data (K-Mean clustering) ___51
# LIST OF ABBREVIATIONS

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