

**DATA QUALITY MINING IMPLEMENTATION
ON SKK MIGAS SISTEM OPERASI TERPADU (SOT) DATA**

By

Bobby Suryajaya
22012208

A thesis submitted to the Faculty of
ENGINEERING AND INFORMATION TECHNOLOGY

in partial fulfillment of the requirements
for the
MASTER'S DEGREE
in

INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY


SWISS GERMAN UNIVERSITY
EduTown BSDCity
Tangerang 15339
INDONESIA

JANUARY 2014
Revision after Thesis Defense on February 18, 2014

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Bobby Suryajaya

Student

Date

Approved by:

Dr. Ir. Mohammad A. Amin Soetomo, M.Sc.

Thesis Advisor

Date

Dr. Harya Widiputra

Thesis Co-Advisor

Date

Alva Erwin, M.Sc.

Thesis Co-Advisor

Date

Dr. Ir. Gembong Baskoro, M.Sc.

Dean

Date

Bobby Suryajaya

ABSTRACT

DATA QUALITY MINING IMPLEMENTATION FOR SKK MIGAS SISTEM OPERASI TERPADU (SOT) DATA

By

Bobby Suryajaya

Dr. Ir. Mohammad A. Amin Soetomo, M.Sc., Advisor

Dr. Harya Widiputra, Co-Advisor

Alva Erwin, M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

SKK Migas implemented SOT (Sistem Operasi Terpadu) to allow SKK Migas system to retrieve specific information from PSC Contractor data source within integrated online system. It is expected that at the completion of SOT program, there will be huge transactional data that will be exchanged on daily basis, and without Data Mining, it will be extremely difficult to perform analysis. Meanwhile, this research will focus on poor data quality that may be resulted from poor data acquisition process, and this research objective is to examine whether data mining techniques can improve SOT Data quality or not by utilizing CRISP-DM methodology. This research found that SOT Data contain some quality issues including missing values, negative values, useless attributes, and outliers that may affect analysis result. Finally, this research proves that Data Quality Mining processes and algorithms can improve SOT Data quality, which expected to contribute to SKK Migas data analysis improvement.

Keywords: Data Quality Mining, DQM, SKK Migas, CRISP-DM, Missing Values, and Outliers



DEDICATION

I dedicate this thesis to my family, Faculty of Engineering and Information Technology in SGU (Swiss German University), and Indonesia Oil & Gas IT society.

Hope this will add more knowledge to Data Mining domain especially the application within Oil & Gas industry.



ACKNOWLEDGMENTS

I would like to thank Allah Jalla wa 'Alaa for the abundance of grace and gift, so I can finish this work as planned and without experiencing significant obstacles.

I also wish to thank everyone for his or her full support and commitment on completion of my thesis, especially:

- My family for their patience and support they had given from the beginning until it is completed.
- My colleagues at SGU MIT Class 2012-2013: Frieda Putri Aryani, Umabala Devarakonda, Fenty Simanjuntak, Aldo Elam Majiah, and Jusak Rahardja.
- My Advisor Dr. Ir. Mohammad A. Amin Soetomo, M.Sc.
- My Co-Advisor Dr. Harya Widiputra.
- My Co-Advisor Alva Erwin, M.Sc.
- SKK Migas Deputy of Internal Control Bpk. Dr. Budi Ibrahim.
- SKK Migas Management Information System Division Head Bpk. Handoyo Budi Santoso.
- SKK Migas Strategic Data & Information Head Bpk. Syukri Waldi.
- SKK Migas Strategic Data Processing Head Ibu Sri Wahyuning Astuti and all her team members.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR	2
ABSTRACT	3
DEDICATION.....	5
ACKNOWLEDGMENTS	6
TABLE OF CONTENTS	7
LIST OF FIGURES	9
LIST OF TABLES.....	10
CHAPTER 1 – INTRODUCTION.....	11
1.1 Background	11
1.2 General Statement of Problem Area	12
1.3 Research Problem	14
1.4 Research Limitations.....	17
1.5 Research Questions	17
1.6 Research Objectives.....	17
1.7 Significance of Study	17
CHAPTER 2 – LITERATURE REVIEW.....	18
2.1 Data Mining Techniques.....	18
2.2 Data Quality	21
2.3 Data Quality Mining	23
2.4 SOT Data	24
2.5 Research on Data Quality	26
2.6 CRISP-DM Methodology	33
2.7 Theoretical Framework	36
CHAPTER 3 – RESEARCH METHODOLOGY	39
3.1 Business Understanding.....	40
3.2 Literature Review.....	40
3.3 Data Understanding.....	41
3.4 Data Preparation.....	41
3.5 Data Cleaning.....	42
3.6 Evaluation	43
CHAPTER 4 – RESEARCH DESIGN & EXPERIMENT	44
4.1 Research Groundwork.....	44

4.2	Data Understanding.....	45
4.3	Data Preparation.....	46
4.4	Modeling.....	48
CHAPTER 5 – RESULT & ANALYSIS		50
5.1	Data Understanding.....	50
5.2	Data Preparation.....	53
5.3	Modeling.....	63
5.4	Result Summary.....	67
5.5	Evaluation Result.....	69
CHAPTER 6 – CONCLUSION & RECOMMENDATION.....		70
6.1	Research Conclusion.....	71
6.2	Recommendation.....	71
6.3	Future Work.....	73
GLOSSARY		74
REFERENCES		75
APPENDIX 1 – RapidMiner Parameter Settings		79
	Data Retrieval Parameter Settings.....	79
	Missing Value Handling Parameter Settings.....	81
	Feature Selection Parameter Settings.....	82
	Outlier Handling Parameter Settings.....	83
	Model Validation Parameter Settings.....	85
APPENDIX 2 – RapidMiner Operator & Algorithm		86
CURRICULUM VITAE.....		95



SWISS GERMAN UNIVERSITY

LIST OF FIGURES

Figures	Page
Figure 1.Research Problem Statement.....	14
Figure 2.Fishbone Analysis of Poor Data Mining Techniques Utilization	16
Figure 3.DIKW Hierarchy (Ackoff, 1989)	18
Figure 4.Knowledge Discovery in Database (KDD) Process (Fayyad et al., 1996)	19
Figure 5.Sample on How Data Mining Classification Works (Ridzuan Daud, 2008)	20
Figure 6.Sample on How Data Mining Classification Works	21
Figure 7.SKK Migas SOT Data Exchange Mechanism (Suryajaya et al., 2013)	25
Figure 8.Isolated points resulted from k-NN method (Aggarwal, 2013)	30
Figure 9.LOF Basic Idea to identify Outliers (Breunig et al., 2000).....	32
Figure 10.DBSCAN Basic Idea (Ester et al., 1996)	33
Figure 11.CRISP-DM 6 Life-Cycle Phases (Wirth & Hipp, 2000).....	35
Figure 12.Theoretical Framework	36
Figure 13.Research Methodology.....	39
Figure 14.Initial DQM Model.....	50
Figure 15.Outliers Detection and Cleaning by Local Outlier Factor (LOF)	60
Figure 16.Outliers Detection and Cleaning by Connectivity-Based Outlier Factor (COF)	61
Figure 17.Outliers Detection by Cluster-Based Local Outlier Factor (CBLOF) with DBSCAN	62
Figure 18.Outliers Detection by Cluster-Based Local Outlier Factor (CBLOF) with DBSCAN from Time Perspective	62
Figure 19.Final Model based on Evaluation Result.....	69

LIST OF TABLES

Table	Page
Table 1.SOT Production Monitoring Data	13
Table 2.SOT September 2013 Metadata (Original).....	51
Table 3.Data Statistics PDEN_SOURCE	52
Table 4.Data Statistics PDEN_TYPE.....	52
Table 5.Data Statistics VOLUME_METHOD	52
Table 6.Data Statistics GAS_VOLUME_OUOM.....	52
Table 7.Data Statistics OIL_VOLUME_OUOM	53
Table 8.Data Statistics WATER_VOLUME_OUOM.....	53
Table 9.Data Statistics ACTIVITY_TYPE	53
Table 10.SOT September 2013 Metadata after Data Selection.....	54
Table 11.Missing Value Statistics before Data Cleaning	55
Table 12.Data Attributes Weight by Gini Index.....	58
Table 13.Data Attributes Weight by Information Gain	58
Table 14.Initial Performance Vector Confusion Matrix.....	64
Table 15.Performance Vector Confusion Matrix LOF with k-NN.....	65
Table 16.Performance Vector Confusion Matrix COF with k-NN	65
Table 17.Performance Vector Confusion Matrix LOF with Naïve Bayes	66
Table 18.Performance Vector Confusion Matrix COF with Naïve Bayes	66
Table 19.Data Collection, Selection, Cleaning, and Feature Selection Result Summary	67
Table 20.Data Cleaning from Outliers Result Summary.....	68
Table 21.Model Validation Result Summary	68