



**AN APPROACH TO IMPROVE DATA QUALITY
FROM VERACITY OF DATA ACCURACY FOR
SENSITIVE COST AND TIME INDICATORS**

Submitted by

Banan Aref Mohammad

Supervisor

Prof. Dr. Mohammed Alfayomi

Co-supervisor

Dr. Wael Alzyadat

**This thesis was submitted in partial fulfillment of the requirements for the
Master's Degree of Science in Software Engineering**

Faculty of Information Technology

ISRA University

December 2018

The undersigned have examined the thesis entitled, An Approach To Improve Data Quality From Veracity Of Data Accuracy For Sensitive Cost And Time Indicators, presented by Banan Aref Mohammad, a candidate for the degree of Master of Science in Software Engineering and hereby certify that it is worthy of acceptance.

5/3/2019

Date



Prof. Dr. Mohammed Alfaimi

5/3/2019

Date



Dr. Wael Zyadaat

5/3/2019

Date



Dr. Thamer Alrousaan

5/3/2019

Date



Dr. Bilal Ibrahim Sowat

DEDICATION

This thesis is dedicated

To my homeland Palestine and its capital city Jerusalem where I will go back one day.

To the kings of my life; my best parents who are the reason of my success in everything with all love and support to achieve my goals.

To my supervisor and co-supervisor for their patience, support and valuable advice and guidance.

To my sisters who shared with me all the steps and details of this work.

To my doctors in master degree study who granted me all their efforts during my study.

To my colleagues during my study for their support and encouragement.

To everyone who helped me even only by some word to convert my dream into reality.

I dedicate this achievement with sincere thanks from my heart

Banan Aref Mohammed

January.2019 .

ACKNOWLEDGMENT

All praise and thanks to Allah the almighty for His guidance and blessings and for granting me knowledge, patience and strength to accomplish this work successfully,

I would also like to kindly present my sincere thanks to my supervisor **Dr. Mohammad Al-Alfayomi**, who has been so helpful and cooperative in giving me support and guidance.

Massive Appreciation I'd like to express with very profound gratitude and personal thanks to my elegant **co-supervisor Dr.Wael Alzyadat** who encouraged and support me and took time to suggest changes, improvements, or clarifications. Although his suggestions caused more work for me, I believe that the end result is the best works I have ever done.

I owe everything to **my family** for being patient and understand while I set the normal flow of my life aside in order to focus on research. This research would have never been possible without the support of a number of people at ISRA University who helped me and provided me information, in addition to answering my questions in ordering to achieve my goal.

They all deserve to be acknowledged as well.

Table of Contents

I

DEDICATION.....	III
ACKNOWLEDGMENT	IV
LIST OF Tables.....	VII
LIST OF FIGURES.....	VIII
Chapter One Introduction.....	1
1.1 Introduction.....	1
1.2 Problem Statement	3
1.3 Research questions	3
1.4 Objectives	3
1.5 Scope of Research.....	3
1.6 Significant.....	3
1.7 Motivation	4
1.8 Justification	4
1.9 Conceptual approach	5
Chapter Two Literature Review	7
2.1 Overview	7
2.2 Related Work	8
2.2.1 Concepts and Definitions.....	8
2.2.2 Similar Approaches	10
2.2.3 Tools.....	17
2.2.4 Relations between Main Factors	18
2.3 Discussion	24
Chapter Three Methodology	26
3.1 Overview	26
3.2 Improving Data Quality from Big Data Characteristics for Sensitive Cost and Time	27
3.2.1 Level1: Generating Volume and Confidence of Cases (Raw Cases).....	28
3.2.2 Level 2: Extracting Value and Context of Cases.....	33
3.2.3 Level 3: Get Best Quality by Choice from High Veracity	34
3.3 Summary	35
Chapter Four Experimental Results and Analysis	38
4.1 Overview	38
4.2 IBM HR Analytics Employee dataset.....	39

4.3 “Deducer” Package in R language tool	39
4.3.1. Initializing and Starting with R	40
4.4 Implementing Improving Data Quality from Big Data Characteristics for Sensitive Cost and Time 40	
4.4.1 Extracting Volume of Data and Getting it into Confidence Level by Sensitive Correlation.....	42
4.2.2. Relation Dependent on Value and Context of Cases.	47
4.5 Summary	51
Chapter Five Conclusion and Future Work.....	53
5.1 Overview	53
5.2 Results	54
5.3 Objectives Achieved.....	54
5.4 Contribution.....	55
5.5 Future Work	55
References.....	56

LIST OF Tables

Table 2.1: Comparison among Approaches.....	15
Table 2.2: Comparison among Tools.....	16
Table 2.3: Comparison among Relations	25
Table 4.1: Spearman’s Correlation Result.....	43
Table 4.2: Results of Filtering Step.....	52

LIST OF FIGURES

Figure 1.1: Iron Triangle.....	2
Figure 1.2: Conceptual Approach.....	4
Figure 2.1 Employee Relation Management	19
Figure 2.2: Employee Relation Quality.....	22
Figure 3.1: Improve Data Quality From Veracity Of Data Accuracy For Sensitive Cost And Time Indicators model.....	27
Figure 3.2: Level 1. Generate Volume and Confidence of Cases	28
Figure 3.3: Level 2. Extracting Value from Context.....	33
Figure 3.4: Level 4. Quality from High Veracity	34
Figure 3.5: Improve Data Quality From Veracity Of Data Accuracy For Sensitive Cost And Time Indicators Deep model.....	36
Figure 4.1: Project Management Iron Triangle	37
Figure 4.2 Implement Improve Data Quality From Veracity Of Data Accuracy For Sensitive Cost And Time Indicators deep model.....	39

Abstract

Big data is a term which describe the characteristics of a dataset, such as volume, value and veracity. There are many challenges which prevent proceeding and working with big data by using traditional techniques to extract value. Project management is a dynamic process that utilizes the appropriate resources of an organization in many phases by measuring in four factor: scope, time, cost and quality.

Improving data quality depends on relations between data and value, which are associated with veracity and accuracy of data and how we can get quality from value. In this research, we attempt to improve data quality from big data characteristics depending on trust of data by working with it in general and especially by using value, volume and veracity of data by finding out correlation statistical analysis results and distance equation. This approach was implemented by using IBM human resource scope with R framework through selecting “Deducer” package from R library.

Implementation and conducting experiment have been carried out by using three main factors: time, cost and scope in two types: product and project, where the strongest relation linking them starts with project scope as the strongest factor followed by cost, product and finally time which is the weakest factor among them. In the final form, we select the best quality using two sides generally: quality degree and middle-quality interval. Especially, relative distance is the strongest factor in the experiment between time and cost, where both sides lead to more trust of data and high accuracy in the chosen process.

Keywords: Big Data Characteristics, Project Management Perspective, Project Management Triangle, Sensitive Rule.