IMPLEMENTATION OF BIG DATA ANALYTIC FOR CUSTOMER-ORIENTED SUPPLY CHAIN MANAGEMENT

Wecka Imam Yudhistyra

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Civil, Transportation and Geo-resources Engineering

Suranaree University of Technology

Academic Year 2020

การประยุกต์แนวคิดของการใช้ฐานข้อมูลขนาดใหญ่เพื่อการวิเคราะห์และจัดการซัพพลายเชนฝั่งลูกค้า

นายเวคก้า อิมัม ยูดิสทีร่า

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

สาขาวิชาวิศวกรรมโยธา ขนส่ง และทรัพยากรธรณี

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2563

IMPLEMENTATION OF BIG DATA ANALYTIC FOR CUSTOMER-ORIENTED SUPPLY CHAIN MANAGEMENT

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Thesis Examining Committee

Thaned Satiennam

(Assoc. Prof. Dr. Thaned Satiennam)

Chairperson

(Prof. Dr. Vatanavongs Ratanavaraha)

Member (Thesis Advisor)

(Dr. I-soon Raungratanaamporn)

Member

(Assoc. Prof. Dr. Wichuda Satiennam)

Member

(Asst. Prof. Dr. Buratin Khampirat)

Member

(Asst. Prof. Dr. Sajjakaj Jomnonkwao)

Member

(Asst. Prof. Dr. Duangdao Watthanakhlang)

Member

(Assoc. Prof. Flt. Lt. Dr. Kontorn Chamniprasart)

Vice Rector for Academic Affairs and Internationalization

Pornsiri Jongkol

(Assoc. Prof. Dr. Pornsiri Jongkol)

Dean of Institute of Engineering

เวคก้า อิมัม ยูดิสทีร่า : การประยุกต์แนวคิดของการใช้ฐานข้อมูลขนาดใหญ่เพื่อการวิเคราะห์และจัดการซัพพลายเชนฝั่งลูกค้า (IMPLEMENTATION OF BIG DATA ANALYTIC FOR CUSTOMER-ORIENTED SUPPLY CHAIN MANAGEMENT) อาจารย์ที่ปรึกษา : ศาสตราจารย์ ดร.วัฒนวงศ์ รัตนวราห, 156 หน้า.

วัตถุประสงค์ของการศึกษาในงานวิจัยนี้ มีวัตถุประสงค์เพื่อการหาแนวทางในการประยุกต์การวิเคราะห์ระบบฐานข้อมูลขนาดใหญ่ เพื่อลดช่องว่างระหว่างขนาดของข้อมูลขนาดใหญ่ที่ใช้ในปัจจุบันและทักษะที่จำเป็นต้องใช้ในการวิเคราะห์ เพื่อส่งเสริมและเพิ่มขีดความสามารถของการตัดสินใจในการพัฒนาการดำเนินการของบริษัทที่เป็นกรณีศึกษา โดยมุ่งเน้นไปที่ส่วนของลูกค้าเป็นสำคัญ ซึ่งในงานวิจัยนี้ได้ดำเนินการศึกษาทั้งสิ้นสองกรณี โดยในงานวิจัยนี้มีลักษณะเป็นวิทยานิพนธ์เชิงวารสาร ได้จัดโครงสร้างของวิทยานิพนธ์ออกเป็นสี่ส่วนได้แก่ ส่วนที่หนึ่ง (บทที่ 2) เป็นการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องเพื่อรวบรวมถึงสถานการณ์ และระบุถึงเหตุและความจำเป็นของการศึกษาในบริบทของการวิเคราะห์ และประยุกต์ใช้ฐานข้อมูลขนาดใหญ่ในภาคโลจิสติกส์และห่วงโซ่อุปทาน และแนวโน้มของการประยุกต์ใช้การวิเคราะห์ดังกล่าวในอนาคต ผ่านการศึกษาด้วยวิธีการศึกษาร่วมข้ามสาขาวิชา ทั้งในด้านของการศึกษาแนวคิด วิธีการทางสถิติที่ใช้ในการวิเคราะห์ และการคาดการณ์แนวโน้มและบทบาทของการใช้ฐานข้อมูลขนาดใหญ่ในอนาคต

ในส่วนที่สอง (บทที่ 3 และบทที่ 4) ประกอบไปด้วยการภิปราย ข้อเสนอแนะและบทสรุปของการประยุกต์ใช้ฐานข้อมูลขนาดใหญ่เพื่อการจัดการซัพพลายเชนฝั่งลูกค้า โดยได้กำหนดเครื่องมือวิจัยและคัดเลือกกรณีศึกษา รวมถึงโครงสร้างของการรวบรวมข้อมูล การแสดงผล และการวิเคราะห์ ซึ่งมีวัตถุประสงค์เพื่อประยุกต์ใช้ฐานข้อมูลขนาดใหญ่ในบริษัทกรณีศึกษาในประเทศอินโดนิเซีย โดยมุ่งเน้นไปที่การค้นหารูปแบบและการได้มาซึ่งความรู้ที่เกิดขึ้นจากการวิเคราะห์ และประยุกต์ใช้ฐานข้อมูลขนาดใหญ่ เพื่อสนับสนุนการตัดสินใจเชิงนโยบายของการดำเนินงาน การอภิปรายผลการศึกษานั้นยังได้ถูกพัฒนา และดำเนินการโดยเน้นการบูรณาการระหว่างฐานข้อมูลขนาดใหญ่ในการดำเนินธุรกิจประจำวันในบริษัท ในบทนี้ผู้ศึกษาได้ใช้แนวคิดของการวิเคราะห์ด้วยวิธีการจัดทำเหมืองข้อมูลเพื่อการวิเคราะห์และใช้ประโยชน์ทางธุรกิจเป็นเครื่องมือในการวิเคราะห์ลูกค้า และดำเนินการเสนอแนะแนวทางในการวิเคราะห์และประยุกต์ใช้ฐานข้อมูลขนาดใหญ่เพื่อเป็นเครื่องมือในการพัฒนากลยุทธ์ขององค์กรภายใต้แนวคิด และวิธีการที่เป็นรูปธรรมและเหมาะสม และในส่วนที่สาม (บทที่ 5) จะเป็นการสรุปผลการศึกษาวิจัยตามวัตถุประสงค์ของการวิจัยและข้อเสนอแนะในการดำเนินการวิจัยต่อไปในอนาคต

สาขาวิชา วิศวกรรมขนส่ง
ปีการศึกษา 2563

ลายมือชื่อนักศึกษา WECKA IMAM YUDHISTYRA
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

WECKA IMAM YUHDISTYRA : IMPLEMENTATION OF BIG DATA ANALYTIC FOR CUSTOMER-ORIENTED SUPPLY CHAIN MANAGEMENT. THESIS ADVISOR : PROF. VATANAVONGS RATANAVARAHA, Ph.D., 156 PP.

BIG DATA/ANALYTICS/CUSTOMER/CLUSTERING/ASSOCIATION RULE MINING/DECISION MAKING

The primary objective of this research is to implement big data analytics to minimize the gap between a large amount of data available now and skill to analyze it. This thesis provides two case studies in implementing big data analytics to help a company strenghthen its decision making ability focused on its customers. Structurally, this thesis consists of four further chapters with three main parts. Part I (Chapter II), I situate the current study related literature and emphasizes the need to investigate current conditions and trends regarding the implementation of BDA in logistics and supply chain research communities. In this regard, I undertook a cross-disciplinary approach, such as a conceptual framework for literature review and statistical analysis for trend prediction of big data research in the future. Combining these insights, Chapter II argues the importance of big data implementation research for researchers around the world to minimize the gaps in the literature that are identified and posed accordingly in this chapter.

Part II (Chapter III and Chapter IV) contains the discussion, recommendation, and conclusion for implementing big data analytics. This includes procedural instruments used in these case studies to collect, visualize, and analyze the data. The main objective of this thesis is to conduct big data analytics focusing on a company in

Indonesia. To be more specific, it is focusing on finding patterns and knowledge from big data available for supporting company decision making. The discussion on the key findings was developed to integrate big data analytics into daily operational business in a company. The methodology used in this chapter is the combination between CRISP-DM and key steps for customer analysis. Finally, the implication and the recommendation for developing strategic planning in a company are drawn based on measurable data and appropriate concepts.

Part III (Chapter V) is the conclusions of the research in this thesis and recommendation for further research. This part is summarized to answer the research objectives and to recommend further research in the future.

School of Transportation Engineering　　Student's Signature__________

Academic Year 2020　　Advisor's Signature__________

Co-advisor's Signature__________

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my supervisor, Professor Vatanavongs Ratanavaraha, Ph.D., who has the substance of a genius: he convincingly guided and encouraged me to be professional and do the right thing even when the road got tough. I would also like to express my sincere gratitude to my co-advisor, Dr. I-soon Raungratanaamporn. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my Ph.D. study. Without their persistent help, the goal of this project would not have been realized.

Besides my advisors, I would like to thank the rest of my thesis committee: Assoc. Prof. Dr. Thaned Satiennam, Assoc. Prof. Dr. Wichuda Satiennam, Asst. Prof. Dr. Buratin Khampirat, Assoc. Prof. Dr. Sajjakaj Jomnonkwao, Asst. Prof. Dr. Duangdao Watthanaklang, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thank also goes to Dr. Sutthipong Meeyai, who provided me an opportunity to join their team as teaching assistance, and who gave access to the laboratory and research facilities. Without his precious support it would not be possible to finish this research.

I thank my fellow lab mates in for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. Also, I thank all my friends in the following institution

Nuttrakit Loacharoen, Pongpat Jodnok, Monthira Phamornmongkhonchai, Wimon Laphrom, Phinyapach Amornpanvattana, Chatchawan Ausrivong, Sum Sonita, Thanapong Champahom, Đissakòón Chonsalasin, Fha Hantanong, Palaphorn Chaimuang, Me-ii Methawadee, Luu Thi Mai Vy, Jubjang Zoe, Chinnakrit Pond, and all of them who could not be mentioned here.

Last but not the least, I would like to thank my family, my parents, and to my brothers and sister for supporting me spiritually throughout writing this thesis and my life in general.

Wecka Imam Yudhistyra

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

LIST OF TABLES (Continued)

Table **Page**

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF ABBREVIATIONS

SCM	=	Supply Chain Management
BDA	=	Big Data Analytics
4IR	=	The Fourth Industrial Revolution
AI	=	Artificial Intelligence
CRISP-DM	=	Cross-Industry Standard Process for Data Mining
CARMA	=	Continuous Association Rule Mining Algorithm
R	=	Recency
F	=	Frequency
M	=	Monetary

CHAPTER I

INTRODUCTION

1.1 Background

In recent year data is everywhere. The evolution of information technology and systems have affected the data which is increasing explosively and growing at a rapid pace (some called the big data or data deluge). Based on the infographic which is distributed from various sources, in 2018 there are 7.593 billion total population; 4.021 billion internet users; 3.196 billion active social media users, 5.135 billion unique mobile users, and 2.958 billion active mobile social media users combined with the statistical information in 2018 from Forbes (Marr, 2018) where there are approximately 2.5 quintillion bytes of data generated every day; on average, more than 40,000 searches every second (or approximately 3.5 billion searches per day) is now processed Google; every minutes in social media-527,760 photos are shared using Snapchat, more than 120 professionals use LinkedIn, 4,146,600 YouTube videos are viewed, 456,000 tweets are sent on Twitter, 46,740 photos are posted using Instagram; some Facebook intriguing statistics-1.5 billion people are busy on Facebook everyday, for example there are more than 307 million users using Facebook in Europe, five new Facebook profiles created every second, more than 300 million pictures get uploaded per day, every minute 510,000 comments are posted and 293,000 statuses are updated; there are 4.7 trillion photo stored every minutes data generated from digital services providers-the Weather Channel gets 18,055,556 forecast requests, the Venmo processes $51,892 peer-to-peer transactions, Spotify adds 13 new songs, Uber riders take 45,788 trips, there are 600 new page edits to Wikipedia; around 33 million voice-first devices are in circulation, 8 million people use voice control every month, the frequency of using voice search queries in Google for 2016 were up 35 times over 2008.

The availability of big data combined with predictive analytics is having an influence on a variety of domains, ranging from baseball where the Oakland A's team (US) utilized unconventional method to build a winning baseball team when the team exerted predictive analytics to spend less money for more accomplishments (Waller & Fawcett, 2013a); Walmart made a contribution to folklore among disaster relief experts by conducting a research on point-of-sale data from shops that located in places where hurricanes were likely to happen. When the logic suggests that deaing with disaster, people would spend money on saws, shovels, and safety equipment, but the fact is that they purchase Pop-Tarts, among other things, in unusually large quantity based on the real data (Waller & Fawcett, 2013a); Google tried to foretell the timing of flu epidemic geographically based on search term frequency (Mayer-Schönberger & Cukier, 2014); Big data are being utilized to transform medical practice, modernize public policy (Mayer-Schönberger & Cukier, 2014); Amazon, the Seattle-based e-commerce giant, currently makes use of Big Data analytics to anticipate the customers' behaviors so that the products could be shipped to them before any decision to buy was made (Marr, 2013); Various medical project based on big data techniques were launched to revolutionize the way for curing cancer (Fry, 2016; Ibnouhsein, Jankowski, Neuberger, & Mathelin, 2018); Enterprises utilized big data analytics to analyze the customers behaviors and changed the interaction between the contact center and their customer (Slider, 2016); Big data is utilized to develop transit planning (Higgins, 2016); A start-up company exerted big data analytics to develop custom data solutions and technology that are used to track key metrics and spot operational inefficiencies in oil and gas industry (Mitchell, 2016); Enterprises transformed the mainstream economy by investing in big data technologies to embrace it and integrate big data-driven initiatives into their core processes and operations to compete in the future (Bean, 2016; Markman, 2016); to the last when big data has deep implications in logistic and supply chain management (SCM) field (Waller & Fawcett, 2013b; Zhong, Newman, Huang, & Lan, 2016). With a lot of data now available, enterprises in almost every industry are concentrating on utilizing data for competitive benefits (Provost & Fawcett, 2013). Waller and Fawcett (2013b) believe that big data and analytics has significant implications in Logistics and Supply Chain discipline.

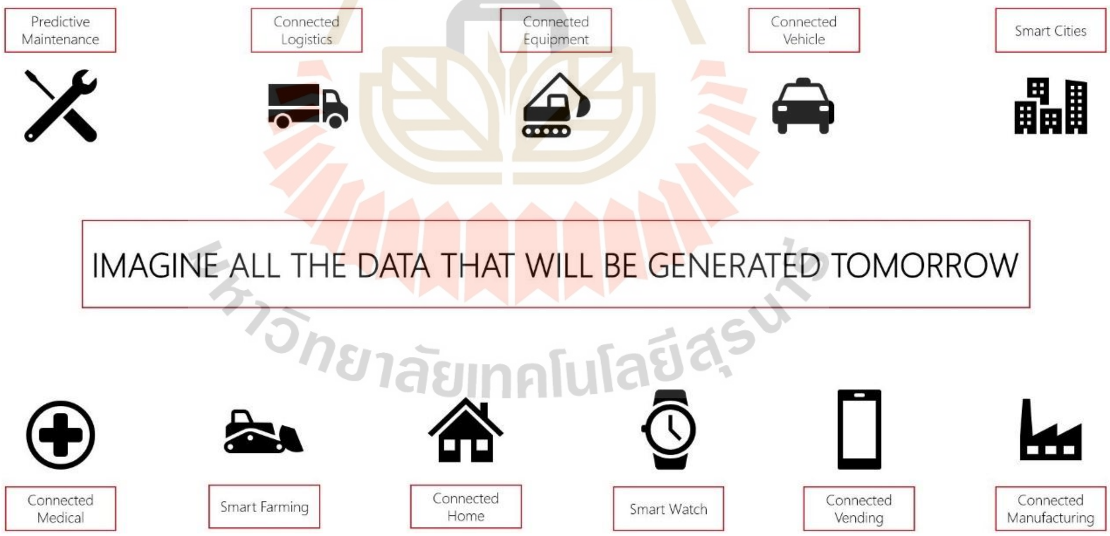


Figure 1.1 Data analysis in the future.

As time goes by, a report from McKinsey (2018), a world-wide famous consultant firm, state that the growth in computer processing power, cloud-storage capacity and usage, and network connectivity has made the current flood of data in most organizations become a tidal wave (i.e. a constant stream of detailed information about personal profiles of customers, sales data annually, product descriptions, process steps, etc.). The data arrive in various formats and from a range of sources, such as Internet-of-Things devices, social-media sites, sales systems, and internal-collaboration systems. Although the rapid increase in some tools and technologies invented to reduce the collection, storage, and evaluation of critical business information, several enterprises are still unclear about the best method to manage these data (Henke & Kaka, 2018). Even in 2014, Walgreens had a $1 Billion forecasting mistakes that made two executives lost their jobs (Trefis, 2014). Furthermore, the main issues underlying this dissertation research are explained as follows. First, as discussed before, managing and analyzing data have always offered the greatest opportunities and greatest benefits for planners, researchers, scientists, and business sectors. Nonetheles, there has not been sufficient research in implementing big data techniques in supply chain and logistics, and the most common result in all areas is that the interest in big data applications has just started, and research on this issue is still scarce(Lamba & Singh, 2017). Second, in view of the fact that the current research paradigm is at the fourth level now where big data is analyzed using various kinds of statistical explorations (Kitchin, 2014), demonstrate the implementation of big data analytics becomes important since it has been enabled novel methods to data generation and analyses to be performed that enable people to inquire information in different ways. Third, by the year of 2020 about 1.7 million Megabytes new information will be produced per second by the people on the planet (Taggart, Anderrson, Kang, & Getty, 2016). In fact, there is a definite shortage of skilled professional people in big data analytics available when there is a big request for data analytics skill at this time, meanwhile the forecast mentions for fantastic increase about 12% into 2024 for data analytics jobs as compared to rise projected to average 6.5% for other jobs (Marr, 2017).

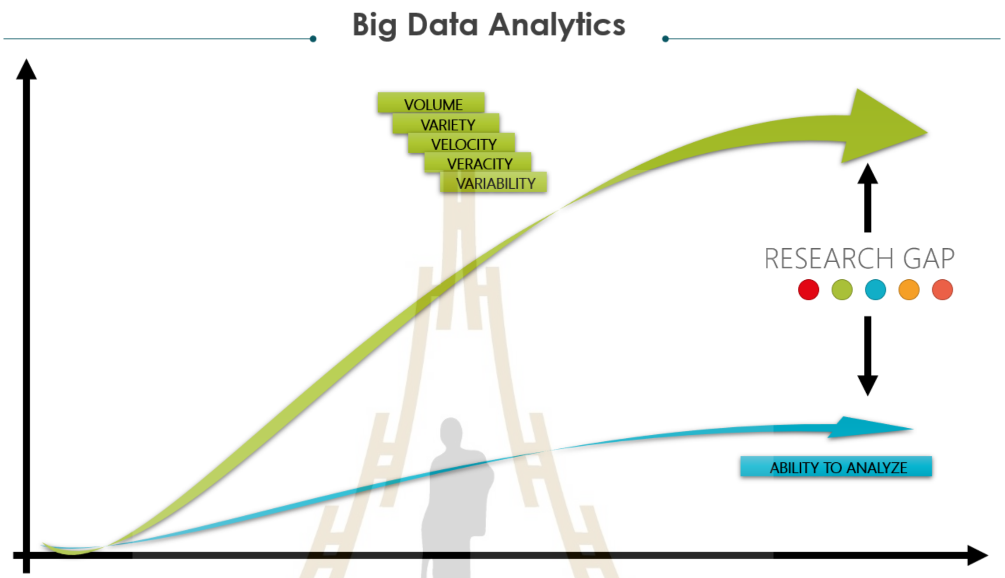


Figure 1.2 Research gap.

1.2 Purpose and Objectives of the Research

The general purposes of this research is to explore, demonstrate, and evaluate the potential of big data techniques implementation in the logistics and supply chain industry. In meeting its objectives, this research is keyed to three issues described as follow:

a) Conduct a comprehensive research and analysis of the current status of big data analytical trends in the logistics and supply chain research communities to support the implementation of data mining and other big data analysis techniques in the logistics and supply chain related fields.

b) Demonstrate the implementation of big data techniques by providing two cases study of the application of big data analytics toward research subjects related customer focused on supply chain management.

c) Evaluate the performance of the result (models and findings) within cases study in logistics and supply chain research area particularly in retail industry.

1.3 Scope of the Research

To explore the implementation of big data analytics in logistics and supply chain area, this research is designed to investigate data analysis techniques that can be utilized to analyze big data. Consequently, several other data analysis techniques and big data technology are outside the scope of this research. The identification of the journals was limited to thirty of the prestigious journals in the logistics and supply chain fields.

Figure 1.3 Scope of the research.

Meanwhile, the scope of this research delimits the intersection of academics, professionals, industries world within logistics, supply chain management, information systems, data science and big data areas as shown in Figure 1.3.

1.4 Significance of the Research

One intended outcome of the study is to reduce the gap between the industries and research communities particularly in logistics and supply chain fields. In other words, the big data analytics discussed in this dissertation research contributes to the development of big data analysis procedure and technique that could help industries and research community to explore the value of their data. This research makes other several contributions as follows:

a) Investigating and providing the statistical and numerical analysis of the implementation of twenty-six big data analytics technique, then performing a time series analysis over the past eight year for the big data scientific article based on the publication from thirty prestigious journals in logistics and supply chain.

b) Providing the application of big data analytics technique to build customers management models that describe the attributes of those customers and build predictive models to predict the future and develop strategies to optimize the management of the customers in a retail industry.

c) Performing an in-depth analysis and comparative study of the descriptive, predictive modelling and other selected data analysis techniques to evaluate each performance of predictive modelling techniques on the case study.

On a practical level, the ultimate goal of the study is to clarify and examine the descriptive and predictive modelling to do with the big data analytics which not only support the conventional prediction but also complete the traditional methods. Specifically, it will contribute to the development of descriptive and predictive analytics procedure with big data analytics to find the appropriate model for different situation and time horizon particularly those associated within customer relationship management for retail industries.

1.5 Contribution of the Research

Since one of the main focus of the of this research is in customers management, it offers some contribution in today's competitive world related to maximize the customer satisfaction, optimize the inventory management, promotion offering, competitive pricing, and implementation of innovative technological solution particularly in retail industries. To be more specific, the contributions of the research are mentioned as follow.

a) This manuscript can provide insight into the current situations and trends of BDA in logistics and supply chain communities around the world and researchers can be aware and realize their current position to minimize the gap between big data available now and the capability to analyze it particularly in a developing country like Indonesia.

b) This manuscript presents analytics with a large amount of data to help a company strengthen its position to stay ahead of the competition.

c) This manuscript minimizes any investment risk on researching and testing the market, product, or ideas.

d) This manuscript facilitates strategic planning based on evidence of BDA to achieve the business goals of a company.

e) This manuscript helps the company in spotting emerging trends on customer needs and demands; and

f) This manuscript has impact on changing seller-buyer relationship/culture in a company.

1.6 Organization of the research

This research is divided into 5 chapters as follows:

Chapter I: The underlying reason of this research, the purpose and objective of this research, the scope of this research, the significant of this research, and the contribution of this research are mentioned in this chapter.

Chapter II: Exploring big data research. This chapter reviews scientific articles published from 2010-2018 in logistic and supply chain community. This chapter also provide insight into the current situations and trends of BDA in logistics and supply chain communities around the world and researchers can be aware and realize their current position to minimize the gap between big data available now and the capability to analyze it particularly in a developing country like Indonesia.

Chapter III: This chapter aims to discover meaningful patterns and ensure high quality of knowledge discovery focused on market basket analysis from the big data available in a company in Indonesia. It can provide new insight into what a company should offer as better incentives to its customers and what strategies a company should develop to allocate the resources appropriately. The combination of CRISP-DM and key step for customer analysis is performed as a methodology in this chapter.

Chapter IV: This chapter focuses on customers to know precisely the characteristics of each customer and know the best way to market the products to the customers immediately by segmenting the customers. The methodology used in this research is CRISP-DM and key steps for customer analysis combined with data mining and behavior scoring method.

Chapter V: Conclusions and recommendations. This chapter summarizes the results from Chapter II to Chapter IV, and identify recommendation for future research.

1.7 References

Bean, R. (2016). [Why Cultural Change Is Necessary For Big Data Adoption].

Fry, E. (2016). [Will Big Data Cure Cancer?].

Henke, N., & Kaka, N. (2018). **Analytics Comes of Age**. Retrieved from London, Mumbai: https://www.mckinsey.com/~/media/McKinsey/Business Functions/c Kinsey Analytics/Our Insights/Analytics comes of age/Analytics-comes-of-age. ashx.

Higgins, N. (2016). [Big Data: Key to Customer Understanding and Service Delivery].

Ibnouhsein, I., Jankowski, S., Neuberger, K., & Mathelin, C. (2018). The Big Data Revolution for Breast Cancer Patients. **European Journal of Breast Health**. 14(2). doi:10.5152/ejbh.2018.0101.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. **Big Data & Society**. 1(1). doi:10.1177/2053951714528481.

Lamba, K., & Singh, S. P. (2017). Big Data in Operations and Supply Chain Management: Current Trends and Future Perspectives. **Production Planning & Control**. 28(11-12): 877-890. doi:10.1080/09537287.2017.1336787.

Markman, J. (2016). [Big Data Unleashes Business Opportunity].

Marr, B. (2013). [The Awesome Ways Big Data Is Used Today To Change Our World].

Marr, B. (2017). The 6 Top Data Jobs In 2018. Retrieved from https://www. forbes. com/sites/bernardmarr/2018/05/09/the-6-top-data-jobs-in-2018/#6f7cf63e430d.

Marr, B. (2018). **How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read**. Forbes. https://www.forbes.com/sites/bernardmarr/ 2018/05/21/how much data do we ceate every day the mind blowing statsevery one should read/#6145c9b560ba.

Mayer-Schönberger, V., & Cukier, K. (2014). **Big Data : a Revolution That will Transform how we Live, Work, and Think**: Eamon Dolan/Mariner Books.

Mitchell, J. (2016). [This Startup Is Using Big Data And Smart Tech To Save The Oil Industry].

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**. 1(1): 51-59. doi:10.1089/big.2013. 1508.

Slider, V. (2016). [Re-thinking Analytics: How big data analytics can revolutionise your contact centre].

Taggart, K., Anderrson, G., Kang, M., & Getty (Writers). (2016). What Exactly Is Big Data? In: Forbes.

Trefis, T. (2014). [Two Key Executives Leave Walgreen Due To a $1 Billion Forecasting Error].

Waller, M. A., & Fawcett, S. E. (2013a). Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. **Journal of Business Logistics**. 34(4): 249-252. doi:10.1111/jbl.12024.

Waller, M. A., & Fawcett, S. E. (2013b). Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management. **Journal of Business Logistics**. 34(2): 77-84. doi:10.1111/jbl. 12010

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big Data for Supply Chain Management in the Service and Manufacturing Sectors: Challenges, Opportunities, and Future Perspectives. **Computers and Industrial Engineering**. 101: 572-591. doi:10.1016/j.cie.2016.07.013.

CHAPTER II

EXPLORING BIG DATA RESEARCH: A Review of Published Articles from 2010 to 2018 related to Logistics and Supply Chains

2.1 Abstract

The emergence of industrial revolution 4.0, digital era, social media networks, ride-sharing transportation networks, video on demand, 5G network technology, and other cutting-edge technologies has transformed the current research paradigm in science into the fourth stage where "big data" is analyzed using various statistical explorations. Conversely, there are still deficiencies in scientific articles exploring "big data" in the logistics and supply chain research communities. Thus, the objective of this study was to conduct a comprehensive review of the trends and the current status of big data analytics (BDA) in the logistics and supply chain research communities. An elaborate examination was performed by analyzing the big data papers published in logistics and supply chain journals. The results have uncovered three things: the changing roles of BDA; some daunting challenges of implementing big data analytics, along with its recommended solutions; and the promising future of logistics and supply chains. All the findings would be helpful for creating a good research design and for implementing big data analytics (BDA) in the related fields.

2.2 Introduction

Today data is everywhere all the time. It is an era where data is influenced by the developments of information technology and systems that increase extensively. Some people call it big data or data deluge (Ghorpade-Aher et al., 2016). The availability of big data has affected many areas, ranging from sports, where the Oakland A's team (United States) used an unconventional approach to build a winning baseball team using predictive analytics to win more games with less budget (Waller & Fawcett, 2013a); precaution case where Google sought to predict the timing of flu outbreaks geographically based on search term frequency (Mayer-Schönberger & Cukier, 2014); to the field of logistics and supply chain management (SCM) when big data has deep implications (Waller & Fawcett, 2013b) (Zhong et al., 2016). In case of the SCM field, big data is providing supplier networks with greater data accuracy, clarity, and insights, leading to more contextual intelligence shared across supply chains (Columbus, 2015).

Other scientific articles reviewing big data in logistics and supply chain exist, and some of them are the following: (1) Tiwari et al. (2018) tried to provide guidance for academicians and practitioners in utilizing BDA in different aspects of supply chains. However, their research solely contributed a general guideline that did not touch the deep aspects of big data utilization, and the result of this research was remarkably narrow since it was highly sensitive to the keyword input to the search; (2) Wang et al. (2016) observed and emphasized that managers must apprehend big data business analytics and supply chain analytics as the strategic assets that must be integrated in all business activities to enable integrated corporate business analytics. However, this research did not insert data from the practitioners and professionals' point of view; (3) Arunachalam et al. (2018) offered a systematic literature review about the capabilities of BDA in supply chain and developed the capabilities for an advance model. The result stimulated the academic researchers to initiate a new empirical research in BDA in the context of supply chain domain and provide a direction for future research. Regrettably, the conclusion was eminently general; (4) Brinch et al. (2018) conducted a Delphi study to gain an understanding about the terminology of big data and identify the application of big data in SCM using adjusted references or the framework of supply chain operations reference. The result of the study showed that the terminology of big data seemed to be more about collecting data rather than managing and utilizing data. Moreover, their research showed that supply chain executives seemed slow to adopt big data. The weakness of this study was on the method of the Delphi study, which was still on small-scale surveys; (5) Lai et al. (2018) discussed the factors that determined the motives of companies in adopting BDA in their daily operations. The empirical results of this study revealed that the benefits and support from the top management can significantly influence the intention to adopt BDA. Since the study of big data was conducted in the early phase, the interpretation of BDA may have multiple diversions from different perspectives, causing some ambiguities in apprehending the meaning of big data.

Meanwhile, the purpose of this manuscript was to explore and analyze current conditions and trends regarding the implementation of BDA in logistics and supply chain research communities. This is important inasmuch as (a) the current research paradigm is at the fourth level where big data is analyzed using various kinds of statistical methods (Kitchin, 2014), this manuscript can provide insight into the current situations and trends of BDA in logistics and supply chain communities around the world; (b) researchers can be aware and realize their current position to minimize the gap between big data available now and the capability to analyze it (Marr, 2017), particularly in ASEAN region (Pujawan, 2016). By 2020, about 1.7 million megabytes of new information will be created per second for every human being on this planet (Taggart et al., 2016). In fact, there will be a definite shortage of skilled professionals in BDA available, while there will be a big demand for data analytics skill at this time. Meanwhile, the forecast calls for fantastic growth estimated to be 12% in 2024 for data analytics jobs as compared with growth projected to average 6.5% for other jobs (Marr, 2017).

2.3 Research Methodology

2.3.1 Data Collection

Because of limited accessibility rights to collect scientific articles, this study focused on the journals indexed by Scopus on Elsevier's database with "logistics" or "supply chain" on its "title." To ensure the quality of the data collection, several journals that had been discontinued were removed, and they were filtered in terms of journal ranking, citations, and relevance to logistics and SCM for each subarea from January 1, 2011, to December 30, 2018. Although the definition of big data has been introduced since 2001, the focus of this study was the papers related to big data published over the past eight years in the journals, and this was in line with the results of a number of previous literature review studies that discovered that scientific articles related to big data were mostly published after 2010 (Tiwari et al., 2018) (Govindan et al., 2018) (Addo-Tenkorang & Helo, 2016) (Barbosa et al., 2018) (Wamba et al., 2018).

2.3.2 Methodology and Method

As explained earlier, this study explored, investigated, and conducted a comprehensive analysis of the current status and trend of the application of BDA techniques in logistics and SCM; therefore, the methodology for this research can be illustrated in Figure 2.1. The first part of the research methodology began with a literature review to reconceptualize the definition of big data, logistics, and supply chains. Rowley and Slack's (2004) framework for literature review was adopted to be the guidance for conducting literature review processes. Likewise, these guidelines were used in recent scientific articles by Chen et al. (2014) and Wang et al. (2016). The second part of the research methodology was the journal searching containing "logistics" or "supply chain" as keyword titles. Next, the journals that had been discontinued and had no citations were removed from the journal searching result. Some journals have changed their names and are still publishing articles; however, there are also a number of journals that have stopped publishing scientific articles. This will help improve the accuracy of the article search process in journals.

The third part of the research methodology was the journal filtering in which this study used 26 BDA techniques listed by the McKinsey Global Institute (Manyika et al., 2011) as shown in Table 2.1 (equipped with the explanation and description of each technique). The fourth part of the research methodology was the paper searching. The searching process in this step (for big data techniques used in each journal) was conducted manually by writing at least one of the 26 big data techniques and several combinations of keywords (including the writing in the title or the abstract) to distinguish the techniques included in the articles. Afterward, the journal was saved and manually reviewed to observe current conditions and trends in the future.

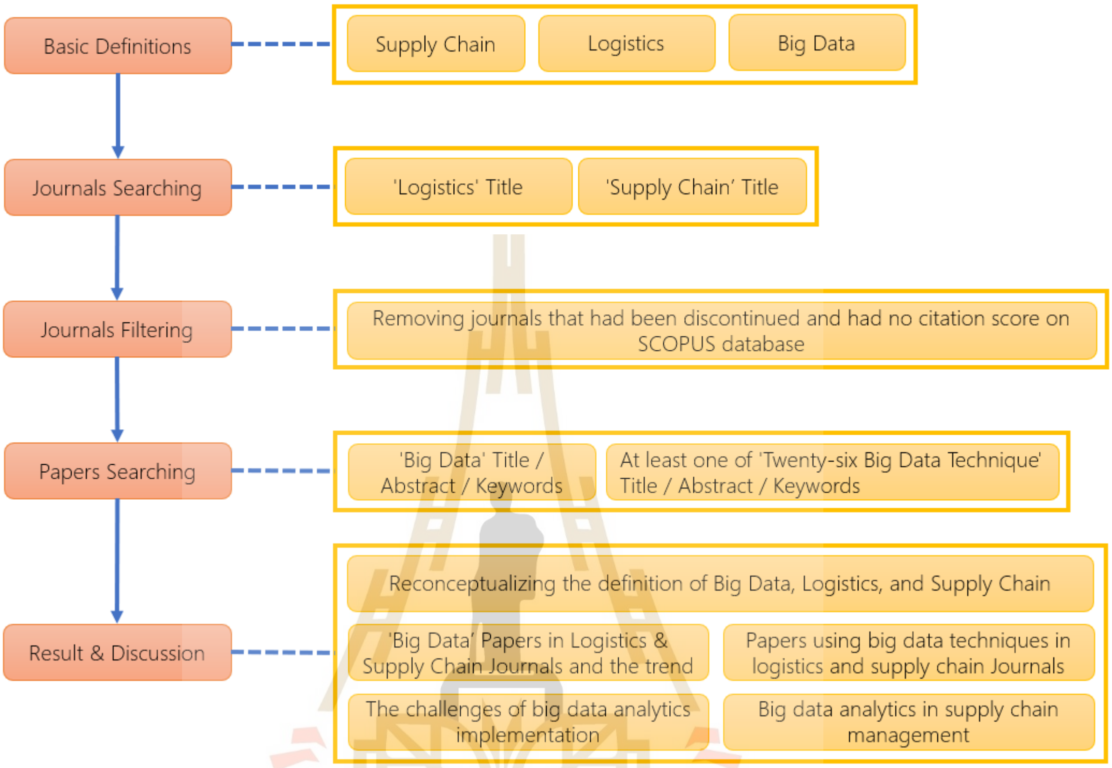


Figure 2.1 Methodology.

Table 2.1 Big data techniques by McKinsey (Manyika et al., 2011).

Technique	Definition
A/B testing	This is often referred as split testing, bucket testing, or A/B/N testing (if using multiple variants in testing). This is a technique for comparing two variants with various tests to be able to see which one has better performance to determine what treatment increases the performance of the variant.
Association rule learning	This is a common technique to find an interesting relationship among many variables in a large database using a variety of algorithms to generate and test the possible rules. It is also referred as market basket analysis used for data mining.
Classification	This is a set of techniques to identify the categories of data point in a collection for targeting categories in which data point belong to, based on data points that have already been categorized.
Cluster analysis	This a statistical method to classify objects into groups of similar objects where the previous characteristics of similarity are unknown.
Crowdsourcing	A combination of "crowd" and "outsourcing," this technique collects data and ideas from a large group of people or community rather than an organization or own employees through an open call, usually through networked media such as the web.
Data fusion and data integration	This is a technique that integrates multiple data sources to obtain more consistent, accurate, and useful information rather than using data provided by individual data source.

Table 2.1 Big data techniques by McKinsey (Manyika et al., 2011) (Cont.).

Technique	**Definition**
Data mining	This is a technique of discovering hidden patterns from a large data set by combining methods from statistics, machine learning, and database management to analyze and extract the pattern.
Ensemble learning	This is a technique that uses multiple learning algorithms and predictive modeling to obtain better predictive performance than can be obtained from any of the constituent models. This is a type of supervised learning.
Genetic algorithms	This is a method for solving and improving both constrained and unconstrained optimization problems and is inspired by the process of natural evolution or "survival of the fittest."
Machine learning	This is a subspecialty in the scientific study of algorithms, statistical models, and computer science (within a field historically called artificial intelligence [AI]) concerned with the design and development to improve performance from a specific task.
Natural language processing	This is a set of techniques from a subfield of information engineering, computer science (within a field historically called AI), and linguistics that use computer algorithms concerned to analyze human (natural) language, in particular, focus on how to use a computer to analyze and interpret a large amount of human (natural) language from a large data set.

Table 2.1 Big data techniques by McKinsey (Manyika et al., 2011) (Cont.).

Technique	Definition
Neural networks	These are powerful computational data models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find and represent patterns in data from complex input/output relationships.
Network analysis	This is a set of techniques that study the graph used to characterize relationships among discrete nodes that have attributes in a graph or a network. Connections among individuals in a community or organization are analyzed in social network analysis.
Optimization	This is a set of numerical and statistical techniques used to redesign complex systems and processes to improve their performance according to one or a number of objective measures (e.g., cost, speed, or reliability).
Pattern recognition	This is often referred as classification techniques with a set of machine learning techniques that assign some sort of output value (or label) to a given input value (or instance) according to a specific algorithm.
Predictive modeling	This is a set of process in which a mathematical model, data mining, and probability are created or chosen to forecast an outcome at its best. Regression is one part of the many techniques of predictive modeling.
Regression	This is a set of statistical techniques often used for forecasting or predicting to determine how the value of the dependent variable changes when one or more independent variables are modified.

Table 2.1 Big data techniques by McKinsey (Manyika et al., 2011) (Cont.).

Technique	Definition
Sentiment analysis	This is often referred as opinion mining, where the application of natural language processing, computational linguistics, and other analytic techniques extract and identify subjective information from source text material.
Signal processing	This is a set of techniques from electrical engineering and applied mathematics to process the “signals” (i.e., radio signals, video, text, sounds, and images) to become meaningful variables that can be extracted and structured before they can be used in data analysis.
Spatial analysis	This is a set of techniques applied to structure at human scale from statistics and analyze the topological, geometric, or geographical properties encoded in a data set.
Statistics	This is a set of techniques including the design of surveys and experiments of the collection, organization, and interpretation of data. This technique overlaps with many other techniques in this research.
Supervised learning	This is a set of machine learning techniques to infer a function or a relationship from the large training data. Classification and support vector machines are examples of these techniques. These are different from unsupervised learning.
Simulation	This is a more human alternative to understand the business problem form modeling the behavior of complex systems that can be used for forecasting, predicting future trends, planning scenarios, and recommending optimum decisions.

Table 2.1 Big data techniques by McKinsey (Manyika et al., 2011) (Cont.).

Technique	Definition
Time series analysis	This is a set of techniques that has both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract the meaningful statistics and other characteristics of the data.
Unsupervised learning	This is a class of machine learning techniques to find a hidden structure or pattern in unlabelled data. Cluster analysis is an example of unsupervised learning (in contrast to supervised learning).
Visualization	This is a general term or technique for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses to help people understand the significance of data by placing them in visual context.

The last part of the research methodology was the result and discussion of this manuscript, which explained the reconceptualization of SCM in the industrial 4.0 era, the statistical data of manuscripts about the current situations and trends with BDA topics, the challenge of BDA implementation, and some knowledge discoveries or insights based on the literature review conducted in this research.

2.4 Result and Discussion

2.4.1 Logistics and Supply Chain

Primarily, it is important to know the definition and the scope of logistics and supply chain in this research since they are often used interchangeably, and many other terms such as material management, demand management, procurement, customer relationship management, manufacture inventory, packaging, and so on have similar definitions. Some claim that SCM is just another name for integration of business logistic management, and the broad scope of SCM has been promoted over the years (i.e., SCM may be concerned with product pricing and manufacturing quality). Although SCM promotes viewing the supply channel with the broadest scope, the reality is that enterprises do not practice this ideal. A single enterprise generally is unable to manage its entire product flow channel from raw materials to customers, and usually, it has a narrow scope. For example, the maximal managerial control that can be expected is over (a) the immediate physical supply channel, which is the time and space gap between the firm's immediate material sources and its processing points; and (b) the distribution channel, which refers to the time and space gap between the firm's processing points and its customers (Ballou, 2007). Because of the similarities in the activities between the two channels, physical supply (referred to material management) and physical distribution combined integrate those activities into business logistics. Afterward, business logistics developed into SCM as shown in Figure 2.2. Other terms such as value nets, value stream, and lean logistics describe similar scope and purpose with SCM.

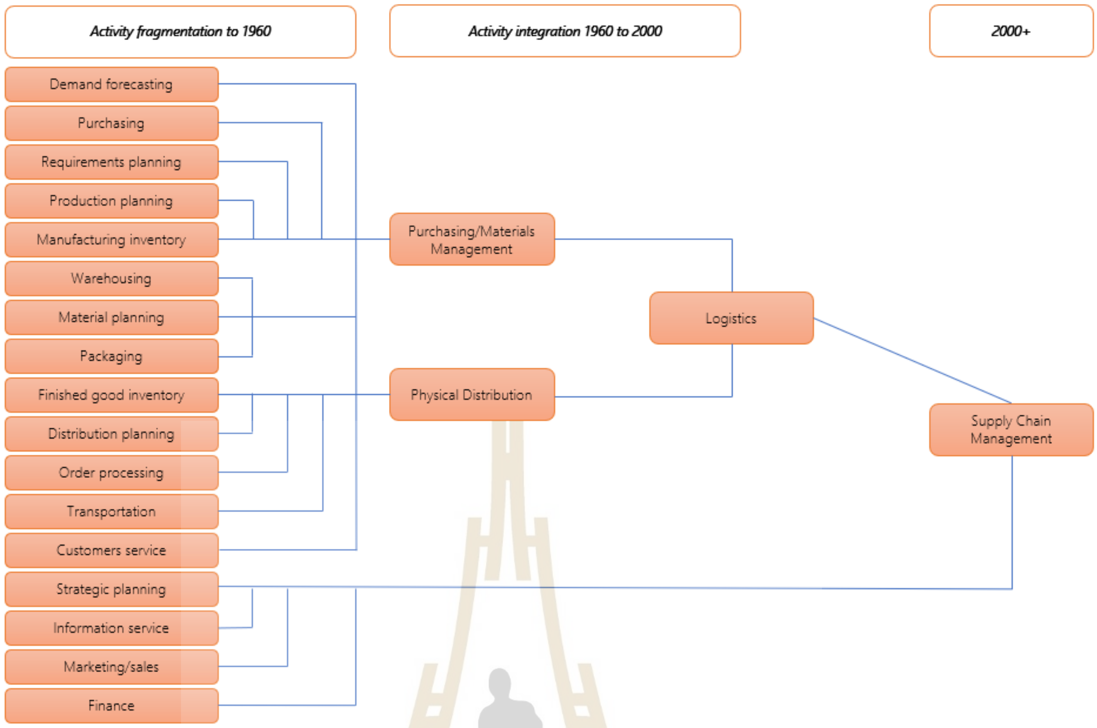


Figure 2.2 Supply chain evolution by Ballou (2007).

Among the previous terms, there were no absolute definitions because the evolution and the background of logistics and SCM come from a set of functional activities (transportation, storage management, information services, etc.). Many manuscripts explained the definition of SCM based on linguistics terminology and a unified definition. In the writing of Jules Dupoit, a French engineer, the idea of trading one cost for another (transportation costs for inventory costs) was evident in the selection between road and water transport (Ballou, 2007).

The fact is that carriage by road being quicker, more reliable and less subject loss or damage, it possesses advantage since businessmen often attach a considerable value. However, it may well be that the saving 0 fr.87 induces the merchant to use the canal; he can buy warehouses and increase his floating capital in order to have a sufficient supply of goods on hand to protect himself against slowness and irregularity of the canal, and if all were told the saving of 0 fr.87 in transport gives him/her advantage of a few centimes, he/she would favor of the new route (Dupuit, 1952).

Around 1961, the first textbook to suggest the benefits of coordinated logistics management was introduced. Logistics is a branch of military science that deals with procuring, maintaining, and transporting materials, personnel, and facilities (Dupuit, 1952) (Pienaar, 2009). This definition put logistics into a military description. A better representation of its definition was described by the Council of Logistics Management, a professional organization of logistics managers, educators, and practitioners formed in 1962 for the purpose of continuing education and fostering the interchange of ideas.

Logistics is that part of the supply chain process that plans, implements, and control the efficient, effective flow and storage of goods, services, and related information from the point of origin to the point of consumption in order to meet customers' requirements (Council of logistics management, 2000).

This definition conveys the idea that product flows should be managed from the point where they exist as raw materials to the point where they are finally discarded (Chen et al., 2014). It is also worthwhile to explore several definitions for the scope and content of logistics and supply chain. Chow and Heaver (1999) described it as groups consisting of producers, suppliers, distributors, retailers and transportation, and other logistics management service providers engaged in supplying goods to consumers (Pienaar, 2009). Ayers (2001) defined it as life cycle processes involving physical goods, information, and financial flows to meet the customer needs with goods and services from suppliers (Ayers, 2001). Grant et al. (2006) characterized it as corporate business processes with integrated systems from end users through suppliers that provide information, goods, and services, which add value for customers (Grant et al., 2006). Mentzer et al. (2001) explained it as a group of entities (i.e., organizations or individuals) directly involved in the supply and distribution flow of goods, services, finance, and information from source to destination (customers) (Mentzer et al., 2001). Hanfield (2017) specified it as the activities to maximize customer value and get a sustainable competitive advantage (Handfield, 2017). LeMay et al. (2017) mentioned it as the coordination of a network through organizations and individuals to get, use, and deliver material goods; to acquire and distribute services; and to make their offerings available to markets, customers, and clients (LeMay et al., 2017). Pienaar W. (2009) specified it as a common description of the integration process that involves organizations to transform raw materials into finished goods and to transport them to the end user (Pienaar, 2009). Some researchers argue that there are many readily available definitions of SCM going beyond the fundamental concept of definition. Meanwhile, the definition of logistics and SCM in this research does not focus on the term in the linguistic definition and refers directly to the definition, which is agreed upon by experts as the following (Rushton et al., 2017).

Logistics = Material Management + Distribution

An extension to the logistics idea includes the supply of raw materials and components as well as the delivery of products to the final customer. Thus, supply chain can be defined as follows.

Supply Chain = Suppliers + Logistics + Customers

In general, it can be said that logistics is the part of SCM that manages the flow of materials and information from the supplier and customers to the end consumers. It includes inventory management, warehousing, marketing/sales, customers, management information systems, technology, transportation, and compliance as its key logistics activities. The activities that constitute the supply chain process vary from enterprise to enterprise, depending on an enterprise's particular organizational structure, management's honest differences of opinion about what constitutes the supply chain for its business, and the importance of individual activities to its operation (Ballou, 2007).

2.4.2 Big Data

Big data is a term that describes the large volume amount of data in business (Jeble et al., 2018). Gartner (at the time, it was named META Group), a consultant firm, defined big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Beyer & Laney, 2012). Later, the IBM scientists added two more dimensions of big data-"high-veracity" (IBM, 2013) and "high-variability" (Andreu-Perez et al., 2015). Currently, the definition of big data has five dimensions. The description and explanation of big data with four dimensions on this research referred to the study of Katal (2013), which has a lot of citations. The description of big data with five dimensions is explained in detail below (Katal et al., 2013) (Andreu-Perez et al., 2015).

a) "Volume"-Currently, data that have reached petabyte volume is already problematic. With the increasing utilization of smartphones, internet and web technology, and wearable devices (such as smartwatches, smart glasses, smart home, etc.), data is predicted to increase to zettabyte volume in the next few years.

b) “Velocity”-The flow of data movement is rapid recently. It also causes challenge for data (previously collected and captured) that change faster than before and continuously in motion.

c) “Variety”-When data are collected not only from one specific source but also from various data formats such as web pages, images, sensors, e-mails, social media, and so on, which are structured and unstructured data, this complicates the conventional analysis technique to analyze big data.

d) “Veracity”-The focus in this dimension is the ambiguity within data (typically from noise and abnormality).

e) “Variability”-This refers to the establishment of a regular contextualization structure of data flows that can be relied upon even in conditions that cannot be predicted to be extreme. This is defined as the need to obtain data that are meaningful considering all possible circumstances.

f) The five “V” pillars mentioned above are the core components of big data that can produce a new “V” pillar called value if organizations have the ability to obtain greater valuable information from big data through deeper insight from superior data analytics (see Part 3.4).

2.4.3 Journal Filtering and Big Data Papers Published in Journals

The searching verification was performed to figure out the list of scientific journals indexed by Scopus on Elsevier’s database with “logistics” and “supply chain” keywords. The journals that had been discontinued and did not have a citation score were excluded as shown in Table 2.2.

Table 2.2 Discontinued journal list.

Journal	Year
Cellular Logistics	2015 to 2016
EURO Journal on Transportation and Logistics	2015
Logistics and Transportation Review*	1978 to 1996*
Naval Research Logistics Quarterly*	1979 to 1986*
EURO Journal on Transportation and Logistics	2015 to 2016
Supply Chain Manufacturing and Logistics	2006
Supply Chain Systems Magazine	2002 to 2006

*Note that the Logistics and Transportation Review journal is continued as Transportation Research Part E: Logistics and Transportation Review, while the Naval Research Logistics Quarterly *is continued as* Naval Research Logistics.

As a result, 23 journals had been chosen, where 5 of 19 search results journals with "logistics" and 2 of 11 journals with "supply chain" keywords had been stopped from Elsevier's database. Among the 23 selected journals indexed by Scopus on Elsevier's database, there were merely 43 published papers related to big data or approximately 0.63% of the total papers (an extensive gap from the total 6,929 papers published in the journals). The journal that published the least scientific articles was the International Journal of Construction Supply Chain Management with 16 papers, while the journal that published the most scientific articles was the Transportation Research, Part E: Logistics and Transportation Review with 1,063 papers. The Journal of Business Logistics and the International of Logistics Management were the journals that published the big data papers in the logistics and supply chain area the most with 11 journals using the term "big data" in their title, abstracts, or keywords. Interestingly, no more than 5% of the total papers were published by the two journals, even though they were the most widely published journals related to big data.

Meanwhile, the authors with the most published papers in the journals were Stanley E. Fawcett and Matthew A. Waller with 4 manuscripts (see Appendix A), and most of the authors came from the United States with 38 authors, followed by the United Kingdom with 15 authors, and Australia with 13 authors. Furthermore, scientific articles were published each year: 2 from 2014 to 2015, 5 in 2016, 3 in 2017, and 25 in 2018. Complete statistical data on manuscripts using big data analysis techniques in selected journals and a comparison between the number of big data articles published and the total number of articles published in selected journals are shown in Figures 2.3, 2.4, and 2.5.

A comparison between the number of articles related to big data and the total number of articles published in logistics and supply chain journals per year showed the massive gap one more time (2 big data papers: 752 papers or 0.26% of total paper published in 2013 and 26 big data papers: 1,158 papers or 2.24% of total paper published in 2018). To analyze the trend toward the big data papers published in logistics and supply chain journals, five time series forecasting techniques were performed (i.e., exponential smoothing, linear trend model, quadratic trend model, exponential trend model, and autoregressive model). Four of the five models (except exponential smoothing) used the consecutive coded value 0 (2013) through 5 (2018) as the X (coded year) variable. The result of these models produced the following forecasting equations as shown in Figure 2.6. The R2 and the adjusted R2 for the quadratic trend model were the highest when they were compared with the other models.

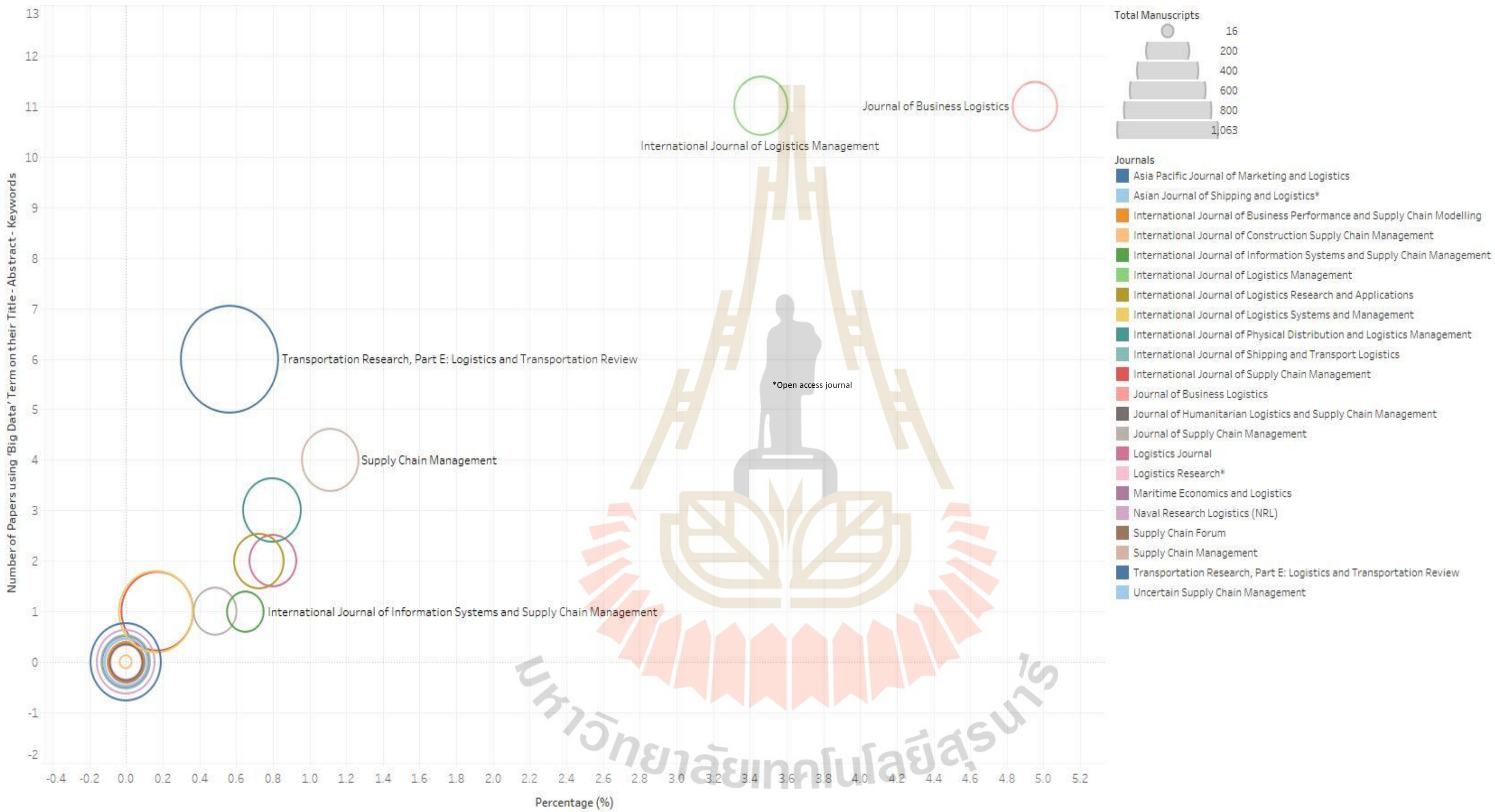


Figure 2.3 Complete statistical data on manuscripts using big data analysis techniques in selected journals.

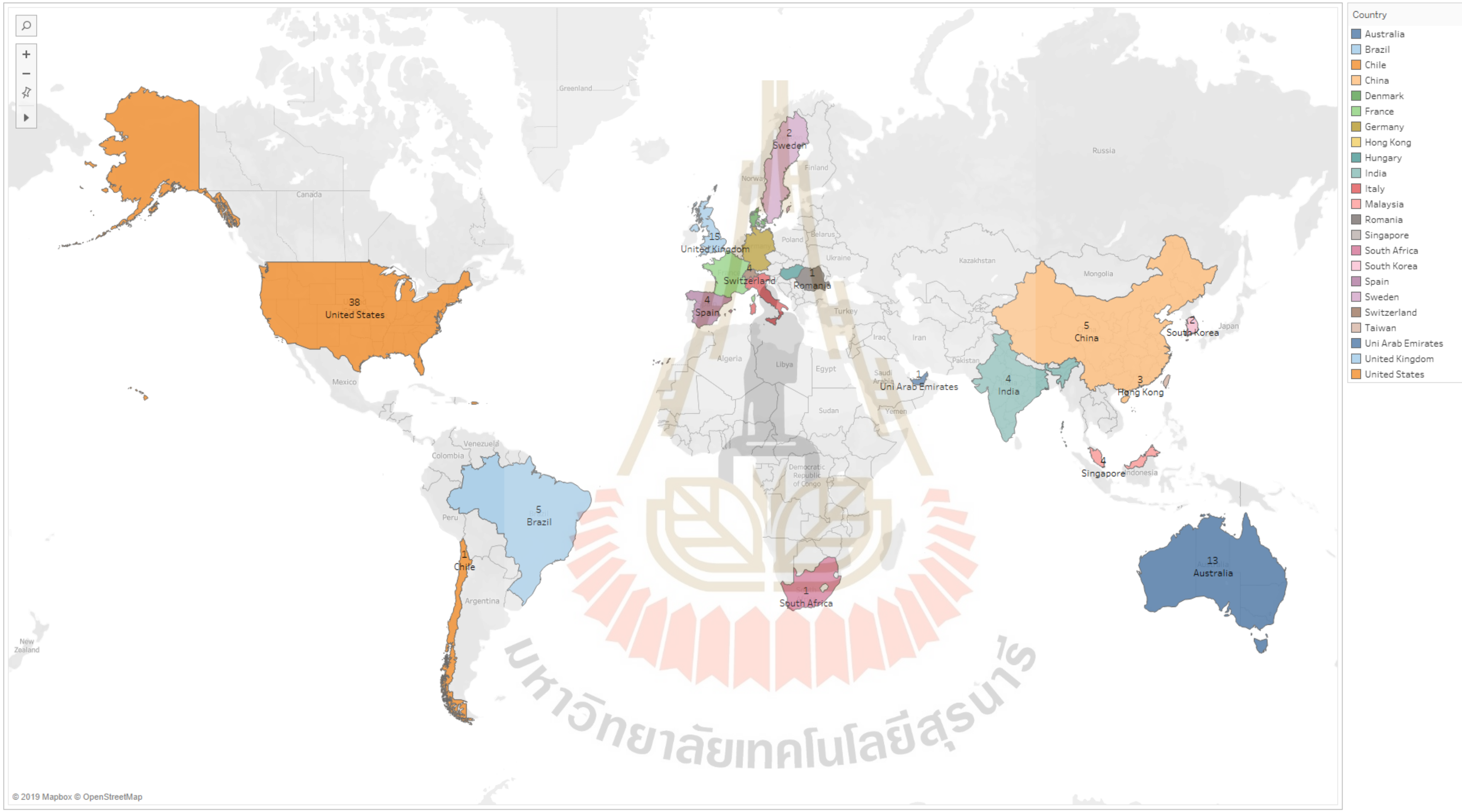


Figure 2.4 Complete statistical data on the distribution of manuscripts using big data analysis techniques in selected journals.

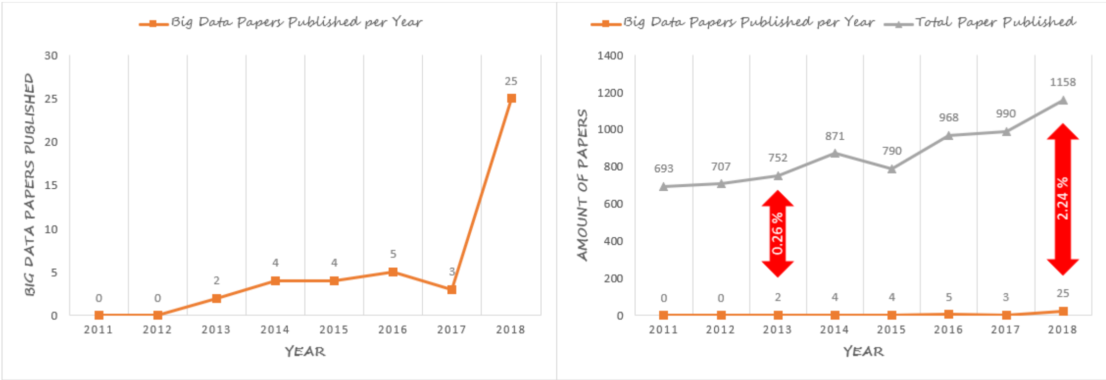
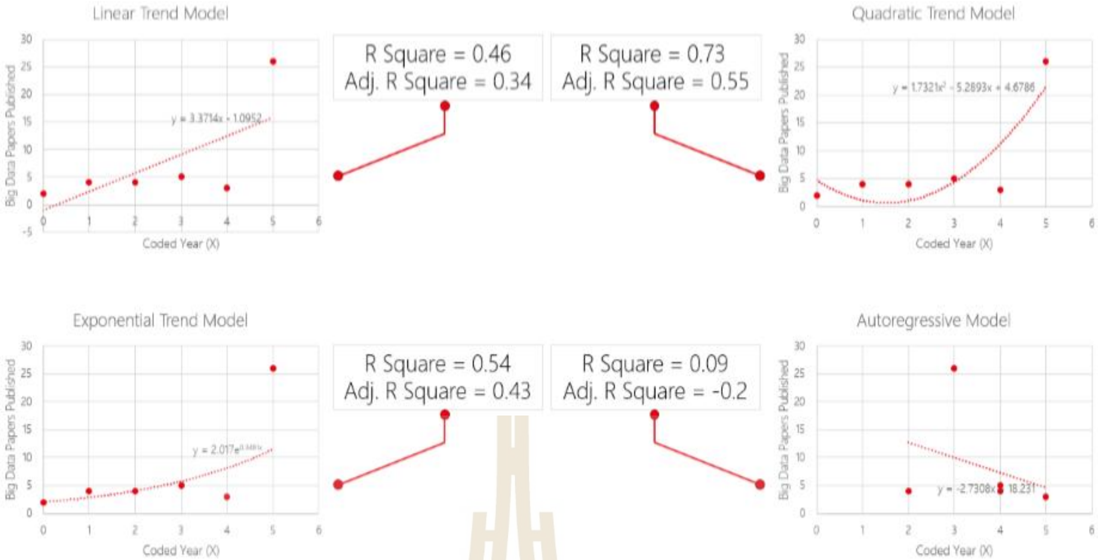


Figure 2.5 Comparison between the number of big data articles published and the total number of articles published in selected journals.

To further verify and to show the actual value of the big data paper published in logistics and supply chain journals, along with the predicted value of the papers, the residual, the error sum of squares (SSE), the standard error of estimates (SYX), and the mean of absolute deviation (MAD) for each of the four models, comparing the magnitude of the residuals in four models was performed. For this time series, the comparison of the SYX and MAD indicates that the first-order autoregressive model provides the poorest fit, and the quadratic trend model is the best model. Using the best model, the result of the analysis for 2019 was 35 papers, which were rather higher than the number of papers from 2011 to 2018 related to big data in logistics and supply chain journals.



		Linear		Quadratic		Exponential		First-Order Autoregressive	
Year	Big Data Papers Published	Predicted	Residual	Predicted	Residual	Predicted	Residual	Predicted	Residual
2013	2	-1.10	-3.10	4.68	2.68	0.30	-1.70	#N/A	#N/A
2014	4	2.28	-1.72	1.12	-2.88	0.46	-3.54	15.50	11.50
2015	4	5.65	1.65	1.03	-2.97	0.61	-3.39	12.77	8.77
2016	5	9.02	4.02	4.40	-0.60	0.76	-4.24	10.04	5.04
2017	3	12.39	9.39	11.24	8.24	0.91	-2.09	7.31	4.31
2018	25	15.76	-10.24	21.54	-4.46	1.06	-24.94	4.58	-21.42
		SSE	224.42	SSE	112.41	SSE	671.28	SSE	712.04
		S_{YX}	7.490	S_{YX}	6.121	S_{YX}	0.290	S_{YX}	10.808
		MAD	5.02	MAD	3.64	MAD	6.65	MAD	10.21

Figure 2.6 Forecasting equations.

Meanwhile, simple exponential smoothing (SES) merely requires deficient data storing, quickness and simplicity in computing, and emphasis on the most up-to-date information that was also exerted to predict the amount of research papers in 2019 related to big data within the scope of logistics and supply chain journals. It is likely that at least 8 papers related to big data will be published in 2019 within the scope of logistics and supply chain journals. Alpha (α) = 0.1 provided better accuracy with its lowest mean squared error (MSE) and MAD score on SES forecasting technique. Every year (from 2013 until the end of 2018), movement on the number of published scientific articles increased (except in 2017). In 2018, there was a drastic rise in the pattern movement with 24 journal papers. The result of the analysis for 2019 was 8 papers, which was higher than the number of papers from 2011 to 2017 related to big data in logistics and supply chain journals. The trend increased from the previous year as shown in Figure 2.7.

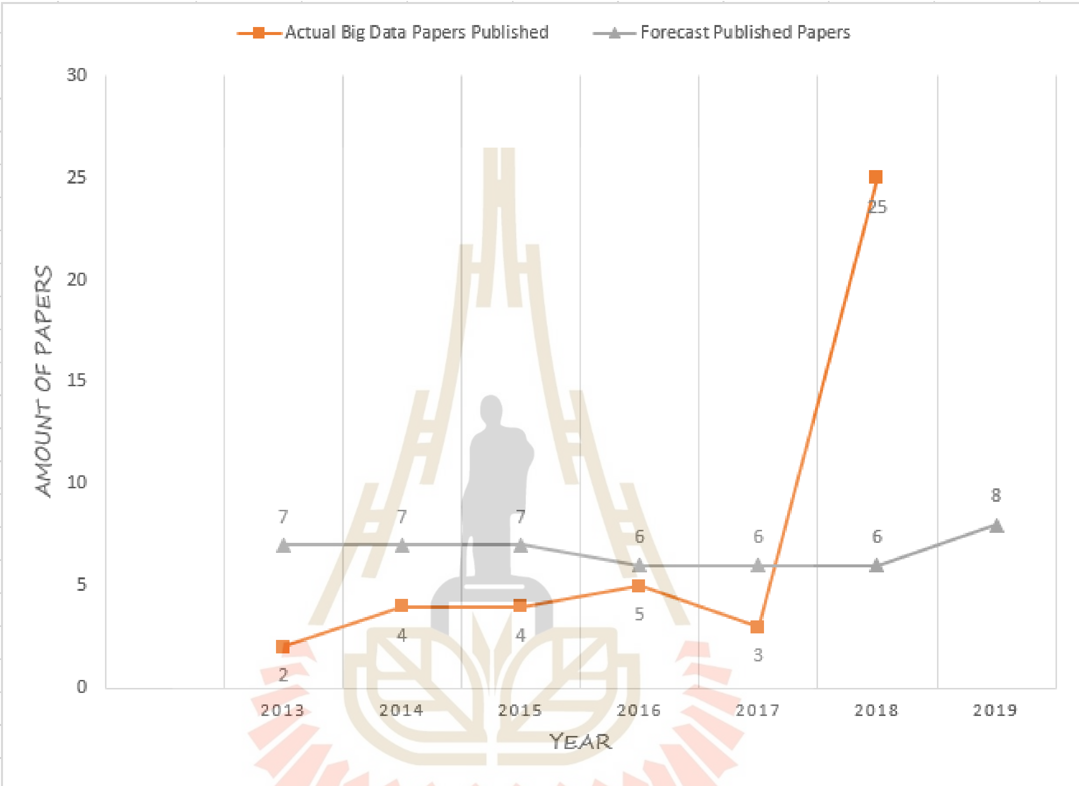


Figure 2.7 Prediction of the big data papers published in 2019 within the scope of logistics and supply chain journals.

Thus, it can be concluded that the research topic related to big data had an extensive gap compared with the other topics within logistics and supply chain journals based on real data and information that had been collected and analyzed.

Furthermore, the result of this study explored the application of each of the 26 big data techniques. It shows that in total, these techniques appeared 1,980 times within the 23 selected journals. It is important to notice that many of the publications applied more than one of these techniques. Table 2.3 presents the selected research areas, the associated journals, and the results of the database search. From the information shown, there were 1,560 scientific articles related to the logistics area and 420 scientific articles related to the supply chain area with a total of about 1,980 scientific articles related to big data techniques used in logistics and supply chain journals. The five most widely used techniques in logistics journals were optimization, simulation, regression, genetic algorithms, and classification (with 1,560 journals). Meanwhile, the five most widely used techniques in supply chain journals were optimization, regression, simulation, classification, and genetic algorithms (with 420 journals). Moreover, Table 2.3 shows the usage of big data techniques per year. Optimization was the most frequently used technique every year in both areas (with 680 scientific articles). From 2011 to the end of 2018, the number of scientific articles that used this technique had increased dramatically. Nevertheless, there were still no scientific articles that used association rule learning, natural language processing, and unsupervised learning.

2.4.4 BDA in SCM

As previously mentioned, "big data" is a term that describes high volume, high variability, high variety, high veracity, and high velocity of data and inundates organizations on a day-to-day basis. The important issue in big data is that it must be combined with a variety of analytical approach (i.e., statistical analysis, optimization, visualization, clustering, etc.) to extract valuable information from the large data set. This combination is referred as BDA as shown in Figure 2.8. The role of BDA in the immense impact of e-commerce, social media, self-driving cars, wearable device, barcode scan, drones, internet, and cloud storage era can transform the

Table 2.3 Amount of papers that used big data techniques from McKinsey per year published in logistics and supply chain journals.

Big Data Techniques	Logistics Journals	Supply Chain Journals	Year 2011	2012	2013	2014	2015	2016	2017	2018
A/B testing	1	0						1		
Association rule learning	0	0								
Classification	70	36	6	9	13	19	10	12	17	20
Cluster analysis	41	13	4	6	7	4	6	9	6	12
Crowd sourcing	1	1						1	1	
Data fusion and data integration	2	2		2				1	1	
Data mining	8	3		4	1		1	2	2	1
Ensemble learning	1	0								1
Genetic algorithms	103	29	16	9	16	14	18	15	18	26
Machine learning	6	0						1	2	3
Natural language processing	0	0								
Network analysis	58	9	4	5	6	11	8	15	6	12
Neural networks	18	4	2	4	2	4		3	5	2
Optimization	570	110	68	69	82	84	79	102	97	99
Pattern recognition	1	1	1							1
Predictive modelling	0	1				1				
Regression	224	105	34	38	36	41	34	47	35	64
Sentiment analysis	3	0						1	1	1
Signal processing	1	1			1		1			
Simulation	363	78	42	49	55	58	45	76	61	55
Spatial analysis	6	1		1	2		1	1	1	1
Statistics	57	19	12	6	9	4	14	6	15	10
Supervised learning	2	0						1		1
Time series analysis	6	1	1	2		1	1	1	1	
Unsupervised learning	0	0								
Visualization	18	6	3	1		4	3	6	5	2

1 102

definition and the management of supply chain. Since the current supply chain cannot be equated with its traditional function (where a single purely operational function was reported to sales or manufacturing and focused on ensuring the supply of production and delivery lines to customers, and it was also combined with other independent SCM functions in several companies), the focus of the SCM function has shifted to advanced processes. First, it enlarges the data set for analysis beyond the traditional enterprise resource planning and supply chain management system internal data. Second, it applies powerful statistical techniques by combining internal and external data sources to analyze. Third, it creates a new approach that improves supply chain decision analysis, strategic choice, and operation optimization.

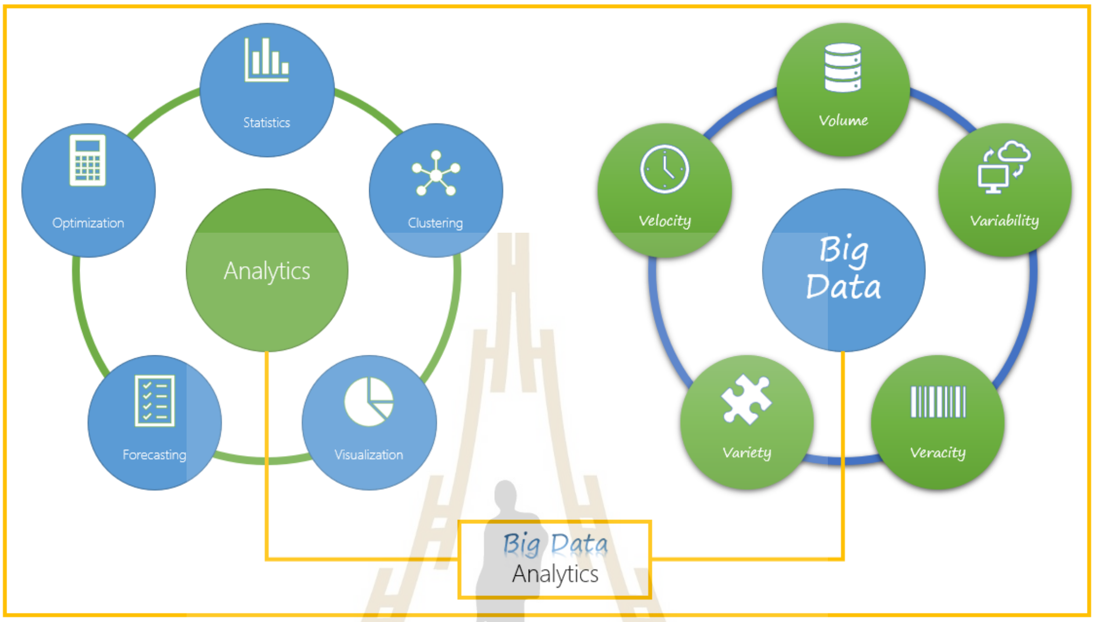


Figure 2.8 BDA.

In relation to the papers publishing the implementation of BDA in logistics and supply chain journals, there were several scientific articles employing big data analytical techniques that were published in the said journals from 2011 to 2018. Based on the data on social media (in these cases, it was Twitter), sentiment analysis techniques, natural language processing, clustering analysis, data visualization, and supervised learning were combined to obtain a rich picture of interaction between the company and customers, to predict market demand, and to find out key issues that affect customer satisfaction (Bhattacharjya et al., 2016) (Bhattacharjya et al., 2018) (Singh et al., 2018). These studies showed the use of a set of tools to gather insights from a large amount of unstructured data from conversations on social media platforms. Bohács and Rinkács (2016) purely performed data visualization to analyze and model the complex systems. To make a good decision in a distributor company, Hamister et al. (2018) integrated data visualization, optimizing modeling, and forecasting safety stock using real data. To improve the value density of raw data is a crucial step that allows analysis to acquire meaningful data, including optimization using data in the real world. Miller et al. (2018) implemented data visualization and statistical approach of big data techniques to devise a framework that examined the services of the supply chain setting of health care with public data regarding hospital-level patient satisfaction.

The deficient number of papers related to the implementation of BDA published in logistics and supply chain journals means there are many opportunities to explore several captivating issues using BDA in the logistics and supply chain areas by the research communities. Some of the main areas of logistics and supply chains that can be explored using BDA are as follows.

a) Production, inventory, and operation planning can be improved using BDA in planning processes and demand capability. For example, the visibility of point of sales data, inventory data, and production planning data can be analyzed to identify mismatch between supply and demand with real-time analysis. By incorporating temperature and data about sunny days, ice cream companies are able to predict the demand more accurately. In addition, other companies are able to perform the similar thing and drive actions such as price changes, the timing of promotions, shaping demand, or the addition of new lines based on the customer behavior in the right place for the right time to realign things with BDA.

b) In workload optimizations and picking zone allocation in warehouses based on efficiency-for example, to improve storage efficiency and picking productivity-a big data 3-D model visualization analysis can be used to model and simulate new configuration. Sensor data in warehouses can be exerted to anticipate real-time transportation requirement.

c) Connecting real-time routing allocation between transportation and warehouses is utilized to improve the operational optimization such as trucks that use consumption fuel analytics to enhance driving efficiency and use GPS to reduce waiting time in warehouses in real time.

d) Routing optimization for transportation vehicles based on traffic congestion, weather, and driver characteristics is easily analyzed with BDA before the era of autonomous truck comes.

e) Real-time out-of-stock detection and prevention for point of sales are aligned with the production planning analysis. The combination between point of sales and consumer behavior data is used to improve inventory management.

f) Consumers' sentiment data on social media can be used to obtain a rich insight into the interaction between enterprise and consumers, and risk management can be used for fraud detection and to provide credit rating to define payment term offered to mitigate the risk.

In the future, the development of automated production at factory, autonomous truck, automated warehouse, shipping rerouting, drone technologies, and AI ecosystems will lift SCM to the next level of management and definition. Automated production machine will provide constant feedback on production capacity based on sales information and consumer behavior. Driverless trucks with live transit location via satellite get in touch with good movements. Operations in warehouses such as picking and transporting good and information flow of status will be handled by machines. Customers will be able to automatically reroute their delivery address to new information on their smartphones. Drones will deliver and pick the goods to the border areas. AI will perform an international network analysis, where, related to law, international transactions, international contracts or agreements, legal systems or blockchain systems, and everything humans perform currently are predicted being replaced by AI in the future. Figure 2.9 illustrates the previous, the current, and the future of supply chain concept.

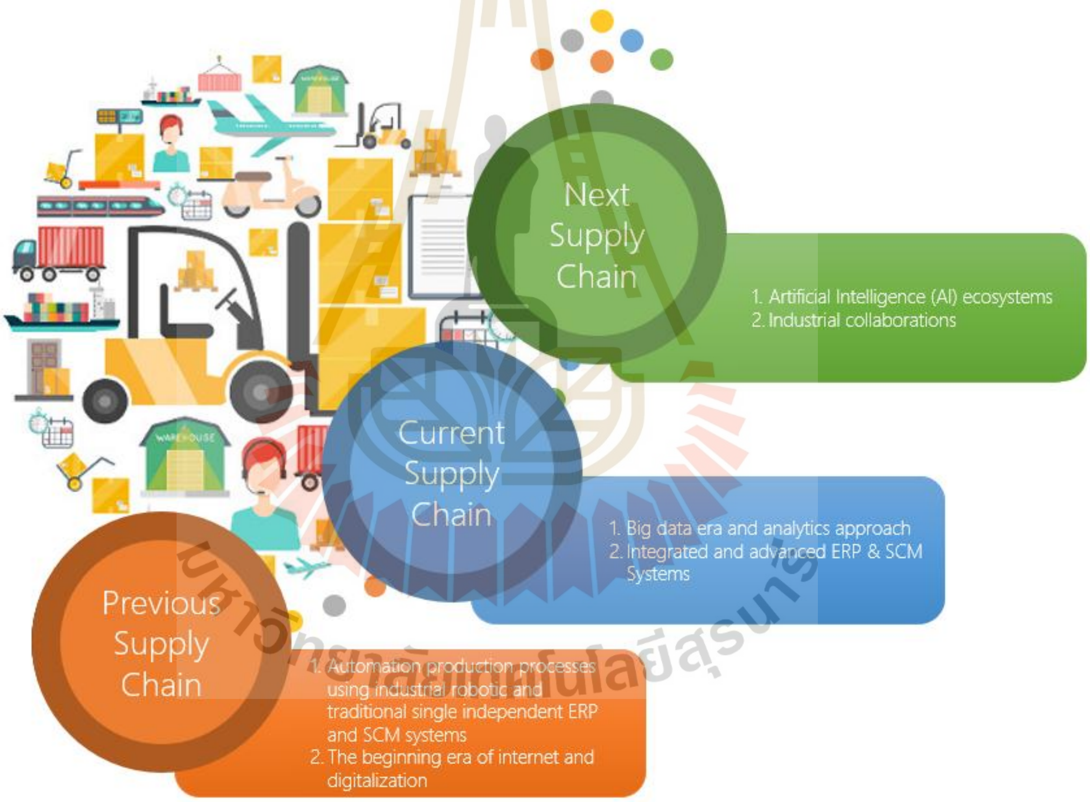


Figure 2.9 SCM Infographic.

2.4.5 Challenges of BDA Implementation

The result of this study also brought up the challenges of big data implementation. As big data starts to grow and evolve, BDA will continue to grow and become important in academic and enterprise areas. Despite the great potential and opportunity of BDA, there are also some great challenges and risks. When the BDA challenges are addressed in a proper manner, the success rate of implementing big data solutions automatically increases. As big data makes its way into academics and particularly enterprises around the world, addressing these challenges are extremely important to unlock the full potential of BDA. Below are some of the main challenges of BDA implementation.

a) Dealing with data growth and amount of accessible data. As big data is continuously growing and expanding, new technologies are being developed every day. The combination of data explosion and technology developments become a big challenge for enterprises to be answered without the introduction of new risks and problems. In addition, the limitless amount of accessible information can become an obstacle for enterprises to deal with because the scale and variety of the data can be too overwhelming to be analyzed and create waste that negatively affects big data usefulness.

b) Synchronization among data source. A lot of different types of software systems from different areas of daily functional activities in enterprises, which are quite challenging to incorporate into analytical insight and meaningful messages, exist.

c) Organizational resistances and culture. Some organizations are still conservative. Most business firms, government institutions, and other organizations may resist big alterations. The biggest impedance to the adoption of BDA is related to cultural change such as organizational alignment, lack of understanding, and managerial change. In an interview, a director of supply chain in automotive industry mentioned, “Koreans, in general, are not familiar with making decision based on decision making quantitative data. Rather, they tend to make important decisions based on their past experience or opinion of someone they trust. This emotional attribute may be an obstacle” (Richey et al., 2016). Attempt to change the doctrine and the culture requires a lot of patience and great persistence.

d) Security and privacy of the data. Soon after the business enterprises discover how to utilize BDA, it brings them a wide range of possibilities and opportunities. However, confidential data, such as unit price data, are important for enterprises, and there are a lot of risks when enterprises share these data to be analyzed. If these kinds of data leak outside the organization, for example, when the company makes a $15 profit from the $25 product, the client will push the company to down the price. Therefore, some confidential information will be difficult to get into big data analysis.

e) Real-time analysis. The challenges of getting important insights through the use of real-time BDA can be an obstacle since the velocity of the data is growing swiftly time by time.

2.5 Conclusion

As the result of this exploratory paper of the current situation related to the implementation of big data techniques within logistics and supply chain journals, there were almost 2,000 scientific articles that used big data techniques within the journals of logistics and supply chain. Furthermore, there were 43 scientific articles that mentioned the context of their scientific research studies within the big data context (where there was a massive gap between big data scientific articles compared with the other topics in the journals). This indicated that the utilization of big data techniques and analysis was sparsely published in the logistics and supply chain journals (within the context of big data). However, a trend analysis showed that the topics related to big data published in logistics and supply chain journals have increased every year (from 2013 to 2018). This also showed that BDA and SCM have received more attention from academics and become a potential topic to be continued into further research, which will be beneficial to many sectors, such as academics and other industries.

Furthermore, BDA in the digital era means that big analytics turn that data into real insight. It is an art that must be understood and implemented. Thus, the principal focus in BDA is not what is being analyzed but how the analysis is being implemented within the context of big data. Based on the reviews and findings of this paper, the following are some points to be noted.

a) There is a lack topic in big data research. There were no adequate research studies in implementing big data techniques in the supply chain and logistics field from January 1, 2011, to December 30, 2018, and the most common finding across all areas is that the interest in big data research is only recent.

b) There is a need to improve and implement research on big data. Based on the statistical result about the current situation and trend in implementing BDA in this research, it is recommended that the future research should focus on the implementation of BDA within specific industries and use real data to reduce the existing gap now. All aspects of SCM must also be studied with real data, especially big data analysis for planning, forecasting, procurement, and material management, as well as big data analysis that supports the alignment, agility, and sustainability of supply chain with integrated systems.

c) Enterprises in ASEAN must assimilate BDA or risk themselves of being left behind. In developed countries, new digital technologies are changing the traditional ways of SCM. Using a digital foundation in place, now enterprises can capture, analyze, integrate, easily access, and interpret high-quality data in real time, which support enterprises to make better decisions. Meanwhile, enterprises in developing countries (i.e., ASEAN) should consider and realize the importance of big data and technology to support and change their business process or face the risk of being left behind.

d) The ability of big data analysis to minimize the existing gap now must be improved. Big data and technology have played important roles in this new era. Since the skills needed for these new roles are not readily available today, the biggest challenge for researchers is to fill those critical roles.

e) The future of SCM is AI. It is clear that the future of SCM is the technology that optimally manages end-to-end workflows and requires little human intervention.

In addition, the limitation of this research was that each of the big data techniques (as previously stated) has overlapped with one another (for example, a combination of statistics and regression analysis constitutes the part of predictive analytics, etc.). Thus, it affects the accuracy of the search results in several aspects. Another limitation was that the sample size examined in the time series forecasting analysis was quite small. It could affect the result of forecasting accuracy in this research.

2.6 References

Addo-Tenkorang, R., & Helo, P. T. (2016). Big Data Applications in Operations/ Supply Chain Management: A Literature Review. **Computers and Industrial Engineering**. 101: 528-543. https://doi.org/10.1016/j.cie.2016.09.023.

Andreu-Perez, J., Poon, C., Merrifield, R., Wong, S., & Yang, G. (2015). Big Data for Health. **IEEE Journal of Biomedical and Health Informatics.** 19(4): 1193-1208. https://doi.org/10.1109/JBHI.2015.2450362.

Arunachalam, D., Kumar, N., & Kawalek, J. P. (2018). Understanding Big Data Analytics Capabilities in Supply Chain Management: Unravelling the Issues, Challenges and Implications for Practice. **Transportation Research Part E: Logistics and Transportation Review**. 114: 416-436. https://doi.org/10.1016/ j.tre.2017.04.001.

Ayers, J. B. (2001). Handbook of supply chain management. St. Lucie Press. https://trove.nla.gov.au/work/20348325?selectedversion=NBD21461646.

Ballou, R. H. (2007). The Evolution and Future of Logistics and Supply Chain Management. **European Business Review**. 19(4): 332-348. https://doi.org/ 10.1108/ 09555340710760152.

Barbosa, M. W., Vicente, A. de la C., Ladeira, M. B., & Oliveira, M. P. V. de. (2018). Managing supply chain resources with Big Data Analytics: a systematic review. **International Journal of Logistics Research and Applications.** 21(3): 177-200. https://doi.org/10.1080/13675567.2017.1369501.

Bean, R. (2016). **Why Cultural Change Is Necessary For Big Data Adoption**. Forbes. https://www.forbes.com/sites/ciocentral/2016/11/08/another-side-of-big-data-big-data-for-social-good-2/#50aad3356628.

Beyer, M. A., & Laney, D. (2012). **The Importance of “Big Data”: A Definition**. In Gartner. https://www.gartner.com/doc/2057415/importance-big-data-definition.

Bhattacharjya, J., Ellison, A. B., Pang, V., & Gezdur, A. (2018). Creation of unstructured big data from customer service: The case of parcel shipping companies on Twitter. **The International Journal of Logistics Management**. 29(2): 739-766. https://doi.org/10.1108/IJLM-06-2017-0157.

Bhattacharjya, J., Ellison, A., & Tripathi, S. (2016). An Exploration of Logistics-Related Customer Service Provision on Twitter: The Case of e-Retailers Article Information. **International Journal of Physical Distribution & Logistics Management.** 46(7): 659-680. https://doi.org/10.1108/IJPDLM-01-2015-0007.

Bohács, G., & Rinkács, A. (2016). Development of a novel material flow simulation model for the integration of spatial and process relevant information. **Logistics Journal**. https://doi.org/10.2195.

Brinch, M., Stentoft, J., Jensen, K., & Rajkumar, C. (2018). Practitioners understanding of big data and its applications in supply chain management. **The International Journal of Logistics Management**. 29(2): 478-484. https:// doi.org/10.1108/ IJLM-05-2017-0115.

Chen, L., Olhager, J., & Tang, O. (2014). Manufacturing Facility Location and Sustainability: A Literature Review and Research Agenda. **International Journal of Production Economics**. 149: 154-163. https://doi.org/10.1016/J.IJPE.2013.05.013.

Columbus, L. (2015). **Ten Ways Big Data Is Revolutionizing Supply Chain Management**. Forbes. https://www.forbes.com/sites/louiscolumbus/2015/07/13/ten-ways-big-data-is-revolutionizing-supply-chain-management/#3258fb3d69f5.

Council of logistics management. (2000). Council of logistics management. In **Encyclopedia of Production and Manufacturing Management**. 134-135. Springer US. https://doi.org/10.1007/1-4020-0612-8_194.

Dupuit, J. (1952). On the measurement of the utility of public works. **International Economic Papers**. 2: 83-110.

Fry, E. (2016). **Will Big Data Cure Cancer?** Fortune. http://fortune.com/2016/11/02/big-data-cure-cancer.

Ghorpade-Aher, J., Pagare, R., Thengade, A., Ghorpade, S., & Kadam, M. (2016). Big Data: The Data Deluge. In **Research Advances in the Integration of Big Data and Smart Computing**. 1-21. https://doi.org/10.4018/978-1-4666-8737-0.ch001.

Govindan, K., Cheng, T., Mishra, N., & Shukla, N. (2018). Big Data Analytics and Application for Logistics and Supply Chain Management. **Transportation Research Part E**. 114: 343-349. https://doi.org/10.1016/j.tre.2018.03.011.

Grant, D. B., Stock, J. R., Ellram, L. M., & Grant, D. (2006). **Fundamentals of logistics management**. McGraw-Hill. https://researchportal.hw.ac.uk/en/publications/fundamentals-of-logistics-management-first-european-edition.

Hamister, J. W., Magazine, M. J., & Polak, G. G. (2018). Integrating Analytics Through the Big Data Information Chain: A Case From Supply Chain Management. **Journal of Business Logistics**. 39(3): 220-230. https://doi.org/10.1111/jbl.12192.

Handfield, R. (2017). **What is Supply Chain Management (SCM)?** NC State University. https://scm.ncsu.edu/scm-articles/article/what-is-supply-chain-management-scm.

Henke, N., & Kaka, N. (2018). Analytics Comes of Age. In **McKinsey Company**. https:// www.mckinsey.com/~/media/McKinsey/Business Functions/McKinsey Analytics/ Our Insights/Analytics comes of age/Analytics-comes-of-age.ashx.

Higgins, N. (2016). **Big Data: Key to Customer Understanding and Service Delivery**. Mass Transit. https://www.masstransitmag.com/technology/article/12264033/big-data-key-to-customer-understanding-and-service-delivery.

IBM. (2013). **The Four V's of Big Data | IBM Big Data & Analytics Hub**. IBM Big Data and Analytics Hub. https://www.ibmbigdatahub.com/infographic/ four-vs-big-data.

Ibnouhsein, I., Jankowski, S., Neuberger, K., & Mathelin, C. (2018). The Big Data Revolution for Breast Cancer Patients. **European Journal of Breast Health**. 14(2). https://doi.org/10.5152/ejbh.2018.0101.

Jeble, S., Kumari, S., & Patil, Y. (2018). Role of Big Data in Decision Making. Operations and Supply Chain Management: **An International Journal**. 11(1). 36. https://doi.org/10.31387/oscm0300198.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, Challenges, Tools and Good Practices. **2013 Sixth International Conference on Contemporary Computing (IC3).** 404-409. https://doi.org/10.1109/IC3.2013.6612229.

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. **Big Data & Society**. 1(1). 205395171452848. https://doi.org/10.1177/2053951714528481.

Lai, Y., Sun, H., & Ren, J. (2018). Understanding the Determinants of Big Data Analytics (BDA) Adoption in Logistics and Supply Chain Management: An Empirical Investigation. **The International Journal of Logistics Management**. 29(2): 478-484. https://doi.org/10.1108/IJLM-06-2017-0153.

Lamba, K., & Singh, S. P. (2017). Big Data in Operations and Supply Chain Management: Current Trends and Future Perspectives. **Production Planning & Control**. 28(11-12): 877-890. https://doi.org/10.1080/09537287.2017.1336787.

LeMay, S., Helms, M. M., Kimball, B., & McMahon, D. (2017). Supply chain management: the elusive concept and definition. **The International Journal of Logistics Management**. 28(4): 1425-1453. https://doi.org/10.1108/IJLM-10-2016-0232.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. In **McKinsey Global Institute**. www.mckinsey.com/mgi.

Markman, J. (2016). **Big Data Unleashes Business Opportunity**. Forbes. https://www.forbes.com/sites/jonmarkman/2016/11/01/big data unleashes business opportunity/#46ce7a6e6c88.

Marr, B. (2013). **The Awesome Ways Big Data Is Used Today To Change Our World**. Linkedin. https://www.linkedin.com/pulse/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world.

Marr, B. (2017). **The 6 Top Data Jobs In 2018**. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/09/the-6-top-data-jobs-in-2018/#6f7cf63e430d.

Marr, B. (2018). **How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read**. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6145c9b560ba.

Mayer-Schönberger, V., & Cukier, K. (2014). **Big Data : a Revolution That will Transform how we Live, Work, and Think**. Eamon Dolan/Mariner Books.

Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., & Zacharia, Z. G. (2001). **Defining Supply Chain Management. Journal of Business Logistics.** 22(2): 1-25. https://doi.org/10.1002/j.2158-1592.2001.tb00001.x.

Miller, J. W., Ganster, D. C., & Griffis, S. E. (2018). Leveraging Big Data to Develop Supply Chain Management Theory: **The Case of Panel Data. Journal of Business Logistics.** 39(3): 182-202. https://doi.org/10.1111/jbl.12188.

Mitchell, J. (2016). **This Startup Is Using Big Data And Smart Tech To Save The Oil Industry**. Forbes. https://www.forbes.com/sites/julianmitchell/2016/10/29/this-startup-is-using-big-data-and-smart-tech-to-save-the-oil-industry/#713f07fd190f.

Pienaar, W. J. (2009). Introduction to business logistics. In **Business logistics management : a supply chain perspective**. 472. Oxford University. https://trove.nla.gov.au/work/37765952?q&versionId=49807648.

Provost, F., & Fawcett, T. (2013). ***Data Science and its Relationship to Big Data and Data-Driven Decision Making***. **Big Data**. 1(1): 51-59. https://doi.org/10.1089/big.2013.1508.

Pujawan, I. N. (2016). Operations and Supply Chain Management: Toward Ten Years of Journey. Operations and Supply Chain Management: **An International Journal, December 2016**. 1. https://doi.org/10.31387/oscm0260175.

Rowley, J., & Slack, F. (2004). Conducting A Literature Review. **Management Research News**. 27(6): 31-39. https://doi.org/10.1108/01409170410784185.

Rushton, A., Croucher, P., & Baker, P. (2017). **The Handbook of Logistics and Distribution Management** (6th ed.). Kogan Page.

Singh, A., Shukla, N., & Mishra, N. (2018). Social media data analytics to improve supply chain management in food industries. **Transportation Research Part E: Logistics and Transportation Review**. 114: 398-415. https://doi.org/10.1016/J.TRE.2017.05.008.

Slider, V. (2016). **Re-thinking Analytics: How big data analytics can revolutionise your contact centre**. Business Tech. https://businesstech.co.za/news/industry-news/140835/re-thinking analytics how big data analytics can revolutionise your contact centre/.

Taggart, K., Anderrson, G., Kang, M., & Getty. (2016). **What Exactly Is Big Data?** (p. Technology). Forbes. https://www.forbes.com/video/4857597029001/#51fa aec f32b8.

Tiwari, S., Wee, H. M., & Daryanto, Y. (2018). Big Data Analytics in Supply Chain Management between 2010 and 2016: Insights to Industries. **Computers & Industrial Engineering**. 115: 319-330. https://doi.org/10.1016/J.CIE.2017.11.017.

Trefis Team. (2014). **Two Key Executives Leave Walgreen Due To a $1 Billion Forecasting Error**. Forbes. https: //www.forbes.com/sites/greatspeculations/2014/08/25/two key executives leave walgreen due to a 1 billion forecasting error/#22 ed8ba484ef.

Waller, M. A., & Fawcett, S. E. (2013a). Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. **Journal of Business Logistics**. 34(4): 249-252. https://doi.org/10.1111/jbl.12024.

Waller, M. A., & Fawcett, S. E. (2013b). Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management. **Journal of Business Logistics.** 34(2): 77-84. https://doi.org/10.1111/jbl.12010.

Wamba, S. F., Gunasekaran, A., Papadopoulos, T., & Ngai, E. (2018). Big data analytics in logistics and supply chain management. **The International Journal of Logistics Management**. 29(2): 739-766. https://doi.org/10.1108/IJLM-02-2018-0026.

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016). Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications. **International Journal of Production Economics**. 176: 98-110. https://doi.org/10.1016/J.IJPE.2016.03.014.

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big Data for Supply Chain Management in the Service and Manufacturing Sectors: Challenges, Opportunities, and Future Perspectives. **Computers and Industrial Engineering**. 101: 572-591. https://doi.org/10.1016/j.cie.2016.07.013.

CHAPTER III

Using Big Data Analytics for Decision Making: Analyzing Customer Behavior using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia

3.1 Abstract

Indonesia is facing many challenges in the fourth industrial revolution (4IR) era. One of them is related to big data technologies and implementation that can be seen clearly from Indonesia Industry Readiness Index (INI) 4.0. Therefore, focusing on implementing big data analytics in a gold, silver, and precious metal trading company is the objective of this manuscript to support daily business operations. To be more specific, the aim is to discover meaningful patterns and ensure high quality of knowledge discovery from the big data available in a company in Indonesia. It is needed to support the Making Indonesia 4.0 as a roadmap to implement industrial digitalization in Indonesia. The methodology used for the big data implementation in this manuscript is the combination of the CRISP-DM framework and key steps for customer analytics. The result of this research is a list of recommendations that facilitate strategic planning based on evidence of measurable big data analytics to achieve the business goals of a company.

3.2 Introduction

The fourth industrial revolution (*forth* 4IR) is a high-tech strategy deriving from technological advancement and disruptive development in the industrial sector worldwide that fundamentally alter the industrial environment work and relate to one another (Dallasega, Rauch, & Linder, 2018; Schwab, 2016c). With technological innovation, it has the potential to raise global income levels and improve the quality of the industrial sector around the world (Schwab, 2016a, 2016b). Some consider it as the integration of emerging technologies such as Internet of Things (IoT), Big Data, and Cloud technologies (Pereira & Romero, 2017). On the other hand, there are scientific arguments that the fourth industrial revolution is not only regarding technology integration but also concerning the whole concepts of 'how to acquire, share, organize data and resource that make and increase the product or service delivery faster, cheaper, more efficiency, more productivity, more effective, and more sustainable' (Benitez, Llorens, & Braojos, 2018; Koh, Orzes, & Jia, 2019; Piccarozzi, Aquilani, & Gatti, 2018). Since technologies integration are considered the core of the 4IR, there some technologies frequently discussed in the literature: IoT, Big Data Analytics (BDA), Cloud, 3D printing, robotic, Artificial Intelligence (AI), Machine Learning (ML), and 5G technology (Kamble, Gunasekaran, & Gawankar, 2018; Piccarozzi et al., 2018).

In response to this, President Joko Widodo has declared 'Making Indonesia 4.0' as a road map to implement industrial digitalization that can positively support industry performances in Indonesia (Deputi Bidang Protokol-Pers-dan Media Sekretariat, 2019). It is followed by Indonesia Industry Readiness Index (INDI) 4.0 by the Ministry of Industry to support the government measuring progresses on 4IR in

Indonesia (Aisyah, 2019). As the interest in the fourth industrial revolution is growing rapidly, this manuscript does not limit the focus on 4IR itself but intends to implement BDA as interconnection in supporting 4IR (Kamarul Bahrin, Othman, Nor Azli, & Talib, 2016). It is essential since the current research paradigm is at the fourth level, where big data is analyzed using various statistical methods for real-world needs in real-time settings (Kitchin, 2014). However, the scientific manuscripts published in the journal related to BDA implementation are still deficient (Fosso Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015; Yudhistyra, Risal, Raungratanaamporn, & Ratanavaraha, 2020) where many manuscripts explain that BDA can improve business strategy (Akter, Wamba, Gunasekaran, Dubey, & Childe, 2016) and organizational performance (Gunasekaran et al., 2017).

Moreover, there is a big gap between big data available now and the capability to analyze it while there will be a big demand for data analytics skills at this time (Marr, 2017, 2018). Thus, this study aims to discover meaningful patterns and ensure high quality of knowledge discovery from the big data available in a company in Indonesia. It can provide new insight into what a company should offer as better incentives to its customers and what strategies a company should develop to allocate the resources appropriately. As a result, the main contributions of this manuscript are mentioned as follows.

a) it presents analytics with a large amount of data to help a company strengthen its position to stay ahead of the competition.

b) it minimizes any investment risk on researching and testing the market, product, or ideas.

c) it facilitates strategic planning based on evidence of BDA to achieve the business goals of a company; and

d) it helps the company in spotting emerging trends on customer needs and demands.

In addition, another contribution of this manuscript is that it helps a company to improve their performances, which are assessed by INDI 4.0 in Indonesia and increase the average score of 2.14 (scale 4) where the lowest score comes from metal companies in Indonesia (Kementrian Perindustrian, 2020).

3.3 Literature Review

3.3.1 Related works

There are few manuscripts published articles in journals related to BDA implementation. For instance, Batarseh and Latif in 2016 and Alani and team in 2016 assessed the quality of service and healthcare organization using BDA in healthcare industries in the United States and Iraq (Batarseh & Latif, 2016); Moyne and Iskandar in 2017 implemented BDA in a manufacturing company (Alani, Ahmed, Majid, & Ahmad, 2018); (Moyne & Iskandar, 2017); Naimur Rahman and his team in 2016 used BDA to predict total electricity forecast in the United States (Naimur Rahman, Esmailpour, & Zhao, 2016); Honarvar and Sami in 2019 tried to find a suitable solution for urban development by using the opportunities of big data and present data related to urban computing with the aim of assessing the knowledge that can be obtained through integration of multiple independent data sources in Smart Cities (Honarvar & Sami, 2019). Meanwhile, the position of this manuscript that makes it different from previous related works is that no study was found in the literature covering BDA in a gold, silver, and precious metal industry. Also, this manuscript explores the customer behaviors and habits based on a large amount of data analysis in conducting the transaction in a developing country like Indonesia.

3.3.2 Frequent patterns, association, and correlation

One famous method to analyze a large amount of data often used by experts and researchers is association rule mining method, a common technique leads to the discovery of associations and correlations among items in a large database using a variety of algorithms to generate and test the possible rules (Manyika et al., 2011; Yudhistyra et al., 2020). A typical example of this method is market basket analysis, an analysis of customers buying habits by finding an association between different items that customers place in their 'shopping basket' (Valarmathi, 2017). Using this analysis, a company can understand the purchase behavior of its customers, which can help a company to create better decision making for supporting its business goals (Kaur & Kang, 2016). For instance, if customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers and at the same time is represented in the following association rule.

$$computer \Rightarrow printer\ [support = 20\%, confidence = 60\%] \quad (1)$$

where a function defining *support* for itemset *x* can be defined as (Lantz, 2015):

$$support(x) = \frac{count(x)}{N} \quad (2)$$

where N is the number of transactions in the dataset and *count*(*x*) is the number of transactions containing itemset *x*, and the confidence is a measurement of its predictive accuracy where the *support* of itemset containing both *x* and *y* divided by the *support* of the itemset only *x* as followed.

$$confidenct(x - y) = \frac{support(x, y)}{support(x)} \quad (3)$$

Rule support and confidence are two measures the usefulness and certainty of discovered rules. A support of 20% means that 20% of all the transactions under analysis show that computers and printers are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the printer. Rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold where users or experts can set it. Additional analysis can be performed to discover the interesting statistical correlation between associated items to get better knowledge. Furthermore, there are various algorithms to perform association rule mining and to compute large datasets such as the Apriori algorithm and the Continuous Association Rule Mining Algorithm (CARMA).

3.3.3 Apriori algorithm

Apriori is an algorithm proposed by Agrawal and Srikant in 1994 for frequent mining itemset for Boolean association rules (Agrawal & Srikant, 1994). The first step of this algorithm simply counts the item occurrences to determine the large l-itemsets. A subsequent pass, say pass *k*, consists of two phases. First, the large itemsets L_{k-i} found in the (*k-1*)-th pass are used to generate the candidate itemsets C_k, using the Apriori-gen function. The Apriori-gen function takes as argument L_{k-1}, the set of all large (*k* - *1*)-itemsets. It returns a superset of the set of all large k-itemsets. Afterward, the database is scanned, and the *support* of candidates in C_k is counted. For fast counting, we need to efficiently determine the candidates in C_k that are contained in a given transaction *t*.

3.3.4 Continuous Association Rule Mining Algorithm (CARMA)

This algorithm needs, at most, two scans of the transaction sequence to produce all large itemset where during the first scan, this algorithm continuously constructs a list of itemsets (a lattice) of all potentially a large itemset. Then, during the second scan, this algorithm determines the precise *support* of each set in the lattice and continuously remove all small itemset (Hidber, 1999).

The more detail explanations about every phase in CARMA algorithm are explained by Huang et. al. (2009) as follows (Huang, Wang, & Shia, 2009). Phase I objective in this algorithm is to produce all potentially large itemsets. Then it is followed by storing the picked itemset and their related value vectors which it is handled by lattice *V*. The value vector is a three-dimensional vector composed of *count* (*v*), *maxMissed* (*v*), and *firstTrans* (*v*) where *count* (*v*) is the number of occurrences of *v* since its insertion in the lattice; *maxMissed* (*v*) is the upper bound on the number of occurrence of *v* before *v* is inserted in the lattice; and *firstTrans* (*v*) is the index of the transaction at which *v* is inserted in the lattice. Both *count* (*v*) and *maxMissed* (*v*) can define the interval of estimated *support* of *v*. The lower bound is $minSupport_{\mathrm{i}}$ (*v*) = $count_{\mathrm{i}}$ (*v*)/*i* and the upper bound is $maxSupport_{\mathrm{i}}$ (v) = ($maxMissed_{\mathrm{i}}$ (*v*) + $count_{\mathrm{i}}$ (*v*))/*i*,*j* is the current transaction number. A *support* threshold can be specified by user. This threshold makes a sequence $\sigma = (\sigma_1, \sigma_2, \sigma_3, \ldots)$. Based on the sequence σ, a new index $[\sigma]_{\mathrm{i}}$ can be constructed, namely the ceiling of *0* up to *i*. $[\sigma]$ is the last monotone decreasing sequence which is up to *i* pointwise greater of equal to σ and *0* otherwise. Assume that *i-1* transactions have been processed and let V_{i-1} be the *support* lattice up to *i-1*. It is reading the *i*-th transaction t_i and update the V_{i-1} to V_i. σ_i is specified for the *i*-th transactions. For any itemset *v*, if it has already been in the lattice *v*, then it updates the count (*v*), increasing it by one. When the set of *v* is not in

V, it considers the following rules to judge whether or not to let it in. If *v* is l-item set then it is directly inserted with count (*v*) = 1, *firstTrans* (*v*) = *i*, *maxMissed* (*v*) = 0. If *v* is *k*-item set (*k*>1), two requirements should be satisfied. The first is that all subsets *w* of *v* have already been contained in *V* and are potentially large. Phase I also includes a prune step-only working every $1/\sigma i$ or every 500 transactions (default)-whichever is the larger.

Phase II removes trivially small itemsets based on the last specified *support* threshold. It rescans the process again and determines the accurate *support* and continues to remove the trivially small dataset. If the user do not need precise *support*, Phase II can be omitted (Hidber, 1999).

3.4 Case Study

The combination of key steps for customer analysis (Yudhistyra, Raungratanaamporn, & Ratanavaraha, 2019) and the CRISP-DM framework (Mariscal, Marbán, & Fernández, 2010; Wirth, 2000) shown in Figure 3.1 is fused to ensure the high quality of knowledge discovery in this case study. The pre-processing steps are business understanding, data understanding, and data preparation; the processing step is the modeling and evaluation; and the post-processing is the final result.

3.4.1 Business understanding

The important factor in this step is the current business situation in the gold, silver, and precious metal trading company, which intends to use BDA to gain knowledge about the behaviors of its customers in conducting a transaction. It is used to develop the company's shelf management and campaign strategies in offering the company's products. To be more detailed, the intention of using BDA in this case study covers a) finding frequent itemset patterns on the company's historical transaction dataset using b) recognizing peculiarities on customers' transactional data; c) identifying useful and actionable patter to be aligned with business strategies.

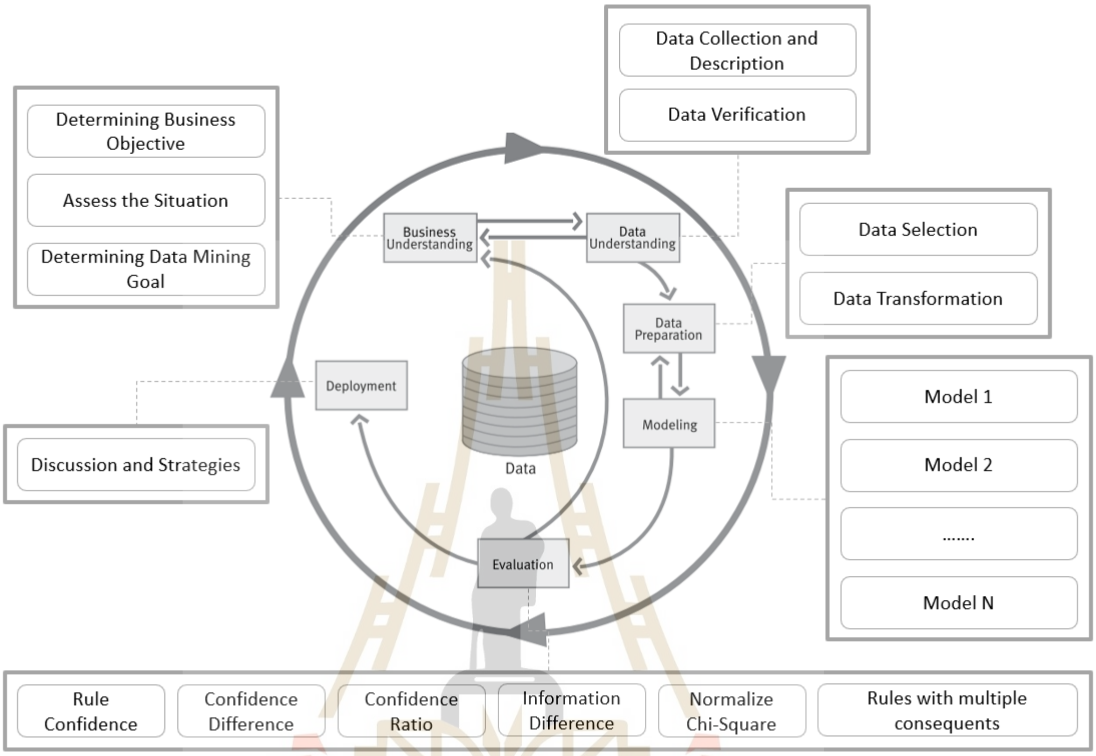


Figure 3.1 Combination of CRISP-DM and keys steps for customers analysis.

Detail fact-findings found to support BDA implementation, such as the resource availability (software, hardware) or legal issues in which the data is allowed to be utilized, had been completed before beginning the project plan. A list of possible project plans to be performed during the rest of the project is listed as follows.

a) A collection of the dataset is collected from various resources available at the company.

b) Association rule mining techniques are utilized to discover meaningful relationships or patterns from the dataset.

c) Data mining software applications such as R, Modeler, Tableau, or RapidMiner are considered to be used for analyzing the dataset and calculating the accuracy of predictive models.

d) The data mining goal criteria are categorized as successful outcomes when the minimum percentage accuracy of the models' percentage is above 50% (moderate fit model) (Field, 2018) .

e) The repetition of the modeling and evaluation process is adjusted based on rule confidence, confidence difference, confidence ratio, information difference, normalize chi-square, and rules with multiple consequents.

3.4.2 Data understanding and preparation

The datasets were provided by the company, which was exported from multiple resources of company database covering July 2010 to October 2019. It contains 3,986,872 observations from 248,856 customers. There are seven variables on the dataset, and the description of datasets is provided by the company shown in Table 3.1 as follows.

Table 3.1 Initial data.

Attribute Name	**Description**
CUSTOMERS_ID	The ID of the customers
TRANS_DATE	The date of the transaction
TRANS_ID	The invoice number of the order transaction
CUSTOMERS_ADDRESS	The address of the customers
DESCRIPTION	Description of the product
QUANTITY_TOTAL	Quantity total of the product
PRICE_TOTAL	Price total of the product

Exploratory data analysis was performed to verify the quality of the data. It covered exploring the structure of the data, measuring central tendency, removing the null of the tuple in the datasets (cleaning), sorting, hashing, exploring the relationship, grouping, aggregation, merging, visualizing the relationship, examining relationship, checking the collinearity, outlier's detection, etc. Afterward, there is a list of some unsatisfied results found in the data quality verification steps as follows.

a) Incorrect data. Some of the data contain common errors, such as error product names and biased customer data.

b) Exogenous force. Some of the data contain a contradiction with the facts and circumstances such as production date or year for some products.

c) Duplication. It is a common mistake made by people when inputting the data, such as similar customer data with different ID recorded in the systems.

d) Inconsistent naming standard. When it is seemingly simple, this kind of disorganization often wreaks havoc on data analysis.

All of these mistakes can lead to measurement errors that cannot be fitted with existing theory. They will inaccurately skew the dataset, resulting in corrupt predictions and poor decision-making because of the unrepresentative dataset. There is some solution to minimize these mistakes such as automate data entry to reduce human error, use consistent and organized naming standard, schedule training events for end-users for the systems to make they understand the value of data, and consider moving to the cloud systems which the systems can see time-stamped logs that users have made throughout the day.

Afterward, the datasets were reduced to be suitable with the variables used for the algorithms based on practical knowledge. The datasets are transformed into 80 variables, as shown in Table 3.2, as follows. The product variables are transformed to become tabular data that is suitable for building association rules mining models in market basket analysis.

Table 3.2 Data transformation.

Attribute name	Description
TRANS_ID	The ID of the transaction
PRODUCT_N (1-79)	The tabular data of products (1-79) that have items represented by separate flags, where each flag field represents the presence or absence of a specific item

3.4.3 Modeling, evaluation, and deployment

After the data preparation completed, initial data analysis for finding associations and checking interrelation among the products in every transaction can be seen in Figure 3.2 as follows.

Figure 3.2 denotes the number of times that the two items occurred in the same transaction. For instance, the number of times that '*product* 8' is purchased with '*product* 11' is 1,007 times. This observation will provide an initial overview of customer behavior in their shopping cart items. To be more specific, the relationship of each transaction is divided into three parts (i.e., strong link, medium link, and weak link) of item frequency in the data, as shown in Figure 3.3 (the thicker the line, the stronger the relationships).

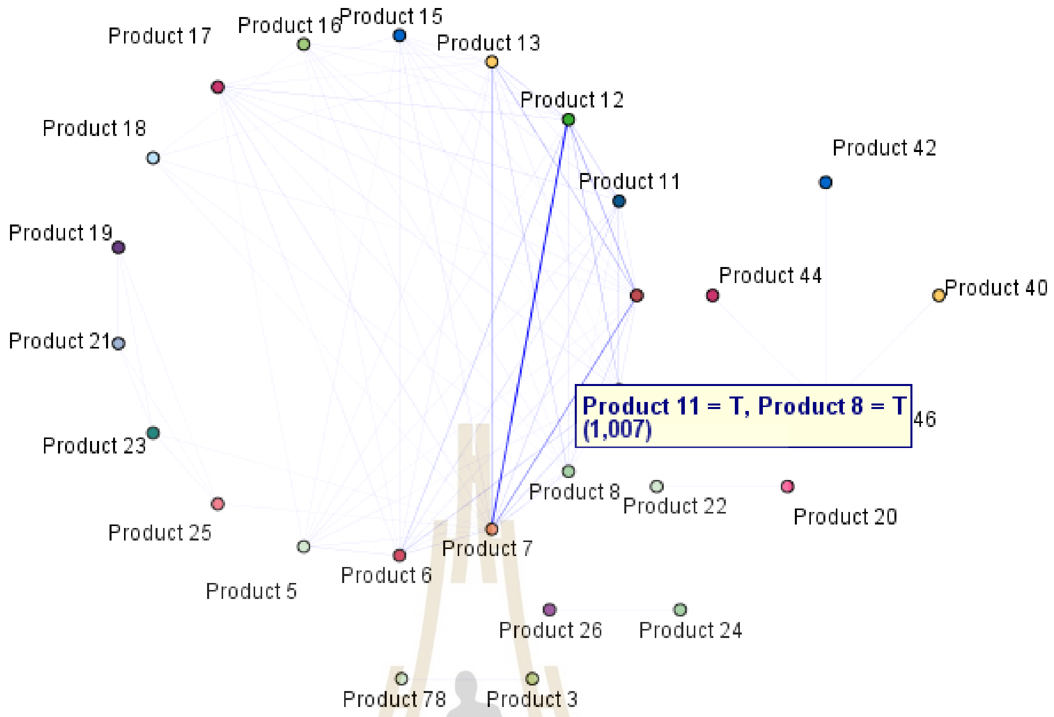


Figure 3.2 A web graph shows the number of times that the two items occurred in the same transaction.

Information from Figure 3.3 shown that some products have a strong relationship based on the company transaction history if the items occurred in the same transaction more than 15,000 times. Some products that have strong relationships are '*product* 7' and '*product* 12' with 38,123 times; '*product* 10' and '*product* 7' with 23,367 times; '*product* 10' and '*product* 12' with 15,530 times; and so forth. The products have a weak relationship if the items occurred in the same transaction less than 5,000 times. Some products that have weak relationships are '*product* 23' and '*product* 7' with 291 times; '*product* 18' and '*product* 8' with 532 times; '*product* 21' and '*product* 25' with 296 times; and so forth. It is also sufficient that there are no customers purchased '*product* 19' and '*product* 18' in the same transaction so that the company should not promote these products at the same time for its customers.

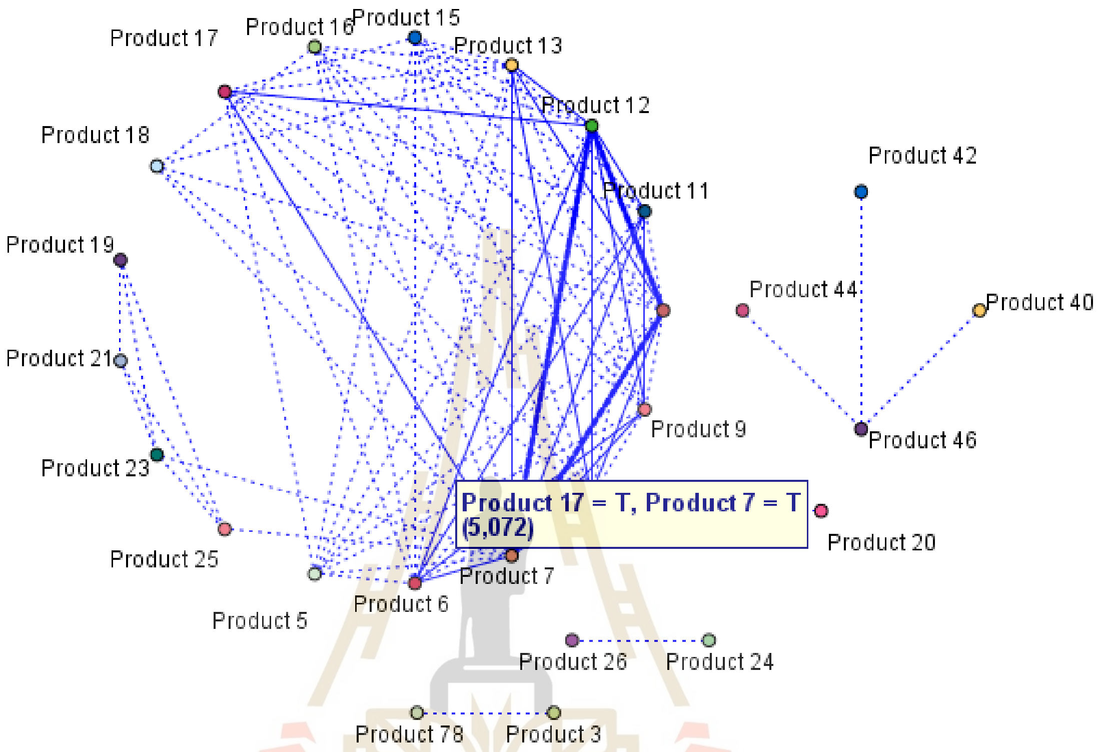


Figure 3.3 Dividing the item frequency relationships.

A predictive model was extracted using an Apriori algorithm to get deeper knowledge and to discover more association rules in the data as a market basket analysis. In this case, if the company wants to send a promotional campaign through multiple channels for existing customers, actionable strategies can be developed better after knowing the recommendations of a strong association in its sales products. Using the Apriori algorithm, there are 21 rules found from the analysis result, as shown in Table 3.3 as follows.

Table 3.3 Rules developed by Apriori algorithm.

Consequent	Antecedent	Support (%)	Confidence (%)	Lift
Product 13	Product 8 Product 10	1.3	63.026	5.547
Product 13	Product 8 Product 7	1.354	58.763	5.171
Product 7	Product 8 Product 10	1.3	63.375	1.905
Product 10	Product 8 Product 7	1.354	60.838	3.58
Product 12	Product 8 Product 7	1.354	52.002	2.152
Product 9	Product 11 Product 6	1.259	57.367	6.674
Product 12	Product 11 Product 9	1.453	50.418	2.087
Product 12	Product 11 Product 6	1.259	51.275	2.122
Product 12	Product 11 Product 7	1.526	56.452	2.337
Product 12	Product 9 Product 7	1.7	59.902	2.48
Product 12	Product 6 Product 7	1.606	61.653	2.552
Product 10	Product 13 Product 12	1.5	61.593	3.624
Product 7	Product 13 Product 10	2.461	50.737	1.525
Product 10	Product 13 Product 7	2.252	55.446	3.262

Table 3.3 Rules developed by Apriori algorithm (Conts).

Consequent	Antecedent	Support (%)	Confidence (%)	Lift
Product 7	Product 13 Product 12	1.5	69.608	2.092
Product 7	Product 10 Product 12	2.816	63.284	1.902
Product 8	Product 13 Product 10 Product 7	1.249	50.632	5.998
Product 8	Product 13 Product 12 Product 7	1.044	51.285	6.076
Product 6	Product 9 Product 12 Product 7	1.018	56.187	6.109
Product 12	Product 13 Product 10 Product 7	1.249	62.756	2.598
Product 10	Product 13 Product 12 Product 7	1.044	75.061	4.417

When constructing the rules, the minimum antecedent *support* of 1% and the minimum rule of *confidence* of 50% were adjusted using this algorithm. Using three ways (*support* percentage, *confidence* percentage, and *lift* value) to measure the association, the first rule on the table above informs as, "if a customer buys '*product* 8' and '*product* 10' (with 1.3% frequency of transaction), they will also buy '*product* 13' (with 63.026% *confidence*). The second rule informs that customers will buy '*product* 13' if they previously buy '*product* 8' and '*product* 7'. This rule covers

1.354 percent of the transactions and is correct in 58.763 percent of purchases involving ‘*product* 13’ (and so on until the last rule in Table 3.3 above). In addition, *lift* values in Table 3.3 above indicate the strength of a rule over the random occurrence of the antecedent item(s) and the *consequent* item(s). It basically mentions the ratio of the *confidence* and the *support* (the strength of any rule). A *lift* value greater than 1 means that the *consequent* item(s) is likely to be bought if the *antecedent* item(s) is bought, while a value less than 1 means that *consequent* item(s) is unlikely to be bought if the *antecedent* item(s) is bought. A larger *lift* value is, therefore, a strong indicator that a rule is important and reflects a true connection between the items.

Furthermore, a CARMA algorithm was also performed to be compared with the Apriori results in order to improve the product recommendations. Another advantage of using CARMA over Apriori is that ‘if rules with many *consequents* are desired’, which it can generate association rules with multiple *consequents*. When constructing the rules, the minimum antecedent *support* of 1% and the minimum rule of *confidence* of 50% were also adjusted using this algorithm. The result, conversely, has no multiple *consequents* rules like it is expected before. There are five rules which are similar to the Apriori rules, as shown in Table 3.4.

Table 3.4 Rules developed by CARMA algorithm.

Consequent	Antecedent	Support (%)	Confidence (%)	Lift
Product 7	Product 12 Product 13	1.5	69.608	2.092
Product 7	Product 10 Product 12	2.816	63.284	1.902
Product 12	Product 7 Product 9	1.7	59.902	2.48
Product 10	Product 13 Product 7	2.252	55.446	3.262
Product 7	Product 10 Product 13	2.461	50.737	1.525

After rerunning the models several times and comparing the results, unique findings are to identify the most useful rules quickly, so the attention is paid directly to the ones with the highest *support*, *confidence*, or *lift* values. The best five rules, according to the *lift* statistic, are denoted in Table 3.5 as follows.

These rules appear to be more sophisticated than those in Table 3.3. The first rule, with a *lift* of about 6.674, implies that customers who buy '*product* 11' and '*product* 6' are six times more likely to buy '*product* 9' than the typical customers. Rule two and rule three are also similar to rule one, which is over six times more likely to be found in the dataset.

Table 3.5 The best five rules according to the lift statistic.

Consequent	Antecedent	Support (%)	Confidence (%)	Lift
Product 9	Product 11 Product 6	1.259	57.367	6.674
Product 6	Product 9 Product 12 Product 7	1.018	56.187	6.109
Product 8	Product 13 Product 12 Product 7	1.044	51.285	6.076
Product 8	Product 13 Product 10 Product 7	1.249	50.632	5.998
Product 13	Product 8 Product 10	1.3	63.026	5.547

Although the *confidence* and the *lift* could give high values, some measurements were conducted to evaluate and compare the associations that provided clear and useful insight. A common approach is to take the association rules and divide them into three categories, as follows (Lantz, 2015).

a) The aim in using association rule mining is to figure out valuable patterns that provide a meaningful insight from datasets that seem obvious once discover.

b) Sometime, the rules are trivial, not worth to be considered as valuable patterns. Although the rule is clear, sometimes it is not appropriate.

c) When the connection between the items is ambiguous, the rule is called inexplicable. It means that figuring out how to use the insight from the patter is absurd.

In this case, suppose that giving the preceding rule, the marketing team has enthusiasm about the possibility of creating an advertisement to promote '*product* 8' and '*product* 10', which are now in season (i.e., Eid Mubarak season for Muslim people or Christmas for Christian people). Before finalizing the campaign, however, the rules reveal that '*product* 8' and '*product* 10' are often purchased with other items, as shown in Table 3.6 as follows.

Table 3.6 An actionable rule.

Consequent	Antecedent	Support (%)	Confidence (%)	Lift
Product 13	Product 8 Product 10	1.3	63.026	5.547
Product 7	Product 8 Product 10	1.3	63.375	1.905

There are two rules involving '*product* 8' and '*product* 10'. A rule with *lift* of 5.547 seems to be interesting enough to be called an actionable rule than another one. Meanwhile, the combination of '*product* 13' and '*product* 12' are also purchased frequently with '*product* 10' with *lift* of 3.624, but the '*product* 7' has lower *lift* score (with *lift* of 2.092) than '*product* 10' as shown in Table 3.7. It is a hidden gem because it has a stronger interrelationship than the ‘*product* 10’ with ‘*product* 13’ and ‘*product* 12’ on the web graph.

Table 3.7 Rules with lower lift score become a hidden gem after evaluation steps.

Consequent	**Antecedent**	**Support (%)**	**Confidence (%)**	**Lift**
Product 10	Product 13 Product 12	1.5	61.593	3.624
Product 7	Product 13 Product 12	1.5	69.608	2.092

Meanwhile, the processing time (elapsed time for the model build) for both algorithms with the same setting have variation when executing the algorithms. In this case, the Apriori algorithm has better performance when it is compared with the CARMA algorithm shown in Figure 3.4 as follows.

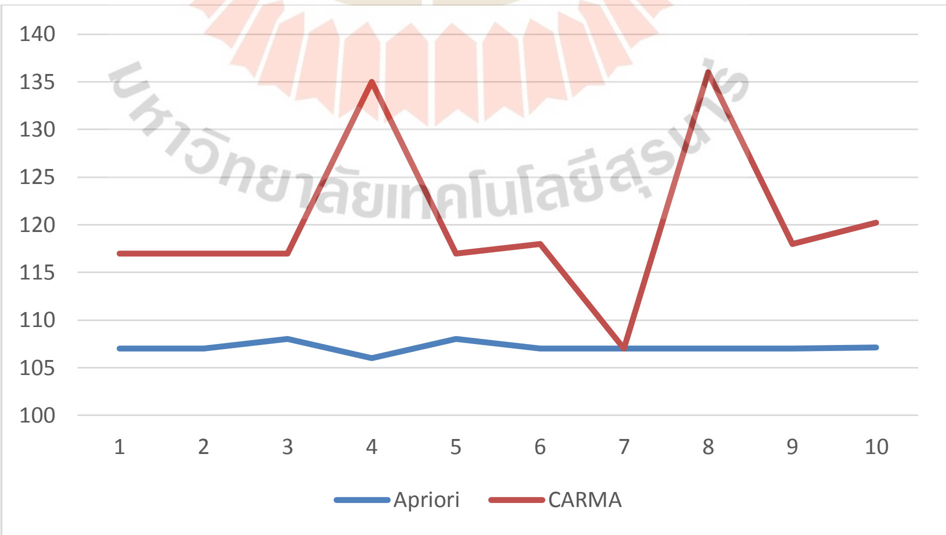


Figure 3.4 Performance comparison between the Apriori and the CARMA algorithm.

between the Apriori and the CARMA algorithm. Information from Figure 3.4 above shows that the range of elapsed time for the model build using the Apriori algorithm is between 106 seconds and 108 seconds (107 seconds for average time). It is faster than the CARMA algorithm, which has ranged between 107 and 136 seconds (120 seconds for average time). Furthermore, the identification of products in this case study depends on the datasets used, and the result of the number of product recommendation has no difference using both Apriori and CARMA algorithm. Although the CARMA algorithm performs faster, the result is not satisfying to satisfy business needs when it is compared to the Apriori algorithm. After finishing the general procedure to create the relevant model, some general strategies that can be proposed listed as followed.

a) Create a new retail store layout to increase sales. Start with attractive window display and design effective shelf display or place the product optimally based on the best model which has been chosen. For example, ‘*product* 9’ should be put near ‘*product* 11’ and ‘*product* 9’. Additional store layout ideas such as have a good lighting, use colour to convince customers (red colour indicates urgency and promotes sales), offer welcoming tones, and speed up the checkout process can be considered as business strategies to be implemented.

b) Use product categorization. Offering a promotional campaign related product in customers basket based on the best model when a customer conducts the transaction with the company;

c) Response campaign documentation. Documenting the promotional campaign whether a customer accepts or reject the promotional campaign for further analysis.

d) Information technology and system upgrade. Know every detail about meaningful orientation to apply these techniques into real-time information systems with integration of logistic and distribution systems, retail systems, customer service systems, transportation, and warehouses systems in the company. In the COVID 19 pandemic era where some countries implement lockdown policy, conventional or traditional business models without sophisticated technology integration will lead companies into bankruptcy.

These strategies will also change the relationship between customers and a company in marketing their products, which can make promotional activities measurable than previously. The important things to be noted is to document the all implementation of business strategies as a whole and hold a sustainable action review with all stakeholders and go through lesson learned.

3.5 Conclusion

Knowledge is power. Using association rules as an unsupervised learning process, a company will have the capability of extracting knowledge from a large database without any prior knowledge of what patterns to seek. After conducting this case study, here is a list of meaningful conclusions from these models and recommendations for companies:

a) This manuscript presents a novel implementation of big data analytics to change the way a company developing a new strategy for its customers, especially in a gold, silver, and precious metal trading company in a developing country like Indonesia.

b) The challenge in implementing big data analytics is that the infrastructure and technology to analyze it. The more data collected by a company, the higher computational technologies and infrastructure are required.

c) With the Apriori algorithm and CARMA algorithm for modeling in this manuscript, a company can explain in detail how the association of each product can be determined. In case of each segmentation using both procedures, the company can promote the possible products to its customers and deliver attractive offers or advanced service as promotional campaign strategies.

In addition, this manuscript is not a theoretical paper but a technical paper which can be useful guidance for managers, professional, or lecturer in implementing BDA. Further research should focus on observing the development of this big data analytics implementation and develop sustainability and real-time model for developing a new model. Once again, if companies really want to improve their business and performance, they should consider implementing big data analytics and embrace its challenges. In this 4IR era, It not only supports decision-makers to make the most valuable decision based on measurable data, but it also can achieve business goals easier and more effectively.

3.6 References

Agrawal, R., & Srikant, R. (1994). **A fast algorithm for mining association rules in image**, Santiago.

Aisyah, R. (2019). **Indonesia rolls out index to assess progress on Industry 4.0**. Retrieved from https://www.thejakartapost.com/news/2019/01/18/indonesia-rolls-out-index-to-assess-progress-on-industry-4-0.html.

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment. **International Journal of Production Economics.** 182: 113-131. doi:10.1016/j.ijpe.2016.08.018.

Alani, A. A., Ahmed, F. D., Majid, M. A., & Ahmad, M. S. (2018). Big Data Analytics for Healthcare Organizations a Case Study of the Iraqi Healthcare Sector. **Advanced Science Letters**. 24(10): 7783-7789. doi:10.1166/asl. 2018.13017.

Batarseh, F. A., & Latif, E. A. (2016). Assessing the Quality of Service Using Big Data Analytics. **Big Data Research**. 4: 13-24. doi:10.1016/j.bdr.2015.10.001.

Benitez, J., Llorens, J., & Braojos, J. (2018). How information technology influences opportunity exploration and exploitation firm's capabilities. **Information & Management**. 55(4): 508-523. doi:10.1016/j.im.2018.03.001.

Dallasega, P., Rauch, E., & Linder, C. (2018). Industry 4.0 as an enabler of proximity for construction supply chains: A systematic literature review. **Computers in Industry.** 99: 205-225. doi:10.1016/j.compind.2018.03.039.

Deputi Bidang Protokol-Pers-dan Media Sekretariat, P. (2019). Presiden Jokowi Bahas Implementasi Peta Jalan Industri 4.0. Retrieved from http://ksp.go.id/ presiden-jokowi-bahas-implementasi-peta-jalan-industri-4-0/index.html.

Field, A. (2018). **Discovering Statistics Using IBM SPSS Statistics** (J. Seaman Ed. 5th Edition ed.): SAGE.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**. 165: 234-246. doi:10.1016/j.ijpe.2014.12.031.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. **Journal of Business Research**. 70: 308-317. doi:10.1016/j.jbusres.2016.08.004.

Hidber, C. (1999, 1999). **Online association rule mining.** New York, USA.

Honarvar, A. R., & Sami, A. (2019). Towards Sustainable Smart City by Particulate Matter Prediction Using Urban Big Data, Excluding Expensive Air Pollution Infrastructures. **Big Data Research**. 17: 56-65. doi:10.1016/j.bdr.2018.05.006.

Huang, Y., Wang, X., & Shia, B.-C. (2009). Efficiency and Consistency Study on Carma. **Paper presented at the 2009 Fifth International Joint Conference on INC, IMS and IDC**.

Kamarul Bahrin, M. A., Othman, M. F., Nor Azli, N. H., & Talib, M. F. (2016). Industry 4.0: A Review on Industrial Automation and Robotic. **Journal Teknologi**. 78(6-13): 2180-3722. doi:10.11113/jt.v78.9285.

Kamble, S. S., Gunasekaran, A., & Gawankar, S. A. (2018). Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. **Process Safety and Environmental Protection**. 117: 408-425. doi:10.1016/j.psep.2018.05.009.

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. **Procedia Computer Science**. 85(Cms): 78-85. doi:10.1016/j.procs.2016.05.180.

Kementrian Perindustrian, R. I. (2020). **Videografis Indonesia Industry 4.0 Readiness Index 4.0 (INDI 4.0)**. Retrieved from https://www.youtube.com/watch?v=Q1q50KdaKx4.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. **Big Data & Society**. 1(1). doi:10.1177/2053951714528481.

Koh, L., Orzes, G., & Jia, F. (2019). The fourth industrial revolution (Industry 4.0): technologies disruption on operations and supply chain management.

International Journal of Operations & Production Management. 39(6/7/8): 817-828. doi:10.1108/IJOPM-08-2019-788.

Lantz, B. (2015). **Machine Learning with R**: Packt Publishing.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). **Big Data: The Next Frontier for Innovation, Competition, and Productivity**. Retrieved from Seoul: www.mckinsey.com/mgi.

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**. 25(2): 137-166. doi:10.1017/S0269888910000032.

Marr, B. (2017). **The 6 Top Data Jobs In 2018**. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/09/the-6-top-data-jobs-in-2018/# 6f7cf63e430d.

Marr, B. (2018). **How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read**. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how much data do we create every day the mind blowing stats everyone should-read/#6145c9b560ba.

Moyne, J., & Iskandar, J. (2017). **Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing. Processes**. 5(3). doi:10.3390/pr5030039.

Naimur Rahman, M., Esmailpour, A., & Zhao, J. (2016). Machine Learning with Big Data An Efficient Electricity Generation Forecasting System. **Big Data Research**. 5: 9-15. doi:10.1016/j.bdr.2016.02.002.

Pereira, A. C., & Romero, F. (2017). A review of the meanings and the implications of the Industry 4.0 concept. **Procedia Manufacturing**. 13: 1206-1214. doi:10.1016/j.promfg.2017.09.032.

Piccarozzi, M., Aquilani, B., & Gatti, C. (2018). Industry 4.0 in Management Studies: A Systematic Literature Review. **Sustainability**. 10(10). doi:10.3390/su10103821.

Schwab, K. (2016a). **The Fourth Industrial Revolution**: **World Economic Forum**.

Schwab, K. (2016b). **The Fourth Industrial Revolution: what it means and how to respond**. Retrieved from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Schwab, K. (2016c). **The Global Competitiveness Report. Retrieved from Geneva**: http://www3.weforum.org/docs/GCR2016-2017/05FullReport/TheGlobal Competi-tivenessReport2016-2017_FINAL.pdf.

Valarmathi, D. A. (2017). Market Basket Analysis for Mobile Showroom. **International Journal for Research in Applied Science and Engineering Technology**. V(X): 1279-1284. doi:10.22214/ijraset.2017.10185.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. **Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining(24959)**. 29-39.

Yudhistyra, W. I., Raungratanaamporn, I. s., & Ratanavaraha, V. (2019, 2019). **Big Data Analytics Techniques Exploration for Customers Analysis to Gain Competitive Advantage**, Bangkok.

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I. s., & Ratanavaraha, V. (2020). Exploring Big Data Research: A Review of Published Articles from 2010 to 2018 Related to Logistics and Supply Chains. **Operations and Supply Chain Management: An International Journal**. 13(2): 134-149. doi:10.31387/oscm0410258.

CHAPTER IV

IMPLEMENTATION OF BIG DATA ANALYTICS IN CUSTOMER FOCUSED SUPPLY CHAIN MANAGEMENT: The Combination of Data Mining and Behavior Scoring for Analyzing the Customers in a Gold, Silver and Precious Metal Trading Company in Indonesia

4.1 Abstract

Making Indonesia 4.0 is a roadmap to implement industrial digitalization in Indonesia, where every business in Indonesia needs to adopt a new approach for supporting their daily business. A few years ago, a business would have gathered information, ran analytics, unearthed information that can be used for future decisions. In the meantime, data is everywhere today (referring to big data), which is possible for businesses to analyze it and get answers from it immediately. Unfortunately, many companies are still having difficulties analyzing and implementing Big Data Analytics (BDA) in Indonesia. Hence, the objective of this manuscript is to conduct BDA implementation in a company in Indonesia. To be more specific, it is focusing on customers to know precisely the characteristics of customers and know the best way to market the products to the customers immediately by segmenting the customers. The methodology used in this research is CRISP-DM and key steps for customer analysis combined with data mining and behavior scoring method. The results of this study are some

implications and recommendations related to the implementation of big data analytics, which will be beneficial to companies and enable them to work faster and stay agile with big data, particularly companies located in a developing country like Indonesia.

4.2 Introduction

Customer Relationship Management (*forth* CRM) is a core part in achieving successful Supply Chain Management (SCM) (Bullington and Bullington, 2005). Therefore, it is important for companies to successfully attract new customers, support existing high-value customers, discover their behavior, and make a considerable effort by satisfying their needs for successful CRM. However, instead of giving the similar enticement for customers equally, companies should select the customers who can bring them benefits in regard to their behaviors, demography, geography, images, or psychographic because they all are unique and not equal (Beane and Ennis, 1987, Dyche, 2001). In this case, customer segmentation can be used to identify and illustrate those customers or '*sets of buyers*', which then become the targeted customers for the company's marketing plans. It is no wonder that many researchers have focused on customer relationship management in various industries, particularly in segmenting customers to strengthen the company's CRM or to develop appropriate promotion strategies for different customer groups or clusters (Wu and Lin, 2005, Yeh et al., 2009). Until now, some existing customers segmentation approaches have been used in the hospital (Chen et al., 2012, Lee, 2012) and healthcare field (Wu et al., 2014, Wei et al., 2012), Fast Moving Goods Company (FMGC) (Chang and Tsai, 2011) or retail stores (You et al., 2015, Abirami and Pattabiraman, 2016, Chen et al., 2009), e-commerce or online stores (Hong and Kim, 2012), global pizza restaurants (Sarvari et al., 2016), five-star hotels (Dursun and Caber, 2016), commercial banks

(Ansari and Riasi, 2016), health and beauty companies (Khajvand et al., 2011), and telecommunication providers (Song et al., 2017). Meanwhile, this study focuses on both the customer data and their transactional behavior based on the data provided by a gold, silver, and precious metal trading company in Indonesia. As a result, the main contributions of this manuscript are:

a) it presents analytics with large amount of data to help a company change the way producing its strategies and strengthen its position to stay ahead of the competition.

b) it minimizes any investment risk on researching and testing the market, product, or ideas.

c) it facilitates useful decision-making models based on evidence from BDA and frameworks that will make company search for opportunities easier and more effective, and company decision making more in touch with reality; and

d) it has impact on changing seller-buyer relationship/culture in a company.

4.3 Literature Review

4.3.1 Related works

Chen et al. (2012) identified patients to identify the demand and adequately distribute the resource in the hospital. They used a clustering model, consisting of two stages, in a target customer segment through initially integrating the Recency-Frequency-Monetary (RFM) attributes and K-Means clustering, and combining the global discretization method and the rough set theory. Lee (2012) analyzed 14,072 patients in a university hospital to find loyal patients going to a medical center frequently and modeled their medical service usage patterns using cluster analysis via a decision tree algorithm. Wu et al. (2014) combined cluster analysis and the Length-Recency-Frequency-Monetary (LFRM) model to analyze patients' values. In a pediatric dental clinic in Taiwan, Self-Organizing Maps (SOM) and K-Means clustering were utilized to divide 1,462 patients into twelve groups. The methodology used by Wu et al. was almost similar to the one used by Wei et al. (2012) to segment the patients into twelve clusters in their research.

Chang and Tsai (2011) offered a model, namely, Group RFM (GRFM). Their model performs purchased item-constrained clustering from both condition of each customer and their purchased items to respond to the market-oriented demands or to determine when is the most suitable time to deliver a specific promotion campaign for customers. You et al. (2015) combined the RFM model, decision tree algorithm, and K-Means clustering for proposing a framework that can identify the potential characteristics of different customer categories. Then this framework can develop suitable strategies that can significantly minimize the company inventory for each customer segment from a Chinese company.

Abirami and Pattabiraman (2016) used association rules mining algorithms and RFM models to analyze and predict customer behaviors based on the data in the retail store. They succeeded in finding out customer shopping habits over time and in achieving useful information on customer purchasing behavior for managerial decision-making. Chen et al. (2009) applied a Sequential Pattern Mining (SPM) method based on the e-commerce dataset. Hong and Kim (2012) identified customer values by segmenting customers into several groups that have similar ideas to buy by applying both SOM and K-Means clustering into a single model based on customer psychographic data. Sarvari et al. (2016) combined association rule mining algorithms and RFM models for segmenting the customers based on demographic and transaction database in a global pizza restaurant.

Dursun and Caber (2016) focused on outlining moneymaking hotel customers by clustering the customers using RFM analysis into eight clusters based on three five-star hotels operating in Antalya, Turkey. Ansari and Riasi (2016) identified the main clusters of bank clients by using two-step scalable clustering to help commercial banks to profile their clients better and design marketing strategies efficiently. Khajvand et al. (2011) estimated customer lifetime value based on RFM analysis and K-Means clustering in a health and beauty company. Song et al. (2017) regularized the clustering outcomes successfully with the RFM model by integrating Multiple Corresponding Analysis (MCA). They also expanded these approaches to many levels for the construction of the time interval of the RFM model. Hwang and Jun (2014) proposed supervised learning paths to tackle the sparsing dataset issue in the training dataset called a cold-start problem.

4.3.2 Recency, Frequency, and Monetary (RFM)

Customer lifetime value for a company refers to the presence of net profit or loss value to the company from a customer based on the transaction of that customer with the company (Jain and Singh, 2002, Gupta and Lehmann, 2003). Generally, it is evaluated using an approach of RFM analysis, which is widely used to measure the value of customers based on their purchasing behavior history (Hu and Yeh, 2014). The terms first defined by Bult and Wansbeek (1995) as (Bult and Wansbeek, 1995): (1) Recency (R) as days since the last purchase; it is identified by analyzing the number of days since the customer's last purchase. Then deduct the most recent purchase data from today to calculate the recency value. (2) Frequency (F) as the total number of transactions made within a certain time. It is identified by analyzing how many times the customers have purchased from the store (for example, the identification of loyal customers can be indicated by high-frequency value). (3) Monetary (M) as the total money spent by the customers during a certain time. If the customer has a high monetary value, it indicates that the company should give more attention on this customer.

4.3.3 TwoStep hierarchical agglomerative clustering

Clustering analysis is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). This agglomerative algorithm measures the distance between all objects and assigns each object to an initial cluster. If there are nominally scale variables, similarity measures such as Tanimoto, Simple-matching, or Russel & Rao coefficients are used to determine the similarity of the objects. If there are metrical valuables (at least interval scale), distance measures are used to determine the dissimilarity. For example, object x and object y are described by $variable_1$, $variable_2$, ... $variable_n$ = $x_1, x_2, \ldots, x_n$ and ($y_1, y_2, \ldots, y_n$). Using the vector components x_i and y_i, the metrics are defined as follows:

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right) \tag{1}$$

considering two components per vector, it is $d = (|x_1 - y_1|^r + |x_2 - y_2|^r)^{\frac{1}{r}}$ with specific value of r, this becomes:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

and

$$d = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3}$$

All of the objects are organized into the form of a tree in the background. If this tree exceeds a specific size, it will be reorganized. Due to the step-by-step analysis of each object, the result depends on the order of the records in the dataset. Important assumptions in using TwoStep hierarchical agglomerative clustering are that the categorical variables are multinomially distributed, and continuous variables are assumed to be normally distributed. Hence, they should be transformed in advance. In case of using a hierarchical clustering algorithm, specifying the number of clusters to be determine is unimportant (Wendler & Gröttrup, 2016). However, Bayesian Information Criterion (BIC), or Akaike's Information Criterion (AIC) can be used to determine the optimal number of clusters in this algorithm (Boone, 2011). The AIC can be calculated as follows:

$$AIC_{model} = -2\mathrm{L} + 2\mathrm{d} \tag{4}$$

where L is the maximized log-likelihood for the model and d is the number of independent parameters to be estimated by the model (Boone, 2011). Meanwhile, the BIC can be calculated as follows:

$$BIC_{model} = -2\mathrm{L} + \ln(N) \times d \tag{5}$$

where N is the number of objects in the dataset (Boone, 2011). Another approach to determine approximate number of cluster can be calculated as follows (Mardia et al., 1994):

$$\sqrt{\frac{number\ of\ objects}{2}} \tag{6}$$

4.3.4 K-Means clustering

Unlike the hierarchical clustering algorithms which compare all objects one-by-one with all other objects (it is time-consuming and especially hard to handle with large dataset), K-Means clustering does not necessarily analyze the distance between all objects. For the first step, the K-Means clustering determines a cluster center within the data and assign each object to the cluster center with the smallest distance. The cluster centers are recalculated, and then optimized by re-arranging some objects. The iteration process ends if it does not improve the quality (i.e. no object is assigned to another cluster) (IBM, 2016). The more detailed steps of this algorithm can be described as follows (IBM, 2016):

a) Specify the number of clusters k.

b) If there are metrical variable, transform it to become 0 and 1 values by using the following formula:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \tag{7}$$

where is the rescaled value of input field x for record i, x_i is the original value of x for record i, x_{min} is the minimum value of x for all record, x_{max} is the maximum value of x for all record.

c) The k centered clusters are defined as follows:

i. In the dataset, the value of the first record is labelled as the initial cluster center.

ii. Distance is calculated from all records to the cluster centers so far labelled.

iii. The values from records with the largest distance to all cluster centers are labelled as a new cluster center.

iv. The process stops if the number of clusters equals the number predefined by the user, i.e. until k cluster centers are defined.

d) The object is assigned to the cluster center with minimal distance by using the squared Euclidean distance between each record or object.

e) Using the 'average' of the object assigned to the cluster, the centered clusters are updated.

f) This process stops when the iteration takes place or there is no change in the centered cluster recalculation.

4.3.5 Cluster Evaluation Algorithms

It is important to note that the clustering methods (TwoStep hierarchical agglomerative and K-Means clustering algorithms) do not provide any help with finding an appropriate description of each cluster. How to best describe each cluster should be figured out by the researchers. For evaluating the clustering models, measures used in this research are described as follows (IBM, 2016)(Kaufman & Rousseeuw, 2005)(Tan et al., 2019).

a) The silhouette coefficient which combines the cluster cohesion concept (contained tightly cohesive cluster) and cluster separation (favoring models which contain highly separated cluster). It is simple the average over all cases of the equation.

$$SC = \frac{1}{N}\sum_{i=1}^{N}\frac{minD_{ij}, j\epsilon C_i - D_{ic_i}}{max(min\{D_{ij}, j\epsilon C_i\}, D_{ic_i})} \tag{8}$$

where $\mathcal{D}_{ij}$ is the distance between case i and the centroid of cluster j, $\mathcal{C}_i$ denotes cluster labels that do not include case i as a member, while c_i is the cluster label that includes case i. If $max(min\{\mathcal{D}_{ij}, j\epsilon\mathcal{C}_i\}, \mathcal{D}_{ic_i})$ equals 0, the Silhouette of case i is not used in the average equation.

b) The Sum of Squares Error (SSE) and Sum of Square Between (SSB) which can be calculated as follows:

$$Average\ SSE = \frac{1}{N}\sum_{j\epsilon C}\sum_{i\epsilon j} D_{ij}^2 \tag{9}$$

and

$$Average\ SSB = \frac{1}{N}\sum_{j\epsilon C} N_j D_j^2 \tag{10}$$

c) The predictor importance which indicates how well the variable can differentiate different cluster. It is displayed in a model nugget which indicates the relative importance of each input for the particular model. When the number of attributes becomes large in the dataset, exploratory analysis of the all the predictors might be infeasible, and concentrating on those with strong relationships with the outcome might be an effective triaging strategy (Kuhn & Johnson, 2013). It is better to rank the predictors in this manner which can be very useful when sifting through large amounts of data. The higher the importance measures (for numeric and discrete variables), the less likely the variation for a variable between cluster is due to some chances and the more likely due to some underlying differences.

4.4 Case Study

In order to achieve a successful project, to ensure high quality of knowledge discovery, to provide means for evaluating the effectiveness of the results, to increase implementation efficiency, and to document the experiences in this research, the process in conducting case study is based on the combination of key steps for customers analysis methodology (Yudhistyra et al., 2019) and the CRIPS-DM (Wirth, 2000)(IBM, 2018)(Chapman et al., 2004) which is an open standard process model used by data mining experts as shown in Figure 4.1 as follows.

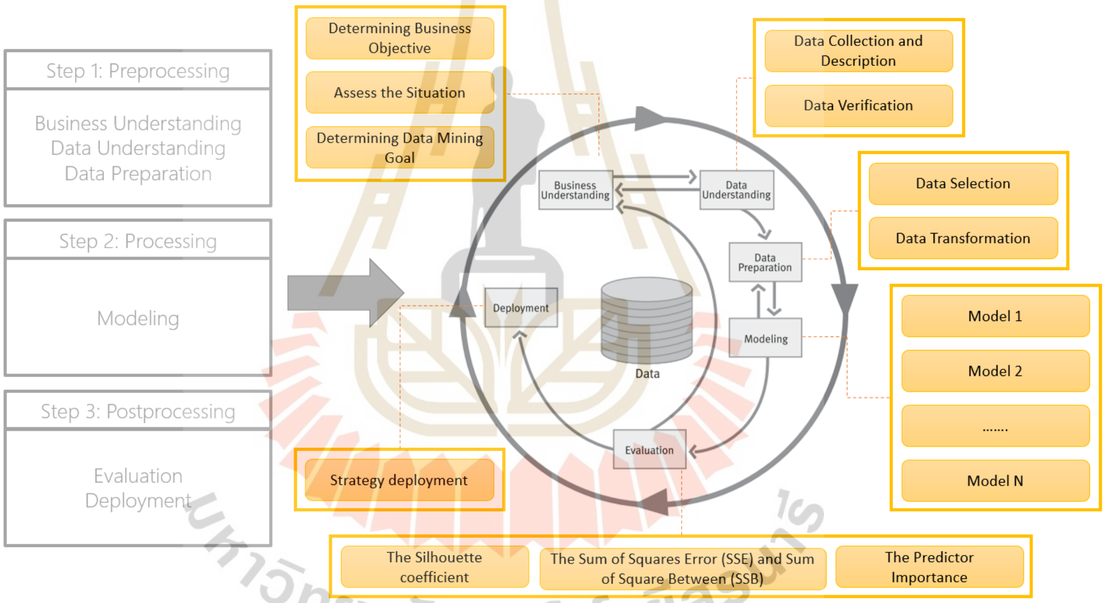


Figure 4.1 The combination of CRISP-DM and key steps for customers analysis.

4.4.1 Business Understanding

The decision-makers in a gold, silver, precious metal trading company in Indonesia intend to develop strategies related to their customers to satisfy customer needs and improve the CRM systems. Until now, the promotional systems used by the company for its customers are only websites and employees' social media such as

Facebook, Twitter, Instagram, or WhatsApp application. Sometimes this company uses newspapers, which can be concluded that this company treats all its customers equally. Sometimes this company uses newspapers, which can be concluded that this company treats all its customers equally. Meanwhile, there are some technological and resource availability constraints that are not compatible with the big data available in the company. All the requirements related to legal and security issues had been resolved before data collection was conducted. Thus, this case study determines the success criteria for developing CRM company strategies as follows: (1) gaining insight from the past with descriptive analytics; and (2) making customer segmentation models based on their purchasing behaviors with data mining and behavior scoring implementation. For achieving data mining goals, there is a list of project plan produced in this case study.

a) Data mining software such as R, Modeler, RapidMiner, and Tableau is considered to be used to analyze and visualize the results.

b) Unsupervised big data analytics techniques such as K-Means and TwoStep hierarchical agglomerative clustering are performed to deals with the process of discovering newer patterns in big data. RFM analysis and value-based segmentation are also performed for the knowledge discovery process.

c) The iteration or repetition of the modeling and evaluation process is adjusted based on the silhouette coefficient, the sum of square and the sum of square between, and the predictor importance. The best model is selected after a lot of repetition with similar results.

4.4.2 Data Understanding and Data Preparation

The initial datasets were collected and exported from the company database (i.e., product database, customer database, transactional data, etc.) covering July 2010 to October 2019. The dataset consists of 3,986,872 observations with 248,856 customers. The descriptions of the data (i.e., primary keys, attribute, and type of the data) are provided by the company shown in Table 4.1 as follows.

Table 4.1 Initial data.

Attribute Name	Description
CUSTOMERS_ID	The ID of the customers
TRANS_DATE	The date of transaction
TRANS_ID	The invoice number of the order transaction
CUSTOMERS_ADDRESS	The address of the customers
DESCRIPTION	Description of the product
QUANTITY_TOTAL	Quantity total of the product
PRICE_TOTAL	Price total of the product

To ensure the quality of the data, the initial data examination (Chatfield, 1995), which has a lot in common with exploratory data analysis (Tuckey, 1993) was performed. It included removing the null of the tuple in the datasets (cleaning), sorting, pivoting, hashing, grouping, aggregation, merging, checking the collinearity, outlier detection, etc. Consequently, there is a list of several unsatisfactory results when data verification steps were conducted, as follows.

a) Duplication and incorrect inputting into the datasets. Users commonly made this mistake when inputting the data, such as error customer names and bias product data.

b) Inconsistent naming standard. It is better to have a consistent naming standard for better analysis.

c) Exogeneous force problem. There was some data that have a contradiction with the facts and reality, such as the production date of each product.

Furthermore, the initial dataset was transformed and prepared into a suitable form for analysis. The detailed processes of data transformation are shown in Figure 4.2 as follows.

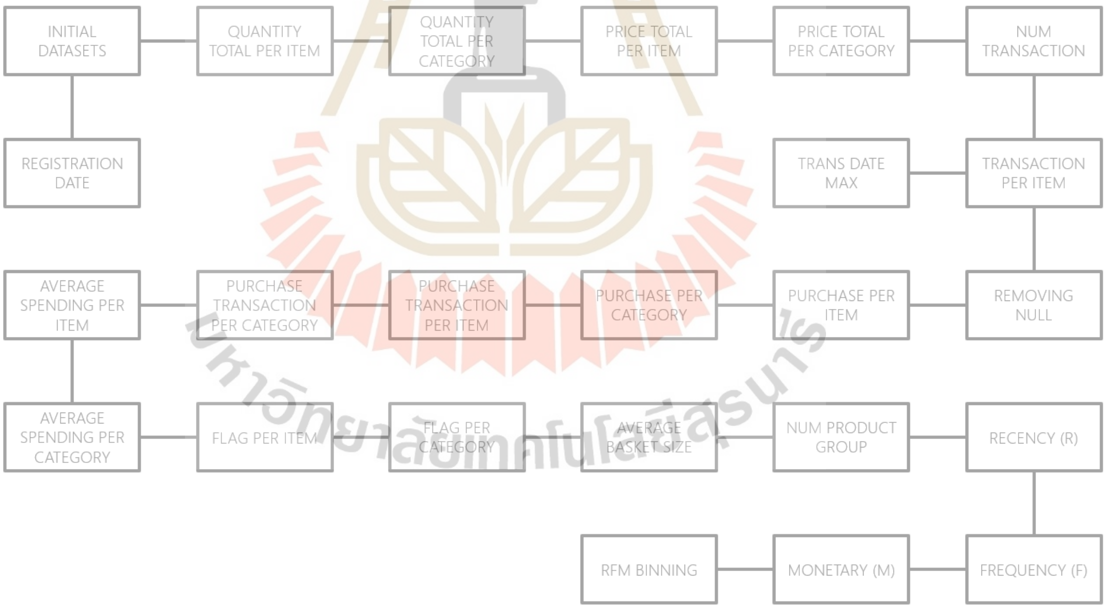


Figure 4.2 Data transformation processes.

First, the initial dataset does not contain the registration date for each customer. The registration date is needed to check the duration of the customers and the company (tenure). To figure out the registration date of each customer, the data of the first transaction made by customers can be chosen as the registration date. Second, the quantity total per item/category is aggregated after the quantity field in the initial dataset is converted from transaction data to tabular data. Third, the price total per item/category is aggregated after the price field in the initial dataset is converted from transaction data to tabular data. Fourth, the number of transactions is converted to flag (binominal) value indicating purchase. Fifth, the total transaction per item is calculated based on the flag value in the fourth step. Six, the maximum transaction date is the last date the customer made a transaction with the company. Seventh, removing the null value. Eighth, purchase per item/category is the relative spending amount per item, which is calculated by dividing each item/category with the total spending amount. Ninth, purchase transaction per item/category is the ration of transaction per item/category, which is calculated by dividing each item/category with the entire transactions. Tenth, the average per item/category is the average spending amount per item/category for each month. To compute the average spending amount every month, each item/category is split by 112 months (from July 2010 to October 2019) for former customers, but it is split with the tenure for the recent customers. Eleventh, flag per item/category is a binomial value field indicating spending for each item/category. Twelfth, the average basket size is an indication of the average spending amount per transaction, which is important to recognize customers who spend a lot per visit. Thirteenth, Recency (R) is the time since the latest sale. Fourteenth, the Frequency (F) of the sale is measured as the average of the transaction for every month through tenure months or 112 months. Fifteenth, the Monetary (M) value is derived by dividing the total spending of each customer by 112 months (tenure). Sixteenth, RFM binning is a procedure to bin the customers into five groups of 20% (quintiles) for each RFM component. Tabel 2 is the description of the transformed dataset. Finally, the transformed dataset contains 119, 905, 962 data, where the attributes of each variable are described in Table 2 as follows.

Table 4.2 Transformed data.

Attribute Name	Description
QUANTITY_TOTAL_PER_ITEM	Quantity total per item
PRICE_TOTAL_PER_ITEM	Price total per item
QUANTITY_TOTAL_PER_CATEGORY	Quantity total per category
PRICE_TOTAL_PER_CATEGORY	Price total per category
TRANSACTIONS_PER_ITEM	Number of transactions per item
NUM_TRANSACTIONS	Total of order transaction where an invoice considers as a transaction
REGISTRATION_DATE	Customer first transaction date as registration date
TENURE	Duration of the customers and the company
PRC_PER_ITEM	Relative spending amount per item
PRC_PER_CATEGORY	Relative spending amount per category
PRC_TRANS_PER-ITEM	Ratio of transaction per item
PRC_TRANS_PER_CATEGORY	Ration of transaction per category
AVG_PER_ITEM	Monthly average spending amount per item
AVG_PER_CATEGORY	Monthly average spending amount per category
FLAG_PER_ITEM	A flag (binomial) value field indicating purchase
FLAG_PER_CATEGORY	A flag (binomial) value field indicating purchase
AVG_BASKET_SIZE	Average spending amount per transaction
RECENCY	Time since last transaction
FREQUENCY	Monthly average number of order transaction
MONETARY	Monthly average purchase amount
NUM_PROD_GROUP	Number of distinct product categories with purchase
TRANS_DATE_MAX	Most recent date of transaction

4.5 Modeling, Evaluation, and Deployment

4.5.1 Descriptive Analytics

Descriptive analytics were performed to look at data, to summarize raw data on data preparation phase, to make it something that is interpretable by humans, and to analyze past events for insights as how to approach the future strategies. Therefore, through extensive data preparation, the detailed raw information form Table 4.2 gives summarized data at a customer level such as the frequency and recency of purchases; the total spending of amount; relative spending amount per product, per category, and per location; average spending amount per transaction tenure or duration of each customers; time since last transaction; monthly average number of transactions; monthly average purchase amount, in total and per group product; average transaction amount per transaction; total number of distinct product groups with purchases, etc. Moreover, an example of a density map visualization of 1.5% customers deployment in an island in Indonesia is shown in Figure 4.3 below (since the company does not allow to publish the real data, the figure is only for illustration and visualization, so it cannot be showed in detail).

There was some information that was regrettably not stored in the company database that would be exceedingly valuable for customer analysis. The datasets did not have important data such as birth date, gender, type of the customers (business-to-business or business-to-customers), number of employees if the type of customers is business to business, customers' social media, etc. It should be noted that data is the gasoline for big data analytics. The more information registered and collected into database systems, the more detailed the characteristics and behaviors of customers can be generated by big data analytics. It is never too late to strengthen the foundation of CRM by developing the current information systems to the next level. Creating many sophisticated applications (i.e., mobile app, web app, or desktop application) that are user-friendly and technologically sophisticated is one of the proper solutions for collecting essential CRM data. In this case, information collected can be transformed into useful knowledge for the company (Luck and Stephenson, 2009).

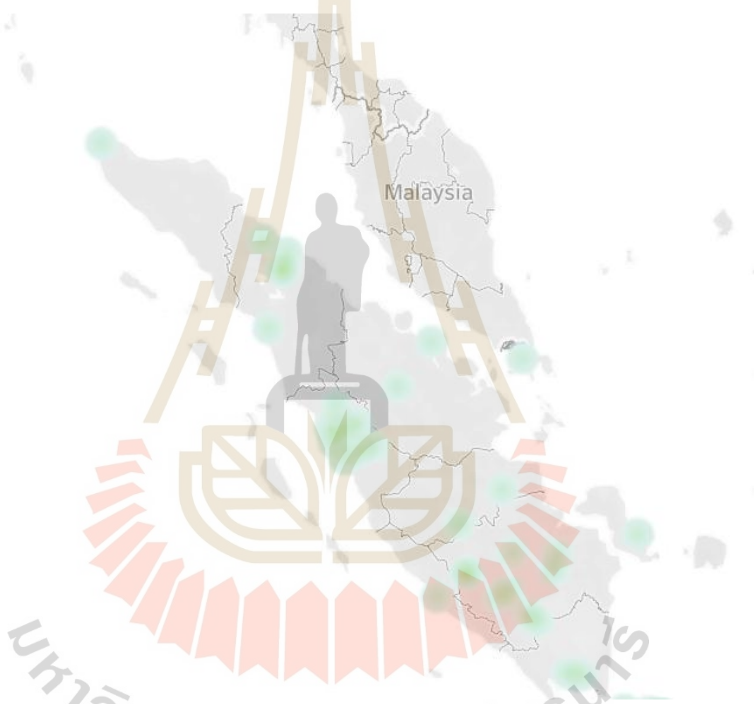


Figure 4.3 An ilustration of a density map of 1.5% customers deployment on an island in Indonesia (for visualization only).

4.5.2 Customers Analytics

As mentioned before, all customers are not equal. Some of the customers are more valuable than others. Identifying the valuable customers and understanding them should be considered as a top priority for developing customers prioritization strategies. It can enable service-level differentiation and prioritization in order to assign each customer to a segment according to his or her value. The significant dimension of the customer's value with the company could be analyzed by Value-based segmentation; RFM analysis; Two-Step clustering; and K-Means clustering.

4.5.2.1 Value Based Segmentation

It was calculated based on monetary value of each customer. The segmentation bands selected were regular customers, bronze customers, silver customers, and gold customers. The conditions for each segment are displayed in Figure 4.4 as follows.

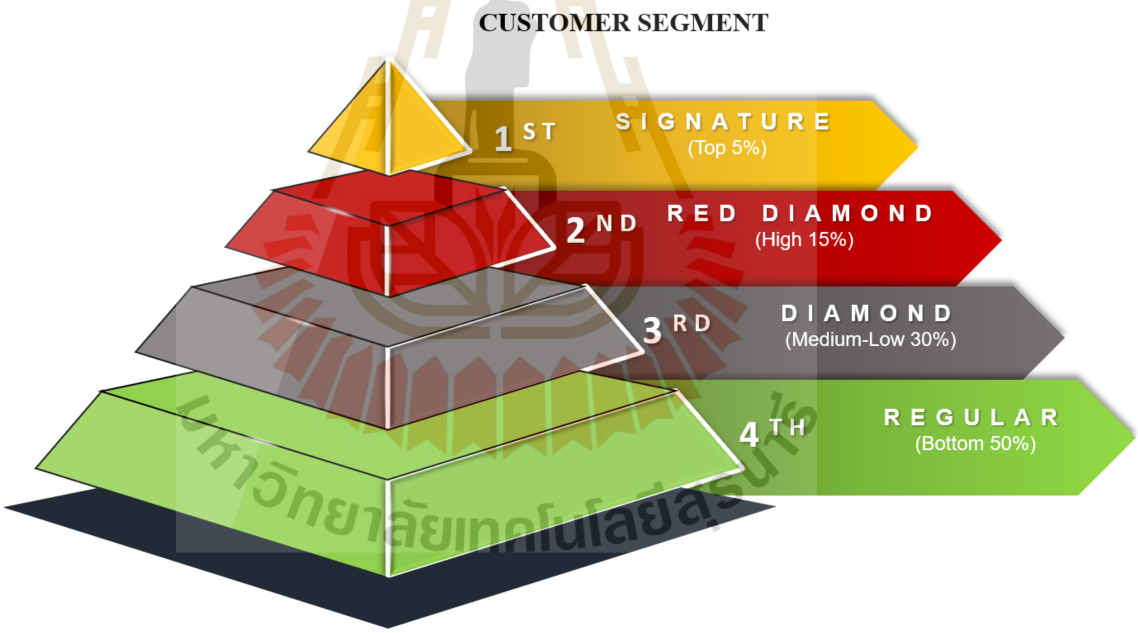


Figure 4.4 Customers segmentation based on their monetary values.

The 'regular' customers (green color) are the undersurface fifty percent of customers with the smallest spending with the company products. Around thirty percent of customers with medium-low total spending are segmented to the 'diamond' segment (grey color). The 'red diamond' customers are the fifteen percent of customers with a high spending. The 'signature' customers are the top five percent of customers with the highest spending. The objective of these segmentations is to acquire the assumed large-scaled difference for every customer in connection with spending money for the approximately 9-years period. The contribution of every segment that has been split to the company revenues is shown in Table 4.3 as follows.

Table 4.3 Contribution of each segment to the total sales amount for approximately nine years period.

		Percentage of customers (%)	**Sum percentage of total purchase amount (%)**
Segments	Signature (Top 5%)	5	75.5
	Red Diamond (High 15%)	15	12.7
	Diamond (Medium-Low 30%)	30	8.5
	Regular (Bottom 50%)	50	3.3
Total		100	100.0

The Table 4.3 above shows the information that almost three-fourth sales arose from the top 5% customers segment for about 75.5% of the total amount spent at the stores. Nevertheless, low-value customers which consist of the mass (bottom 50%) segment provided the lowest 3.3% of the total sales. The identification of value segments is prominent for the business which can provide valuable help for company in setting the appropriate action according to each customer's value.

4.5.2.2 RFM Analysis

Implementing RFM analysis is used to extend the value segmentation where the monetary value information is examined along with recency and frequency of purchase. The customers are assigned to cells labelled as 'since when (R)', 'how much (M)', and 'how often (F)' they purchase. These actual values (RFM values) for those customers are arranged and ranked into smaller number of RFM values intervals (is a number from 1-5, known as RFM score) in order to minimize the effects of minor observation errors. The distribution of customers into RFM cells is shown in Figure 4.5 below.

According to the above jitter scatter plot, the analysis about the relationship of customers with company can be assessed. In this case, the signature (the most valuable) customer segment is the customers with the RFM values equal to 555-means that the customers recency value is 5, frequency value is 5, and monetary value is 5. This analysis gives information about 30% of the total revenue at the store was spent by 10.04% of gold customers segment. Next analysis explained in the next section is the analysis of TwoStep hierarchical agglomerative clustering and K-Means clustering which is a collection of many different multivariate statistical methods for segmenting customers.

Figure 4.5 RFM jitter scatter plot.

4.5.2.3 TwoStep hierarchical agglomerative clustering

After previously dividing customers into four segments or clusters (gold, silver, bronze, and regular customers) calculated based only on monetary value, identifying segment of customers was also conducted using TwoStep hierarchical agglomerative clustering. Using this algorithm, the customers were assigned to a cluster, based on similarity or dissimilarity/distance measures of their three individual characteristics (R, F, and M scores). The importance of each continuous variable was calculated with an F-test (based on theory the three variables are important). As well as the three variables used were not independent and were not normally distributed (even after it was transformed in advance with log transformation, exponential transformation, square root transformation, inverse, so forth), Euclidian distance was used for calculating the similarity or dissimilarity/ distances between the new clusters and all other objects. In addition, the advantage of using a TwoStep node is that it determines the number of clusters automatically. Nevertheless, it is also vital to assess the clusters based on silhouette values depending on the number of cluster in detail for characteristics checking of each cluster as shown in Table 4.4 as follows.

Table 4.4 Dependency of silhouette measure and the number of cluster determined by TwoStep.

Number of clusters	Silhouette measure of cohesion and separation
2	0.587
3	0.4426
4	0.3996
5	0.4536
6	0.4624
7	0.4127
8	0.4096
9	0.431
10	0.441
11	0.4345
12	0.4295
13	0.4367

The table 4.4 above shows that the best model assessed by 0.587 of silhouette measure is the model with two clusters. However, the characteristics with two clusters was too simple to be described (for example only a customer group that has high RFM score and another group that has low RFM score). In addition, using a simple regression function indicates that for every additional number of cluster, the quality of the clustering impairs by -0.0059, in term of the silhouette measure of cohesion and separation as shown in Figure 4.6 as follow.

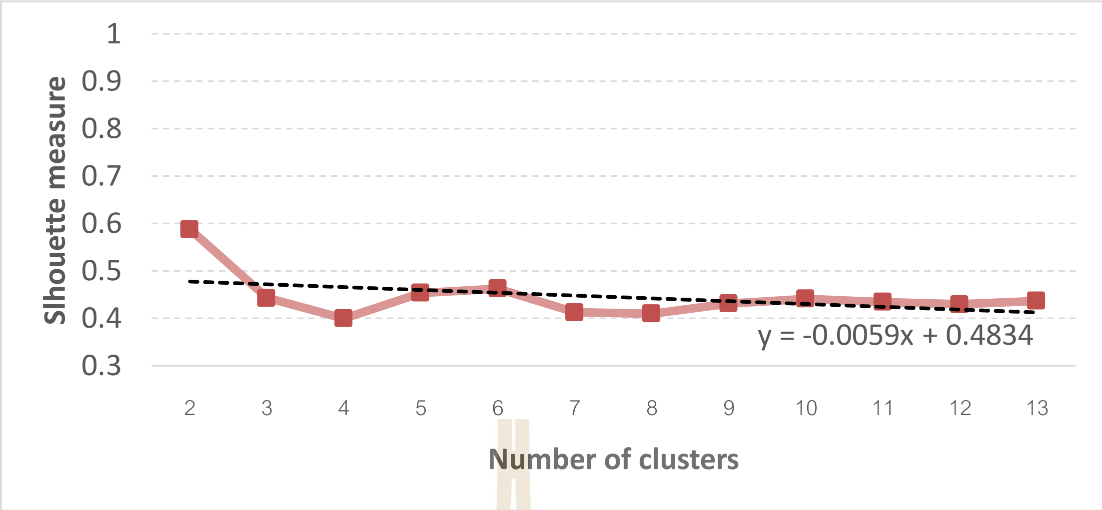


Figure 4.6 Graph of silhouette measure vs number of clusters determined by TwoStep.

So, when using TwoStep, the better option is looking for each characteristic of each cluster result and modifying the number of clusters based also on the background of application of the clustering algorithm to determine the right cluster. After the evaluation process, the model proposes to segment the customers is the model with six clusters shown in Figure 4.7 below.

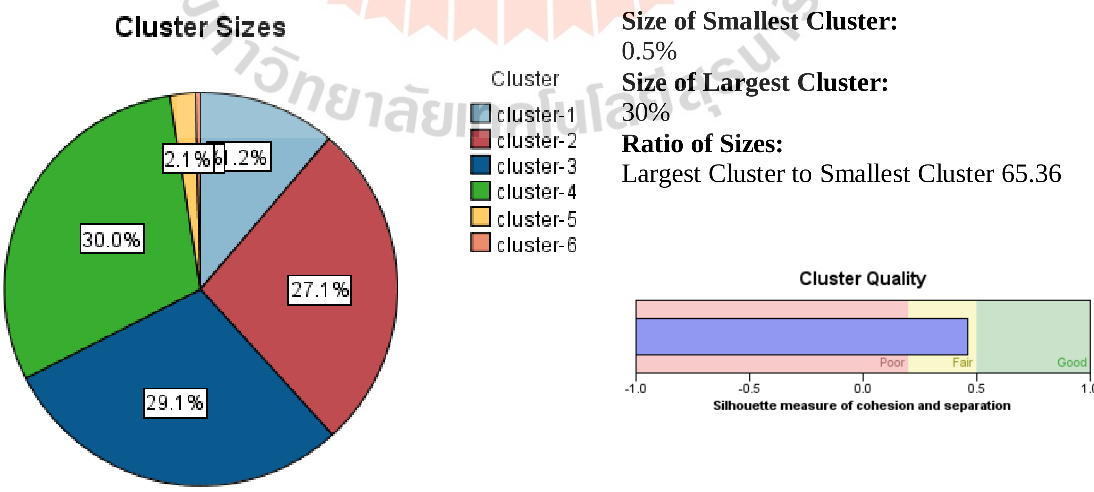


Figure 4.7 TwoStep hierarchical agglomerative clustering model with six clusters.

The above clustering model informs that the size smallest cluster equal to 0.5% of total customers, the size of largest cluster equal to 30% of total customers, and the ratio of sizes is 65.36. Moreover, the interpretation of the determined cluster is described in Table 4.5 as follows.

Table 4.5 Cluster description of the customer segmentation by TwoStep.

Cluster number	Percentage (%)	Characteristics of customers segment
1	11.2	Customer segment with an average frequency tile5 of 2.14, monetary tile5 of 3.95, and recency tile5 of 1.72.
2	27.1	Customer segment with an average frequency tile5 of 3.22, monetary tile5 of 2.47, and recency tile5 of 3.35.
3	29.1	Customer segment with an average frequency tile5 of 1.40, monetary tile5 of 1.55, and recency tile5 of 1.55.
4	30.0	Customer segment with an average frequency tile5 of 4.62, monetary tile5 of 4.40, and recency tile5 of 4.58.
5	2.1	Customer segment with an average frequency tile5 of 4.21, monetary tile5 of 4.77, and recency tile5 of 2.43.
6	0.5	Customer segment with an average frequency tile5 of 1.86, monetary tile5 of 3.16, and recency tile5 of 4.64.

As shown in Table 4.5, the customers assigned to cluster number 5 are customers who spend the most and have above average frequency, but have not returned for a long time. Offering these customers newer product and talking to them may help the company to bring back these customers. Cluster number 4 can be described as the champion segment that bought recently, bought often, and spent good money with the company. This customer segment should be engaged and asked for product reviewing. Cluster number 6 with recent customers who spent above average should be offered the awareness of the company brand that distinguish the company product from its competitors. Cluster number 1 is the customer segment who spent big money but not often with the company and long time ago. The customers in this cluster should be reconnected and be recommended the company popular products. Cluster number 2 is the customers with above average recency and frequency values but they did not spend too much. Making limited times offers and recommending products based on past purchased should be conducted by the company. The customers assigned to cluster number 3 are uninteresting to the company with lowest recency, frequency, and monetary scores. The customers in this segment should be revived their interest with reach out campaign or be ignored otherwise.

4.5.2.4 K-Means clustering

Using K-Means clustering, the rule of thumb was not used because it delivered too much clusters with number of million objects and made each cluster harder to be described. The three variables used here do not allow the implementation that precision. Therefore, starting with a lower number and deciding to use two till eleven clusters cluster became a good option. The model summary for each cluster is shown in Table 4.6 as follow.

Table 4.6 Dependency of silhouette measure and the number of cluster determined by K-Means.

Number of clusters	Silhouette measure of cohesion and separation
2	0.6004
3	0.4966
4	0.4964
5	0.4803
6	0.4627
7	0.4595
8	0.4528
9	0.4378
10	0.4698
11	0.4486
12	0.4477
13	0.4572

The result based on the Table 4.6 shows that the best model using this algorithm is the model with two clusters. A more detailed analysis was conducted to be more precise with the result since the model with two clusters were too general to be described. Moreover, a simple regression function is also performed to determine the optimal number of clusters as shown in Figure 4.7 as follows.

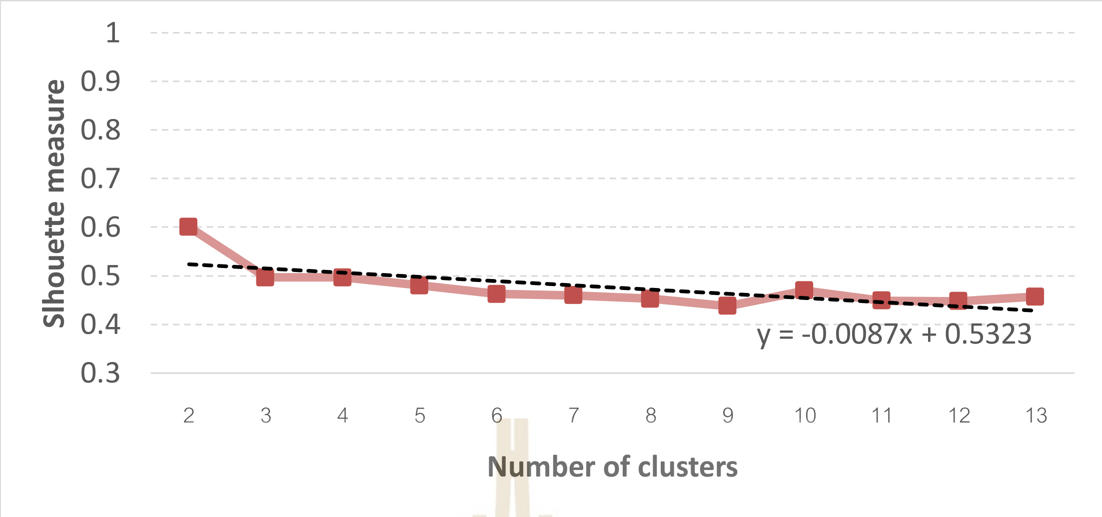


Figure 4.8 Graph of silhouette measure vs number of clusters determined by K-Means.

The graph for the K-Means demonstrates that for every additional number of cluster, the quality of the clustering impairs by -0.0087, in terms of the silhouette measure of cohesion and separation. First, a model with six clusters was assessed and checked. Unfortunately, there were many customers with lowest recency, frequency, and monetary scores. A model with four clusters was hard to be described based on their characteristics. After comparing the different models based on their silhouette values and assessing the results for each cluster based on practical and theoretical knowledge, the best model for this algorithm is the model with five clusters as shown in Figure 4.8 as follows.

The information from Figure 4.8 shows that the smallest cluster with the K-Means clustering is 12.5% of the total customers; the largest cluster is 41.2% of the total customers, and the ratio of the largest cluster to the smallest cluster is 2.6. The characteristics of each segment is shown in Table 4.7 as follows.

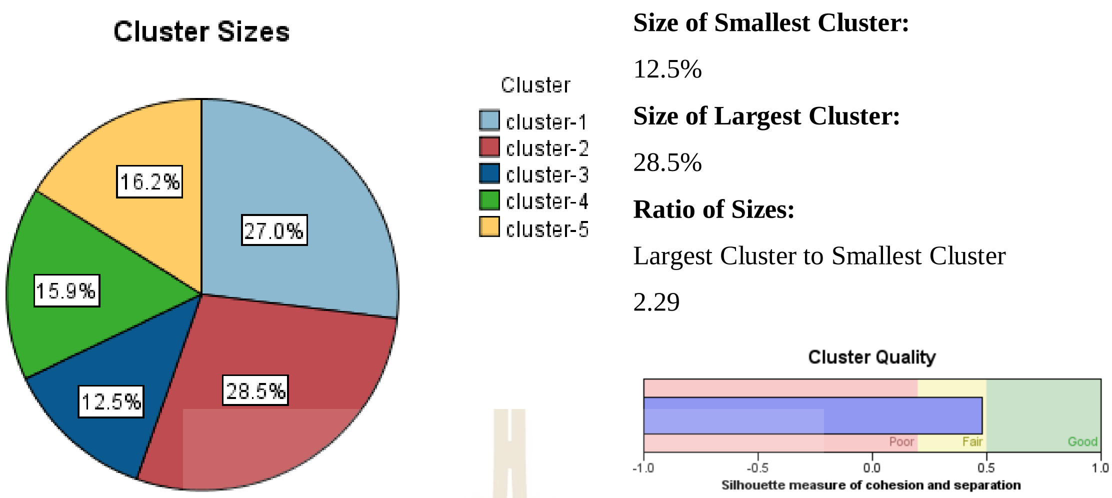


Figure 4.9 K-Means clustering model with five clusters.

Table 4.7 Cluster description of the customer segmentation by K-Means.

Cluster number	**Percentage (%)**	**Characteristics of customers segment**
1	27.0	Customer segment with an average frequency tile5 of 1.33, monetary tile5 of 1.59, and recency tile5 of 1.45.
2	28.5	Customer segment with an average frequency tile5 of 4.62, monetary tile5 of 4.54, and recency tile5 of 4.57.
3	12.5	Customer segment with an average frequency tile5 of 2.78, monetary tile5 of 1.54, and recency tile5 of 3.10.
4	15.9	Customer segment with an average frequency tile5 of 3.72, monetary tile5 of 2.89, and recency tile5 of 3.75.
5	16.2	Customer segment with an average frequency tile5 of 2.39, monetary tile5 of 3.87, and recency tile5 of 1.99.

Based on this information, cluster number 2 is the most valuable cluster with an average of 4.62 frequency, 4.54 monetary, and 4.57 recency scores. The customers in this cluster should be rewarded, and they can be an early adopter for new products. On the other hand, cluster number 1 is the cluster that is unattractive to the company with low recency, frequency, and monetary scores. Offering other relevant products and special discounts based on their purchase history can become good solutions for the customers in this cluster. The customers in clusters number 4 can be described as promising customers who have above average frequency and recency but have not spent much. Creating brand awareness (consistently building the company brand and strengthening its associations in the mind of company customers) is important to help the company business to stand out above its competitors. Customers in cluster number 5 who spent big money but may not have bought recently need to be sent a personalized email to reconnect and provided helpful resources. Cluster number 3 is a customer group that has a below-average frequency and monetary. Sharing valuable resources and recommending popular products is important to reconnect with them. The company will lose them if not reactive.

Furthermore, the challenge for clustering is to cluster the valid customer segment based on data distribution and its characteristics. Outliers may have significant importance with clustering analysis. Discovering these outliers is remarkably non-trivial, and eliminating them sometimes is not necessarily covetable, but it can influence the model results. With TwoStep algorithm and K-Means algorithm for producing models in this manuscript, a company can explain in detail how the connection of two objects can be decided. In case of each segmentation using both procedures, the company can anticipate the potential requirement of a customer group and send fascinating deals or advanced services. To be more specific, some relevant recommendations and implications that will be extremely important for the company are listed as follows.

a) More focus on specific customers and establish appropriate service options for 'priority' customers. Invest resources to give more attention to the most valuable customers and match the need of each customer segmentation.

b) Offer customers surprise gifts and services. For example, sending a handwritten note to its customers which shows that the company values its customers.

c) Send surveys. A quick online survey for its customers is important to improve Key Performance Indicator (KPI) in the company, get positive feedback, and learn from customer complaints.

d) Create a frequent communication calendar. It is important for a company to keep in touch with its customers, to educate its customers, to improve its customer knowledge, and to give special offers at regular intervals. For example, the company anniversary can be selected as a date every year to create a frequent communication calendar with its customers.

e) Provide training for end-users and audit the information systems. It will be important to develop new CRM systems (such as mobile application, chatbots messenger, blockchain, cloud technology, Internet of Things (IoT), etc.) to succeed in today's business environment and get a deeper insight in customer behavior.

f) Develop E-Gold money systems. This could be the most innovative strategy to be implemented by this company. The concept is similar to a normal Automated Teller Machine (ATM) that enables customers to perform financial transactions such as cash withdrawal, deposits, fund transfer, or account information inquires at any time without direct interaction with the company staff. The difference in the concept is in the currency used for transactions. In the current bank system, customers use the national currency (for example, Rupiah in Indonesia). Meanwhile, this system uses the gold price standard to conduct transactions. For example, a customer can withdraw money worth 1 gram of gold price from an ATM machine (about 1 million Rupiah in cash). The advantage of this system is that customer savings are not affected by fluctuations in national currency exchange rates which are often fluctuating.

4.6 Conclusion

Direct experience implementing big data analytics in a gold, silver, and precious metal company is an extremely valuable lesson. This does not only reduce the gap between the big data available now and the ability to analyze it, particularly in a developing country like Indonesia. During the implementation of big data analytics in this case study, there is a list of important notes that will be useful to support the continuous sustainable improvement in the industry 4.0 era.

a) Reduce the bureaucratical procedure to collect big data in the company. There were many challenges to get and collect the data from the company with many bureaucratical procedures to be passed. Minimizing bureaucratical procedures in the company will also help researchers to easily collect the data and give external insights that can be compared with internal knowledge to enrich the point of view before the stakeholders make business decisions.

b) Embrace the infrastructure, technology, and digital transformation challenges. The larger the dataset is analyzed, the higher the technological requirements are needed. Each developed country has at least one ‘big tech’ company such as the United States has Amazon, Microsoft, or Google; South Korea has Samsung; France has Schneider Electric, or Germany has SAP that routinely deals with big data. It is the time for a developing country like Indonesia or other developing countries to have at least a company that can develop technology and infrastructure like AI, big data, cloud computing infrastructure and security, quantum computing, Augmented Reality (AR)/Virtual Reality (VR), blockchain, cognitive computing, and IoT data management. It is massively important for the success of Making Indonesia 4.0 in Indonesia to counterbalance developed countries and to welcome industry 5.0 in the future.

c) Legal issues cause by technology development. Technology innovation should be aligned with the practice of law, particularly with the rise of the internet and big data. Where now most of every activity can be done virtually, electronically, or digitally, the law requirement should adapt to regulate and protect legal issues (for example, the privacy law, copyright law, or patent law). These issues can be overcome by a joint cooperation between governments and industries or private organizations.

In addition, this manuscript is not a theoretical paper that refers to new or established principles related to a specific field of knowledge. It is a technical paper that has a purpose to increase the store of human knowledge and to be a useful guidance for practitioners or professionals, students, or lecturers who have an interest in big data analytics. Further research should focus more on observing the development of the implementation of this big data analysis to develop sustainability and real-time models that support day-to-day business in the company.

4.7 References

Ansari, A., & Riasi, A. (2016). Taxonomy of Marketing Strategies Using Bank Customers' Clustering. **International Journal of Business and Management**. 11(7): 106. https://doi.org/10.5539/ijbm.v11n7p106.

Beane, T. P., & Ennis, D. M. (1987). Market Segmentation: A Review. **European Journal of Marketing**. 21(5): 20-42. https://doi.org/10.1108/EUM0000000004695.

Boone, D. S. (2011). Determination of the Number of Clusters in a Data Set. **International Journal of Strategic Decision Sciences**. 2(4): 1-13. https://doi.org/ 10.4018/jsds.2011100101.

Bullington, K. E., & Bullington, S. F. (2005). Stronger supply chain relationships: learning from research on strong families. **Supply Chain Management: An International Journal**. 10(3): 192-197. https://doi.org/10.1108/13598540510606241.

Bult, J. R., & Wansbeek, T. (1995). Optimal Selection for Direct Mail. **Marketing Science**. 14(4): 378-394. https://doi.org/10.1287/mksc.14.4.378.

Chang, H.-C., & Tsai, H.-P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. **Expert Systems with Applications**. 38(12): 14499-14513. https://doi.org/10.1016/j.eswa.2011.05.034.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2004). **CRISP-DM 1.0.** IBM. ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf%5Cnpapers3://publication/uuid/8209D32B-855C-47AD-9C5D-9BBBF10698AB.

Chatfield, C. (1995). **Problem Solving: A Statistician's Guide** (2nd ed.). Chapman and Hall.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. **Electronic Commerce Research and Applications**. 8(5). 241-251. https://doi.org/10.1016/j.elerap.2009.03.002.

Chen, Y.-S., Cheng, C.-H., Lai, C.-J., Hsu, C.-Y., & Syu, H.-J. (2012). Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment. **Computers in Biology and Medicine.** 42(2): 213-221. https://doi.org/10.1016/j.compbiomed.2011.11.010.

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. **Tourism Management Perspectives**. 18: 153-160. https://doi.org/10.1016/j.tmp.2016.03.001.

Dyché, J. (2001). CRM Handbook: A Business Guide to Customer Relationship Management. In **Addison Wesley**.

Gupta, S., & Lehmann, D. R. (2003). Customers as assets. **Journal of Interactive Marketing**. 17(1): 9-24. https://doi.org/10.1002/dir.10045.

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. **Expert Systems with Applications**. 39(2): 2127-2131. https://doi.org/10.1016/j.eswa.2011.07.114.

Hu, Y.-H., & Yeh, T.-W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. **Knowledge-Based Systems**. 61: 76-88. https://doi.org/10.1016/j.knosys.2014.02.009.

IBM. (2016). **IBM SPSS Modeler 18.0 Algorithms Guide**. IBM Corp.

IBM. (2018). **IBM SPSS Modeler CRISP-DM Guide**. IBM. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_18.0.0/modeler_crispdm_ddita/model er_crispdm_ddita-gentopic1.html.

Jain, D., & Singh, S. S. (2002). Customer lifetime value research in marketing: A review and future directions. **Journal of Interactive Marketing.** 16(2): 34-46. https://doi.org/10.1002/dir.10032.

Kaufman, L., & Rousseeuw, P. J. (2005). **Finding Groups in Data: An Introduction to Cluster Analysis**. John Wiley & Sons.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science. 3: 57-63. https://doi.org/10.1016/j.procs.2010.12.011.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. In **Applied Predictive Modeling**. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3.

Lee, E. W. (2012). Data Mining Application in Customer Relationship Management for Hospital Inpatients. **Healthcare Informatics Research**. 18(3): 178. https://doi.org/10.4258/hir.2012.18.3.178.

Luck, D., & Stephenson, M. L. (2009). An Evaluation of the Significant Elements of Customer Relationship Management within the Hotel Industry. **Tourism Today**. 9: 7-26.

Mardia, K. V, Kent, T. J., & Bibby J M. (1994). **Multivariate Analysis**.

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. **Kybernetes**. 45(7): 1129-1157. https://doi.org/10.1108/K-07-2015-0180.

Song, M., Zhao, X., E, H., & Ou, Z. (2017). Statistics-based CRM approach via time series segmenting RFM on large scale data. **Knowledge-Based Systems**. 132:

21-29. https://doi.org/10.1016/j.knosys.2017.05.027.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining. In **Data Mining and Data Warehousing** (2nd ed.). 17-27. Cambridge University Press. https://doi.org/10.1017/9781108635592.003.

Tuckey, J. W. (1993). **Exploratory Data Analysis: Past, Present, and Future** (Issue 302).

Vijayakumar, V., & Neelanarayanan, V. (2016). Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16’). In V. Vijayakumar & V. Neelanarayanan (Eds.), **Smart Innovation, Systems and Technologies** (Vol. 49). Springer International Publishing. https://doi.org/10.1007/978-3-319-30348-2.

Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children’s dental clinic. **Expert Systems with Applications**. 39(5): 5529-5533. https://doi.org/10.1016/j.eswa.2011.11.066.

Wendler, T., & Gröttrup, S. (2016). Data Mining with SPSS Modeler. In **Data Mining with SPSS Modeler**. Springer International Publishing. https://doi.org/10.1007/978-3-319-28709-6.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. **Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining**. 24959: 29-39. https://doi.org/10.1.1.198.5133.

Wu, H.-H., Lin, S.-Y., & Liu, C.-W. (2014). Analyzing Patients’ Values by Applying Cluster Analysis and LRFM Model in a Pediatric Dental Clinic in Taiwan. **The**

Scientific World Journal. 2014: 1-7. https://doi.org/10.1155/2014/685495.

Wu, J., & Lin, Z. (2005). Research on customer segmentation model by clustering. Proceedings of the 7th International Conference on Electronic Commerce-ICEC ’05. 113: 316. https://doi.org/10.1145/1089551.1089610.

Yeh, I.-C., Yang, K.-J., & Ting, T.-M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. **Expert Systems with Applications.** 36(3): 5866-5871. https://doi.org/10.1016/j.eswa.2008.07.018.

You, Z., Si, Y.-W., Zhang, D., Zeng, X., Leung, S. C. H., & Li, T. (2015). A decision-making framework for precision marketing. **Expert Systems with Applications**. 42(7): 3357-3367. https://doi.org/10.1016/j.eswa.2014.12.022.

Yudhistyra, W. I., Raungratanaamporn, I., & Ratanavaraha, V. (2019). Big Data Analytics Techniques Exploration for Customers Analysis to Gain Competitive Advantage. **The 16th Pacific Regional Science Conference Organization (PRSCO) Summer Institute**. 58-63.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

The conclusion of this study is summarized to answer the research objectives mentioned in Chapter I. First, conducting a comprehensive analysis of the current status of big data analytical trends (in logistic and supply chain journals and research communities) gives the result that topics related to the BDA implementation are extremely deficient. Second, direct experience in implementing BDA in a company constitute the most important lesson learned that could give researchers, professionals or practitioners, and students guidance and a new point of view for BDA implementation. Third, Evaluating the models used in BDA implementation is performed with a lot of iteration to select the best model before suggesting a new policy and recommendation for the company.

Some conclusions and recommendations have been mentioned in each chapter (Chapter II, Chapter III, and Chapter IV) including recommendations for the company (Chapter II and Chapter III). In addition, the results from Chapter II (market basket analysis) can be combined with the results from Chapter III (customer segmentation) to develop better decision making by stakeholder in a company. For example, Chapter II identifies the habits of the customers when they purchase product from the company. Meanwhile, a company should not treat all the customer equally since around 75% of the revenue comes from 5% customers only (mentioned in Chapter III). It means that these 5% customers are more profitable to the company than other customers. Giving more attention and priority to these 5% customers (such as offering surprise gift and extraordinary service) are important to keep these customers loyal to the company.

A perfect example of what can be achieved through customer loyalty can be seen from Apple and Toyota (including Lexus) brands. Most of products from both brands are successful because they focus on sweating their brand through creativity, innovation, and dedication for their existing customers. Here are of some hidden benefits that loyal customers can give to the company:

1. Loyal customers are often willing to buy more. They are usually not influenced by the price of company products because they trust the brand that they are buying from.

2. Loyal customers refer other people to your products and your company. The power 'word of mouth' referrals and endorsement from loyal customers will economize the advertising and marketing budget to acquire new customers.

3. Loyal customers give feedback and help your company grows. They understand your product and can give feedbacks for improvement. It is a critical element to develop company products.

4. Good relationship between loyal customers and the company will make everything easier for both side because they are mutual beneficial. To company, being being relational makes everything easier because loyal customers are less likely to fly off the handle at the first sign of distress.

Furthermore, there is a list of recommendations for developing further research to increase the storage of human knowledge, particularly related to customer-focused supply chain management.

1. Social media analytics. Every customer has social media. Gathering and analyzing data from customer social media such as Facebook, Instagram, LinkedIn, Twitter, etc., will help companies to track and to measure online conversations about products and companies. By paying attention to social media analytics, and measuring the insight from social media for better solution also allow companies to see where strategic opportunity exists, and where threats need to be nullified with adjustment to the strategy.

2. Market research. Big Data Analytics (BDA) is a megatrend, but it is not to replace market research. Regardless the amount of the data, nothing can be compared to the insight gained simply from talking and surveying to the others. This is where the value and the real power of market research lie in the big data world. BDA and market research should be used to complement each other to be extremely effective in understanding customer behavior.

3. Blockchain technology and Internet of Things (IoT). As future technologies, blockchain and IoT will revolutionize the way everything works, including the way people connect to the internet. Both technologies can be used to develop new supply chain concept which involves lots of parties across different location, to analyze new data and information, to protect the amount of data and makes it more manageable.

APPENDIX I

LEGAL DOCUMENTS

บันทึกข้อความ
มหาวิทยาลัยเทคโนโลยีสุรนารี

สาขาวิชาวิศวกรรมขนส่ง
สำนักวิชาวิศวกรรมศาสตร์
รับที่ 951
วันที่ 17 DEC 2019
เวลา 8.53

หน่วยงาน ฝ่ายมาตรฐานและเครือข่ายวิจัย สถาบันวิจัยและพัฒนา โทรศัพท์ 4757 โทรสาร 4750
ที่ อว 7421(4)/558 วันที่ 11 ธันวาคม 2562
เรื่อง เอกสารรับรองโครงการวิจัยในมนุษย์ (EC-62-99)

เรียน ศาสตราจารย์ ดร.วัฒนวงศ์ รัตนวราห
สำนักวิชาวิศวกรรมศาสตร์

ตามที่ท่านได้ส่งเอกสารขออนุมัติการทำวิจัยในมนุษย์ เรื่อง "การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน" (EC-62-99) มาเพื่อขอรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ นั้น

คณะกรรมการจริยธรรมการวิจัยในมนุษย์ ได้พิจารณาแล้ว เห็นว่าโครงการวิจัยดังกล่าวเป็นงานวิจัยที่ไม่ขัดต่อหลักจริยธรรมสากล และเป็นไปตามคำประกาศเฮลซิงกิ จึงสมควรให้ดำเนินการวิจัยในขอบข่ายของโครงการที่เสนอได้ ตาม COA No. 92/2562 ตั้งแต่วันที่ 10 ธันวาคม 2562 หมดอายุวันที่ 9 ธันวาคม 2563 ทั้งนี้ หากเอกสารรับรองโครงการหมดอายุเกิน 120 วัน โดยไม่มีการแจ้งต่ออายุ จะทำการปิดโครงการอัตโนมัติ หากผู้วิจัยจะดำเนินการวิจัยต่อ จะต้องยื่นขอตามระบบอีกครั้ง โปรดอ่านคำเตือนและปฏิบัติตามอย่างเคร่งครัดหลังจากได้รับเอกสารรับรองที่ระบุด้านหลังเอกสารรับรองโครงการที่แนบมาพร้อมบันทึกนี้ **โดยขอให้ท่านแจ้งปิดโครงการตามแบบฟอร์ม AF/01/13/01.0 เมื่อดำเนินการโครงการวิจัยเสร็จสิ้นแล้ว** สามารถดาวน์โหลดแบบฟอร์มได้ที่ http://ec.sut.ac.th/index.php

จึงเรียนมาเพื่อโปรดทราบและดำเนินการต่อไป

(ผู้ช่วยศาสตราจารย์ แพทย์หญิงพรทิพย์ นิ่มขุนทด)
ประธานคณะกรรมการจริยธรรมการวิจัยในมนุษย์

TO: บาร์.

COA No. 92/2562

คณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีสุรนารี

เอกสารรับรองโครงการวิจัยในมนุษย์

คณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีสุรนารี ดำเนินการให้การรับรองการพิจารณาจริยธรรมแบบเร่งรัดโครงการวิจัยตามแนวทางหลักจริยธรรมการวิจัยในมนุษย์ที่เป็นมาตรฐานสากล ได้แก่ Declaration of Helsinki, The Belmont Report, CIOMS Guideline, International Conference on Harmonization in Good Clinical Practice (ICH-GCP) and 45CFR 46.101(b)

โครงการ	: การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน
รหัสโครงการ	: EC-62-99
ชื่อหัวหน้าโครงการ	: ศาสตราจารย์ ดร.วัฒนวงศ์ รัตนวราห
สังกัด	: สำนักวิชาวิศวกรรมศาสตร์
วิธีทบทวน	: Expedited
รายงานความก้าวหน้า	: ส่งรายงานความก้าวหน้าอย่างน้อย 1ครั้ง/ปี หรือส่งรายงานฉบับสมบูรณ์หากดำเนินโครงการเสร็จสิ้นก่อน 1 ปี
เอกสารรับรอง	: ข้อเสนอโครงการ, เอกสารชี้แจงผู้เข้าร่วมการวิจัย,หนังสือแสดงเจตนายินยอม แบบสอบถาม (version 1.0, 7 ตุลาคม 2562)

ลงชื่อ...

(ผู้ช่วยศาสตราจารย์ แพทย์หญิงพรทิพย์ นิ่มขุนทด)

ประธานคณะกรรมการจริยธรรมการวิจัยในมนุษย์

มหาวิทยาลัยเทคโนโลยีสุรนารี

วันที่รับรอง : 10 ธันวาคม 2562

วันหมดอายุ : 9 ธันวาคม 2563

ทั้งนี้ การรับรองนี้มีเงื่อนไขดังที่ระบุไว้ด้านหลังทุกข้อ (ดูด้านหลังของเอกสารรับรองโครงการวิจัย)

ผ่านการพิจารณาจาก ...กรรมการจริยธรรมการวิจัยในมนุษ ...วิทยาลัยเทคโนโลยีสุรนารี แล้ว

ผู้ร่วมโครงการวิจัย

-

นักวิจัยทุกท่านที่ผ่านการรับรองจริยธรรมการวิจัยต้องปฏิบัติดังต่อไปนี้

1. ดำเนินการวิจัยตามที่ระบุไว้ในโครงการวิจัยอย่างเคร่งครัด
2. ใช้เอกสารแนะนำอาสาสมัคร ใบยินยอม (และเอกสารเชิญเข้าร่วมวิจัยหรือใบโฆษณา (ถ้ามี)) แบบสัมภาษณ์ และ/หรือแบบสอบถาม เฉพาะที่มีตราประทับของคณะกรรมการจริยธรรมการวิจัยในมนุษย์เท่านั้น
3. รายงานเหตุการณ์ไม่พึงประสงค์ร้ายแรงที่เกิดขึ้นหรือการเปลี่ยนแปลงกิจกรรมวิจัยใด ๆ ต่อคณะกรรมการจริยธรรมการวิจัยในมนุษย์ ภายในระยะเวลาที่กำหนดในวิธีดำเนินการมาตรฐาน (SOPs)
4. ส่งรายงานความก้าวหน้าต่อคณะกรรมการจริยธรรมการวิจัยในมนุษย์ ตามเวลาที่กำหนดหรือเมื่อได้รับการร้องขอ
5. หากการวิจัยไม่สามารถดำเนินการเสร็จสิ้นภายในกำหนด ผู้วิจัยต้องยื่นขออนุมัติใหม่ก่อน อย่างน้อย 30 วัน
6. เอกสารทุกฉบับที่ได้รับการรับรองครั้งนี้หมดอายุตามอายุของโครงการวิจัยที่ได้รับการรับรองก่อนหน้านี้ (รหัสโครงการ EC-62-99)

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในม
มหาวิทยาลัยเทคโนโลยีสุรนารี แล่

COA No. 92/2562

Human Research Ethics Committee, Suranaree University of Technology

Certificate of Approval

Ethics Committee for Researches Involving Human Subjects, Suranaree University of Technology, Nakhon Ratchasima, Thailand, has Expedited the following study which is to be carried out in compliance with the International guidelines for human research protection as Declaration of Helsinki, The Belmont Report, CIOMS Guideline, International Conference on Harmonization in Good Clinical Practice (ICH-GCP)

Title of Project	: Implementation of Big Data Analytics in Logistics And Supply Chain
Project Code	: EC-62-99
Principal Investigator	: Prof. Dr. Vatanavongs Ratanavaraha
Department	: Institute of Engineering
Review Method	: Expedited
Continuing Report	: At least once annually or submit the final report if finished
Document Reviewed	: Protocol, Information Sheet, Informed Consent Questionnaire (version 1.0, 7 October 2019)

Signature..Chairman

(Asst. Prof. Porntip Nimkuntod, MD)

Human Researches Ethics Committee, Suranaree University of Technology

Date of Approval : 10 December 2019

Approval Expiry Date : 9 December 2020

Approval is granted subject to the following conditions : (see back of this Certificate)

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในมนุษย์
มหาวิทยาลัยเทคโนโลยีสุรนารี แล้ว

Human Research Ethic Committee of Suranaree University of Technology (HREC-SUT)

มหาวิทยาลัยเทคโนโลยีสุรนารี
Suranaree University of Technology

111 ถ.มหาวิทยาลัย ต.สุรนารี อ.เมือง จ.นครราชสีมา 30000 Tel. 0-4422-3000 Fax. 0-4422-407
111 University Avenue, Sub District Suranaree, Muang District, Nakhon Ratchasima 30000, Thailar

Co-investigator

-

All approved investigators must comply with the following conditions:

1. Strictly conduct the research as stated by the protocol.

2. Use only the information sheet, consent form (and recruitment material, if any), interview outlines and/or questionnaires bearing the ethics committee seal of approval.

3. Report to the Ethics Committee any serious adverse event or any changes in the research activity within the timeframe started in the standard operating procedures.

4. Provide progress reports to the Ethics Committee within the specified time period or upon request

5. If the study cannot be finished within the expired date of the approval certificated, the investigator is obliged to reapply for approval at least 30 days to the expiration date.

6. The expiry date of every approved document is based on the expiration date of the origin approved protocol (Protocol Number EC-62-99)

Human Research Ethic Committee of Suranaree University of Technology (HREC-SUT)

ผ่านการพิจารณาจาก
คณะกรรมการจริย... การวิจัยในมนุษ
มหาวิทยาลัย... สุรนารี แล้ว

AF/15-08/01.0	EC-62-0099

สถาบันวิจัยและพัฒ
รับที่ 2366/256
วันที่ 18 พ.ย.
เวลา 16:04

บันทึกข้อความ

มหาวิทยาลัยเทคโนโลยีสุรนารี

หน่วยงาน สาขาวิชาวิศวกรรมขนส่ง สำนักวิชาวิศวกรรมศาสตร์ **โทรศัพท์** 4238 **โทรสาร** 4220

ที่ ศธ 7414 (4)/438 **วันที่** 15 พ.ย. 2562

เรื่อง ขออนุมัติทำการวิจัยในมนุษย์และขอรับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์

เรียน ผู้อำนวยการสถาบันวิจัยและพัฒนา

ด้วยข้าพเจ้า **ศ. ดร.วัฒนวงศ์ รัตนวราห** อาจารย์ประจำสาขาวิชา**วิศวกรรมขนส่ง** สำนักวิชา**วิศวกรรมศาสตร์** มีความประสงค์จะทำโครงการวิจัย เรื่อง **การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน** (Implementation of Big Data Analytics in Logistics And Supply Chain) เพื่อ **เพื่อทำการวิจัย ไม่มีแหล่งทุนสนับสนุนการทำวิจัย** ประเภทโครงการวิจัย **เชิงปริมาณ** โดยได้แนบเอกสารประกอบการพิจารณามาพร้อมดังนี้

รายการเอกสาร
1. แบบเสนอโครงการวิจัยเพื่อขอการรับรอง (Protocol) docx
2. แบบเสนอโครงการวิจัยเพื่อขอการรับรอง (Protocol) pdf
3. ข้อเสนอโครงการวิจัยฉบับเต็ม (Full proposal)
4. ประวัติส่วนตัว/ผลงานของเจ้าของโครงการ
5. เอกสารผ่านการอบรมจริยธรรมการวิจัยของผู้วิจัยและผู้เข้าร่วมโครงการวิจัย
6. แบบประเมินโครงการวิจัยด้วยตัวเอง (Self-Assessment Form)
7. เอกสารชี้แจงผู้เข้าร่วมการวิจัย (Participant information sheet)
8. หนังสือแสดงเจตนายินยอมเข้าร่วมการวิจัย (Informed consent form)

(ลงชื่อ)..................หัวหน้าโครงการวิจัย
(ศ. ดร.วัฒนวงศ์ รัตนวราห)
วันที่ 14 เดือน พ.ย. 2562 พ.ศ. ....

(ลงชื่อ)..................หัวหน้าสาขา
(ศ. ดร.วัฒนวงศ์ รัตนวราห)
วันที่ 14 เดือน พ.ย. 2562 พ.ศ. ....

(ลงชื่อ)..................หัวหน้าสถานวิจัย
รองศาสตราจารย์ ดร.ชาญชัย ทองโสภา
หัวหน้าสถานวิจัย สำนักวิชาวิศวกรรมศาสตร์
วันที่ ...... เดือน 15 พ.ย. 2562 พ.ศ. ....

(ลงชื่อ)..................คณบดี
(รศ. ดร.กนต์ธร ชำนิประศาสน์)
วันที่ ...... เดือน 15 พ.ย. 2562 พ.ศ. ....

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในมนุษย์
มหาวิทยาลัยเทคโนโลยีสุรนารี แล้ว

แบบเสนอเพื่อขอรับการพิจารณาด้านจริยธรรมของการวิจัยในมนุษย์
สำหรับโครงการวิจัยด้านสังคมศาสตร์/มานุษยวิทยา (Social/Anthropological study)

1. **ชื่อโครงการวิจัย** การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน (Implementation of Big Data Analytics in Logistics And Supply Chain)

2. **หัวหน้าโครงการวิจัยและหน่วยงานที่สังกัด** วัฒนวงศ์ รัตนวราหะ (Vatanavongs Ratanavaraha)

   สังกัด (สาขา/สำนักวิชา) วิศวกรรมขนส่ง / สำนักวิชาวิศวกรรมศาสตร์

   ตำแหน่งวิชาการ ศาตราจารย์

   สถานที่ทำงาน/ติดต่อ สาขาวิศวกรรมขนส่ง/044-224-238

   หมายเลขโทรศัพท์ที่ติดต่อได้ทั้งในและนอกเวลาราชการ 08-9477-5805

   E-mail address vatanavongs@g.sut.ac.th

3. **ผู้ร่วมโครงการวิจัยและหน่วยงานที่สังกัด**

   -

4. **ความสำคัญของปัญหาที่ทำการวิจัย**

ในปัจจุบันนวัตกรรมด้านเทคโนโลยีทำให้การเก็บข้อมูลทำได้อย่างสมบูรณ์มากยิ่งขึ้น และกำลังถูกนำไปใช้อย่างแพร่หลาย ซึ่งข้อมูลชุดนั้นอาจถูกเรียกว่า "Big data" ยกตัวอย่างเช่น ในปัจจุบัน 3.196 พันล้านคนเป็นผู้ใช้ social median, 5.135 พันล้านคนมีโทรศัพท์มือถือ และ ใน google มีการค้นหา 40,000 ครั้ง ในทุกๆ วินาที (3.5 พ้านล้านครั้งต่อวัน)

สำหรับการนำไปประยุตกใช้ ยกตัวอย่างเช่น บริษัทวอล์มาร์ท (Walmart) ข้อมูลการขายย้อนหลังเพื่อคาดการณ์ความน่าจะเป็นของการซื้อของลูกค้าในอนาคต ซึ่งสามารถบอกได้ว่าควรจะขายสินค้าประเภทใดมากหรือน้อย โดยจำแนกตามช่วงเวลา บริษัทอเมซอน (Amazon) ใช้ข้อมูลการขายย้อนหลังเพื่อคาดการณ์ความต้องการของผู้บริโภคในอนาคต โดยเอาผลลัพท์นี้ไปคาดการณ์การขนส่งเพื่อให้มีต้นทุนที่ต่ำที่สุด

ดังนั้นการวิเคราะห์ข้อมูลบน Big data สามารถทำอะไรได้หลายอย่าง และสามารถทำได้ในหลายอุตสาหกรรม (Fig.1)

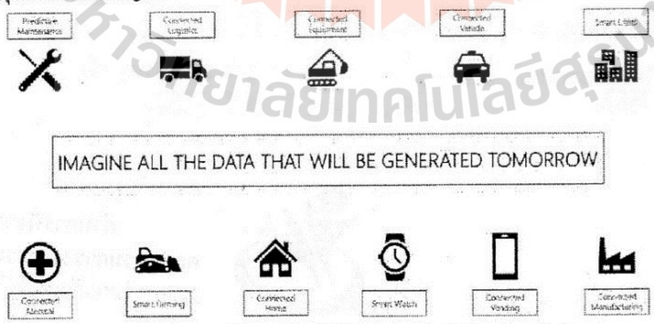


Fig.1 Data analysis in the future

การจัดการและการวิเคราะห์ข้อมูลเป็นทางเลือกที่ดีและเป็นประโยชน์อย่างยิ่งสำหรับผู้วางแผน นักวิจัย นักวิชาการ และผู้ประกอบการ อย่างไรก็ตามยังไม่เคยมีการวิจัยใดที่ประยุกต์ใช้เทคนิคนี้ ในงานการจัดการโลจิสติกส์และห่วงโซ่อุปทาน ซึ่งถือเป็นส่วนสำคัญในการดำเนินไปของเศรษฐกิจทั้งแบบจุลภาคและมหภาค ดังนั้นการวิจัยนี้จึงต้องการที่จะเติมเต็มการศึกษาในส่วนของการประยุกต์การวิเคราะห์ข้อมูลขนาดใหญ่ โดยเจาะจงไปที่อุตสาหกรรมโลจิสติกส์และห่วงโซ่อุปทาน (Fig.2)

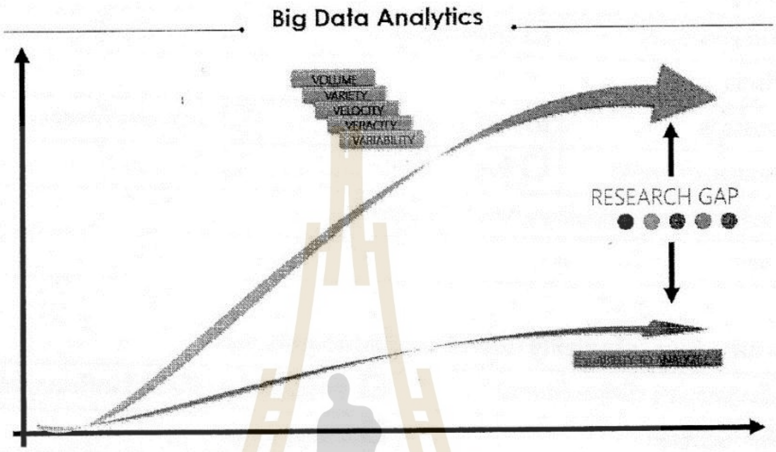


Fig.2 Research gap

5. วัตถุประสงค์ของโครงการ (ระบุให้ชัดเจน)

วัตถุประสงค์ทั่วไปของการศึกษานี้คือการอธิบายและประเมินของการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อใช้สำหรับงานการจัดการโลจิสติกส์และห่วงโซ่อุปทาน โดยมีรายละเอียดตามประเด็นย่อย ดังนี้

- พัฒนารูปแบบการวิเคราะห์ข้อมูลขนาดใหญ่ในปัจจุบัน เพื่อประยุต์ใช้ในงานการจัดการโลจิสติกส์และห่วงโซ่อุปทาน
- นำผลการวิเคราะห์ข้อมูลขนาดใหญ่ไปใช้ในอุสาหกรรมห่วงโซ่อุปทาน
- ประเมินความสามามารถของการวิเคราะห์ด้วยข้อมูลขนาดใหญ่ จำแนกตามกรณีศึกษา

6. ประโยชน์ของโครงการนี้ เมื่อเสร็จสมบูรณ์แล้วจะเป็นประโยชน์อย่างเป็นรูปธรรมอย่างไรบ้าง

การศึกษานี้มุ่งเน้นที่จะเติมเต็มการศึกษาในด้านโลจิสติกส์และห่วงโซ่อุปทาน ซึ่งเป็นการนำเอาข้อมูลขนาดใหญ่มาใช้ให้เกิดประโยชน์สูงสุด สำหรับประโยชน์ของงานนี้ ได้แก่

- วิเคราะห์ข้อมูลในด้านสถิติซึ่งอยู่ในรูปแบบของข้อมูลแบบตัวเลข และประยุตก์ใช้เทคนิคการวิเคราะห์ทั้งหมด 26 วิธี รวมไปถึงการศึกษาและวิเคราะห์บทความวิจัยที่ถูกตีพิมพ์ในกลุ่มของการศึกษาโลจิสิติกส์และห่วงโซ่อุปทาน ตามแนวโน้มเวลา (Time series)

- นำข้อมูลขนาดใหญ่มาวิเคราะห์เพื่อคาดการณ์ความน่าใจเป็นในอนาคต เช่น แนวโน้มของการซื้อขายสินค้าในอนาคตเป็นต้น และนำไปผลลัพท์ที่ได้ไปเสนอแนะในเชิงนโยบายสำหรับผู้ประกอบการ เพื่อเพิ่มผลกำไล และกำหนดแนวทางกลยุทธ์ของการดำเนินธุรกิจในอนาคต
- เปรียบเทียบวิธีวิเคราะห์ข้อมูลขนาดใหญ่ เพื่อประเมินความสามารถในการคาดการณ์ข้อมูล โดยจำแนกตามกรณีศึกษา

7. วิธีการศึกษา (Methodology) ที่ใช้ในการวิจัย (สามารถเลือกได้มากกว่าหนึ่งวิธี)

☐ เชิงคุณภาพ		
☐ Phenomenology	☐ Ethnography	☐ Grounded Theory
☐ เชิงปริมาณ		
☐ เชิงบรรยาย	☑ การศึกษาความสัมพันธ์	☐ การทดลอง/กึ่งทดลอง
☐ Systematic reviews	☐ อื่นๆ	
☐ Action Research/ Participatory Action Research		
☐ อื่นๆ ระบุ		

8. วิธีการรวบรวมข้อมูล/เครื่องมือที่ใช้ (โปรดแนบมาเพื่อประกอบพิจารณา)

☐ การใช้แบบสอบถามชนิดตอบด้วยตนเอง	☐ การสัมภาษณ์แบบมีโครงสร้างหรือแบบกึ่งโครงสร้าง
☐ การสัมภาษณ์เชิงลึก	☐ การสนทนากลุ่ม
☐ การสังเกต ระบุ...(เช่น แบบมีส่วนร่วม, แบบไม่มีส่วนร่วม)	☑ อื่นๆ ใช้ข้อมูลทุติยภูมิ

......การรวบรวมข้อมูลของการศึกษานี้ เป็นการเก็บรายละเอียดการซื้อขายและข้อมูลส่วนบุคคลของลูกค้าที่ซื้อสินค้าประเภท ทองคำแท่ง หรือโลหะมีค่า ของบริษัทแห่งหนึ่งในประเทศอินโดนีเซีย โดยได้ถูกเก็บรวบรวมผ่านระบบการซื้อขายของบริษัท และฐานข้อมูลรายละเอียดส่วนตัวของลูกค้า เช่น หมายเลขในกำกับภาษี วันที่ของการโอนเงิน วันเกิดของลูกค้า อายุ เพศ ที่อยู่ จำนวนบุตร การใช้ Social media เป็นต้น

9. ระบุวิธีการวัดผล/การวิเคราะห์ผลการวิจัย/สถิติที่ใช้ (Outcome measurement / Data Analysis)

วิธีในการวิเคราะห์ผล จะใช้แบบจำลองคณิตศาสตร์เพื่อทำนายความสัมพันธ์ระหว่าง ปัจจัยลักษณะของลูกค้าที่ส่งผลต่อการเลือกซื้อสินค้าประเภทโลหะมีค่า ซึ่งวิธีดังกล่าวประกอบด้วย

1. วิธีทำเหมืองข้อมูล (Data mining)
2. วิธี (Machine leaning)
3. วิธีแบบจำลองถดถอย (Regression model)
4. วิธีวิเคราะห์เชิงพื้นที่ (Spatial analysis)
5. วิธี Data Visualization

10. เหตุผลและความจำเป็นที่ต้องดำเนินการวิจัยในอาสาสมัครกลุ่มนี้ (ระบุความเป็นมา/ปัญหาวิจัย (อย่างย่อพร้อมระบุเอกสารอ้างอิง)

อาสาสมัครกลุ่มนี้ เป็นบุคคลที่ซื้อสินค้าโลหะมีค่า เช่น ทองคำ ของบริษัทแห่งหนึ่งในประเทศอินโนนีเซีย ซึ่งเป็นบริษัทที่พร้อมจะให้ข้อมูลกับผู้วิจัยเพื่อนำมาใช้ศึกษาแนวโน้มของการ ซื้อ-ขาย สินค้า ที่มีความสัมพันธ์กับปัจจัยต่างๆ ของตัวบุคคล โดยการประยุกต์นำเอาการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อนำมาสรุปเป็นแนวทางการปรับปรุงด้านกลยุทธ์การขายของบริษัท และยังสามารถใช้เป็นแนวทางในการศึกษาด้าน big data analysis ต่อไปได้ในอนาคต **Fig.3** แสดงองค์ประกอบของการศึกษา

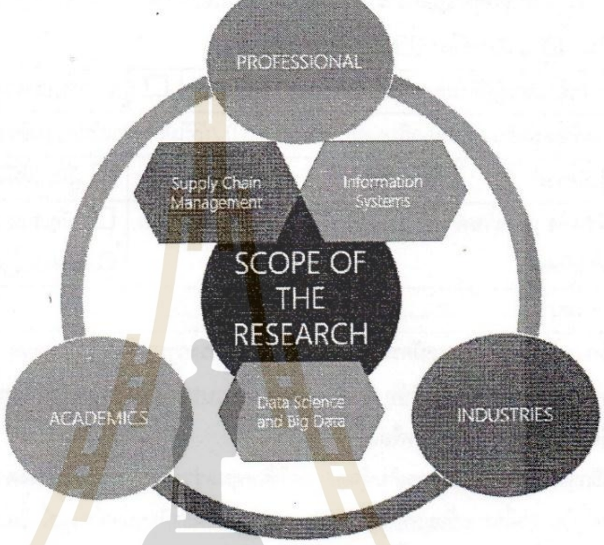


Fig.3 องค์ประกอบของการศึกษา

11. กลุ่มประชากรอาสาสมัคร

11.1 จำนวนกี่คน ระบุเหตุผลการได้มาซึ่งขนาดตัวอย่างที่เหมาะสม การคำนวณขนาดตัวอย่าง

เป็นข้อมูลที่ซื้อจากบริษัท UBPP Logam Mulia PT. Antam TBK. Public-listed company ซึ่งเป็นลูกค้าที่สามารถซื้อขาย-สินค้าของบริษัทในช่วงเวลา ตั้งแต่ ปี.พ.ศ. 2551 – พ.ศ. 2562 สำหรับจำนวนตัวอย่าง เมื่อพิจารณาหลักการของ big data analysis จะพบว่าขึ้นอยู่กับจำนวนข้อมูลที่สามารถหาได้ ซึ่งหากข้อมูลมีจำนวนมาก ยิ่งจะทำให้ผลการวิเคราะห์มีความสมจริงมากยิ่งขึ้น จากงานวิจัยส่วนใหญ่จะใช้จำนวนตัวอย่างมากกว่า 20,000 ข้อมูล

11.2 ระบุคุณสมบัติของอาสาสมัคร มีวิธีการคัดเลือกผู้เข้าร่วมโครงการอย่างไร

เป็นข้อมูลลูกค้าที่ซื้อสินค้าโลหะมีค่าของบริษัทแห่งหนึ่งในห้วงเวลา 10 ปี ย้อนหลัง

11.3 ระบุวิธีการแบ่งกลุ่มอาสาสมัครเป็นกลุ่มทดลองและกลุ่มควบคุม ไม่มี

11.4 การคัดเลือกผู้เข้าร่วมการวิจัย (Subject selection and allocation)

ก. เกณฑ์การคัดเลือกผู้เข้าร่วมการวิจัย (Inclusion criteria)

เป็นข้อมูลลูกค้าที่ซื้อสินค้าโลหะมีค่าของบริษัทแห่งหนึ่งในห้วงเวลา 10 ปี ย้อนหลัง

ข. เกณฑ์การคัดออกผู้เข้าร่วมการวิจัย (Exclusion criteria)

ข้อมูลของลูกค้าที่ไม่สมบูรณ์ ไม่ครบตามที่กำหนด

ค. เกณฑ์การถอนผู้เข้าร่วมการวิจัยหรือยุติการเข้าร่วมการวิจัย (Withdrawal or termination criteria)

ข้อมูลของลูกค้าที่ไม่สมบูรณ์ ไม่ครบตามที่กำหนด

ง. การจัดผู้เข้าร่วมการวิจัยเข้ากลุ่ม (Subject allocation) ไม่มี

11.5 มีการใช้อาสาสมัครกลุ่มเปราะบาง (ซึ่งเป็นกลุ่มที่ไม่สามารถตัดสินใจเองได้ในภาวะสำคัญ) เหล่านี้หรือไม่ ☑ ไม่เกี่ยวข้อง ☐ เกี่ยวข้อง* ระบุ

☐ ทารก/เด็กเล็ก/ผู้ที่ยังไม่บรรลุนิติภาวะ (อายุ <18 ปี)**			☐ ผู้พิการหรือมีความบกพร่องทางสมอง/จิตใจ
☐ ผู้ป่วยห้องฉุกเฉิน หรือหออภิบาลผู้ป่วยหนัก (ICU) /ผู้ที่ไม่สามารถให้ความยินยอมด้วยตนเอง			
☐ หญิงมีครรภ์	☐ ผู้พิการ	☐ ผู้สูงอายุ	☐ ผู้ป่วยที่มีโรคเรื้อรัง
☐ ผู้ต้องขัง แรงงานต่างด้าว ในบางกรณีอาจรวมทั้งผู้ด้อยโอกาสทางสังคม			☐ นักเรียน / นักศึกษา/ทหารเกณฑ์/ผู้ใต้บังคับบัญชา
☐ อื่นๆ ระบุ			

11.6 วิธีการเข้าถึงกลุ่มอาสาสมัครที่ต้องการให้เข้าร่วมโครงการ สถานที่เก็บข้อมูล

เป็นการซื้อขายข้อมูลจากบริษัท ซึ่งได้มีการถูกเก็บรวบรวมไว้ในฐานของมูลของบริษัท

11.7 มีค่าตอบแทนหรือรางวัลหรือไม่ : ไม่มี

12. อธิบายวิธีการศึกษาทดลอง การเก็บข้อมูลและให้เหตุผลว่าทำไมการศึกษานี้จึงความเสี่ยงต่ำ

การศึกษานี้เป็นการซื้อขายข้อมูลทางธุรกิจ ซึ่งเป็นข้อมูลทางธรุกิจขนาดใหญ่เพื่อวิเคราะห์ข้อมูลในภาพรวม ดังนั้นถึงแม้ว่าจะมีข้อมูลส่วนบุคคลของลูกค้าแต่ละคน แต่ไม่มีการเปิดเผยข้อมูลรายบุคคลแต่อย่างใด โดยการเก็บข้อมูลเป็นการเก็บโดยพนักงานของบริษัทและถูกรวบรวมไว้ในฐานของมูลของบริษัท ซึ่งลูกค้าได้ให้ข้อมูลด้วยความเต็มใจ และไม่มากไปกว่าความเสี่ยงในชีวิตประจำวัน

13. วิธีการเชิญชวนเข้าร่วมโครงการด้วยการลงลายมือชื่อยินยอมหรือด้วยวาจา

ไม่มีการเชิญชวนอาสาสมัคร แต่เป็นการซื้อข้อมูลจากบริษัท

14. อธิบายกระบวนการขอความยินยอมอาสาสมัคร

ไม่มีการเชิญชวน ไม่มีการลงนามยินยอม เป็นการวิเคราะห์ข้อมูลทางธุรกิจในภาพรวม

15. ประโยชน์ต่ออาสาสมัครและชุมชนที่เข้าร่วมการวิจัย รวมทั้งการสร้างความเข้มแข็งแก่ชุมชนอย่างไร

ไม่มีประโยชน์ต่ออาสาสมัครที่เข้าร่วมการวิจัย เป็นการวิเคราะห์ภาพรวมเพื่อทดสอบโมเดล

16. ผลกระทบที่อาจเกิดแก่อาสาสมัครและ/หรือชุมชนที่เข้าร่วมการวิจัย เช่น

ก. ระบุความเสี่ยงที่อาจเกิดขึ้นกับอาสาสมัคร ต่อร่างกาย จิตใจ สังคม เศรษฐกิจ และผู้วิจัยเตรียมการป้องกันไม่ให้เกิด ผลเสีย หรือเตรียมการแก้ไขไว้อย่างไร

ความเสี่ยงที่อาจเกิดขึ้นกับอาสาสมัคร คือ การที่ข้อมูลการซื้อขายจะถูกเปิดเผย เช่น วันที่การซื้อขาย จำนวนเงิน ประเภทของสินค้า เป็นต้น จนอาจทำให้เกิดความเสี่ยงต่อสภาพจิตใจ สังคม ขึ้นได้ โดยการ

เตรียมการป้องกันมิให้เกิดผลเสีย คือการปกป้องข้อมูลชุดนี้มิให้ถูกเผยแพร่ต่อสาธารณะ และรายงานการวิจัยเป็นภาพรวม

ข. ในกรณีที่มีผลกระทบต่อชุมชน ผู้วิจัยมีวิธีการเข้าถึง หารือต่อชุมชนอย่างไร

ติดต่อบริษัทเพื่อซื้อขายข้อมูล

17. วิธีปฏิบัติที่ใช้ในการวิจัยเพื่อปกป้องความลับของอาสาสมัครหรือชุมชนทำอย่างไร

ใช้รหัสแทนข้อมูลส่วนบุคคลของอาสาสมัคร

หากมีการบันทึกข้อมูลดังกล่าวข้างต้น เฉพาะหัวหน้าโครงการวิจัยที่เข้าถึงข้อมูลได้ หากเสร็จสิ้นโครงการจะทำลายข้อมูล**ทิ้งภายใน 1 ปี**

18. รายละเอียดงบประมาณ/แหล่งทุน

ไม่มีแหล่งทุน

19. ระยะเวลาการดำเนินการ โครงการวิจัยนี้

ก. ระยะเวลาดำเนินการทั้งโครงการ 2 ปี (วิเคราะห์ผล ธันวาคม พ.ศ. 2562 – ธันวาคม 2564)

ข. ข้อมูลที่นำมาวิเคราะห์ เดือน มกราคม พ.ศ 2551 เสร็จสิ้นเดือน ธันวาคม พ.ศ 2562

20. ประสบการณ์ด้านจริยธรรมการวิจัย (โปรดแนบหลักฐานการอบรมทุกคน หากเป็นนักศึกษาให้แนบหลักฐานของอาจารย์ที่ปรึกษาด้วย)

1) **ชื่อผู้วิจัย ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห**

หลักสูตร/ชื่อหัวข้อการอบรม หลักสูตรอบรมออนไลน์ด้านจริยธรรมการวิจัยในมนุษย์ จัดโดย สำนักงานคณะกรรมการวิจัยแห่งชาติ (วช.) ร่วมกับชมรมจริยธรรมในคนในประเทศไทย

22. ข้อสัญญา

1. ข้าพเจ้าและคณะผู้วิจัยดังมีรายนามและได้ลงชื่อไว้ในเอกสารนี้ จะดำเนินการวิจัยตามที่ระบุไว้ในโครงการวิจัยฉบับที่ได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีสุรนารี และได้ขอความยินยอมจากผู้เข้าร่วมการวิจัยอย่างถูกต้องตามหลักจริยธรรมการวิจัยในมนุษย์ ดังที่ได้ระบุไว้ในแบบเสนอโครงการวิจัย โดยจะให้ความเคารพในศักดิ์ศรี สิทธิ และคำนึงถึงความปลอดภัยและสวัสดิภาพของผู้เข้าร่วมการวิจัยเป็นสำคัญ
2. หากมีความจำเป็นต้องปรับแก้ไขโครงการวิจัย ข้าพเจ้าจะแจ้งคณะกรรมการจริยธรรมฯ เพื่อขอการรับรองก่อนเริ่มดำเนินการตามที่ต้องการปรับเปลี่ยนทุกครั้ง และหากการปรับโครงการวิจัยมีผลกระทบต่อผู้เข้าร่วมการวิจัย ข้าพเจ้าจะแจ้งการปรับเปลี่ยนและขอความยินยอมจากผู้ที่ได้เข้าร่วมการวิจัยแล้วทุกครั้ง
3. ข้าพเจ้าจะรายงานเหตุการณ์ไม่พึงประสงค์/เหตุการณ์ที่ไม่สามารถคาดการณ์ได้ล่วงหน้าในระหว่างการวิจัย ตามระเบียบของคณะกรรมการจริยธรรมฯ ภายในเวลาที่กำหนด และจะดำเนินการแก้ไขเหตุการณ์ไม่พึงประสงค์ที่เกิดขึ้นระหว่างการวิจัยอย่างเต็มความสามารถ
4. ข้าพเจ้าและคณะผู้วิจัยมีความรู้ความเข้าใจในกระบวนการวิจัยที่เสนอมาอย่างดีทุกขั้นตอน และมีความสามารถในการแก้ไขปัญหา หรือเหตุการณ์ไม่พึงประสงค์ที่อาจจะเกิดขึ้นในระหว่างการวิจัย เพื่อความปลอดภัยและสวัสดิภาพของผู้เข้าร่วมการวิจัยได้เป็นอย่างดี

5. เมื่อทำการวิจัยเสร็จสิ้น ข้าพเจ้าจะสรุปการดำเนินงานและแจ้งปิดโครงการวิจัย และหากการวิจัยใช้เวลาเกินกว่า 1 ปี ข้าพเจ้าจะรายงานความคืบหน้าของโครงการพร้อมทั้งขอต่ออายุการรับรองก่อนครบกำหนดอายุของเอกสารรับรองที่ได้รับ

ข้าพเจ้าขอรับรองว่าข้อความข้างต้นเป็นความจริง และเข้าใจความหมายโดยชัดเจนทุกประการ

ลงชื่อ.. ลงชื่อ..

(..) (ศ.ดร. [illegible])

อาจารย์ที่ปรึกษาโครงการ หัวหน้าโครงการวิจัย

กรณีหัวหน้าโครงการวิจัยเป็นนักศึกษา

ลงชื่อ ....

(..)

ผู้ร่วมโครงการวิจัย

โครงการวิจัยนี้ได้ผ่านความเห็นชอบจากหน่วยงานต้นสังกัดแล้ว

ลงชื่อ..

(ศ.ดร. [illegible])

หัวหน้าสาขา

ลงชื่อ..

(..)

หัวหน้าสถานวิจัย

ลงชื่อ..

(รองศาสตราจารย์ เรืออากาศเอก ดร.กนต์ธร ชำนิประศาสน์)

คณบดีสำนักวิชาวิศวกรรมศาสตร์

คณบดี

แบบประวัติส่วนตัว

อาจารย์ ศ.ดร.วัฒนวงศ์ รัตนวราห

Prof.Vatanavongs Ratanavaraha, (Ph.D.)

การศึกษา/คุณวุฒิ

ปริญญาเอก : Ph. D. (Transportation Engineering), Vanderbilt University (พ.ศ. 2542)

ปริญญาโท : M. Eng. (Transportation Engineering), สถาบันเทคโนโลยีแห่งเอเชีย (พ.ศ. 2538)

ปริญญาตรี : วศ. บ. (วิศวกรรมโยธา), มหาวิทยาลัยเทคโนโลยีสุรนารี (พ.ศ. 2555)

: วศ. บ. (โยธา), จุฬาลงกรณ์มหาวิทยาลัย (พ.ศ. 2535)

ตำแหน่งปัจจุบัน : อาจารย์ประจำสาขาวิชาวิศวกรรมขนส่ง

ผลงานทางวิชาการ /ผลงานวิจัย

<u>1) วารสารวิชาการระดับนานาชาติ (International Journal)</u>

Champahom T., Jomnonkwao S., Karoonsoontawong A., Hantanong N., Beeharry R., and **Ratanavaraha V.** (2019). "Modelling User Perception of Bus Service Quaility: Case Study of Mauritius". Songklanakarin Journal of Science and Technology, (In Press) (Scopus)

Siridhara S., Paonoi W., Meeyai S., and **Ratanavaraha V.** (2019). "Container transport mode choice analysis with binary logit model case study: Northeastern Thailand". Engineering and Applied Science Research. (Accepted) (Scopus)

Paonoi W., Siridhara S., Meeyai S., Beeharry R., and **Ratanavaraha V.** (2019). "Mode Choice Models for Agricultural Products by Freight Rail Transport in the Northeast of Thailand". Suranaree Journal of Science and Technology. (Accepted) (Scopus)

Sangphong O., Siridhara S. and **Ratanavaraha V.** (2019). "Analysis of appropriate overtaking position under equal block lengths". Lowland Technology International, (Accepted) (Scopus)

Watthanaklang D. and **Ratanavaraha V.** (2019). "The development of guidelines on promoting bicycle use tourism in Khao yai national park". Lowland Technology International, (Accepted) (Scopus)

Champahom T., Goodary R., Beeharry R., Jomnonkwao S. and **Ratanavaraha V.** (2019). "Evaluation user's satisfaction of bus service in Mauritius: Decision tree approach". Lowland Technology International, (Accepted) (Scopus)

Amprayn C., Wissawapisal K., **Ratanavaraha V.** and Chatpattananan V. (2019). "The relationship between area of arterial roads and related factors in city plan". Lowland Technology International, (Accepted) (Scopus)

Watthanaklang D., **Ratanavaraha V.**, Jomnonkwao S., Boonyoo T. and Nambulee W. (2019). "The study of willingness to pay for bicycle hire service at tourist attractions in Thailand". Lowland Technology International, (Accepted) (Scopus)

Ngamsom D., Jomnonkwao S., Satthamnuwong N., Boonyoo T. and **Ratanavaraha V.** (2019). "Reviewing the effectiveness of public transport accessibility for person with reduced mobility: A case study of airport terminal". Lowland Technology International, (Accepted) (Scopus)

Siridhara, S., Paonoi, W., Meeyai, S. and **Ratanavaraha, V.**, (2019) "Container transport mode choice analysis with binary logit model case study: Northeastern Thailand", Engineering and Applied Science Research, Accepted.

Mon, E. E., Jomnonkwao, S., Khampirat, B., Satiennam, T., & **Ratanavaraha, V.** (2019). Estimating the willingness to pay and the value of fatality risk reduction for car drivers in Myanmar. *Case Studies on Transport Policy.*

Johnstone, L., & **Ratanavaraha, V.** (2019). Green cargo movement, locality: mekong region across localities. *International journal of geomate, 16*(56), 110-117.

Watanya, N., Sajjakaj, J., Siradol, S., **Vatanavongs, R.**, Ampol, K., & Beeharry, R. (2018). The intercity bus passenger' s locus of control with regard to seat belt use intention. *Suranaree Journal of Science & Technology, 25*(3), 235-246.

Nambulee, W., Jomnonkwao, S., Siridhara, S., & **Ratanavaraha, V.** (2018). Modelling of seat belt use intention for intercity buses based on health belief model. *Transportmetrica A: Transport Science*, 1-19.

Champahom, T., Jomnonkwao, S., Satiennam, T., Suesat, N., & **Ratanavaraha, V.** (2019). Modeling of safety helmet use intention among students in urban and rural Thailand based on the theory of planned behavior and Locus of Control. *The Social Science Journal.*

Mon, E. E., Jomnonkwao, S., Khampirat, B., Satiennam, W., & **Ratanavaraha, V.** (2018). Willingness to pay for mortality risk reduction for traffic accidents in Myanmar. *Accident Analysis & Prevention, 118*, 18-28.

Mon, E. E., Jomnonkwao, S., Khampirat, B., & **Ratanavaraha, V.** (2018). Myanmar Motorbike Riders' Willingness to Pay for Fatality Risk Reduction. *Suranaree Journal of Science and Technology, 25*(2), 131-142.

Johnstone, L., & **Ratanavaraha, V.** (2017). Green Freight Movement: The Dilemma of the Shifting of Road Freight to Alternatives. *Transportation Research Procedia, 21*, 154-168.

Watthanaklang, D., **Ratanavaraha, V.**, Chatpattananan, V., & Jomnonkwao, S. (2016). Measuring the motivation to ride bicycles for tourism through a comparison of tourist attractions. *Transport Policy, 52*, 153-163.

Ratanavaraha, V., Jomnonkwao, S., Khampirat, B., Watthanaklang, D., & Iamtrakul, P. (2016). The complex relationship between school policy, service quality, satisfaction, and loyalty for educational tour bus services: A multilevel modeling approach. *Transport Policy, 45*, 116-126.

Jomnonkwao, S., & **Ratanavaraha, V.** (2016). Measurement modelling of the perceived service quality of a sightseeing bus service: An application of hierarchical confirmatory factor analysis. *Transport Policy, 45*, 240-252.

Jomnonkwao, S., Sangphong, O., Khampirat, B., Siridhara, S., & **Ratanavaraha, V**. (2016). Public transport promotion policy on campus: evidence from Suranaree University in Thailand. *Public Transport, 8*(2), 185-203.

Jomnonkwao, S., **Ratanavaraha, V.**, Khampirat, B., Meeyai, S., & Watthanaklang, D. (2015). Factors influencing customer loyalty to educational tour buses and measurement invariance across urban and rural zones. *Transportmetrica A: Transport Science*, 1-27.

Kassawat, S., Sarapirome., and **Ratanavaraha V.** (2015). Integration of Spatial Models for Web-based Risk Assessment of Road Accident. *Walailak Journal of Science and Technology (WJST), 12*(8), 671 - 679.

Ratanavaraha, V., & Jomnonkwao, S. (2015). Trends in Thailand CO2 emissions in the transportation sector and Policy Mitigation. *Transport Policy, 41*, 136-146.

Ratanavaraha, V., & Jomnonkwao, S. (2015). The efficiency testing of Shoulder Rumble Strips (SRS) on noise for alerting drivers in Thailand: a comparison among three types of SRS. *Advances in Transportation Studies: an international Journal, 36*, 109-118.

Jomnonkwao, S., Siridhara, S., & **Ratanavaraha, V.** (2015). Awareness status of sightseeing bus entrepreneurs: A case study in rural areas of Thailand. *Lowland Technology International, 17*(1), 47-52.

Ratanavaraha, V., & Jomnonkwao, S. (2014). Model of users' expectations of drivers of sightseeing buses: confirmatory factor analysis. *Transport Policy, 36*, 253-262.

Ratanavaraha V., and Suangka S. (2014). Impacts of Accident Severity Factors and Loss Values of Crashes on Expressways in Thailand. *IATSS Research, 37*(2), 130-136.

Sudadet S., **Ratanavaraha V.**, and Siridhara S. (2014), Travel Choice Model for Bus and Other Modes in Ubon Ratchathani Municipality and Amphoe Warin Chamrap by Using Stated Preferences Data. *Journal of Society for Transportation and Traffic Studies, 5*(2), 1 - 11.

Ratanavaraha V., and Watthanaklang D. (2013). The effectiveness of temporary traffic calming devices on reducing speeds of traffic flow in school zones. *Indian Journal of Science and Technology*, 6(5), 4478 - 4484.

Ratanavaraha V., and Watthanaklang D. (2013). Road Safety Audit: Identification of Bus Hazardous Location in Thailand. *Indian Journal of Science and Technology, 6*(8), 5126 - 5133.

Ratanavaraha, V., & Jomnonkwao, S. (2013). Community participation and behavioral changes of helmet use in Thailand. *Transport Policy, 25*, 111-118.

Luadsakul C., and **Ratanavaraha V.** (2013). The Study of Walkability Index: A Case Study in Nakhon Ratchasima Province. *International Journal of Emerging Technology and Advanced Engineering, 3*(3), 471 - 476.

Kasemsuk C., and **Ratanavaraha V.** (2012). Participatory Communication and Information Management for Community Development: Thai Rural Community Perspectives. *China-USA Business Review. 11*(11), 1432-1440.

Amprayn C., and **Ratanavaraha V.** (2011), The Study on Trip Generation Model of Residential Building in Bangkok Area. *Journal of Society for Transportation and Traffic Studies, 2*(4), 10 - 15.

Chatpattananan V., and **Ratanavaraha V. (2011),** An M/M/m Analysis of Passport Checking Points in Suvanabhum Airport. *Proceeding of the 3rd International Conference on Logistics and Transport*, Maldives.

Chatpattananan V., and **Ratanavaraha V. (2011),** A Survey on Mass Rapid Transit Types Suitable for Bangkok Metropolitan. *Proceeding of the 3rd International Conference on Logistics and Transport*, Maldives.

Limanond, T., Jomnonkwao, S., Watthanaklang, D., **Ratanavaraha, V.**, & Siridhara, S. (2011). How vehicle ownership affect time utilization on study, leisure, social activities, and academic performance of university students? A case study of engineering freshmen in a rural university in Thailand. *Transport Policy, 18*(5), 719-726.

Witchayapong P., **Ratanavaraha V.,** and Amprayn C. (2011), The Application of Accident Clock for the Identification of Hazardous Locations : A Case Study of Nakhon Ratchasima Province. *Journal of Society for Transportation and Traffic Studies*, *2*(3), 46 - 55.

Ratanavaraha V., Siridhara S., and Tippayawong K. (2011). A Study of Road Safety Management in School Zone in Thailand. *Proceeding of the Eastern Asia Society for Transportation Studies, 8*, 2011, Jeju, South Korea.

Ratanavaraha V., and Kasemsuk C. (2010), The Study of Helmet Use Behavior of Motorcyclists in Nakhon Pathom Province by Using Community Participation. *Journal of Architectural/Planning Research and Studies, 7* (1), 73-86.

Prabjabok P., Lertworawanich P., and **Ratanavaraha V.** (2010). The Increment of Noise from the Truck while Running through Profile Marking Shoulder Rumble Strips. *Proceeding of the 15th International Conference of Hong Kong Society for Transportation Studies (HKSTS),* Hongkong.

Ratanavaraha V., Witchayaphong P., Lakhawattana N., and Chaiwong K. (2010). Comparing the Performance of Wearing Helmet Behavior Model While Driving Motorcycle by Binary Logistic Regression Analysis Method and Learning Vector Quantization of Artificial Neural Network. *Proceeding of the International Conference on Software,* Knowledge, Information Management and Application, Paro, Bhutan.

Ratanavaraha V., Lertworawanich P., Prabjabok P., Taesiri Y. (2010). The Study on the Effectiveness of Profile Marking Shoulder Rumble Strip to Prevent the Truck Driver Falls Asleep. *Proceeding of the 7th Asia Pacific Conference on Transportation and Environment*, Semarang, Indonesia.

Kasemsuk C., **Ratanavaraha V.** (2010). Community Participation: A Methodology for Encouraging Helmet Use of Motorcyclist. *Proceeding of the International Conference on Sustainable Community Development*, Putrajaya, Malaysia.

Ratanavaraha V. (2009). A Syudy of Safety at U-Turn in Bangkok and Surrounding Area. *Proceeding of the 6th Asia Pacific Conference on Transportation and Environment*, Shanghai, China.

Ratanavaraha V. (2009). Travel Delay Cost due to Road Accident for Songkhla Province of Thailand. *Proceeding of the Eastern Asia Society for Transportation Studies*, *7*, 2009, Surabaya, Indonesia.

Ratanavaraha V., Taneerananon P. and Thongchim P. (2007). Travel Delay Cost Case Study : Khon Kaen Province, Thailand. *Proceeding of the 5th Asia Pacific Conference on Transportation and Environment,* Singapore.

Ratanavaraha V. and Ampray C. (2005). A Predictive Trip Generation Model for Hypermarkets in Bangkok. *Proceeding of the 4th Asia Pacific Conference on Transportation and Environment,* Xi'An, China.

Ratanavaraha V. and Ampray C. (2003). Causative Highway Accident Factors of the Expressway System in Thailand. *the Journal of Eastern Asia Society for Transportation Studies.*

Ratanavaraha V. and Pawan P. (2002). Prediction on Accident Factors of the Expressway System in Thailand, Proceeding of the 9th World Congress on Intelligent Transportation Systems, Chicago, Illinois.

Ratanavaraha V. (2001). Identification of Hazardous Location. *Proceeding of the 4th SEA Regional Conference on Higher Engineering Education Network*, Kuala Lumpur, Malaysia.

Ratanavaraha V. (2001). How Causative Highway Accident Factors Vary as Different Identification of Hazardous Location Methods. *Proceeding of the 3rd Regional Symposium on Infrastructure Development in Civil Engineering RSID-3,* Tokyo, Japan.

Chen J.H., Yamploy K. and **Ratanavaraha V.** (1996). Modal Choice Model between car and bus a case study of Thailand, *the Journal of Eastern Asia Society for Transportation Studies*, *1*, 681.

2) วารสารวิชาการระดับนาชาติ (National Journal)

ดิสกุล ชลศาลาสินธุ์ ศิรดล ศิริธร **วัฒนวงศ์ รัตนวราห** สัจจากาจ จอมโนนเขวา ดวงดาว วัฒนากลาง. (2560). การศึกษาปัจจัยที่มีผลต่อความตั้งใจในการสวมหมวกนิรภัย โดยทฤษฎีแบบแผนความเชื่อด้านสุขภาพ. *วารสารวิชาการครุศาสตร์อุตสาหกรรม (Technical Education Journal King Mongkut's University of Technology North Bangkok), 8*(1), 26-35.

ภาณุพงษ์ ประจงค้า **วัฒนวงศ์ รัตนวราห** สัจจากาจ จอมโนนเขวา และดวงดาว วัฒนกลาง. (2560). แบบจำลองการวัดคุณภาพบริการของรถทัศนาจรสำหรับนักท่องเที่ยวชาวต่างชาติ. *วารสารวิชาการครุศาสตร์อุตสาหกรรม (Technical Education Journal King Mongkut's University of Technology North Bangkok), 8*(1), 36-45.

ณัฐิญา วงละคร, **วัฒนวงศ์ รัตนวราห**, กาญจน์กรอง สุอังคะ, และสัจจากาจ จอมโนนเขวา (2559). ปัจจัยด้านแรงจูงใจในการเดินทางท่องเที่ยวของผู้สูงอายุในประเทศไทย. *วารสารวิชาการครุศาสตร์อุตสาหกรรม พระจอมเกล้าพระนครเหนือ (Technical Education Journal King Mongkut's University of Technology North Bangkok)*, 7(2), 220-230.

Chermkhunthod C., **Ratanavaraha V.,** and Limanond T. (2012). The Study of Accessibility Level for Public Transportation by Using the Application of GIS: A Case Study in Nakorn Ratchasima Province. *Proceeding of the 8th Conference of the National Conventional on Transportation Engineering*, Chonburi, Thailand.

Prabjabok P., Pueboobpaphan R., and **Ratanavaraha V.** (2012). The Study of the Effectiveness of Shoulder Rumble Strip on Motorway, *Proceeding of the 8th Conference of the National Conventional on Transportation Engineering*, Chonburi, Thailand.

Uttra S., and **Ratanavaraha V.** (2012). The Application of Temporary Traffic Calming Devices at School Zone: A Case Study in Nakorn Ratchasima Province. *Proceeding of the 8th Conference of the National Conventional on Transportation Engineering*, Chonburi, Thailand.

3) การประชุมวิชาการระดับนานาชาติ (International Conference)

Jomnonkwao, S., & **Ratanavaraha, V.** (2013). What Factors Are Related to the Analysis of Customer Loyalty toward Non-Fixed Route Bus Services? *Paper presented at the International Conference on Tourism and Hospitality Management 2013*, Colombo, 8th-10th December, Sri Lanka.

4) การประชุมวิชาการระดับชาติ (National Conference)

นราธิป อยู่สำแดงกิจ, **วัฒนวงศ์ รัตนวราห,** ไอศูรย์ เรื่องรัตนอัมพร, และ สัจจากาจ จอมโนนเขวา, (2560). การพัฒนาแบบจำลองเพื่อพยากรณ์ความพึงพอใจจากการให้บริการรถทัศนศึกษา. *การประชุมวิชาการและนำเสนอผลงานวิจัยระดับชาติ ราชธานีวิชาการ ครั้งที่ 2 "การวิจัย 4.0 เพื่อการพัฒนาประเทศสู่ความมั่นคง มั่งคั่ง และยั่งยืน"*, 26 - 27 กรกฎาคม 2560, อุบลราชธานี.

ชุติมา เจิมขุนทด, สวลี อุตรา, **วัฒนวงศ์ รัตนวราห** และ ถิรยุทธ ลิมานนท์ (2553). การปรับปรุงความปลอดภัยบนท้องถนนในพื้นที่มหาวิทยาลัยเทคโนโลยีสุรนารี. *งานสัมมนาวิชาการด้านความปลอดภัยทางถนนระดับชาติ ภาคกลาง – ภาคตะวันออก.* 2 – 3 สิงหาคม 2553. กรุงเทพมหานคร.

วัฒนวงศ์ รัตนวราห. (2551). การศึกษาความปลอดภัยบริเวณจุดกลับรถในกรุงเทพมหานครและจังหวัดใกล้เคียง. *การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 13*, ชลบุรี.

วัฒนวงศ์ รัตนวราห. (2551). การศึกษามูลค่าความล่าช้าในการเดินทางอันเนื่องมาจากการเกิดอุบัติเหตุทางถนนของจังหวัด ลำปาง. *การประชุม 10 ปีวิชาการแม่ฟ้าหลวง.* 26-28 พ.ย. 51. มหาวิทยาลัยแม่ฟ้าหลวง, เชียงราย.

วัฒนวงศ์ รัตนวราห และ ชิษณุ อัมพรายน์. (2550). การพิสูจน์จุดอันตรายต่อการเกิดอุบัติเหตุบริเวณถนนรามอินทราและถนนแจ้งวัฒนะ. *การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 12*, พิษณุโลก.

วัฒนวงศ์ รัตนวราห และ ชิษณุ อัมพรายน์. (2548). การศึกษาการเกิดและการดึงดูดการเดินทางของร้านค้าปลีกขนาดใหญ่ในเขตกรุงเทพมหานคร. *การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 10*, ชลบุรี.

วัฒนวงศ์ รัตนวราห. (2545). สาเหตุการเกิดอุบัติเหตุบนทางด่วนเฉลิมมหานคร. *การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 10*, ขอนแก่น.

วัฒนวงศ์ รัตนวราห. (2544). การหาต้นเหตุของการเกิดอุบัติเหตุบนทางด่วน กรณีศึกษาของประเทศไทย. กรุงเทพ: มหาวิทยาลัยศรีปทุม, 2544.

วัฒนวงศ์ รัตนวราห. (2543). การหาต้นเหตุของการเกิดอุบัติเหตุ. *การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 6*, เพชรบุรี.

AF/02-08/01.0

Suranaree University of Technology Institutional Ethics Committee		Self-Assessment Form for PI
Protocol No.	Title	(ไทย) การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน (English) Implementation of Big Data Analytics in Logistics And Supply Chain
Investigator's name and qualification: ศาสตราจารย์ ดร.วัฒนวงศ์ รัตนวราห สาขาวิชาวิศวกรรมขนส่ง สำนักวิชาวิศวกรรมศาสตร์		☐ Exemption ☐ Expedited review ☐ Full board review

Item for Assessment	A	B	NA	A=appropriate, B=in appropriate, NA= Not applicable หมายถึงไม่มีหรือไม่เกี่ยวข้อง
Protocol				Opinion/Suggestion
1. Research value/merit	✓			
2. Research validity				
2.1 Good rationale	✓			
2.2 Appropriate design and methodology	✓			
2.3 Sample size consideration	✓			
2.4 Statistical analysis	✓			
3. Inclusion/exclusion criteria				
3.1 Assure fair selection	✓			
3.2 Answer research question	✓			
3.3 Concern about risk group			✓	
4. Risk (to whom...)			✓	
5. Benefit (to whom...........ด้านสังคมและชุมชน............)	✓			
6. Vulnerability			✓	
7. Additional safeguard				
7.1 Appropriate recruitment			✓	
7.2 Adequate informed consent process			✓	
7.3 Acceptable treatment available			✓	
8. MTA/CTA (Material Transfer Agreement/Clinical Trial Agreement)			✓	
9. Others (Advertising, CRF, etc.)			✓	
ICF (Informed consent form) **ส่วนประกอบของเอกสาร** (ICH GCP 4.8.10)				Opinion/Suggestion
1. เอกสารข้อมูลคำชี้แจง/อธิบายสำหรับอาสาสมัครที่เข้าร่วมการวิจัย				
1.1 หัวข้อเรื่องที่จะทำการวิจัย	✓			
1.2 ภาษาที่เข้าใจง่าย	✓			
1.3 มีข้อความระบุว่าเป็นงานวิจัย	✓			
1.4 เหตุผลที่อาสาสมัครได้รับเชิญให้เข้าร่วมในโครงการวิจัย				
1.5 วัตถุประสงค์ของโครงการวิจัย	✓			
1.6 จำนวนอาสาสมัครที่เข้าร่วมโครงการวิจัย				
1.7 วิธีดำเนินการที่จะปฏิบัติต่อผู้เข้าร่วมวิจัย	✓			
1.8 ระยะเวลาที่อาสาสมัครแต่ละมนุษย์จะต้องอยู่ในโครงการวิจัย				
1.9 ผลประโยชน์ที่คาดว่าน่าจะเกิดขึ้นจากการวิจัยต่ออาสาสมัครโดยตรงและ/หรือประโยชน์ต่อชุมชน/สังคม/เกิดความรู้ใหม่	✓			

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในมนุษย์

Item for Assessment	A	B	NA	A=appropriate, B=in appropriate, NA= Not applicable หมายถึงไม่มีหรือไม่เกี่ยวข้อง
1.11 ทางเลือกหรือกระบวนการรักษาอื่นๆ ในกรณีที่อาสาสมัครไม่เข้าร่วมโครงการวิจัย			✓	
1.12 การให้เงินชดเชยค่าเดินทาง การเสียเวลา ความไม่สะดวกไม่สบาย และรายได้ที่เสียไปจากการที่อาสาสมัครเข้าร่วมการวิจัย วิธีการให้และเวลาที่ให้			✓	
1.13 การให้การรักษาพยาบาลหรือค่าชดเชย เมื่อมีความเสียหายหรืออันตรายที่เกิดจากการวิจัย			✓	
1.14 แหล่งเงินทุนวิจัยและสถาบันที่ร่วมการทำวิจัย			✓	
1.15 การวิจัยทางพันธุศาสตร์จะต้องมีการขอความยินยอมและมีการให้คำปรึกษาเกี่ยวกับ genetic counseling			✓	
1.16 การขอเก็บตัวอย่างที่เหลือจากการวิจัยและระยะเวลาที่เก็บเพื่อตรวจเพิ่มเติมในอนาคต หรือเพื่อการศึกษาใหม่ในอนาคต ต้องมีการขอความยินยอมเพื่อเก็บตัวอย่างที่เหลือ แต่การใช้ตัวอย่างนั้นจะต้องยื่นเรื่องให้คณะกรรมการจริยธรรมพิจารณา			✓	
1.17 บุคคลและหมายเลขโทรศัพท์ที่สามารถติดต่อได้ตลอด 24 ชม. ในกรณีที่อาสาสมัครเกิดเหตุไม่พึงประสงค์			✓	
1.18 หมายเลขโทรศัพท์สำนักงานคณะกรรมการพิจารณาจริยธรรมการวิจัยที่อาสาสมัครสามารถติดต่อกรณีมีข้อร้องเรียน			✓	
1.19 มีเอกสารข้อมูลฯ ฉบับที่เหมาะสมสำหรับเด็กอายุ 7-12 ปี			✓	
2. หนังสือแสดงเจตนายินยอมเข้าร่วมการวิจัย (Consent Form)				
2.1 มีข้อความ "อาสาสมัครมีอิสระที่จะปฏิเสธหรือถอนตัวจากโครงการวิจัยเมื่อใดก็ได้โดยไม่มีผลใดๆ ต่อการรักษาพยาบาลที่ควรจะได้รับตามมาตรฐาน หรือสูญเสียผลประโยชน์ใดๆ ที่พึงจะได้รับตามสิทธิ"			✓	
2.2 มาตรการการรักษาความลับของข้อมูลเกี่ยวกับอาสาสมัคร			✓	
2.3 ความเหมาะสมของการลงนามโดยผู้เข้าร่วมการวิจัย และ/หรือผู้แทนโดยชอบด้วยกฎหมาย			✓	
2.4 ความเหมาะสมของการแสดงความยินยอมของผู้เข้าร่วมการวิจัยที่ไม่สามารถอ่านและเขียนได้			✓	
2.5 ความเหมาะสมของกระบวนการขอ assent และการลงนามสำหรับเด็กอายุ 7-12 ปี			✓	

Decision: Risk/Benefit Category
☑ Research involving not greater than minimal risk (การวิจัยที่เกี่ยวข้องกับความเสี่ยงเพียงเล็กน้อย)
☐ Research involving greater than minimal risk but presenting the prospect of direct benefit to the individual subject (การวิจัยที่เกี่ยวข้องกับความเสี่ยงมากกว่าปกติแต่ได้แสดงถึงประโยชน์ต่ออาสาสมัครโดยตรงในอนาคต)
☐ Research involving greater than minimal risk and no prospect of direct benefit to individual subject, but likely to yield generalized knowledge about the subject's disorder or condition (การวิจัยที่เกี่ยวข้องกับความเสี่ยงมากกว่าปกติและไม่ได้แสดงถึงประโยชน์ต่ออาสาสมัครโดยตรงในอนาคต แต่มีความเป็นไปได้ที่จะนำความรู้เกี่ยวกับเรื่องความผิดปกติหรือภาวะของโรคของอาสาสมัครไปใช้กับผู้ป่วยมนุษย์อื่นๆ ได้
☐ Research not otherwise approvable which presents an opportunity to understand, prevent or alleviate a serious problem affecting the health or welfare of children (การวิจัยที่มีนัยยะหนึ่งที่สามารถพิสูจน์ได้ถึงโอกาสที่จะเข้าใจ ป้องกันหรือบรรเทาปัญหาร้ายแรงที่มีผลกระทบต่อสุขภาพหรือสวัสดิภาพความเป็นอยู่ที่ดีของเด็ก

Investigator's Signature ....

(ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห)

Date........../................../..........

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในม
มหาวิทยาลัยเทคโนโลยีสุรนารี แล

AF/12-08/01.0

Suranaree University of Technology Institutional Ethics Committee	**ข้อมูลคำอธิบายสำหรับผู้เข้าร่วมในโครงการวิจัย** (Information Sheet for Research Participant)

สำหรับโครงการวิจัยทางสังคมศาสตร์ที่ใช้แบบสอบถามตอบด้วยตนเอง

เรียน......ผู้เกี่ยวข้องกับการศึกษานี้.........................

เนื่องด้วย ข้าพเจ้า **ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห** กำลังดำเนินการวิจัย เรื่อง การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน

โดยมีวัตถุประสงค์ของการวิจัย

วัตถุประสงค์ทั่วไปของการศึกษานี้คือการอธิบายและประเมินของการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อใช้สำหรับงานการจัดการโลจิสติกส์และห่วงโซ่อุปทาน โดยมีรายละเอียดตามประเด็นย่อย ดังนี้

- พัฒนารูปแบบการวิเคราะห์ข้อมูลขนาดใหญ่ในปัจจุบัน เพื่อประยุตใช้ในงานการจัดการโลจิสติกส์และห่วงโซ่อุปทาน
- นำผลการวิเคราะห์ข้อมูลขนาดใหญ่ไปใช้ในอุสาหกรรมห่วงโซ่อุปทาน
- ประเมินความสามามารถของการวิเคราะห์ด้วยข้อมูลขนาดใหญ่ จำแนกตามกรณีศึกษา

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าจะได้รับความร่วมมือจากท่านเป็นอย่างดี และขอขอบพระคุณเป็นอย่างสูงมา ณ โอกาสนี้ หากท่านมีข้อสงสัยเกี่ยวกับงานวิจัย โปรดติดต่อได้ที่ ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห สาขาวิชาวิศวกรรมขนส่ง สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี โทรศัพท์ : 044- 224238

หากท่านมีปัญหาสงสัยเกี่ยวกับสิทธิของท่านขณะเข้าร่วมการวิจัยนี้ ต้องการทราบข้อมูลเพิ่มเติม โปรดสอบถามได้ที่ "สำนักงานคณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีสุรนารี โทร. 044-224757

ขอขอบพระคุณอย่างสูง

(ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห)

Version 2.0 (วันที่ 18 ตุลาคม 2561)

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยในมนุษย์
มหาวิทยาลัยเทคโนโลยีสุรนารี แล้ว

AF/16-08/01.0

Suranaree University of Technology Institutional Ethics Committee	**แบบขอรับการยกเว้นการขอความยินยอมจากอาสาสมัคร** (Waiver of Informed Consent)

1	ชื่อโครงการวิจัย การประยุกต์ใช้การวิเคราะห์ข้อมูลขนาดใหญ่ในงานจัดการโลจิสติกส์และห่วงโซ่อุปทาน
2	ชื่อหัวหน้าโครงการวิจัย ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห
3	ชนิดของการยกเว้นการขอความยินยอมจากอาสาสมัคร โปรดเลือกตอบเฉพาะ 3.1 หรือ 3.2 ข้อใดข้อหนึ่งเท่านั้น
3.1	ยกเว้นการลงนามเป็นลายลักษณ์อักษรในแบบยินยอมของอาสาสมัครบางคนหรือทั้งหมด (Waiver of documentation of consent) และโปรดแสดงเหตุผลในการขอยกเว้น
	1) การวิจัยมีความเสี่ยงต่ออาสาสมัครไม่มากเกินกว่าความเสี่ยงที่อาสาสมัครจะได้รับใน การดำเนินกิจวัตรประจำวัน เพราะ <u>เพราะใช้ข้อมูลทุติยภูมิซึ่งมิได้สำรวจจากผู้เข้าร่วมโครงการโดยตรง และจะไม่ปรากฏข้อมูลส่วนบุคคลของผู้ที่เกี่ยวข้องในการศึกษานี้</u> ดังนั้นจึงไม่จำเป็นต้องขอความยินยอมเป็นลายลักษณ์อักษร (เช่นหัตถการที่เกี่ยวข้องกับการตรวจวินิจฉัยและการรักษาเป็นต้น) (21 CFR 56.109(c); 45 CFR 46.117(c) (2)).
	2) การลงนามเป็นลายลักษณ์อักษรในแบบยินยอมของอาสาสมัครเป็นข้อมูลเดียวที่เชื่อมโยงระหว่างตัวตนของอาสาสมัครกับการวิจัย และความเสี่ยงหลักของการวิจัยทำให้อาสาสมัครตกอยู่ในภาวะอันตราย หากมีการเปิดเผยความลับของอาสาสมัคร เพราะ..........-... (45 CFR 46.117(c)(1)).
3.2	ขอยกเว้นการขอความยินยอมจากอาสาสมัคร (Waiver of Informed Consent) (45 CFR 46.116(d)).
	1) การวิจัยมีความเสี่ยงต่ออาสาสมัครไม่มากเกินกว่าความเสี่ยงที่อาสาสมัครจะได้รับในการดำเนินกิจวัตรประจำวัน เพราะ..
	2) การยกเว้นการขอความยินยอมจากอาสาสมัครจะไม่ส่งผลกระทบต่อสิทธิและความเป็นอยู่ที่ดีของอาสาสมัคร เพราะ..
	3) ผู้วิจัยไม่สามารถทำวิจัยได้หากไม่ยกเว้นการขอความยินยอมจากอาสาสมัคร เพราะ ....
	4) อาสาสมัครจะได้รับแจ้งเพิ่มเติมเกี่ยวกับการวิจัยหรือไม่ และได้รับข้อมูลอย่างไร ....

<u>หมายเหตุ</u> : การขอยกเว้นการขอความยินยอมจากอาสาสมัครไม่สามารถกระทำได้หากเป็นโครงการวิจัย ที่เกี่ยวข้องยาหรือเครื่องมือแพทย์ที่อยู่ในระหว่างการวิจัยเพื่อขอขึ้นทะเบียนยาขององค์กรอาหารและยาของประเทศสหรัฐอเมริกา

ลงชื่อ..

(ศาสตราจารย์ ดร. วัฒนวงศ์ รัตนวราห)

หัวหน้าโครงการวิจัย

วันที่..

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัย
มหาวิทยาลัยเทคโนโลยีสุรนา

Certificate of Completion

National Research Council of Thailand (NRCT) and Forum for Ethical Review Committee in Thailand (FERCIT)

Certify that

Vatanavongs Ratanavaraha

Has completed the ON-LINE RESEARCH ETHICS TRAINING

Course หลักสูตรหลักจริยธรรมการวิจัยในมนุษย์ สำหรับนักศึกษา/นักวิจัย

Date approved (20/03/2562)

Date expired (20/03/2565)

S. Songsivilai

(Professor Dr.Sirirurg Songsivilai)
Secretary-General
National Research Council of Thailand

http://ohrs.nrct.go.th/lms/reportcertcompletion?uid=ODkyMw==&typeID=MQ==

ผ่านการพิจารณาจาก

IMPLEMENTATION OF BIG DATA ANALYTICS IN LOGISTICS AND SUPPLY CHAIN MANAGEMENT

Vatanavongs Ratanavaraha

Transportation Engineering
Suranaree University of Technology
Academic Year 2018/2019

ผ่านการพิจารณาจาก
คณะกรรมการจริยธรรมการวิจัยใ
มหาวิทยาลัยเทคโนโลยีสุรนารี

APPENDIX II

LIST OF AUTHOR PUBLICATIONS

Yudhistyra, W. I., Raungratanaamporn, I.-s., & Ratanavaraha, V. (2019, 2019). *Big* **Data Analytics Techniques Exploration for Customers Analysis to Gain Competitive Advantage**. Paper presented at the 16th Pacific Regional Science Conference Organization (PRSCO2019) Summer Institute Bangkok.

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I.-s., & Ratanavaraha, V. (2020a). Exploring Big Data Research: A Review of Published Articles from 2010 to 2018 Related to Logistics and Supply Chains. **Operations and Supply Chain Management: An International Journal**. 13(2): 134-149. doi:https://doi.org/10.31387/oscm0410258 (Indexed by ISI Web of Science and SCOPUS).

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I.-s., & Ratanavaraha, V. (2020b). **Implementation of Big Data Analytic: Customers Analyzing using an Association Rule Modeling in a Gold, Silver, and Precious Metal Trading Company in Indonesia**. Paper presented at the 2020 International Conference on Big Data in Management (ICBDM 2020), Manchester. (Indexed by Ei Compendex and SCOPUS and submitted to be reviewed by Thomson Reuters Conference Proceedings Citation Index (ISI Web of Science)).

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I.-s., & Ratanavaraha, V. (2020c). Implementation of Big Data Analytics in Customers Focused Supply Chain Management: The Combination of Data Mining and Behavior Scoring for Analyzing the Customers in a Gold, Silver, and Precious Metal Trading Company in Indonesia. Industrial Engineering and Management Systems Journal. (Indexed by ISI Web of Science and SCOPUS).

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I.-s., & Ratanavaraha, V. (2020). Using Big Data Analytics for Decision Making: Analyzing Customer Behavior using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia. **International Journal of Data Science**. 1(2): 51-71. doi:https://doi.org/10.18517/ijods.1.2.57-71.2020.

BIOGRAPHY

Mr. Wecka Imam Yudhistyra completed his previous study in electrical engineering and information technology from Gadjah Mada University in Yogyakarta, Indonesia. His research interests focus on implementation of big data analytics in various sectors particularly supply chain management. He enjoys playing music, reading books, and learning how to take advantage of information technology and systems so they can add value to their businesses, products and services.