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 صحينة، أفوض جامعة الاسراء للدراسات العليا بتزويد نسخ ورقية من رسالتي ورقيا والكترونيا للمكتبات او المنظمات او الهيئات والمؤسسات المعنية بالابحاث والدراسات العيا عند طلبها.

الاسم: عرين متعب ناصر اخو صحينة

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

٢٠١٩/٢/٣

Date

Dr. Wael Al-Zyadat

٢٠١٩/٢/٢

Date

٢٠١٩/٢/٢

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

LIST OF TABLES

LIST OF FIGURES

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 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.