BIG DATA MATURITY MODEL: ASSESSING ORGANIZATIONAL READINESS FOR THE NEW ERA OF DATA ANALYTICS

Doctor of Business Administration Dissertation

By

Abdulkhaliq Alharthi

In Partial Fulfillment

of the Requirements for the Degree

of Doctor of Business Administration

Abu Dhabi University

College of Business Administration

Abu Dhabi, UAE

June, 2018

Big Data Maturity Model: Assessing Organizational Readiness for the New Era of Data Analytics

Abdulkhaliq Alharthi
ID Number: 1012622

Approved:

Gabriele Piccoli
Edward G. Schlieder Endowed Chair of Information Sciences
Louisiana State University
Committee Chair

Vladyslav Krotov
Associate Professor of MIS and TSM
Murray State University
Committee Member

Professor Chris Pitelis
Dean, College of Business
Abu Dhabi University

ABSTRACT

Today, in the era of Big Data, many organizations around the world realize that the ability to analyze and use big and complex data sets and data streams is becoming one of the most important sources of competitive advantage. Not surprisingly, Big Data capabilities are becoming essential for the survival and success for both private and public organizations. In order for organizations to take full advantage of opportunities that Big Data offers, they need to reach a new level of maturity with respect to data analytics. Some researchers view Big Data as a natural extension of Business Intelligence (BI). Because of that, the current Big Data maturity models are still grounded in the BI era. But the models that have been used for evaluating organizational maturity in relation to BI cannot be used to evaluate organization maturity in relation to Big Data due to the difference in the underlying processes, methodologies, and technologies. This dissertation highlights the theoretical differences between BI and Big Data and empirically explores the key Big Data capabilities via a newly proposed Big Data Maturity Model populated by specific Big Data capabilities at various levels of maturity.

The main goal of this research it to develop a practical model that is suitable for evaluating maturity of various organizations in relation to Big Data. The development of the maturity model goes through two main phases. The first

phase is the literature review phase. The literature review highlights the limitations of existing maturity models in relation to Big Data and then uses the Socio-Technical Perspective (STP) and the Resource-Based View (RBV) as its primary theoretical lenses to identify the key capabilities required for working with Big Data. These capabilities fall into the people, technology, and organizational domain. The second phase involves using the developed theoretical model to empirically explore Big Data capabilities among various organizations. The preliminary results suggest that the Big Data model can be used to discriminate among organizations at various levels of maturity in relation to Big Data. In addition to proposing a preliminary theoretical model for understanding what it takes to build Big Data capabilities, this research also provides practical guidance to organizations wishing to improve their Big Data capabilities for the purpose of improving their organizational performance and gaining competitive advantage over their rivals.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Vlad Krotov, for his continuous support of my doctoral studies and dissertation research - for his patience, motivation, and immense knowledge. His guidance helped me tremendously to finish this research. I could not have imagined having a better advisor and mentor for my doctoral studies.

I also extend my gratitude to my Dissertation Committee Chair, Dr. Gabriele, for his insightful comments, guidance, and, most importantly, the smart questions he raised in relation to my dissertation research. These questions motivated me to improve my work and expand my research perspectives.

My special thanks go to my brother Salem for encouraging and supporting me throughout my studies. He is always besides me when I need him. My sincere gratefulness and love goes to my sister and to my wife and kids for being always behind all successes I have achieved, including this great achievement.

And finally, I would like to take a moment to spare a thought for those whom are not now with us on this earth. My Father and my Mother, through whose being and passing, I learnt the meaning of life and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BD	Big Data
BDC	Big Data Capability
BDMM	Big Data Maturity Model
BI	Business Intelligence
BICC	Business Intelligence Competency Centre
CAO	Chief Analytics Officer
CDO	Chief Data Officer
CIO	Chief Information Officer
CTO	Chief Technology Officer
DAS	Direct-Attached Storage
DS	Data Scientist
DW	Data Warehouses
GB	Gigabytes
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
ICT	Information and Communication Technologies
IoT	Internet of Things
KPI	Key Performance Indicator
NAS	Network-Attached Storage
NoSQL	Not Only Structured Query Language

PMO	Program Management Office
RBV	Resource-Based View
RDBMS	Relational Database Management System
RFID	Radio Frequency Identification Technology
ROI	Return on Investment
SAN	Storage Area Networks
SCA	Sustained Competitive Advantage
SDLC	Systems Development Life Cycle
STP	Socio-Technical Perspective
SQL	Structured Query Language
WWW	World Wide Web
ZB	Zettabytes

CHAPTER 1

INTRODUCTION

Recent advancements in Information and Communication Technologies (ICT) as well as ever increasing affordability and ubiquity of networks and electronic devices, have resulted in a massive increase in the volume of digital data from various sources and in different formats. This volume is measured today in Zettabytes (ZB) – a measure equal to one trillion gigabytes (GB) and equivalent to data storage capacity of about 250 billion DVDs. The World Wide Web (WWW) alone was estimated to contain 0.5 ZB of data in 2009 (Fan & Bifet, 2013). This amount of data is available from more than one trillion WWW pages currently accessible. The total amount of digital data in the whole world reached 1.8 ZB by 2011 and will be 50 times bigger by 2020 (Jeon, 2012). According to the 2011 McKinsey Global Report, 90% of digital data currently available has been created in the last two years (Gobble, 2013). Many organizations across the globe realize the importance of Big Data as a source of competitive advantage. For example, participants of the 2012 World Economic Forum in Davos, Switzerland declared that Big Data is increasingly becoming a crucial economic asset, similar in significance to currency and gold (Johnson, 2012). The term "Big Data" is often used to refer to this massive volume of digital data available to organizations and individuals today. However, in addition to this vast volume, the Big Data concept also includes the variety and velocity of data, forming the so-called “3V’s” definition of Big Data. In this study, Big Data (BD) is defined as a set of technologies and practices used to manage and analyze data characterized by vast volume, velocity and variety for the purpose of improving organizational performance.

1.1 Research Problem

Despite the fact that Big Data spurred a relatively new (and quite fashionable) movement, it did not emerge “overnight”. It can be argued that Big Data is a natural extension of the Business Intelligence (BI) movement that emerged in the 1990’s. However, the Big Data is bigger, more complex and comes in an endless variety of formats. This puts Big Data outside of the usual BI technologies and capabilities. Therefore, analyzing and exploiting Big Data requires new types of IT infrastructure, analytical tools, and managerial approaches. Thus, handling Big Data requires new organizational resources and capabilities.

Due to the apparent absence of maturity models adequately reflecting the new trends in Big Data, several Maturity Models used earlier for BI and Data Warehousing were reviewed. None of these models was found to be suitable to assess organizational capabilities in relation to Big Data. The review of these models identified a substantial gap in the literature in relation to Big Data that this research aims to close.

1.2 Use of Maturity Model

A Big Data Maturity Model is necessary to assess an organization's readiness to use Big Data analytics and provide guidance to organizations in improving their maturity in relation Big Data. In order for organizations to take advantage of the opportunities offered by the Big Data analytics, they need to identify challenges linked to Big Data and develop appropriate capabilities to overcome these challenges and use Big Data for improving organizational performance. This kind of information can be provided through use of appropriate maturity model.

1.3 Research Objectives and Contribution

The aim of this research is to develop a Big Data Maturity Model that would take account of the new developments in relation to Big Data. The proposed model will help to evaluate the level of maturity of organizations with respect to their readiness to exploit the Big Data opportunity. The level of maturity of an organization can be measured by assessing how regularly and rigorously organizations perform certain behaviours. Overall, the model, when used as a benchmarking tool, can help organizations assesses their capabilities in relation to Big Data, enhance schedule and process consistency, help managers improve organizations capabilities in relation to Big data, and gain competitive advantage via Big Data use.

The theoretical development of the Maturity Model goes through the following phases:

- Review of several Maturity Models, to identify the gaps and limitations in the existing models regards Big Data.
- Develop a framework for Big Data Capability using Socio-Technical Perspective (STP) theory and Resource-Based View (RBV) of the organization to identify key resources and capabilities required for building Big Data Capability.

So in addition to adopting a clear theoretical stance, this study also attempts to address the gaps and limitations of existing maturity models. Based on the literature presented in Chapter 2 and on the Big Data Capability (BDC) frameworks discussed in Chapter 3, a new Maturity Model for Big Data is developed. This model is referred to as the Big Data Maturity Model (BDMM). The model consists of four levels: Unaware, Data Processing, BI and Big Data. The model is also populated with specific capabilities falling into the people, technology, and organizational domains.

1.4 Dissertation Structure

The remainder of this dissertation is organized as follows: Chapter 2, the Literature Review, discusses the evolution of data processing through 20th and 21st century in addition to an overview of Big Data and its concepts. The chapter concludes with a discussion on the need for the BDMM. Chapter 3 presents the STP and RBV as the two main theoretical lenses used for building the BDMM. Chapter 4 discusses the data and methodology for providing a preliminary validation of the proposed BDMM via an exploratory study. Chapter 5 presents the findings of this exploration, followed by a discussion on implications and suggestions for future research in Chapter-6. Finally, Chapter 7 presents the main conclusions of this study.

CHAPTER 2

LITERATURE REVIEW

The Literature Review chapter starts by discussing the evolution of data processing through 20th and 21st century. This evolution has gone through four main different eras, starting with Primary Data Processing Era and ending with the Era of Big Data. The various approaches to data processing identified through this chronological order are used as a foundation for developing maturity levels in the proposed Big Data Maturity Model at a later stage of this research. Then, an overview of Big Data is given with some details on the main differences between Big Data and Business Intelligence. The overview also gives a brief on various definitions of Big Data and sheds some lights on the main opportunities and challenges related to Big Data. The Literature Review chapter concludes by providing details on the concept of a maturity model and its practical value. The discussion includes a summary of several maturity models used primarily for BI. It is then argued that the reviewed maturity models are poorly suited for the era of Big Data.

Overall, this chapter demonstrates the need for a new maturity model that will appropriately assess organizational capabilities' and provide guidance in the era of Big Data, Moreover, the content of this chapter is used as a base for developing a framework for Big Data Capabilities and for developing the Big Data Maturity Model in Chapter 3.

2.1 Evolution of Data Processing

Since the very beginning of the computer era, data processing, data management, and data analytics have been driven by the advancements in ICT's described by Moore's Law (i.e. the number of transistors in a dense integrated circuit doubles approximately every two years). These advancements took the shape of four main "waves" or "eras" of computing: Primary Data Processing Era, Advanced Data Processing Era, Business Intelligence Era, and the Era of Big Data. The Primary Data Processing Era began with emergence of computers and was phased out by 1960's. Then, the Advanced Data Processing Era started with the introduction of transistors and magnetic tapes and continued until the 1990's. The Advanced Data Processing Era was driven by relational databases. It is believed that E. F. Codd at IBM invented the term ”relational database” in 1970. Oracle Corporation used the "relational database" term to describe its first commercial Relational Data Base Management System (RDBMS) in 1979. This release was followed by IBM's own RDBMS released in 1981 called SQL/DS. Starting from 1990, a new movement in data analytics, known as Business Intelligence (BI), became popular. The development in data processing during the BI era was largely driven by the advancements in desktop computing, client-server architecture, and the Internet. BI continued as the main movement in data management and data analytics until 2005, when a new concept of data management and analysis called Big Data was introduced. This overview of the evolution of data processing is presented in this section in order to provide a historical perspective on data analytics and allow for a better understanding of how the Big Data movement differs from other previous forms of data analytics.

2.1.1 Primary Data Processing Era: Before 1960's

Vacuum tubes and punch cards were the most important technologies for electronic computing in this era until approximately 1960's. Vacuum tubes were made of two or more electrodes in a sealed container. Early digital computers used vacuum tubes as electronic switching elements for implementing computer logic and performing computations. Punch cards were used as the main medium for data entry, data storage, and programming code loading. The punch cards were made of stiff paper that could be punched in predefined positions. The holes were then used to represent a digit - an alphanumeric or special character in accordance with the encoding system used. One of the first successful applications of the punch cards system was during the US Census of 1890, when 62.5 Million punched cards were processed in three years. This was much faster than the data processing for the 1880 Census – it required about seven years to process the results. The cards during the 1890's Census were punched by census takers in predetermined location representing the answers. Thereafter, punch cards were processed using a device that could detect holes and translate them into electrical signals (Cortada, 2000).

2.1.2 Advanced Data Processing Era: 1960 to 1990

During this period of time, computer processing power and storage capacity witnessed significant advancements due to the invention of transistors and magnetic tape. Transistors are made of semiconductor material. Just like vacuum tubes, transistors were used to amplify and switch electronic signals. Transistors completely replaced vacuum tubes and changed the field of electronics. Transistors facilitated the invention of modern radios, TVs, mobile phones, computers, etc. Nowadays transistors exist in modern electronic devices in the form of integrated circuits. Compared to vacuum tubes, transistors are cheaper, smaller, more energy efficient, more reliable, and can be integrated in very small circuits. Today, an integrated circuit may contain billions of transistors. These technical and economic advantages of transistors helped increase processing power of computers tremendously.

Parallel to the development in transistor technology, magnetic tape replaced the punch cards as the data storage medium. Magnetic tape was invented for recording sound based on magnetic wire recording technology. The magnetic tapes used long stripes coated with a magnetic oxide layer to record audio, video, and data. A motor was used to move the magnetic tape past the tape “heads”. These tape heads converted electrical signals to magnetic signals and vice versa in order to read or write audio, video, or data as the magnetic tape moved through them. Magnetic tapes allowed computers to have larger storage capacity at a lower cost. Moreover, data could be stored for a longer period of time and accessed or edited with relative ease.

Processing and storing of massive volumes of data became possible due to the advancements in transistor and magnetic tape technologies. Powerful mainframe computers were created based on these data technologies. Mainframes were characterized by their large size, high processing power, and big storage capacity. Mainframes made possible processing of massive volumes of data for large corporations and government entities. Data management and analytics with the help of databases, statistical methods, and data mining witnessed significant development as well (Chen, Chiang, & Storey, 2012; Cortada, 2000; Davenport & Dyché, 2013).

In 1970, E. F. Codd, while working at IBM, invented the term "relational database". The data model used for relational databases relies using a unique key for each row of data in a table. Row in any table can be linked to other rows using the unique key known as a “foreign key”. This simple feature of relational databases became central for the Relational Database Management Systems (RDBMS) – an integrated data management solutions relying on the relational data model. The language used for querying and maintaining the RDBMS was called SQL (Structured Query Language). The first commercial RDBMS was released in 1979 by Relational Software (now Oracle Corporation). Two years later, in 1981, IBM released its commercial RDBMS called SQL/DS.

Parallel to that, during the 1980s, data processing was moved off mainframes to a smaller architecture called minicomputers and microcomputers, paving the way for desktop computing. This change in architecture allowed for a more cost efficient alternative to mainframe-based architectures. During this period a variety of tools were developed for data processing and analysis, including optimization and simulation, artificial intelligence, statistical software packages, spreadsheet, in-database analytics, etc. But computers and software applications used during this period for data processing were still somewhat expensive and rather simple compared to today's computers and applications (Gorry & Morton, 1989; Picciano, 2012; Power, 2007).

2.1.3 Business Intelligence Era: 1990-2005

The 1990s saw the developments of the Internet and the World Wide Web (WWW) – probably the most significant technology developments of the 20th century. Development and deployment of advanced web-based analytical tools had witnessed a steady growth. Some well-known organizations (e.g. Google, Yahoo, Amazon, eBay, etc.) developed their own advanced tools for data collection, storage, and analysis of web data (Bhargava & Power, 2001; Doan, Ramakrishnan, & Halevy, 2011; O'reilly, 2009; Powell, 2001). The practice of capturing and analysing business data became known as Business Intelligence (BI). The term BI was arguably coined by a Gartner consultant and became very popular in 1990s in the business community. BI is commonly used to describe tools and techniques that utilize data to improve business decision making. It was often used interchangeably with data-driven Decision Support Systems (DSS). But DSS mainly focused on analysing data and transforming it into meaningful information for decision makers (Hribar, 2010; Power, 2007).

2.1.4 Big Data Era: 2005 and Onwards

At the end of the 20th century, more than 75% of data was stored on paper, film, and analog media, such as vinyl disks and magnetic tape (Cukier & Mayer-Schoenberger, 2013). However, due to the advancements in ICTs and data storage technologies, the proportion and volume of digital data has been increasing rapidly. Nowadays, 98% of data is stored in digital format.

The vast growth of digital data was driven, in part, by the heavy Internet use and online activities such as online gaming, video streaming, and cloud computing. These ever-growing uses of the Internet and WWW produced complex and vast amount of data. The data produced could not be managed or manipulated any longer using the traditional BI technologies and tools. To address this problem, in 2003, Google, together with many other companies, began experimenting with an algorithm capable of managing and processing large volumes of complex and unstructured data in real time. The algorithm was later called "MapReduce". The MapReduce algorithm uses parallel processing for analysing large data sets across large number of cheap computers or servers. The MapReduce algorithm has two main phases: *map* and *reduce*. The first phase, "map", splits data into smaller subsets and distributes it over different nodes called "slave nodes". In these slave nodes the data is processed; the results are passed back to the "master node". During the second, "reduce" phase, the master node collects all the returned results and combines them into useful output that can be analysed further. The MapReduce is considered to be one of the first, real Big Data tools. The

MapReduce algorithm drove the invention of Hadoop in 2006, which became a synonym to Big Data. Since then, Big Data started to receive more attention from researchers, practitioners, and vendors due to its promise to revolutionize the way business is conducted.

2.2 Big Data Overview

2.2.1 Definition of Big Data

The term Big Data is used to describe massive volumes of data and information produced by human activity that are hard to manage using conventional data analysis tools and technologies (Marshall, 2012; Lewis, Zamith, & Hermida, 2013). Despite different views on what precisely is meant by Big Data, the definition by Gartner is widely used. According to the definition, Big Data is characterized by the “3 V’s”: volume, variety, and velocity. Each of these dimensions is described in more detail below:

- *Volume* refers to the vast amount of structured and unstructured data that is hard to collect, manage and analyze with the traditional tools and technologies.
- *Variety* refers to the fact that the data comes from various sources and in different formats, such as text, images, audio, video, GPS data, sensor data, web data, social media data, etc.
- *Velocity* refers to the fact that many types of data nowadays are generated, modified, processed, analysed, or visualized in real time.

Big Data is mostly comprised of the massive amount of unstructured data, like texts, images, videos, sensor data, etc. Unstructured data comprises more than 80% of digital data stored today (Sint, Schaffert, Stroka, & Ferstl, 2009). Unfortunately, unstructured data has never been managed or analysed before on such scale due to a number of technological limitations. One of the best examples of unstructured data that has been gaining substantial attention recently is sensor data. The sensor data has been growing exponentially due to the widespread deployment of sensors in many objects and environments. Sensors are now available in mobile handsets, cars, smart systems, cameras, automatic doors, electrical stairs, elevators, traffic signals, factories, airports, etc. Data collected from sensors can be used to monitor pressure, temperature, speed, motion, location, quantity, power, etc. Collected data can be used for different purposes, such as operation management, supply chain management, asset management, health monitoring, traffic control, etc. Volume of sensor data is expected to grow more than 10 times faster than data from Twitter, Facebook, YouTube, and other social media feeds in the coming years. Given this projected growth, Big Data tools should be developed and utilized with the sensor data in mind in order to take advantage of it (Ferguson, 2012; Sun & Heller, 2012; Vorhies, 2013).

Following are some of the most common definitions for Big Data in the literature:

1. In 2001, Gartner Inc. proposed one of the most cited definitions for Big Data. The definition states that “Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Glossary, 2013). The definition is known as the “three V's” of Big Data: Volume, Velocity and Variety. "The 3V's" are used by researchers and practitioners alike to define Big Data (Gandomi & Haider, 2015).
2. IBM added a fourth V to the "three V's": Veracity. The term Veracity was introduced to capture such Big Data issues as uncertainty and abnormality in data (Ward & Barker, 2013).
3. SAS added two additional dimensions to "the 3 V's": Variability and Complexity. Variability refers to inconsistent data flows. Complexity reflects issues related to complexity of dealing with data from multiple sources and finding relationships between them (SAS, 2016).

4. Oracle stresses that Big Data is not a new movement but rather is an evolution in existing tools and technologies with the addition of different and bigger data sources and new, innovative data technologies. Oracle refers to traditional enterprise data, machine-generated or sensor data, and social media data as Big Data. The company recommends that enterprises develop their IT infrastructure to handle these new types of data and integrate them with traditional enterprise data to be analysed. Oracle's view on Big Data also focuses on the technologies used to manage and analyse this data, such as NoSQL, Hadoop, and R (Dijcks, 2012; Ward & Barker, 2013).
5. Microsoft defines Big Data as the process of applying significant computing power to massive and complex datasets. The process should utilize advanced techniques and technologies for analytics like machine learning and artificial intelligence (Ward & Barker, 2013).

The term "Big Data" often refers not only to data, but also to the technologies and processes for analyzing, processing and managing massive and complex datasets.

2.2.2 Big Data Technologies

Some important technologies of Big Data related to these "3 V's" characteristics are MapReduce, NoSQL, and In-Memory processing. Each technology is predominantly related to one of the 3V's as shown in Figure 2-1.

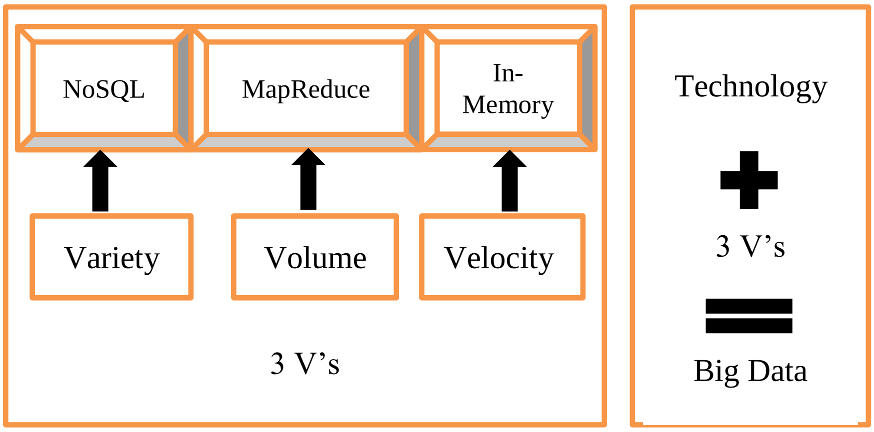

Figure 2-1. Big Data Technologies and 3 V's

NoSQL technology addresses the variety of data and inadequacies of relational databases for handling truly Big Data**.** NoSQL stands for "non SQL" or, as some suggest, "Not only SQL". NoSQL databases are not limited to the structured data that can be accessed using the traditional Structured Query Language (SQL); they can handle unstructured data that is beyond the capability of traditional relational database tools. Unstructured data includes such formats as texts, images, videos, sensor data, social network, etc. The MapReduce technology allows to deal with the issues arising from massive volumes of data via fault tolerant parallel processing done across a large number of cheap, commodity computers or servers. In-Memory technology mainly addresses the velocity of data. In-Memory concept is based on a simple idea to load (or cache) data into Random Access Memory (RAM) and process it there instead of using hard disks. The key difference is that data cached from a hard drive is typically small and specific to certain requirements. Under the In-Memory approach, the whole data needed for analysis is cached or available in RAM. Since data is kept "in-memory", the response time of any calculation is faster compared to traditional approaches to data loading. This increased speed makes In-Memory technologies suitable for real-time data processing and analytics (Yellowfin, 2010).

Of course, the Big Data innovations are not limited to these three areas of technologies. But, together, MapReduce, NoSQL and In-Memory technologies are core enablers of Big Data analytics. It can be argued that these technologies have led to the emergence of data-driven organizations such as Google, Yahoo, Amazon, eBay and many others (Chen et al., 2012; Davenport & Dyché, 2013).

2.2.3 Big Data vs. Business Intelligence

A discussion on Big Data would not be complete without a conceptual reflection on how Big Data differs from an earlier data analysis movement called Business Intelligence. On the surface, Big Data Analytics and Business Intelligence appear to be similar. Yet, three major differences between these two movements can be identified based on the characteristics of data used. The underlying data is different along the dimensions of the 3Vs: volume, velocity and variety. Table 2-1 provides a brief summary of these differences. Big Data technologies overcome the limitations of BI in at least three major areas: Volume, Velocity and Variety. Differences between BI and Big Data Analytics along the lines of each of these three dimensions are discussed in detail below.

Table 2-1 Big Data vs. Business Intelligence

Characteristics	Big Data	Business Intelligence
Volume	Infinite	Finite
Velocity	Real time	Offline
Variety	Unstructured	Structured

In general, Big Data technologies are capable of managing and analyzing virtually infinite volumes of data through the use of MapReduce algorithms and parallel processing architectures using cheap commodity computers or servers, as it was discussed earlier. Most BI technologies have limitations when it comes to dealing with massive volumes of data. BI "fixes" or solutions for increasing storage capacity and processing power tend to be expensive and cumbersome.

BI and Big Data technologies also differ on the dimension of velocity. BI tools and technologies had been developed for handling and analyzing static data, or data "at rest". Thus, traditional BI technologies can hardly be used for analyzing massive data in real time. In contrast to that, Big Data technologies have been developed during the era of streaming media and other real-time data. Because of that, Big Data technologies are capable of managing and analyzing real time data through In-Memory technology and other innovative solutions in virtually real time. This stands in sharp contrast with the BI paradigm.

Last difference between BI and Big Data lies in the capability of Big Data tools to manage and analyze unstructured data. Unstructured data cannot be sorted, managed, or analysed in the same way as structured data. Therefore, new tools and processes were developed for Big Data projects to allow handling unstructured data, such as data in NoSQL structural form. Unstructured data can hardly be accommodated by BI tools built around the relational data model.

2.2.4 Big Data Opportunities

Organizations across the globe are increasingly realizing that the ability to analyze and use big and complex data sets could be one of the most important enablers of competitive advantage for the twenty first century. Big Data analytics has the potential to deliver better customer experience, enhance internal efficiency, and, ultimately, improve profitability. Organizations can use Big Data to get “smart” and innovative in ways that were not possible before the advent of the “Zettabyte era” (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). The need to leverage Big Data analytics is driving the -business of today and tomorrow. In fact, according to some pundits, companies could go out of business if they do not take advantage of Big Data analytics (Johnson, 2012; Loebbecke, Bienert, & Sunyaev, 2013).

A survey of nearly 3,000 executives from more than 30 industries in 100 countries conducted by the IBM Institute and MIT Sloan Management Review journal revealed that top performing organizations are more advanced in data analytics compared to low performing organizations. The leaders in top performing organizations tend to make their decisions based on data - what is known today as data-driven of fact-based approach to decision making (LaValle et al., 2011).

The following case studies illustrate how large organizations across the globe have leveraged Big Data for improving their operational efficiency and competitiveness:

- Etihad Airways: Based in Abu Dhabi, Etihad flies to more than 89 destinations across the globe and carries 10 million passengers annually. Etihad uses Big Data for exploiting new revenue opportunities and optimizing its pricing and product offering strategies. Etihad also uses Big Data for their fleet management. For this purpose, it uses hundreds of sensors on planes to capture data related to these planes in real time. The arrangement is similar to the Internet of Things (IoT) architecture. The real time data provided by the sensors is analysed and used to control the entire fleet of airplanes. This helps the Etihad to predict maintenance and take proactive measures to fix problems before they happen. All this results in substantial costs savings for the company (Rijmenam, 2014). Walt Disney: Every year, about 100 million people visit Walt Disney's parks around the world. The company introduced a park entry pass in the form of a bracelet equipped with radio-frequency identification technology (RFID). The bracelets, called MagicBands, offer visitor many benefits, such as jumping the queues and pre-booking rides. At the same time the MagicBand records the visitor movements and allows collecting large amounts of complex and valuable data about visitors. Disney has been using its Big Data platform based on Hadoop since 2011. The Disney uses these Big Data tools to analyze the data collected about customers' behaviours (e.g. purchasing history, waiting time, preferences, etc.) to improve their operations and marketing offers. The use of Big Data enables Disney to provide a pleasant experience for park visitors, attracting more customers and increasing their revenue (Rijmenam, 2014).

- United Parcel Service (UPS): UPS tracks data on 16.3 million packages per day for 8.8 million customers. The company uses telematics sensors in over 46,000 delivery vehicles to generate a vast amount of data about delivery trucks' speed, location, and performance. Big Data analytics is subsequently used to optimize truck routs. But cutting millions of miles from daily truck routes, the company saves on fuel. A similar project is under study for implementation to optimize movements of the company's fleet of aircraft flights per day (Davenport & Dyché, 2013).
- Bank of America: This large American bank collects Big Data from three different areas: data on transactions, data about customers, and unstructured data. Instead of using data samples for research and development purposes, the bank collects and analyses all the data available on its 50 million customers. This data analysis is used to customize the bank's offers to specific customer segment and, by doing so, become a more customer oriented bank. As a result of these data analytics projects, customer satisfaction at Bank of America went up substantially (Davenport & Dyché, 2013).

- General Electric (GE): GE has more than 12,000 gas turbine jet engines in the aircraft fleet they service. Thousands of sensors are used to monitor the condition of these turbines. Senior executives at GE see clear benefits of data analytics for the company operation and services. The data collected from gas turbines allows GE to prevent technical problems before they happen and optimize their turbine service (Davenport & Dyché, 2013).
- CenterPoint Energy: CenterPoint is a Houston-based Energy Company. It serves more than five million electric and natural gas customers. The company uses smart meters connected via a network to improve its electric power operations and responsiveness of its customer service. These smart meters are streaming vast amount of data every minute over the network. The Big Data platform helps analyze data in near-real time, offering great value to the company in terms of faster response to problems. The company can even predict and prevent potential problems with energy supply before customers call their helpline. The Big Data platform succeeded to transform data into valuable information which results in faster services, enhanced energy efficiency, and improved customer satisfaction. (IBM, 2013a).

- Dublin City Council: The council is responsible to deliver housing, water, and transport services to approximately 1.2 million residents of the city of Dublin in Ireland. For example, there are about 1000 public buses that need to be managed by the city to ensure smooth operations of the public bus services. The Dublin City Council sought to equip buses with GPS and collect the geospatial data these sensors in real time. The data collected is used to optimize bus routes. Using a Big Data solution for data collection and analysis resulted in a better utilization of resources: optimized bus routes help the council save energy and decrease the level of air pollution (IBM, 2013b).
- The Daimler Group: The Daimler Group is the company behind Mercedes-Benz Cars, Daimler Trucks, Mercedes-Benz Vans, Daimler Buses, and Daimler Financial Services. Daimler produces around 10,000 cylinder heads daily that are used in car production. During the manufacturing process Daimler continuously collects more than 500 data attributes of cylinder heads including dimensions, temperature, and many others. Daimler has implemented a comprehensive IBM SPSS solution to continuously monitor the data collected during the cylinder production process. The IBM SPSS solution helps the company predict errors and correct variances before cylinder-head defects happen. In two years since the start of the initiative, the efficiency of Daimler in cylinder production has increased by about 25% (IBM, 2014).

2.2.5 Challenges of Big Data

Despite the numerous potential benefits that Big Data analytics can offer, organizations typically need to overcome certain barriers (or challenges) in relation to Big Data. These barriers range from developing new employee skills and upgrading IT infrastructure to instilling new management practices or organizational culture across the entire organization (Brown, Chui, & Manyika, 2011). These challenges are discussed below and summarized in Table 2-2. Barriers to Big Data and Potential Solutions

2.2.5.1 Infrastructure Readiness

Development of IT infrastructure for Big Data analytics requires significant investments in software and hardware that can support analysis of hundreds of millions records in real time (Loebbecke et al., 2013). Most existing data processing technologies were not designed to meet the growing requirement of Big Data analytics. While such established technologies as cloud and heterogeneous computing are believed to have solutions for Big Data, these technologies often fail when a large amount of data is transmitted between computing nodes. For example, 1,000 cloud nodes each processing a petabyte (one petabyte equals one million gigabytes) would take 750 days (at 15 MB/s) and cost around $6,120,000. Thus, cloud and heterogeneous computing may not help organizations avoid bottlenecks and inadequate software processing power (Trelles, Prins, Snir, & Jansen, 2011).

2.2.5.2 Complexity

An organization wishing to exploit the power of Big Data may face significant problems as this data is often stored in many different formats, including unstructured databases (Douglas, 2013). Moreover, the volume of data is increasing day by day, making the task of handling data from various sources and in different format even more challenging (Johnson, 2012). The specific factors that contribute to the challenges associated with data complexity are described as follow:

- **Rate of data growth.** Most organizations do not have a plan in place to address the problem of large rate of data growth at a reasonable cost; most organizations simply opt to delete old data at some point.
- **Multiple sources of data**. The challenge of combining data from different sources arises from the different ways data is organized in various sources (the so-called "semantic conflict" between various data sources). For example "earnings" may mean "profit" in one dataset while representing "sales" in a different one (Ray, Bandyopadhyay, & Pal, 2009)
- **Multiple formats of data**. Nowadays, data is stored in many formats, including structured and unstructured formats. Examples of unstructured data include text documents, SMS, images, videos, audio, emails, massages, transactions and so on. Structured data is typically organized in relational databases. The challenge of managing and analyzing different formats of data is often beyond the capability of many organizations.

2.2.5.3 Lack of Big Data Skills

Currently, the lack of people with Big Data analytics skills is one of the major challenges facing organizations. It is predicted that the shortage in the number of skilled people worldwide who are able to do advanced analytics with Big Data will be between 120,000 and 190,000, according to the U.S. Department of Labour (Douglas, 2013). Lack of data analytics skills among existing employees may increase data entry errors, result in placing information in the wrong record, and increase chances of losing valuable information (Hoffman & Podgurski, 2013). To make things worse, many organizations do not have the technology to recognize and recover missing or erroneous data (Schouten, 2013).

2.2.5.4 Privacy

Concerns over privacy often hinder adoption and use of Big Data analytics within organizations. Big Data Analytics often employs personal data that was gathered for a specific purpose that was collected for an entirely different purpose. This phenomenon is also known as a score creep (Douglas, 2013). Combining personal information with other data sources can create numerous challenges related to ethics, such as the potential of leaking private information about the individual (e.g. medical records, financial situation, embarrassing behaviour, family relationships, etc.). In many cases organizations are not explicit about how they use their customer data. For example, Twitter, Facebook, and WhatsApp do not seem to have clear and easily accessible policies describing their activities in relation to customer's data. These organizations do not specify in their policies how exactly customer data will be used. At some point Twitter was suspected of selling large amount of Tweets to a Big Data dealer. Similarly, WhatsApp is facing accusation of using customer's data for commercial purposes and violating customer rights to privacy (Rijmenam, 2014).

2.2.5.5 Cultural Barriers

It is observed that most obstacles that organizations face in relation to Big Data are related to organizational culture and not to data or technology. Many organizations are not prepared to change their organizational practices in order to accommodate Big Data analytics and simply lack understanding of how data analytics can improve the business (LaValle et al., 2011). In a study conducted about organizational and technology management practices at 330 public companies in North America, it was discovered that many organizations were not ready to embrace data for decision making (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). Moreover, the results also reveal that the cultural challenges, such as lack of senior management support for using Big Data analytics, were even greater than the technical challenges. Traditionally, most organizations depend on senior management to plan and develop strategies and not on available organizational data. Organizations which fail to figure out how to combine senior executives' expertise with data may be left behind by their rivals.

Table 2-2. Barriers to Big Data and Potential Solutions

Barrier	Literature Sources	Possible Solutions
1. Technological Barriers		
Infrastructure Readiness	Trelles et al., 2011	Commodity hardware should be used to enhance processing power and storage capacity.
Complexity of Data	Douglas, 2013; Johnson, 2012	Specialized software tools and algorithms should be used, such as MapReduce and Hadoop, to store, manage, and analyse complex data in a more efficient, reliable, fast and economical manner (Rijmenam, 2014).
2. Human Barriers		
Lack of Skills	Douglas, 2013;Hoffman & Podgurski, 2013; Schouten, 2013	Organizations should collaborate with educational institutions to align their educational curriculum with the industry requirements for big data skills. Industry and academic institutions should collaborate on providing practical training to address missing skills in the field of data analytics and big data (Miller, 2014).
Privacy	Douglas, 2013; Rijmenam, 2014	Legislative bodies are required to endorse laws that protect individuals from the misuse of data (Schadt, 2012). Organizations need to accommodate this legislation as well as incorporate general best practices for handling sensitive customer data into their policies and operations.
3. Organizational Barriers		
Culture	LaValle et al., 2011; McAfee, et al., 2012	Successful cultural change can be achieved by having a clear organizational vision in relation to big data (Rogers et al., 2006).

2.3 The Need for a Maturity Model for Big Data

In order to take advantage of the opportunities offered by the Big Data analytics, organizations need to identify challenges linked to Big Data and develop appropriate capabilities to overcome these challenges and limitations. So there is a clear need for a Big Data Maturity Model (BDMM) that can be used for assessing an organization's readiness to use Big Data analytics and guiding organizations in improving their maturity in relation to Big Data. Due to the apparent absence of maturity models reflecting the new trends in Big Data, a survey was completed on several Maturity Models used earlier for BI and Data Warehousing. None of these were found suitable to assess organizations in relation to Big Data (even the ones which were posited as focused on Big Data) The review of these models identified substantial gaps in relation to Big Data tools, technologies, and processes. Detailed results of this analysis are presented in the next section.

Given this gap, this research aims to develop a Maturity Model for Big Data which overcomes the limitations of existing maturity models. The proposed BDMM will help organizations assess their readiness to use Big Data and provide guidance on how to enhance their capabilities in relation to Big Data. This can help organizations improve their overall competitiveness by leveraging Big Data.

In order to understand the concept of Maturity Models, an overview of origins of Maturity Models is presented in the next section. After that, an overview of several Maturity Models is presented in tabular format with analysis and brief summary.

2.4 Origins of Maturity Models: Capability Maturity Model

From a historical perspective, it may be argued that BI Maturity Models (BIMM) were derived from the Capability Maturity Model (CMM) (Tan, Sim, & Yeoh, 2011). The CMM model originally developed by the Software Engineering Institute at Carnegie Mellon University in the United States. It has been used for software development and in other fields to assess and improve the process, specifically which related to software development. Maturity Models help identify strengths and weaknesses in software development capability in an organization and provide guidelines for improvement and enhancement (Hribar, 2010; Tan et al., 2011) .

2.5 Review of BI Maturity Models

BIMM's have been developed to measure and identify the maturity level of an organization in relation to BI and determine where it is now and where it should be in the future. The BIMM's normally start from "ground zero" architectures and progress to sophisticated and mature BI architectures (Lahrmann, Marx, Winter, & Wortmann, 2010). This section provides review of 20 BIMM's spanning the period of time from 2001 to 2013. The review of is summarized in Table 2-3. The table gives brief description for each model with a focus on motivation of each model and provides summary of model limitations. The analysis and review in Table 2-3 is adopted from different studies, namely Lahrmann et al., (2010), Shaaban et. al. (2011), and Thamir et al., (2013).

Table 2-3. BI Maturity Models Overview

Model (author, year)	Model Focus	What is Missing?
BI maturity model (Stock, 2013)	Return on investment (ROI), program management office (PMO), management support, data quality	Organization culture, skills and competencies
Enterprise BI maturity model (Chuah & Wong, 2012)	Business alignment, data management, information quality, skills, competencies, redundancy of data	Training, skills, management support and infrastructure
Impact- oriented BI mm Lahrmann et al (2011)	Architecture, analytics tools, technical capability, social capability, data quality, integration and governance	Organization structure, culture , training, skills and management support
American SAP user group hawking et al (2010)	Business needs, KPIs, BI competency centre, standards and processes, data analytics, applications	External environment, training, skills, management support, data management
BI development model (Sacu and Spruit, 2010)	BI implementation, culture, analysis tools, data analytics, data type, data sources and granularity level	Business alignment, training, skills and management support
Teradata' s BI and DW MM Miller et al (2009)	Strategy, business alignment, ROI, training, project management, data governance and acquisition techniques, data quality and management	Organization structure, training, skills, management support and staff
TDWI'S BI Maturity Model(Eckerson, 2009)	Business and organizational view, BI functionality, infrastructure, analytics tools and data architecture	Strategy, training, skills, management support and staff
Hewlett package BI Maturity Model (HP,2009)	Business alignment, BI competency centre, project management, PMO, management support, skills, infrastructure, data quality, data governance	Organization structure, organization culture , staff, data warehousing and analytics
BI MM Steria Mummert Consulting (SMC, 2009)	Business alignment, analytical saturation, business relevance, organizational structure, ROI, it architecture, data management	Organization culture, training, skills, management support, staff

Model (author, year)	Model Focus	What is Missing?
Gartner Maturity Model for BI and performance management (Rayner and Schlegel, 2008)	Alignment between BI and performance management, strategies and business goals, BI competency centre, data policies, management support, it architecture, data consistency, data quality	Organization structure, staff, training, skills, management support, technical aspects
SAS information evolution model (sas,2009)	Information management, alignment between dimensions, BICC implementation, information skills, training, culture, sharing information, information architecture	Organization structure related to BI, staff, tools and technologies required for BI implementation
BI maturity hierarchy (Deng,200 7)	Knowledge management, organization strategy, efficiency of reporting, data warehousing, ROI, experience, perception, infrastructure, data quality, data analytics and data integration.	Organization structure, training, skills and management support
Analytical capability Maturity Model (Davenport & harries, 2007)	Organization strategy, customers, markets, competitor, management support, culture, architecture, data quality, data integration	Organization structure, organization culture, ROI
Infrastructure optimization Maturity Model (Microsoft, 2007)	Culture, ROI, mobility, infrastructure, data mining, data warehousing, data types, data integration, efficiency of reporting and analytics	Organization structure, organization culture and environment, training, skills, management support
Enterprise data management Maturity Model(fisher,2007)	ROI, skills and competencies, management support, data governance, data management, data quality and risks.	Organization structure, strategy, human factors like training, skills and management support
Business information Maturity Model(William and William, 2007)	Business alignment, strategic position, portfolio management, information culture, technical readiness	Training, skills, management support, tools and technologies, data management and utilization

Model (author, year)	Model Focus	What is Missing?
Data warehousing process matur ity (Sen et al.,2006)	Business rules, skills, training, culture, sharing information, BI applications, it infrastructure, data quality, data definition, size and architecture and reliability of data	Organization structure, organization strategy, staff
Amr research's BI/performance management (Hagerty, 2006)	KPIs, organizational strategies, project management, management support, culture, performance management, data source	Training, skills, management support
Ladder of BI (Cates et al.,2005)	PMO role, it governance, management support, business roles, infrastructure, information analysis, processes, data source and data quality	Organization culture and environment, training, skills and management support
Data warehousing stages of growth (Watson et al., 2001)	ROI, staff, applications, data utilization, and structure.	Organization structure, organization culture, strategy, training, skills and management support

Table 2-3 analysis indicates the majority of BIMM's focus on IT components, such as IT infrastructure, applications, and data management (Lahrmann et al., 2010). There is less focus on other important components, such as the ones related to people or organizations. The existing BIMM are especially non-accommodating of important organizational elements when assessing the maturity of organizations in relation to BI and Big Data. Examples of important and overlooked organizational components are strategy structures and organizational culture. The same thing can be said about important human elements, such as users' skills and expertise, top management support, etc. More details of gap analysis is provided in section 3.3.1

CHAPTER 3

THEORY

The analysis of BI maturity models in the literature review chapter highlighted the limitations of the BIMM's in relation to Big Data. This chapter uses the Socio-Technical Perspective (STP) and the Resource-Based View (RBV) of the firm as theoretical perspectives to identify the necessary capabilities and resources for organizations to build Big Data Capabilities.

The STP allows viewing organizations from social and technical aspects. Thus, the framework focuses on people and organization capabilities in relation to Big Data analytics (and not just technology capabilities). The RBV focuses on the internal resources for achieving sustained competitive advantage. Based on the RBV, resources can be classified as tangible and intangible assets that can be used to outperform other organizations. Examples of tangible assets may include products, services, etc., while intangible assets might include skills, processes, reputation, etc. So, the theory helps to identify the key internal resources that form the necessary capabilities for Big Data.

The gap analysis of BIMM and the STP and RBV theories are used to develop a framework for Big Data Capabilities. The proposed framework is presented in

Section 3.3 and is used as a holistic approach for assessing organizations capabilities in relation to Big Data. The Big Data Capability (BDC) framework consists of three main areas: People Capabilities, Organizational Capabilities, and Technology Capabilities. Each capability area is supported by several specific resources related to Big Data.

Furthermore, in order for organizations to be able to determine their maturity level with respect to Big Data capabilities, we have developed a maturity model for Big Data called Big Data Maturity Model (BDMM). The BDMM is comprised from four levels derived from the chronological progress of Data Processing discussed in the literature review chapter. Thus, each level of the BDMM represents a historical period in data processing. The four maturity levels are called: Unaware, Dada Processing (DP), Business Intelligence (BI), and Big Data (BD). The proposed BDMM assumes organization capabilities improve over time through continuous improvement within the organization which is measured through regularity (process level) and rigor (organizational level) of behaviours related to Big Data capabilities. Further details of Big Data Capability framework and BDMM are presented below.

3.1 Socio-Technical Perspective

Developing capabilities in relation to Big Data can be viewed as an IT-dependent strategic initiative aimed at improving organizational performance via deployment of an information system relying on Big Data (Piccoli & Ives, 2005). Implementing this initiative requires attention to three distinct capabilities: people, technology, and organization. Thus, this study uses the socio-technical view on Information Systems as its broad theoretical lens (Alter, 2013; Bostrom & Heinen, 1977; Keen, 1981; Lucas, 1971).

The main reason for selecting this meta-theory is that it allows for a more holistic approach to implementing Big Data initiatives in organizations. Developing, implementing, and operating information systems within organizations is an “intensely political as well as technical process” (Keen, 1981). IT-dependent strategic initiatives often fail because managers ignore organizational behaviour problems during the design and the operation of computer-based information systems (Lucas, 1971). Thus, a broad, STP is required for implementing a comprehensive initiative aimed at improving organizational Big Data capability.

Under STP, an organization is viewed as a combination of social and technical systems. The technical system “is concerned with the processes, tasks, and technology needed to transform inputs to outputs” (Bostrom & Heinen, 1977, p. 17). The social system “is concerned with the attributes of people (e.g., attitudes, skills, values), the relationships among people, reward systems, and authority structures” (Bostrom & Heinen, 1977, p. 17). An information system is produced and operated as a result of the interaction between these two systems. Because of that, “any design or redesign of a work system should deal with both systems in an integrated form” (Bostrom & Heinen, 1977, p. 18). Consistent with this theoretical lens, this study focuses on people, technology, and organization capabilities in relation to Big Data analytics.

3.2 Resource-Based View

The RBV explains how internal resources, including tangible and intangible ones, impact organizational performance. RBV stresses the importance of internal resources for achieving sustained competitive advantage (Barney, 1991; Wernerfelt, 1984). This is in contrast to the environmental approach to strategy which relies on examination of external factors (i.e. threats and opportunities) in determining sustainability of competitive advantage (Porter, 1980).

A fundamental assumption of RBV theory is organizations should have resources that are heterogeneous and immobile, so they may achieve sustained competitive advantages. Resources can be classified as tangible and intangible assets that an organization can use to outperform other organizations. Tangible assets may include products, services, etc., while intangible assets might include skills, processes, reputation, etc. If an organization possesses resources that are heterogeneous and cannot be moved or transferred to any other organization then these resources are said to be immobile resources. Immobile resources are typically intangible resources, such as brands, processes, knowledge or intellectual property. These resources cannot be easily replicated by other organizations (Barney, 1991; Barney & Arikan, 2001; Wernerfelt, 1984).

Using heterogeneous and immobile resources that are not concurrently being employed by any present or potential competitors and cannot be duplicated allows an organization to have a sustained competitive advantage. Attributes of resources that hold the potential of Sustained Competitive Advantages (SCA) are: valuable, rare, imperfectly imitable and non-substitutable (Barney, 1991):

- Valuable – resources are valuable if they allow an organization to increase the value offered to the customers or improve its operational efficiency and effectiveness. Valuable resources alone may not be sufficient to act as a source of competitive advantages but may need other attributes like rareness, inimitability and/or non-substitutability.
- Rare – valuable resources are considered rare when they are possessed only by one or few companies.
- Inimitable – valuable and rare resources can be considered imperfectly imitable (or inimitable) if competitors are not able to duplicate them perfectly. A company can achieve sustained competitive advantage if the valuable and rare resources they have are costly to imitate for a rival.
- Non-substitutable – even if a resource is valuable, rare, and imperfectly imitable but can be substituted, then the resource may not be a source of competitive advantages. Therefore, it is necessary to have a resource that holds this attribute, beside other attributes.

Having a resource that meets some of the aforementioned attributes is necessary but not sufficient for achieving (SCA). Alternatively, resource that displays each of the four attributes is a confirmed source of a SCA. The organization that holds SCA resource should be capable of exploiting it, protecting it and carefully maintaining it for sustainability.

Table 3-1 identifies several resources in three main areas that can be potentially used as a source of SCA for Big Data:

Table 3-1. Resources in Three Main Areas: People, Technology and Organization

Area	Resource	Description
People	Use	Support and proficiency using Big Data
	Skills	Skills and knowledge in developing and using advanced analytics applications for Big Data
	Expertise	IT skills in building hardware and software platform for Big Data
Technology	Presentation	Report quality, format, visualization techniques, dashboard, delivery method and other techniques used to present and deliver information to the end user
	Analytics	The tools used for Big Data manipulation and analysis
	Platform	Collective set of hardware and software required for Big Data. A platform may consist of data storage devices, network elements, transmission links, processing systems, etc.
Organization	Strategy	The high level plan for Big Data analytics implementation and resources allocation
	Process	Specific ways in which Big Data analytics is used and implemented in an organization
	Structure	How organization integrates Big Data analytics in its structure and how responsibilities are allocated to departments or organizational units for different functions related to Big Data analytics
	Culture	How people act toward Big Data analytics with respect to their different experiences, values, education, etc.

The next section discusses development of the Big Data framework based on the theories presented in this section.

3.3 A Framework for Big Data Capability

3.3.1 Gap Analysis of BIMM's in Theory & Literature Review

The review of existing BIMM's (see Table 2-3) reveals that the focus of these models is predominantly on BI technology components. They tend to overlook other important elements, such as people skills, organization structure, and many other "soft" components. They also often fail to include the current Big Data tools and technologies, such as Hadoop and NoSQL.

Relying on the STP and the key elements of RBV, Big Data capabilities presented here take into consideration not only technology aspects but also people and organization sides. Technology capabilities refer to the organization capacity to deploy hardware and software platforms for Big Data technology. People capabilities include people skills and expertise required to build and operate the Big Data solutions. Organizational capabilities refer to the organizational capacity to develop and promote organization strategy, process, structure, and culture to adapt and use Big Data (Barney, 1991; Wernerfelt, 1984).

Limitations of existing BIMM's from a theoretical perspective can also be classified into these three categories: technology, people, and organization. Looking at existing BIMM's from the technology perspective, it appears that current BIMM's assume that organizations are dealing with a limited volume of data in a structured format. Thus, existing BIMM's cannot be used to measure the maturity in relation to Big Data. This is due to the fact that Big Data is characterized by its capability to manage and analyse infinite volumes of unstructured data. In addition to that, from a practical perspective, BI systems were not designed (from hardware and software perspectives) to deal with real time data or data generated from complex sources such as sensors; and this limitation is reflected in existing BIMM too.

From the perspective of the people dimension, people working or dealing with Big Data have different skills and capabilities than those working with data processing using traditional BI tools. For example, Big Data technology experts should have good skills in dealing with distributed platforms and parallel processing. Moreover, they should have a good knowledge of solutions specific to Big Data, such as Hadoop, MapReduce, and NoSQL. This is again indicates the limitation of many existing BIMM's in relation to Big Data, as these skills are often not included in these models

Finally, from the organizational perspective, it appears that Big Data teams and initiatives are separated from IT departments in many organizations due to the difference in technology and people skills. In other words, it is assumed that a Big Data system can somehow work as a standalone system, without the support and interoperability with other IT systems. In addition, the processes for developing Big Data systems and managing Big Data require an organizational culture different from the one that can accommodate BI. Based on these broad dimensions, we conclude that most existing BIMM's are not suitable for Big Data. To develop a proper maturity model for Big Data, it is necessary to develop a new framework for Big Data capabilities. The framework is explained in the following section.

3.3.2 Development of Big Data Capabilities Framework.

In order to form a holistic vision on Big Data capabilities, a Big Data Capabilities (BDC) framework is proposed (see Figure 3-1). The framework presents major capabilities of Big Data and is used as a basis for building the BDMM. The framework consists of three areas of capabilities: people, technology and organization. Each capability area consists of several dimensions representing various resources as shown in Figure 3-1. This framework has been developed to address the gap in the reviewed BIMM's and to account for Big Data characteristics, namely volume, velocity and variety (the 3Vs). The framework shall be used to provide guidelines for organizations to build their capability related to Big Data.

PEOPLE CAPABILITY	ORGANIZATION CAPABILITY	TECHNOLOGY CAPABILITY
BUSINESS USERS	STRATEGY	DATA VISUALIZATION
DATA ANALYSTS	PROCESS	DATA ANALYTICS
TECHNOLOGY EXPERTS	STRUCTURE	BIG DATA PLATFORM
	CULTURE	

Figure 3-1: Big Data Capability (BDC) Extended Framework

3.3.3 People Capability

The people capability is the most important and critical capability. The importance of people stems from the fact that the humans are the ones who create and manage the organization with its resources and capabilities, including technological capabilities ((Davenport, Harris, De Long, & Jacobson, 2001). In relation to Big Data, people in an organization can be classified into three groups: business users, data analysts, and technology experts. Explanations for each of the three groups and their roles in building Big Data capabilities are provided below.

3.3.3.1 Business Users

Business users refer to the people who use Big Data to get information, monitor the status or make their decision based on data. Business users can be classified into two groups: managers and employees. Further details are provided for each group below.

3.3.3.1.1 Managers

The technical challenges in building Big Data technology are many; however, the managerial challenges of adapting and using Big Data seem to be of greater impact than the technical challenges (McAfee, et al., 2012). Managers are responsible for creating organizational policies, fostering organizational culture and making the strategic decisions. These managerial competencies or responsibilities are important for enhancing Big Data capabilities. Managers do not need to be experts in data analytics or Big Data technologies, but need to have a general knowledge of the technologies behind Big Data in order to understand how Big Data analytics can be used for improving organizational performance.

Organizations which lack the tools and technologies to support Big Data are typically headed by managers with poor understanding of Big Data and low interest in using it. Managers in these organizations often base decision on their experience and intuition and do not believe that their knowledge and experience can be outweighed or replaced by technology or data analytics (Davenport, et al., 2001; McAfee, et al., 2012).

Alternatively, managers in organizations driven by Big Data enabled have more interest in data and understand the importance of using it in decision making. Such managers believe in data analytics and demonstrate excellent abilities in making good decisions based on data. This data-driven decision-making is often supported by the manager’s own domain skills and knowledge. These “data driven” managers are key to enhancing Big Data capability in their organizations (Davenport, et al., 2001).

The following competencies are required for managers to enhance the use of Big Data within their organizations (Davenport & harries, 2007; McAfee, et al., 2012; Rijmenam, 2014):

- A clear vision, goals, and strategies in relation to Big Data
- Commitment to use and enhance the Big Data capability
- Knowledge of Big Data tools
- Knowledge of data sources and types
- Determination to make decisions based on data, even if it goes against intuition or previously made decisions
- Knowledge of the internal and external factors influencing their organization
- A sense of fairness that allows treating people based on their capabilities and performance

Smart managers look at Big Data tools not only as analysis tools, but also as an opportunity for achieving SCA. With the help of Big Data, managers can know more about their businesses and transform this knowledge into better decisions and improved organizational performance (McAfee, et al., 2012).

Recently, a role of Chief Data Officer (CDO) has been created to manage Big Data initiatives within organizations. The protection of customers’ privacy and data security are also among the responsibilities of the CDO. Furthermore, the CDO responsibilities involve the following:

- Promoting the data culture in the organization
- Setting the strategies for data management, analysis, and use
- Using data to develop new revenue streams and enhance the "bottom line"

Some of the most important skills and qualifications for the CDO are listed here (Rijmenam, 2014):

- Knowledge of Big Data technologies and solutions, such as Hadoop and MapReduce
- Experience in managing Big Data technical teams across multiple departments
- Understanding of different data modelling techniques and presentation methods
- Experience in managing Data Analysts and Data Scientists
- Business skills for managing internal and external business issues related to Big Data

3.3.3.1.2 Employees

Employees come from many levels and units of an organization. Most are directly affected by the Big Data analytics and, in their turn, also have a direct impact on Big Data use in organizations. Having Big Data analytics embedded into an organization may result in good and timely decisions by employees in all levels and units. In their turn, employees may enhance the organizational Big Data capabilities if they possess the following characteristics Davenport & harries, 2007; McAfee, et al., 2012; Rijmenam, 2014):

- Interests aligned with the organization data's strategy and policy
- Appreciation of the usefulness of Big Data analytics in decision making process
- Willingness to improve their ability to use Big Data

At the end of the day, employees need to be persuaded in the importance of Big Data use and satisfied with the benefits achieved with the help of Big Data. They also need to get trained and empowered to appreciate the benefits of the Big Data. For instance, HR employees can be trained to use Big Data in the selection and recruitment process. Once they are influenced by the Big Data advantages, they will contribute to the enhancement of Big Data in an organization.

3.3.3.1.3 Data Analysts

Data Analysts are analytics professionals involved in retrieving and modelling data, analyzing the results, and communicating findings to others. A Data Analyst needs to have special skills and competencies for transforming data into useful information. Key skills required for a Data Analyst are as follows (Davenport & Harris, 2007; McAfee, et al., 2012; Rijmenam, 2014):

- Capability to work with large data in various formats for the purpose of discovering useful insights
- Knowledge of Big Data platforms, such as Hadoop and MapReduce
- Knowledge of specific tools and applications used for Big Data
- Knowledge and skills in relation to SQL and NoSQL databases
- Understanding programming languages and statistical tools used in Big Data analytics
- Knowledge in developing advanced models based on data
- Ability to see through the data and analyze it to find useful results
- Ability to present findings and complex ideas in a useful and convenient way

Relatively recently, the term “Data Scientist” became popular to describe Data Analysts who have a mixture of advanced skills, such as data transformation and manipulation, research design and execution, statistics, programming languages (e.g. Pig, R, Python, Java, etc.), as well as “soft skills” such as communication and knowledge about business. On the organizational side, people working in this position should have good knowledge of business, speak the same language as executives, and be able to win trust within their organizations. Organizational decision makers may reject any recommendation coming from a Data Analyst or Data Scientist if he or she feels that his or her authority and input were not respected. Thus, Data Scientists should be able to play politics and provide solutions or results as recommendations or options rather than providing “ready-made” decisions. At the same time, a Data Analyst or Data Scientist should maintain good relationships with the technology experts to rely on their expertise and users to close the gaps between different parties within the organization (McAfee, et al., 2012). The people with these skills are currently difficult to find in the labour market, and that is why Data Scientist is a popular profession nowadays. Many organizations often rely on a team consisting of several people to do this job.

3.3.3.2 Technology Experts

A Technology Expert is a technical specialist possessing significant proficiency and knowledge in programming and a wide range of hardware and software technologies related to Big Data. In the recent years, the Big Data technologies have witnessed great advancements. These advancements have largely been driven by cost-effective open source tools and platforms such as Hadoop. The Big Data technology is considered a new development in business, which needs specialized technology skills and competencies that are difficult to find in an average organization. The needed technical knowledge for this role is largely related to the hands-on knowledge of hardware and software used to extract, manipulate, analyze, and present Big Data.

A Technology Expert is responsible for identifying the technical requirements related to Big Data that help organization pursue its Big Data strategies. In general, a Technology Expert should be capable of selecting a hardware platform, designing the network architecture, and selecting the software applications required for deployment of Big Data solutions. On a micro level, a Technology Expert should be able to: (Davenport, et al., 2001; McAfee, et al., 2012; Rijmenam, 2014):

- Understand various hardware platforms, including mainframes, distributed platforms, desktops, and mobile devices
- Design highly scalable distributed systems architectures
- Work with the main Big Data platforms, such as Hadoop and MapReduce
- Build large and scalable distributed data processing systems based on Hadoop and Hive using programming languages, such as Java, C++, or Python
- Understand databases and "data in motion" based on Relational Database Management Systems (RDBMS) and "Not Only SQL" (NoSQL) database architectures
- Recommend technical solutions to the data security and privacy problems arising from the use of Big Data

3.3.4 Technology Capability

When people talk about Big Data capabilities, they usually refer to the technologies behind Big Data and but not the organizational or human competencies. The reason is that the technology is the prevailing capability in the minds of technical specialists and is usually defined in terms of hardware and software tools. Big Data technology has been developed due to the need for new technology capable of handling massive volume of unstructured data. Older technologies comprising the traditional movement related to Business Intelligence (BI) and Data Warehousing (DW) were not capable of dealing with the "3 V's" of Big Data. Thus, the Big Data technology capability is comprised of the following three layers:

- Big Data Platform
- Data Analytics Layer
- Data Visualization Layer

Each of these layers is explained in more details below.

3.3.4.1 Big Data Platform

Building a new and independent hardware platform for Big Data analytics is the best option for new organizations, but not for well-established organizations with existing IT systems. Though Big Data platform can work independently from the existing IT infrastructure, it is often not practical for many organizations to run an additional system in parallel specifically for Big Data. A survey conducted in 2013 of 20 large US organizations showed that none of them had implemented their Big Data architecture separately from their existing IT infrastructure (Davenport & Dyché, 2013). For many valid reasons, organizations prefer to find a suitable solution that allows integration of a Big Data platform with an existing IT systems, such as a BI platform or a Data Warehouse (Davenport & Dyché, 2013; Sun & Heller, 2012).

The Big Data platform includes two layers: hardware layer and software layer. This classification is suggested for a better understanding of Big Data platforms. The hardware platform is, for some reason, often ignored or insufficiently discussed in the literature on Big Data. Each of these sub-layers is explained in more detail in the sections below.

3.3.4.1.1 Hardware

Big Data analytics is often designed to run on high performance commodity hardware characterized by availability, interoperability, and low cost (Fujitsu, 2014; Hortonworks, 2014). Some parameters that need to be considered while developing a hardware platform for Big Data are: processing power, storage capacity, and networking bandwidth. All these are essential for handling large data sets. Other non-data parameters that should be taken into account include power, cooling, space, and weight of components. The following section provides an explanation on the requirements for an efficient and effective Big Data platform.

3.3.4.1.1.1 Processing Power

Using high performance processors is important for building an effective platform for Big Data. The processors required for Big Data platform need to meet the challenges of the “3 V’s”: volume, velocity, and variety. They should be capable of processing massive volumes of data at a high speed. Also, they need to be protected against service interruption or failure (Vasudeva, 2012;

Davenport & Dyché, 2013; Hortonworks, 2014; Guide, 2013; Sun & Heller, 2012).

Big Data analysis often makes use of the distributed processing architecture which allows distributing data across large number of commodity servers to process data in parallel. The commodity servers are usually classified into two categories: (1) slave nodes and (2) master nodes. The master nodes are used to manage data, while the salve nodes are used to process data. This type of architecture is scalable: it allows adding additional nodes to increase the processing power if the processing load is growing. For Big Data platforms it is recommended to use cost-effective, medium clock speed processors for slave nodes and avoid idle capacity in the nodes. To improve the processing power, it is suggested to provide sufficient memory with the option of adding more memory in future. But for master nodes it is recommended to use higher clock speed processors for handling messaging traffic that tends to grow exponentially with the increased processing load (Hortonworks, 2014).

3.3.4.1.1.2 Storage Capacity

The storage for Big Data platform should possess capability to store very large and complex data sources. Therefore, storage capacity, availability, performance, and integrity need to be carefully observed as they play very important roles in Big Data networks. In fact, the performance of the storage system is extremely important. High availability of a storage system is essential to avoid downtime or service interruption that can jeopardize data availability.

Data integrity is also an important factor. Big Data storage should support automatic data replication across multiple nodes to eliminate the need for backup. Furthermore, the storage needs to support advanced capabilities, such as compression and automated filtering of data. Moreover, advanced security features such as data encryption and access control are also needed (Vasudeva, 2012; Davenport & Dyché, 2013; Fujitsu, 2014; Hortonworks, 2014; Guide, 2013; Sun & Heller, 2012).

Different types of storage can be used in Big Data architecture, such as Storage Area Networks (SANs), Network-Attached Storage (NAS), or Direct-Attached Storage (DAS). Some organizations prefer direct-attached storage (DAS) for Big Data architectures over the external network storage (NAS or SAN). DAS has direct links to its nodes and does not use the organization's existing network. Because of that, it is faster, and, in many ways, simpler to setup. However, there are few disadvantages of DAS. First, it cannot be managed via the existing network. Moreover, DAS requires dedicated servers which are often underutilized. Finally, DAS configuration does not provide economies of scale via sharing the storage with other applications (Fujitsu, 2014; Hortonworks, 2014).

3.3.4.1.1.3 Network Bandwidth

Usually, Big Data networks include many servers and storage systems connected via Ethernet or Fibre networks. The performance and efficiency of the network depends mainly on two factors: network bandwidth and availability. Higher network bandwidth improves the efficiency of data exchange across the network and allows servers to communicate with each other faster. For Big Data, the network should be extremely fast, efficient, reliable, and scalable to cope with fast data growth and the resulting network expansion. In addition to that, network adapters and connections should support the high throughput and bandwidth associated with high volumes of data moving across the servers. In order to maintain the network reliability and performance stability, it is recommended to use tools for monitoring the network utilization and stability. Furthermore, the network availability needs to be protected against interruption by connecting it in a redundant and secure way (Fujitsu, 2014; Hortonworks, 2014).

3.3.4.1.1.4 Other Components (Power, Cooling, Weight and Space)

The power consumption and cooling requirements are a major concern when designing large networks for Big Data. To effectively utilize power and cooling systems, it is important to calculate the actual hardware requirements rather than going for the maximum loads design. The weight and space of Big Data systems and racks are also important factors that need to be considered when designing the Big Data platform. The weight and space of a racks and systems need to be within the capacity of the floor and it is recommend to reserve sufficient space and weight capacity to accommodate more racks in future (Hortonworks, 2014).

3.3.4.1.2 Software

Due to increasing amount of Web data and Internet activities after year 2000, Google and other Internet-based organization encountered difficulties in indexing their digital content. As a part of a solution to this issue, Google Labs published a paper in December 2004 on a new algorithm called MapReduce. The MapReduce was defined as a model used for parallel processing of large data sets across large number of commodity computers (nodes). The invented algorithm helped Google to rewrite search index system and analyze the Web content and user behaviour in a more efficient, reliable, faster, and cheaper way (Dean & Ghemawat, 2008; Vance, 2009; Rijmenam, 2014).

The MapReduce algorithm attracted Doug Cutting, who decided to create his own version of the technology that would allow applications based on the MapReduce algorithm to run on commodity hardware. The technology was named Hadoop (after Cutting son's toy elephant). Hadoop is now considered a synonym for Big Data. The Hadoop was initially adopted by Yahoo and later by a few companies which started their business around it. In 2008, Hadoop was released as an open-source tool. Today Hadoop is managed by the non-profit foundation called Apache Software Foundation (ASF).

Hadoop is emerging as one of the top innovative technologies of Big Data. It has drawn extensive attention from the Big Data community due to its features and advantages. Hadoop is synonymous with Big Data nowadays. It is the most widely used product in this field. Without Hadoop the Big Data movements would have taken much longer time to reach the masses (Collett, 2011; Guide, 2013; Vance, 2009; Rijmenam, 2014). Some of the main features or components of Hadoop are as follows:

- MapReduce: a programming model for processing large sets of data in parallel
- HDFS: a distributed file system that can store all kinds of data without prior organization
- YARN: a resource management framework
- Pig: a high-level data-flow language and execution framework for parallel computing
- Hive: a data warehousing tool that provides data summarization and ad hoc querying features
- HBase: a distributed database that supports structured data storage for large tables
- Zookeeper: an application that coordinates distributed processes
- Cassandra: a multi-master database with no single points of failure
- Ambari: a Web-based interface for managing Hadoop services and components
- Flume: a software that collects, aggregates and moves data into HDFS.
- Sqoop: a connection tool that moves data between Hadoop and relational databases

Despite the drawbacks of open source tools (such as lack of documentation, training, and support), Hadoop has gained increasing attention for the advantages and features it has. Some advantage of Hadoop is that it is free and can be easily downloaded from the Internet. Hadoop is continuously developed and improved by the community that was born around it. In addition to that, Hadoop has a scalable (i.e. new nodes can be added as needed), flexible (i.e. can absorb any type of data from any source), cost effective (i.e. using commodity computers) and reliable (i.e. the system redirects work to another node when any node fails). architecture These features of Hadoop make it cheap, efficient, and reliable when compared to proprietary Big Data products (Davenport & Dyché, 2013; Rijmenam, 2014).

Besides the open source version of Hadoop (e.g. Apache Hadoop), several commercial distributions of Hadoop are available in the marketplace. Moreover, a number of organizations have created their own Hadoop distributions. With Hadoop distributions from other software vendors, customers receive additional software components, tools, training, documentation, and other services. The most recognized enterprise versions include Cloudera, MapR, and Hortonworks. In addition to the mentioned distributions, IBM and EMC have introduced their own Hadoop distribution. Such companies as Microsoft, Teradata, and Oracle provide support for other Big Data platforms like Cloudera and Hortonworks (Henschen, 2015).

3.3.4.2 Data Analytics Layer

The data analytics nowadays is more sophisticated smarter and relatively more affordable than ever before due to a substantial drop in the cost of data storage and processing power. Moreover, the recent developments in analytic tools lifted them into new level of performance and accessibility. The advancements in analytical tools, especially the ones suited for Big Data, made the results of analytics more accurate (since bigger datasets tend to produce better, more accurate results).

In general, data analytics is typically defined as knowledge discovery from data. There are different techniques and tools used for managing and analyzing data. However, due to technology challenges and complexity, a small number of organizations use data analytics tools to enhance organizational performance. Even with those who use data analytics tools, not all of them are successful in deriving benefits from data. There are certain sectors and types of businesses that gain benefits from data analytics more often than others, such as healthcare, transport, merchandising, services and utilities (Ahuja & Moore, 2013; Courtney, 2012).

Because Big Data contains messy, unstructured, and often low quality data, the analytics tools and techniques need to be "fault tolerant". The data quality processes used in BI, for example, are not suitable for Big Data because they create the risk of stripping out valuable data from unstructured data. So, attempts to replicate and scale the traditional BI technologies for Big Data are often not practical (Davenport, Barth, & Bean, 2012; Ferguson, 2012; Russom, 2011).

3.3.4.2.1 Types of Data Analytics

Data analytics methods and techniques can be broken down into three categories: statistical analysis, data mining, and machine learning. Statistical analysis uses various models and methods to summarize data and extract predictions from it. Some of the well-known statistical methods include: ANOVA, t-test, regression analysis, correlation, chi-squared test, and factor analysis. For example, ANOVA (Analysis of Variance) is used when we compare the means of multiple samples (one can use the t-test to compare the means of two samples only). An example of ANOVA is to test the effect of three kind of food on 15 persons. First, we divide them into three groups. Each group of five persons will be assigned one kind of food. The changes in their weight will be recorded after a certain period. Then we will discover whether the effect of food on weight is significantly different by comparing the mean weight within each of the three groups.

Data mining uses statistical analysis and a variety of other techniques (such as clustering, classification, regression, etc.) to discover patterns in data. Data mining is also called "Knowledge Discovery in Databases" or KDD. The main objective of data mining is to extract information from data and use that information for something useful. For example, Bank X process thousands of transactions daily for its customers. The transactions data is collected and stored in a central database. The bank can analyze this data to understand better their customer behaviour. Based on this improved understanding, the bank can develop new marketing strategies and techniques. This can help the bank to improve its business performance and enhance customer loyalty.

Machine Learning is another popular analytical technique. Machine Learning usually refers to a set of algorithms that allow computers to learn from data and make decisions based on this data (Ahuja & Moore, 2013). Machine learning is more into prediction and learning from experience rather than following a pre-programmed analytical or decision heuristic (Simon, 2013). For example, machine learning can be applied for traffic light at a busy intersection. The traffic light may be operated based on past traffic patterns. Another example of machine learning is prediction of whether it will rain tomorrow or not based on meteorological data.

Overall, most analytical methods fall into the following categories: descriptive, predictive, and prescriptive. Descriptive statistics is one of oldest methods used to explain how an organization performs based on historical and transactional data. In the predictive analysis, organizations use the historical data to predict the future using some advanced techniques and tools. The prescriptive analytics uses models to identify ideal actions or decisions that should be taken to get best results. The prescriptive analytics have received increased emphasis in recent times. The prescriptive analytics uses Big Data not only to tell what might happen but why it will happen. Consequently, prescriptive analytics focuses on actions and decisions that should be taken in the future. This type of analytics usually utilizes mathematics, business rule algorithms, machine learning, and computational modelling techniques. Google self-driving car is a good example of using prescriptive analytics. The car takes decisions based on numerous data points, calculations, and predictions of future outcomes that frequently change. The car should make correct decision in proper time to avoid an accident.

Many of the types of analytics discussed here are used together. When it comes to addressing important business problems, different types of analytics should be implemented to obtain a complete understanding of what happened, what might happen, and what should be done (Davenport & Dyché, 2013; Rijmenam, 2014).

3.3.4.2.2 Analytical Tools and Applications

Which software application is used for analytics depends on several factors. The first factor is the availability of the right skills among Data Scientists or Data Analysts to use an application. Then, the organization should identify whether a suitable "off the shelf" software application is available or they need to develop a customized solution for their business. Finally, having a single or few analytics applications is preferred over big number of software tools from different vendors, as this minimizes the complexity and maintenance costs (Davenport & Harris, 2007)

Typical analytical applications include tools for statistical analysis, data mining, simulation, and optimization. Such traditional software tools as SAS, SPSS, R, and Matlab support sophisticated analysis methods, but were originally designed to work with BI era scenarios. On the other hand, tools which use large scale processing platforms such as Hadoop/MapReduce and NoSQL often do not support advanced forms of analysis. Therefore, there is a necessity to bridge the gap between Big Data platforms and analytical tools by extending the features of Big Data platforms to support sophisticated analysis. Alternatively, software vendors of proprietary BI tools often develop extensions that enable these tools to be used in for analysing Big Data (Madden, 2012).

3.3.4.3 Data Visualization Layer

Visualizing large, unstructured datasets is an important Big Data capability. Different visualization tools have been developed to deal with unstructured data. They use different techniques to present unstructured data. For example, some of the techniques recognize patterns and discover abnormalities in unstructured data based on the shape, colour, or thicknesses detected in data visualizations (Russom, 2011; Rijmenam, 2014).

Data visualization tools act as an interface which users normally use to get results from sophisticated data analysis methods. The data visualizations should present the information in a simple, graphical display which allows users to drill down further if more details are needed. For example, some advanced data visualization tools are interactive: they change according to the data and analysis type to present the most suitable and simple graphic display. The graphic display helps to simplify information and analysis. Sometimes a simple graph offers more valuable information than advanced data analysis and visualization tools. Therefore, many data visualization tools have been developed with simplicity in mind (Davenport & Dyché, 2013; Davenport & Harris, 2007; Rijmenam, 2014).

3.3.5 Organizational Capability

Organizational capabilities in relation to Big Data consist of several dimensions that are important to enhance and encourage the Big Data use within the organization. These dimensions include: strategy, process, structure, and culture. These dimensions are assumed to be the main pillars for building organizational capability in relation to Big Data.

3.3.5.1 Strategy

Strategy comprises setting goals and determining what resources and actions will be required to achieve these goals (Mintzberg & Quinn, 1992). Strategy can be considered a high level plan or a vision which senior management adopts for their organization. Having a clear and well-defined strategy for the transformation from the current BI capabilities towards Big Data capabilities is important for a successful implementation of Big Data initiatives in organizations. Strategy has two major stages: development and implementation. Strategy development (or formulation) for Big Data requires a broad knowledge about solutions available in the market for Big Data implementation. Development or formulation should involve identifying activities, listing the necessary resources, and developing schedules for implementing Big Data initiatives. Strategy implementation involves allocation and mobilization of resources, and execution of activities planned within the allocated budget and time frame. Using proper tools and methodologies for development and implementation of Big Data strategy can contribute to improvement in Big Data capabilities (Davenport, et al., 2001).

3.3.5.2 Processes

Processes are collections of specific activities, procedures, and methods that organizations employ to manage their important actions and decisions. Processes can be described in the form of a flowchart containing various activities and decision points. Development of comprehensive and consistent processes for Big Data is important for successful implementation of a Big Data strategy. Big Data processes need to be developed specifically to fit the organizational information systems and culture.

Processes can be classified into different groups according to the level of organizational structure and the specific functional area that they belong to. Some of the most common processes which can be developed for Big Data are: procurements process, development process, and decision making process. The procurement of Big Data technology and resources may include several stages to help organizations benchmark their procurement activity against best practice references. These stages may comprise needs analysis, finance approval, procurement plan, supplier selection, contract management, and procurement assessment.

The development process in the context of Big Data might employ the Systems Development Life Cycle (SDLC), which is used to describe processes for planning, creating, testing, and deploying information systems (Blanchard & Fabrycky, 2006). Consistent with SDLC, the main steps involved in the Big Data systems development life cycle are: requirement analysis, systems design, system development, integration and testing, implementation, operation, and maintenance. During the operation and maintenance process, the system should be evaluated to ensure it is continuously updated and properly maintained.

The decision making process in relation to Big Data should rely on the classic steps of decision making, such as collecting information about a problem area, identify alternatives, specify criteria for selecting the best option, choosing among the alternatives, and implementing or making a decision. Decision making process in Big Data organizations can be automated, at least partially, using modern Decision Support Systems (DSS).

Well-designed processes are important for successful implementation and use of Big Data. Alternatively, lack of well-defined and rigorously enforced process in relation to Big Data would make it hard for organizations to achieve tangible performance results with the help of Big Data initiatives (Davenport, et al., 2001).

3.3.5.3 Structure

Organization structure is often defined in terms of the management levels (e.g. starting with CEO position down to field managers and supervisors) in an organization and the links among those levels involving the lines of authority and accountability. For organizations developing Big Data capabilities, organizational structure should include people accountable for Big Data analytics. In the traditional structure, organizations keep the data analytics teams in IT departments. But, it is recommended to have Big Data activities separated from the IT department so that the teams responsible for Big Data have enough authority for decision making and innovation within the organization (Davenport & Dyché, 2013; Pearson & Wegener, 2013).

Such Big Data leaders as Amazon and LinkedIn have created separate entities within their organization to oversee Big Data analytics. These entities are often headed by a Chief Analytics Officer (CAO). Other organizations have created the Chief Data Officer (CDO) position to be the primary "owner" of Big Data (Pearson & Wegener, 2013; Rijmenam, 2014). The creation of such separate Big Data entities (e.g. departments, sections, or units) within an organization furnished with the necessary resources and authority can be helpful for building and supporting Big Data capabilities across the organization.

3.3.5.4 Culture

Organizational culture is a collection of expectations, experiences, and values within an organization that influence people thinking, observations, assumptions, understanding, and decision making in different situations (Needle, 2010; Ravasi & Schultz, 2006). Organizational culture may be reflected in such artifacts as vision, mission, beliefs, information systems, language, etc. The culture which is built on flexibility, continuous improvement, and acceptance of change is more likely to be conducive of building Big Data capabilities in an organization (Needle, 2010; Ravasi.& Schultz, 2006).

In a survey conducted by Tata Consultancy Services (TCS, 2013) among 1,217 companies in nine countries covering the US, Europe, Asia-Pacific and Latin America, it was found that the highest rated challenges for implementing Big Data initiatives were related to cultural and not technological issues. Cultural issues restrict many organizations from reaping the benefits of Big Data. Management support and encouragement for Big Data implementation is an important element for removing the cultural barriers towards the Big Data. Organizations with an organizational culture supportive of data analytics are expected to have more acceptance and easier change to Big Data than organizations with neutral or negative attitude towards data analytics (Davenport, et al., 2001).

3.4 Big Data Maturity Model

In the light of the Big Data capabilities discussed earlier, this section proposes a Big Data Maturity Model (BDMM) that can be used to assess organizational maturity in relation to Big Data analytics and record the different stages of maturity organizations reach as their Big Data capabilities improve over time. As explained previously, Big Data capabilities are classified into the following three areas: people, organization, and technology. Each capability is represented by a set of processes or behaviors performed with certain levels of regularity and rigor. The organization performs behaviour regularly if it follows a strict schedule. It does not matter how frequently as long as it adheres to the schedule (Piccoli, Carroll, & Hall, 2011). And organization performs the behaviour rigorously if it follows formal processes or procedures rather than performing the activity as it see fit (Piccoli, et al., 2011). The regularity and rigor of these behaviors discriminates among the organizations at different levels of Big Data maturity. As an organization performs these behaviors more regularly and rigorously, its maturity increases. Thus, the proposed BDMM assumes that the organizational capabilities can improve overtime through conscious continuous improvement.

As can be seen in Figure 3-2, Big Data Maturity improves via the following stages: Unaware, Data Processing (DP), Business Intelligence (BI), and Big Data Analytics. Each of these levels is discussed in more detail in the sections below.

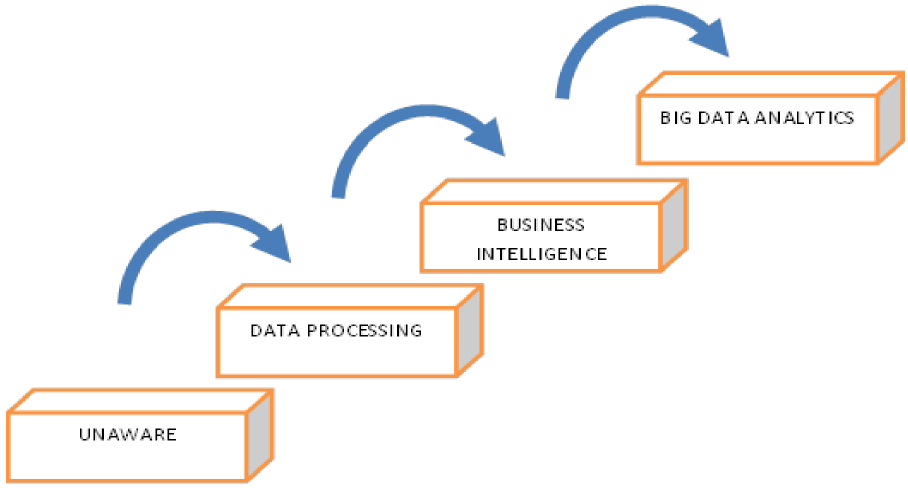

Figure 3-2: Big Data Maturity Model Levels

3.4.1 Unaware

This is a “residual level” for “catching” all organizations that do not fall into a particular stage of data analytics. This is the stage where an organization has no systematic organizational, technological, or people capabilities in relation to data analytics. There may be individual analytics initiatives and breakthroughs, but on an “ad-hoc” basis and without no consistent results.

3.4.2 Data Processing

The key characteristic of Data Processing level is that the organization in this stage has developed basic, primitive types of processes and procedures in the area of data analytics. The focus is mainly on the use of desktop computers by individuals to analyze data at rest using basic tools and statistical techniques (e.g. descriptive) s to provide a rather historical (meaning looking into the past) understanding of business performance. The organization in this level starts recognizing Key Performance Indicators (KPIs) and uses them to produce dashboards so that business performance information is clearly presented. However, inconsistencies in metrics, practices, and overall goals among various business units or departments are very common. The organization in this level starts to invest into BI, but users are often not skilled enough in the field of data analytics in order to benefit from BI systems.

3.4.3 Business Intelligence

The key characteristic of the Business Intelligence level is that organizational processes and procedures in relation to data analytics are well organized and consistently performed. The focus in this level is shifted to enhance BI competency & usage of existing BI assets. The Chief Information Officer (CIO) drives organization strategy for data analytics and BI development. Experts from business and IT are normally joined together in order to fulfil the needs of business users. Information is available to all employees of the company and the usage of BI is usually extended to suppliers, business partners, and, sometimes, to customers. Data analytics in this level involves applying machine learning and predictive modelling tools run on high performance computing servers or mainframes for providing insights on future performance. Data analytics is widely used in respected within an organization. It is used for exploiting opportunities, competing with other organizations, and for strategic decision-making. Users in this level are trained for data analytics and are able to use their skills effectively for strategic and operational decisions.

3.4.4 Big Data Analytics

The key characteristic of the Big Data Analytics level is that the organization systematically and constantly involves in quantitative and qualitative data gathering and analytical behaviour. At this level, data analytics is integrated across all areas of the business and becomes an essential part of the corporate culture. The focus in this level is shifted to enhancing Big Data competencies for analyzing unstructured data in real time. The organization in this level institutionalizes the process of continuous learning through testing and experimentation and continuously improves its Big Data processes based on formal, quantitative feedback obtained. The organization automates decision making process to take immediate action usually guided by automated KPIs related to business performance, resilience, and security. Scenarios and simulation tools are used to let Data Analysts explore future alternatives and identify both positive and negative impacts of decisions to be made in one part of the organization. Users are well trained in analytics and have access to data, information, and tools necessary for improving business performance with analytics. Moreover, the organization collaborates with universities to stay abreast with the latest developments in data analytics and influence data analytics curriculum.

Table 3-2 provides a summary of BDMM levels and preliminary list of behaviors.

Table 3-2. Overview of BDMM Levels and Behaviours

Level	Behaviour
Unaware	This is a residual level where organizations have no systematic competency in relation to data analytics. There may be individual data analytics initiatives, but on an ad-hoce basis and with no consistent results.
Data Processing	At this level organizations have started developing formal processes and procedures in the area of data analytics. The organizations still depend on individual's competencies to analyze data. They use basic tools and statistics to provide historical view of business performance. Inconsistency in metrics or goals between business units or departments is very common.
Business Intelligence	At this level organizational processes and procedures are well organized and consistently performed in relation to data analytics. The CIO in this level drives organizational strategy for data analytics and BI development. Machine learning and predictive modelling tools used constantly to provide insight on future performance.
Big Data	At this level an organization systematically and constantly evolves and improves in quantitative and qualitative data analysis and related processes. Data analytics is integrated across all areas of the business and become essential part of the corporate culture. Big Data assets and tools (such as Hadoop and NoSQL) are used to manage and analyze unstructured and real time data. Decision making process is automated and usually guided by automated KPIs. Scenarios and simulation tools are used by Data Analysts on the top of traditional analysis techniques.

3.5 Implementation of Big Data Maturity Model

For evaluating individual behaviors or processes within BDMM, a two dimensional sliding scale is used for capturing (1) regularity and (2) rigor with which each behaviour is performed. Each scale has four categories with specific cut-off points: Unaware, Low (L), Medium (M) and High (H). BDMM four levels are presented on the scale with the following assumptions: Unaware level exists between 0 and L on regularity scale and on rigor scale; the Data Processing stage involves a low regularity and low rigor (L-L); The Business intelligence stage involves medium regularity and medium rigor (M-M) of behaviors; the Big Data stage involves a high level of regularity and rigor (H-H) with which these behaviors are performed. These categories are reflected in Figure 3-3.

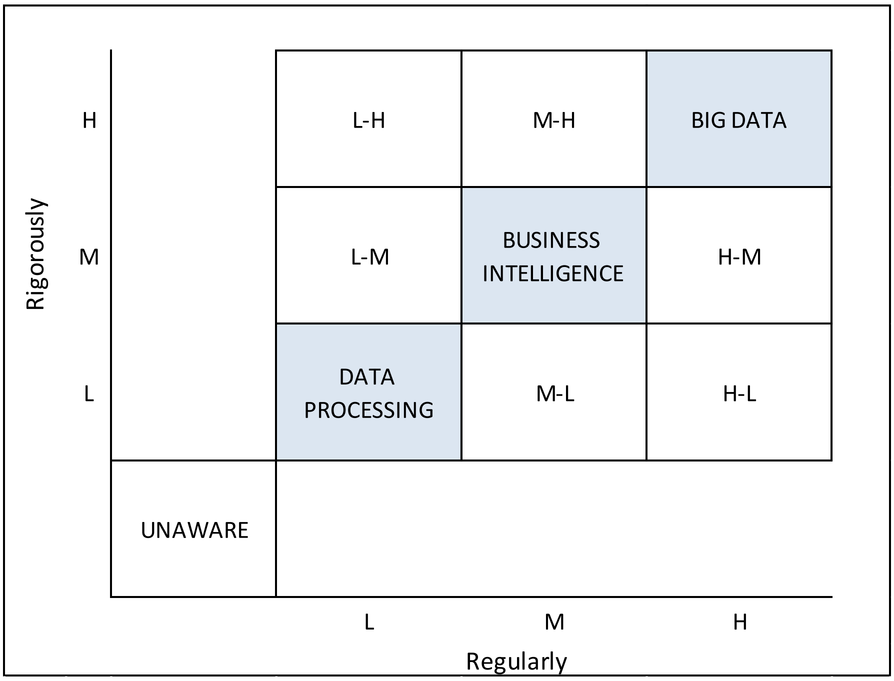

Figure 3-3. Big Data Maturity Model (BDMM)

CHAPTER 4

DATA AND METHODOLOGY

Chapter 4 presents the research methodology used in this study. The purpose of the study is reiterated first to provide a better understanding of how the BDMM was operationalized. Then the quantitative approach employed by this study is briefly discussed. The process of behaviour generation and validation is outlined as an input to the subsequent data collection with the help of a quantitative, online survey. It is also explained how the selection of participants was done in line with the research objective to ensure that all respondents have knowledge and experience related to Big Data analytics. An overview of the analytical techniques applied to the collected data is presented at the end of this chapter.

4.1 Study Purpose

The study utilizes the previously developed BDC framework as conceptual lens for operationalizing Big Data maturity via the BDMM. In line with the socio-technical view, BDC is comprised of people, organizational, and technological capabilities (see Figure 4-1)

Figure 4-1: Big Data Capability (BDC) Framework

A capability is defined as an organizational capacity to perform certain actions. Behaviours represent tangible representation of this capacity (Piccoli, et al., 2011). Thus, this research assesses organization behaviours in the following areas:

- People Behaviours (PB)
- Organizational Behaviours (OB)
- Technological Behaviours (TB)

Organizations are expected to perform behaviours within the broad areas listed above with certain levels of regularity and rigor, depending on their level of Big Data maturity. As previously explained, mature organizations are expected to perform behaviours in all three areas with high level of regularity and rigor, while organizations with low maturity will perform behaviours with low level of regularity and rigor (i.e. in an ad-hoc or informal manner). Therefore, organizational maturity levels can be measured by confirming how regularly and rigorously it performs a predefined set of behaviours.

4.2 Study Approach

This study followed a quantitative approach using a survey to numerically record and describe behaviours of participants in the three capability areas of Big Data: people, organization, and technology. The reason for the survey-based, quantitative approach is that it is highly structured and provides a direct and objective way of obtaining information in relation to Big Data It also enables a broad, exploratory approach as it allows comparison of a diverse set of behaviours among different organizations on a uniform, numerical scale. Accordingly, information is collected about the same behaviours from many organizations across various industries and countries. The survey questions are mostly designed to explore various levels of maturity with respect to Big Data. The survey is prepared in an online format and designed to be anonymous to protect the identity of participants. The data collected does not include personal information.

4.3 Behaviours Generation

In our effort to operationalize and validate the proposed BDMM, a set of behaviours are developed to represent different levels of maturity with respect to Big Data. These behaviours are developed based on the literature review and reflect the adopted theoretical stance involving the STP and RBV. Initially, a list of 84 behaviours was developed across the three capability areas: people, organisation, and technology (see Table 4-1). This initial list of behaviours was sent via email to three experts working in the fields of

Information Systems and Data Analytics for them to discuss and review the list.

Table 4-1. Initially Identified Behaviours

	People	Organization	Technology	TOTAL
Behaviours	30	24	30	84

The experts were requested to assess appropriateness and relevance of behaviours in each of the three capability areas and to review the wording of questions representing the behaviours. Great attention was devoted to question wording to make them easier to understand and answer. The result of the review and validation after several iterations was 45 validated behaviours broken down into the three categories as follows: (6), organizational (16), and technological (23) (see Table 4-2).

Table 4-2. Validated Behaviours

	People	Organization	Technology	TOTAL
Behaviours	6	16	23	45

The dropped behaviours are potentially useful for understanding Big Data maturity with more granularity. However, it is very important to ensure the survey is succinct so that the time required to fill in the survey does not affect the participants' willingness to complete it as this may impact the quality of collected data. A lengthy survey could lead to a situation where survey participants lose concentration while completing it. Based on the expert review the 45 behaviours were selected for the survey (see Appendix A for the full list).

4.4 Survey Design

The survey sign was designed through several iterations and was guided mainly by the approach to survey design in Piccoli, et al. (2011). Three draft surveys were created before the survey was finalized.

The final survey includes two parts. The first part is for capturing demographic information, such as position of participants within the organization, location of organization's headquarters, size of organization, year organization founded, industry organization belongs to, and whether the organization conducts business online or not. The first page of this part contains the required informed consent statements (see Appendix B). Responses were received only from participants who voluntarily agreed to participate. In addition to that, "disqualification logic" logic was added to the survey to exclude participants whose title did not match any of the following titles: CIO, CTO, CDO, or Director Data Science (DDS).

The second part of the survey starts out by giving information and instructions on completing the remaining part of the survey (see Appendix C). This part is divided into three sections: People Behaviours (PB), Organizational Behaviours (OB), and Technology Behaviours (TB). For each question, respondents were asked to report the degree of regularity and rigor with which their organization performs the behaviours. A continuous scale ranging from 0 to 100 is used for measuring regularity and rigor, with such anchors as "not at all", "somewhat", and "extremely" (see Figure 4-2).

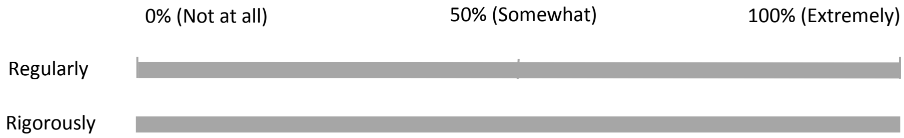

Figure 4-2. Scale of Regularity and Rigor

“Extremely regularly” indicates the organization follows a strict schedule. It does not matter how frequently. If the organization does not perform a particular activity, the response should be 0% (not at all). “Extremely rigorously” means the organization follows formal processes or procedures rather than performing the activity as it sees fit. If activity is performed on a completely “ad hoc” basis while formal policies and procedures (if any) in relation to this activity tend to be ignored, then the answer should be 0% (“not at all”). However, even if respondent intended to select 0% for a behaviour, he or she still needed to use a slider to select 0%. Otherwise the system will prompt that the question was not answered. This was done to ensure that none of the questions were omitted and the level of 0% was not selected unintentionally.

To reduce response error and identify question ambiguity, a pilot survey was conducted with 10 persons: two research advisors and eight work colleagues. The pilot provided valuable feedback on how the survey was perceived and understood by the pilot participants. This resulted in better explanation of informed consent statements (see Appendix B) and some modifications to the instructions page (see Appendix C). In addition, the pilot resulted in layout changes that drove part 2 to be divided into three sections for people, organization, and technology (as opposed to having all questions in one section).

4.5 Participants

The selection of participants was influenced by the requirements of this research. The criterion for selecting participants was developed in line with the research objectives and to ensure that all respondents have reasonable knowledge and experience in Big Data analytics. The study employed a convenience sample provided by a well-known research panel company. There was only one selection criterion - the participant must hold one of the following titles: Chief Information Officer (CIO), Chief Technology Officer (CTO), Chief Data Officer (CDO), or Senior Director (including Director for Data Science). It is assumed that people in these positions have a holistic view on data analytics within their organizations.

4.6 Data Collection

Survey was prepared using an online survey service: Survey-Gizmo. A well-known third-party research panel supplier was enlisted to distribute the survey online and collect the data. The project manager from the panel research supplier was cooperative when asked to distribute and collect the survey twice based on refinement of the survey.

A total of 528 responses were received. The completed surveys which meet the accepted selection criterion (title of participant) were 246. The completed forms which failed to meet the accepted selection criteria were 119. All are excluded from the data analysis. In addition, 163 incomplete forms were excluded. A small number of completed forms did have missing values due to the survey length and complexity, however have been accepted as is. An overview of survey responses is given in Table 4-3.

Table 4-3. Total Number of Participants

Status	Total
Complete	246
Not meeting Criteria	119
Disqualified (Incomplete)	163
Grand Total	528

CDO	11
CIO	82
CTO	42
Director Data science	111
Grand Total	246

As it was mentioned previously, the data collected does not contain personal information that can be used to identify survey respondents and their organizations. A clearance from the university's Institutional Review Board (IRB) and the Human Subjects Committee was received to use the survey for data collection.

4.7 Data Analysis

Data collected from the final set of completed surveys was coded and entered into SPSS for statistical analysis. Later, Microsoft Excel was used for additional statistical analysis. The following types of analysis were performed:

- Analysing the demographics of survey respondents by providing descriptive statistics for each of the major demographic categories.. Results of demographic questions analysis are used at a later stage to explore the relationship between certain behaviours and specific demographic characteristics of participants and their organizations. Analysis of demographic questions also provides information on any biases or limitations in the sample: it helps to see whether the collected sample is representative of the intended population or not.

- Analysing responses to questions from Part 2 using descriptive statistics. Descriptive statistics have been applied for each question and for each category (people, organization, and technology). The descriptive statistical analysis has been applied separately for both regularity and rigor scale. This type of analysis provides a base for comparison of averages between different behaviours and between groups of behaviours belonging to one area. The boxplot was used to graphically illustrate numerical data through quartiles and for comparing averages and variance between different data grouping. In addition to that, reliability of scales was assessed with the help of Cronbach's Alpha. The Cronbach's Alpha is a measure of internal consistency or coefficient of reliability. A reliability coefficient of .70 or higher is considered acceptable in most social science research situations, proposing that the items have relatively high internal consistency.
- Applying Clustering Analysis (CA) to discover clusters or structures in data. CA is an exploratory data analysis tool. It helps to categorize different cases into clusters in a way that the degree of relationship between two cases in one cluster is maximized. In other words, clustering is utilized for combining cases of similar type into separate categories. Cluster analysis simply discovers structures in data without explaining why these structures exist. This is suitable given the exploratory nature of

this study. Since the model is hierarchical in nature, a hierarchical clustering approach was applied to explore the relationship between behaviours and the organization maturity level. Under a hierarchical model, an organization is required to enhance behaviours in one level before it can move to a greater level of maturity. In other words, organizational maturity changes from one level to the next when the organization exhibits an advanced degree of behaviours in the current level. Ward’s clustering method was used to combine cases having similar responses. First, this involves calculating the mean value for each response in the three capability areas: people, organization, and technology. Then clustering analysis is applied to all responses. The result of clustering is depicted in the form of a dendrogram using the so-called Ward Linkage.

CHAPTER 5

SURVEY RESULTS

The main purpose of this chapter is to empirically explore organizational capabilities and maturity in relation to Big Data. This is achieved with the help of the following steps. First, the main research objectives and an overview of the methodology are provided to enhance understanding of the results among the readers. Second, an overview of survey results using descriptive statistics and clustering output is used to analyze the responses. Consistency and reliability of survey items are also provided. Third, findings of this analysis are related to the proposed BDMM.

5.1 Survey Objective

The survey is used mainly to explore the capabilities of organizations in relation to Big Data. These capabilities are operationalized as specific behaviours. These behaviours capture tangible, real-world representations of organizational Big Data capabilities. Organizations are expected to perform Big Data behaviours in relation to people (PB), the organization as a whole (OB), and also behaviours in relation to technology (TB). Each behaviour performed with a certain level of regularity and rigor - depending on the level of organizational maturity in relation to Big Data. Mature organizations perform all three of these behaviours with high level of regularity and rigor. Low maturity organizations perform these behaviours at a low level of regularity and rigor (i.e. in an ad-hoc or informal manner). Therefore, the level of maturity of an organization can be measured by assessing how regularly and rigorously organizations perform relevant behaviours across the three general dimensions listed above.

5.2 Part 1: Profile of Respondents and Organizations

The first part of the survey consists of six questions designed to collect details about the survey participants and their organizations. Results are provided below for each question.

Q1: In what region is your organization currently headquartered?

The sample consists of 246 participants representing different organizations located in various countries. Majority or 88% of organizations are headquartered in the US or Canada. The remaining organizations are distributed as follows: 7% in the Middle East, 2% in Latin America, 2% in the Caribbean, 2% in Europe, and 1% in Africa. Figure 5-1 depicts the geographic distribution of organizations which participated in this study.

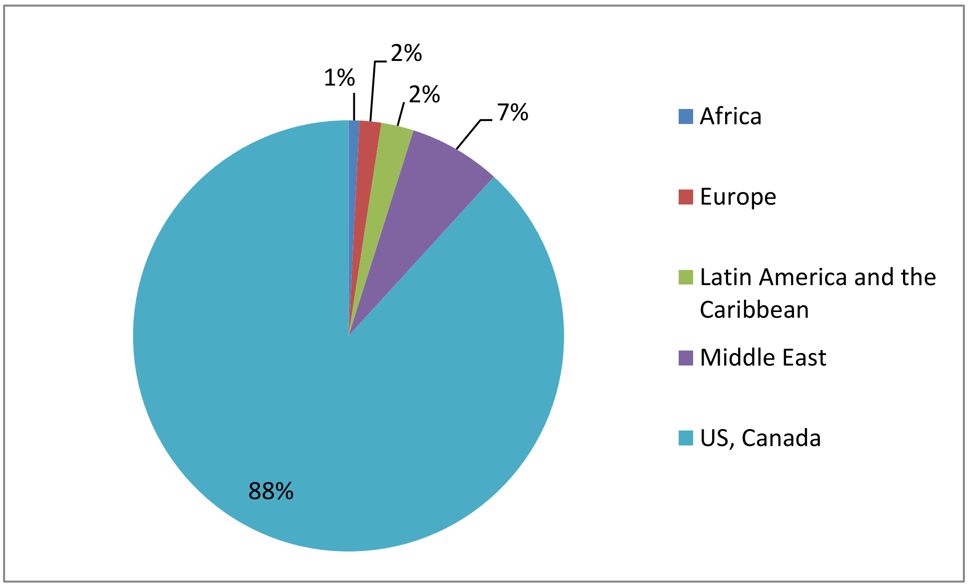

Figure 5-1. In what region is your organization currently headquartered

Q2: What is your title/position within the organization?

Participants are limited to people holding one of the following titles: CTO, CIO, CDO or DSS as it was decided that other participants may not possess adequate information in relation to Big Data behaviours within their respective organizations. 45% of participants identify themselves as DDS and 33% as CIO. Remaining participants identify themselves as CTO and CDO (17% and 5% respectively).

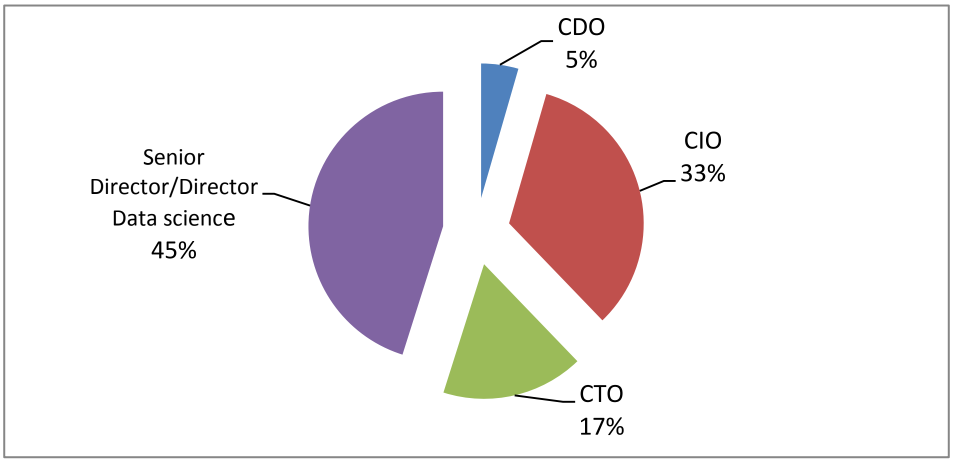

Figure 5-2. What is your title/position within the organization?

Q3: Please indicate the size of your organization?

The study included both small (<50 employees) and large (>2500 employees) organizations. Other factors, such as revenues, were not used as size variables.

Figure 5-3 indicates that organizations of 250 – 1000 staff represent 35% of the organizations surveyed. Larger organizations (1000 – 2500 and 2500+ employees) represent 25% and 20% respectively. Small (< 50 employees) and very small (<10 employees) organisations combined represent <4%. Figure 5-4 shows the count of organizations.

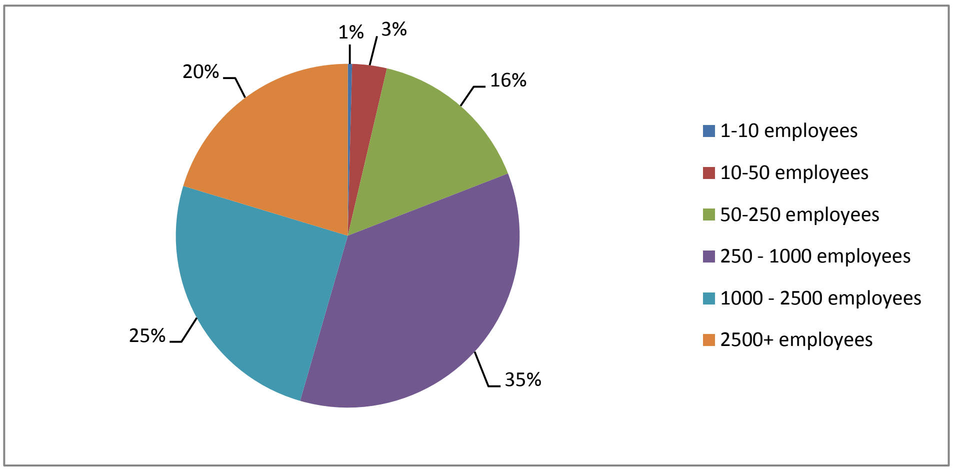

Figure 5-3. Please indicate the size of your organization?

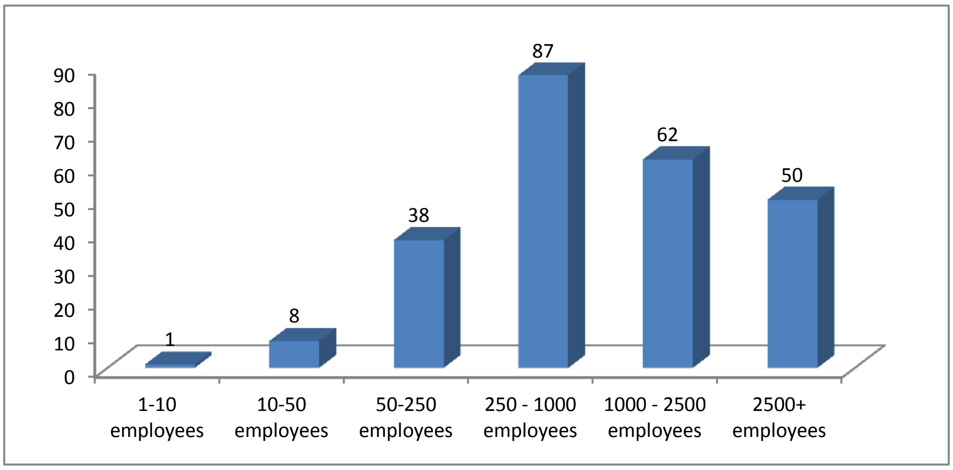

Figure 5-4 Count of Organizations in each Segment

Q4: In what year was your organization founded?

Figure 5-5 shows that most organizations surveyed are quite young with 67% founded after 1990. Correspondingly, 33% were founded before 1990. Figure 5-6 shows the count of organizations by the decade when they were founded.

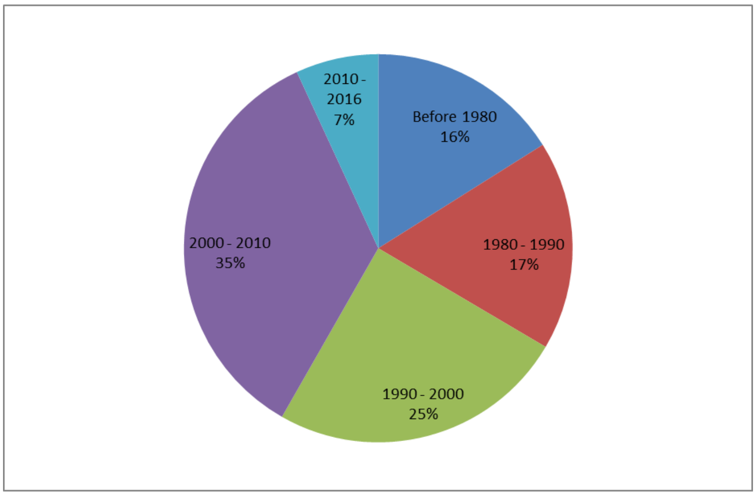

Figure 5-5 In what year was your organization founded?

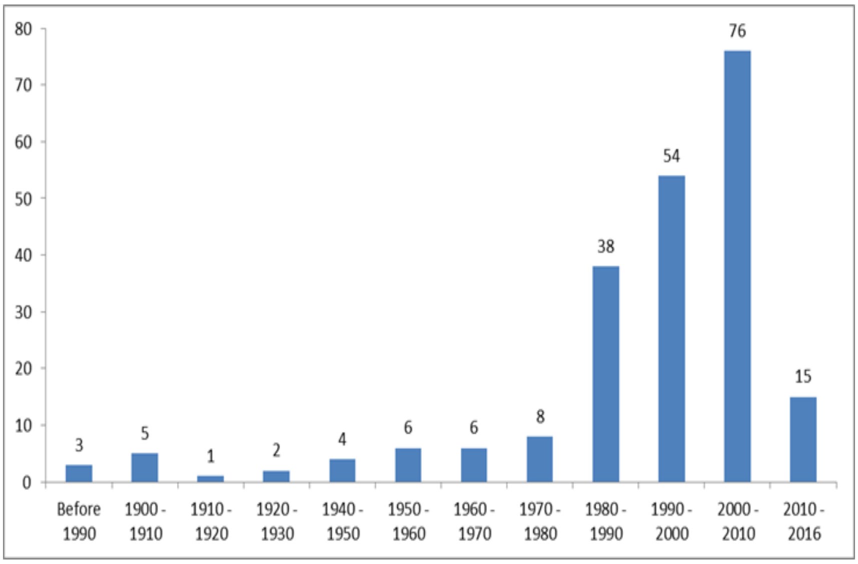

Figure 5-6 Count of Organizations by Decade

Q5: What industry does your organization belong to?

Surveyed participants belong to organizations in different industries, such as technology, healthcare, transportation, manufacturing, retail, education, energy, finance, etc. Almost 33% of organizations surveyed belong to the technology sector, making it the most represented industry in the sample. Less than 1% of organizations are in the government sector, making it the least represented industry. Figure 5-7 shows the proportions for all industries.

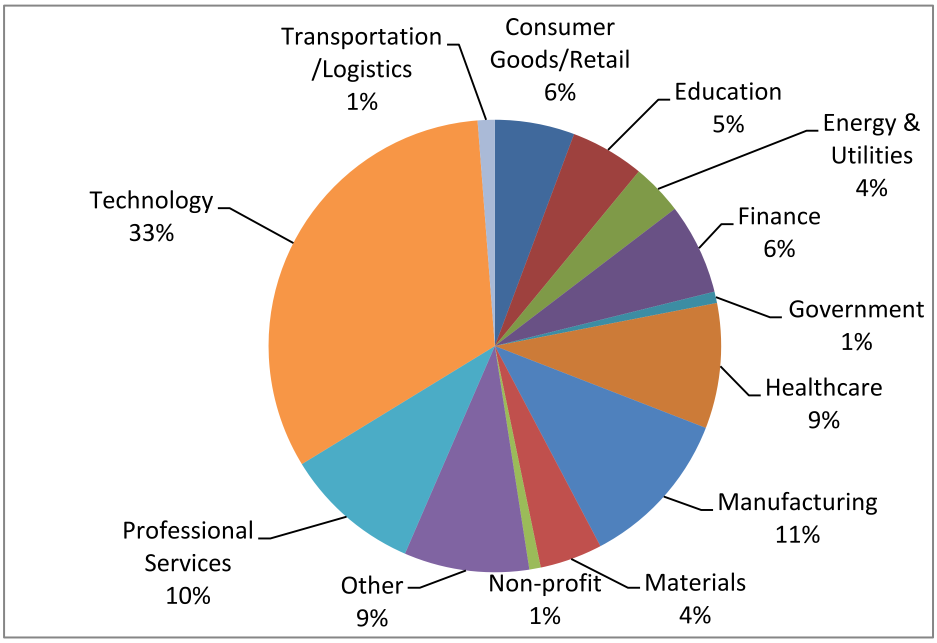

Figure 5-7. What industry does your organization belongs to?

Q6: Does your organization conduct business online?

The last question in Part 1 relates to whether an organization is engaged in online business. As shown in Figure 5-8, 95% of organizations surveyed conduct business online

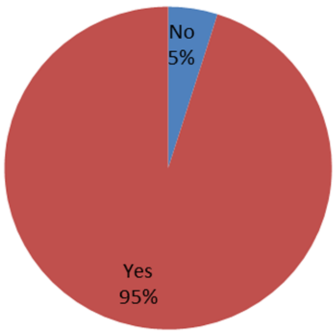

Figure 5-8. Does your organization conduct business online?

5.2.1 Summary of Part 1

In summary, results suggest that majority of participants are working in organizations headquartered in the US and Canada. 60 % of organizations survived are of medium size (250 – 2500 employees). Organizations founded after 1990 make 67% of the sample. Some 33% of the organizations work in the technology sector. Finally, 95% of the organizations surveyed conduct online business.

5.3 Part 2: Behaviours Analysis

The second part of survey consists of 45 questions. All these behaviours are related to Big Data capabilities in the following three areas: PB, OB, and TB.

For each question, participants were asked to report the degree of regularity and rigor with which their organization performs the desired behaviour using a sliding scale control. The control represents a continuous scale ranging from 0 to 100. It is used for capturing responses with anchors “not at all”, “somewhat”, and “extremely” (see Figure 5-9). This scale is believed to be more precise than Likert type scales because the latter scales focus on a limited number of choices. This can lead to a situation where concepts are not precisely measured. Continuous scales also allow for advanced and robust statistical analyses assuming continuous scales (Treiblmaier & Filzmoser , 2011).

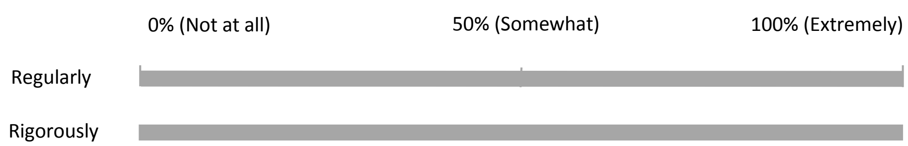

Figure 5-9. Regularly/Rigorously Scale

The responses on this scale are interpreted as follows. “Extremely regularly” implies the organization follows a strict schedule in relation to a particular behaviour. It does not matter how frequently the behaviour is performed as long a schedule is adhered to. Organizations that do not perform a particular behaviour at all, then response should be 0% (not at all). Similarly, “extremely rigorously” implies organizations follow formal processes or procedures rather than disorganized ad-hoc methods in performing behaviours. If performed on a completely "ad hoc" basis while formal policies and procedures (if any) are ignored, the response should be 0% (not at all). All these interpretations of the scales were provided to the survey respondents together with the survey.

5.3.1 Examination of Behaviours

The average (mean) score of regularity and rigor with which organizations perform various behaviours related to Big Data are provided and discussed below by area: People Behaviours (PB), Organization Behaviours (OB), and Technology Behaviours (TB).

5.3.1.1 People Behaviours (PB)

As shown in Table 5-1, the PB most regularly performed is PB3 with 75.57%, while the PB most rigorously performed is PB4 with 72%. This shows the average organization is less than 25% away from developing a strict schedule for performing PB3 and less than 28% away from perfectly following formal processes for performing PB4. The PB least regularly performed is PB1 with 70.72%, while the PB least rigorously performed is PB6 with 69%. Overall, the difference between the lowest scores and the highest scores is too minor to discriminate between behaviours based on this scale.

Table 5-1. Average Scores: People Behaviours (PB)

Code	Behaviour	Regularly		Rigorously	
		Mean	Std. Dev.	Mean	Std. Dev.
PB1	We appoint specific individuals in our organization to carry out data analytics initiatives in their respective organizational units.	70.72	20.18	70.38	21.63
PB2	We train employees on how to use data analytics.	74.45	21.19	71.06	22.35
PB3	We encourage employees to use data analytics in their work.	75.57	20.73	70.23	22.66
PB4	We have a Chief Information Officer (CIO), Chief Data Analytics Officer (CDO), or a Chief Data Scientist (CDS) who is involved in developing organizational strategies.	74.26	22.35	72.48	23.62
PB5	Our employees collaborate with universities to remain well-informed about developments in data analytics and to influence the data analytics curriculum.	71.11	23.47	69.44	25.04
PB6	Our employees participate in specialized groups, organizations, or conferences devoted to disseminating knowledge on data analytics.	72.23	22.12	69.16	24.26

The overall PB average values of regularity and rigor are 73.06% and 70.46% respectively (see Figure 5-10). Accordingly, PB's of the average organization are closer to perfectly following strict schedules and strict formal processes (100% on regularity and rigor scales) rather than performing these behaviours without rigor and regularity.

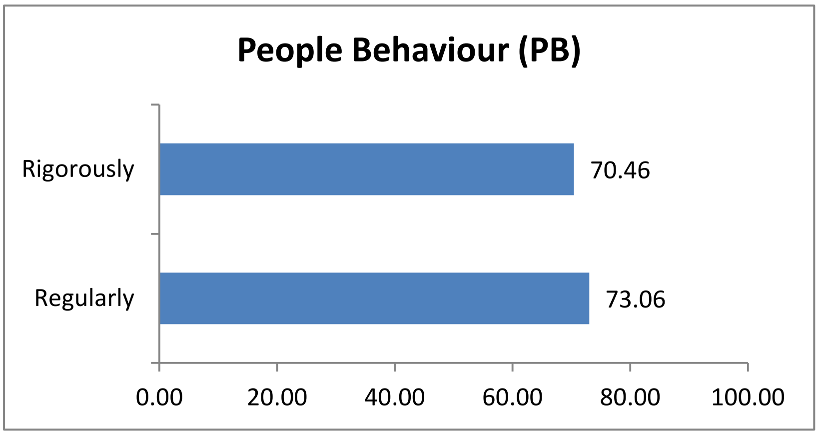

Figure 5-10. People Behaviours

5.3.1.2 Organization Behaviours (OB)

The OB most regularly performed is OB7 with 76%, while the OB most rigorously performed is OB2 with 73.74% (see Table 5-2). The highest score on the regularity scale implies the organization is less than 24% away from perfectly following a strict schedule for OB7. The highest score on rigor scale implies that the average organization is less than 27% away from perfectly following formal processes to perform OB2. The OB least regularly performed is OB8 with 73.17%, while the OB least rigorously performed is OB14 with 69.8%. Overall, the lowest and highest scores on regularity and rigor scales show minor differences between behaviours.

Table 5-2. Average Scores: Organization Behaviours (OB)

Code	Behaviour	Regularly		Rigorously	
		Mean	Std. Dev.	Mean	Std. Dev.
OB1	We use data analytics for organizational decision making.	75.46	20.22	73.56	21.87
OB2	We use data analytics in several departments or organizational units.	76.22	19.98	73.74	20.62
OB3	We integrate data across multiple departments and make the data available to our management.	75.54	19.91	73.18	21.23
OB4	We perform data analytics across all organizational sub-units or sub-systems in an integrated fashion.	75.21	20.17	71.22	23.27
OB5	We develop analytical models for improving organizational performance, managing risks, and finding new opportunities.	75.22	20.23	72.30	23.28
OB6	We utilize data analytics to predict future performance of our organization by developing various future scenarios that can be changed or customized using model parameters.	73.75	21.52	71.52	24.05
OB7	We develop decision making processes and policies based on data analytics.	76.24	20.31	71.95	22.96
OB8	Data analytics is performed in our organization in a centralized fashion (i.e. there is a formal or ad-hoc unit responsible for data analytics across the entire organization)	73.17	23.50	70.69	23.51
OB9	We integrate data analytics in our organizational processes.	74.48	20.65	71.59	22.58
OB10	We formulate and implement data analytics goals and objectives for our organization.	74.67	21.32	71.95	22.22
OB11	We use a formal approach to establishing and reviewing data analytics processes and policies.	74.16	20.77	71.95	21.65
OB12	We use data analytics for improving organizational performance.	74.91	21.73	73.16	21.42
OB13	We integrate data analytics in our organizational governance structure and governance processes.	74.50	20.65	70.10	23.40
OB14	We use data analytics to automate decision-making processes.	73.18	22.29	69.80	23.39
OB15	We use data analytics for competing with other organizations.	75.04	20.60	71.00	23.22
OB16	We use data analytics to create and exploit new opportunities.	75.33	19.80	72.75	21.69

The overall OB average values of regularity and rigor is 74.82% and 71.90% respectively (see Figure 5-11). Accordingly, OB's of the average organization are closer to perfectly following strict schedules and strict formal processes

(100% on regularity and rigor scales) rather than not at all. This pattern is similar to that of PB.

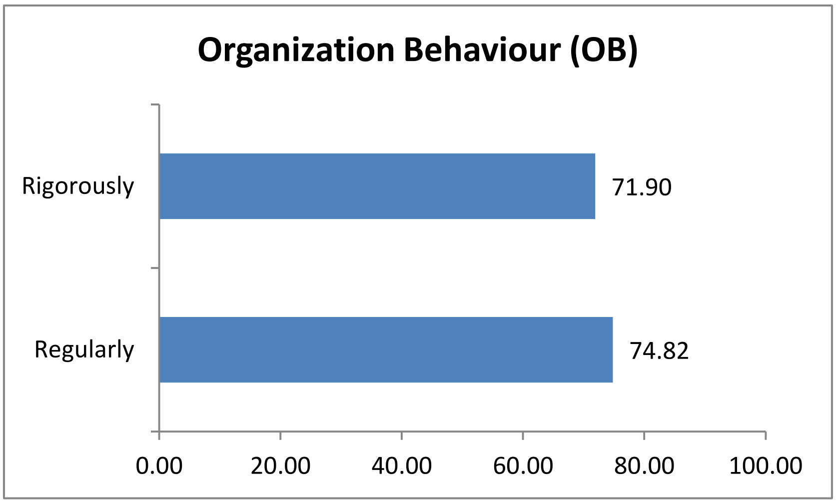

Figure 5-11. Organization Behaviours

5.3.1.3 Technology Behaviours (TB)

The TB most regularly and rigorously performed is TB2 with 75.81% and 72.67% respectively (see Table 5-3). The highest score on the regularity scale implies the organization is less than 25% away from perfectly following a strict schedule for performing TB2. The highest score on rigor scale implies the average organization is less than 28% away from perfectly following formal processes to perform TB2. The TB least regularly performed is TB18 with 70.82%, while the TB least rigorously performed is TB14 with 69.18%. Overall, the lowest and highest scores on regularity and rigor scales suggest minor differences between behaviours in this general category.

Table 5-3. Average Scores: Technology Behaviours (TB)

Code	Behaviour	Regularly		Rigorously	
		Mean	Std. Dev.	Mean	Std. Dev.
TB1	We use traditional spreadsheet tools (e.g. Excel) for data management, analysis, and reporting.	73.74	20.39	72.36	20.85
TB2	We analyse structured data.	75.81	19.94	72.67	22.20
TB3	We analyse archival data (i.e. data at rest).	74.79	20.91	71.85	22.84
TB4	We use data analytics to analyse past organizational problems and to get a historical overview of our organizational performance.	74.32	21.52	71.05	23.15
TB5	We employ desktop computers for data management and storage.	74.06	21.75	71.60	23.21
TB6	We utilize Business Intelligence (BI) tools (e.g. SAP BI, IBM BI, Microsoft SQL Server - BI, etc.) to manage and analyse data.	75.38	22.54	72.20	23.36
TB7	We analyse semi-structured data.	73.47	22.55	71.45	23.90
TB8	We analyse real time or nearly real time data.	74.30	21.38	70.90	23.52
TB9	We perform predictive modelling on our data.	73.80	21.57	70.16	24.54
TB10	We perform data mining on our data.	72.39	23.25	70.96	23.38
TB11	We employ servers/mainframes for data storage and management.	75.26	22.22	71.66	23.15
TB12	34) We employ Business Intelligence (BI) tools to generate reports.	73.75	22.37	71.72	23.00
TB13	We use a data centre as our main hardware/software platform for Business Intelligence (BI).	74.48	21.02	71.85	23.98
TB14	We utilize Hadoop/MapReduce and NoSQL databases (e.g. MongoDB, Apache Cassandra, etc.) to manage and analyse a variety of data	70.98	24.64	69.18	24.52
TB15	We analyse various types of data (e.g. transactional data, clickstream data, sensor data, videos, pictures, social media, text data, etc.)	73.07	22.44	70.91	23.36
TB16	We utilize in-memory and streaming database technology (e.g. SAP HANA, IBM Streaming Analytics, etc.) to analyse real time data.	72.34	23.41	70.24	23.98
TB17	We employ cloud solutions to store and manage a virtually unlimited volume of data.	74.04	22.61	72.05	23.46
TB18	40) We use scalable commodity servers to store and manage data.	72.29	23.77	69.68	24.32
TB19	We employ specialized reporting and data visualization tools that can be integrated with Hadoop or NoSQL platforms.	70.82	24.57	69.30	24.45

TB20	We perform traditional analysis techniques on our data, such as descriptive statistics, segmentation, sensitivity analysis, etc.	73.45	22.04	71.78	22.38
TB21	We perform advanced analysis techniques on our data, such as simulations, optimization, and predictive modelling.	73.22	21.45	71.65	22.50
TB22	We develop interactive or customizable reports for data presentation.	74.63	21.12	72.23	22.32
TB23	We utilize innovative approaches and new data sources in our data analysis (e.g. real-time analytics, video and voice analytics, image analytics, streaming analytics, geospatial analytics, sensor data analytics, text mining, machine learning, etc.)	72.55	21.86	71.14	23.05

The TB average values of regularity and rigor is 73.61% and 71.24% respectively (see Figure 5-12). Accordingly, TB's of the average organization are closer to perfectly following strict schedules and strict formal processes (100% on regularity and rigor scales) rather than not at all. This pattern is similar to that of PB and OB.

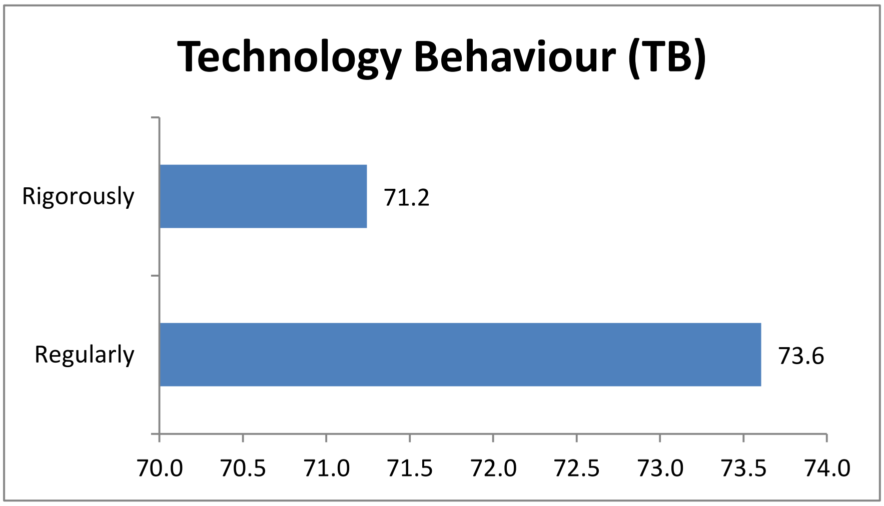

Figure 5-12 Technology Behaviours

5.3.2 Investigation Difference between Behaviours

The differences between PB, OB, and TB are small for both regularity and rigor, as shown in Figure 5-13. Results imply there are no systematic differences between regularity and rigor. The Bivariate Pearson Inter – item correlation test is performed to formally confirm that there are no systematic differences for either regularly or rigorously measured PB, OB, or TB. Results confirm a high correlation between all three. The significance level is less than .01 (see Table 5-4 and Table 5-5).

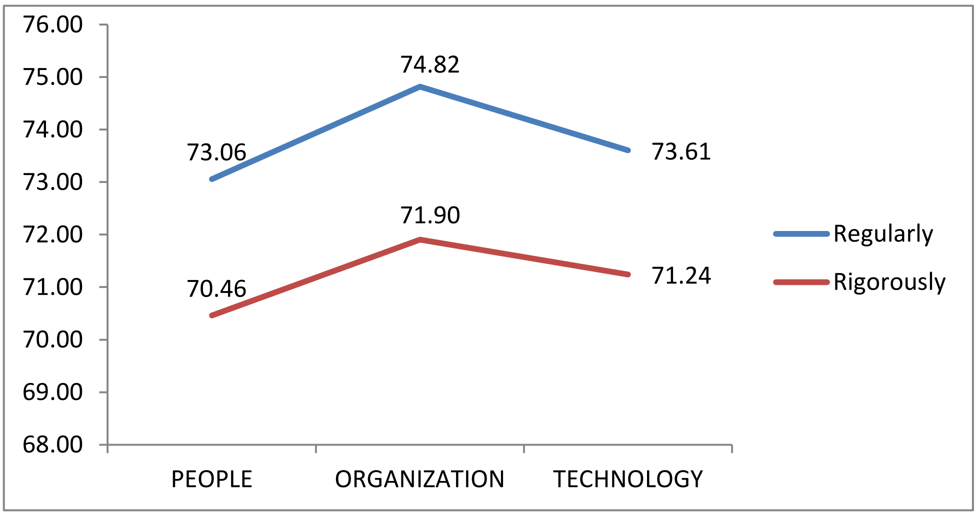

Figure 5-13. Difference between behaviours by area

Table 5-4. Regularly Inter-Items Corrrelation

Correlations- Regularity

		PB	OB	TB
PB	Pearson Correlation	1	.910**	.843**
	Sig. (2-tailed)		.000	.000
	N	246	246	246
OB	Pearson Correlation	.910**	1	.901**
	Sig. (2-tailed)	.000		.000
	N	246	246	246
TB	Pearson Correlation	.843**	.901**	1
	Sig. (2-tailed)	.000	.000	
	N	246	246	246

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5-5. Rigorously Inter-Items Corrrelation

Correlations-Rigor

		PB	OB	TB
PB	Pearson Correlation	1	.902**	.832**
	Sig. (2-tailed)		.000	.000
	N	246	246	246
OB	Pearson Correlation	.902**	1	.916**
	Sig. (2-tailed)	.000		.000
	N	246	246	246
TB	Pearson Correlation	.832**	.916**	1
	Sig. (2-tailed)	.000	.000	
	N	246	246	246

**. Correlation is significant at the 0.01 level (2-tailed).

5.3.3 Regularity vs. Rigor

For each behaviour the average response on the rigor scale is less than that on the regularity scale as shown in Figure 5-14. A correlation analysis is performed to see whether regularity and rigor are correlated for individual behaviours. Also, the paired sample t – test is performed to determine whether the mean value difference between regularity and rigor is zero. Results imply there is correlation between regularity and rigor in the three areas (see Table 5-6). Also t is greater than 2.8 for all three areas (see Table 5-7), which indicates a statistically significant difference between regularity and rigor. Accordingly, it is inferred that organizations prioritize regularity over rigor.

We speculate that this can be explained by the fact that organizations are generally driven by actions and accomplishment rather than how rigorously these actions were.

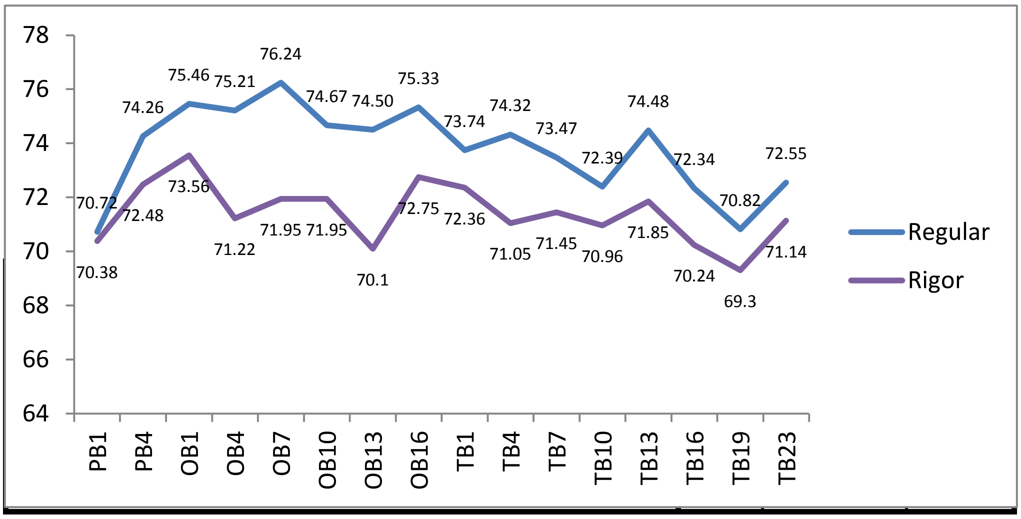

Figure 5-14 Regularly Vs. Rigorously

Table 5-7. Paired Sample t-Test

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	PB “Regular” - PB “Rigorous”	2.59	14.428	.919	.787	4.41	2.82	245	.005
Pair 2	OB “Regular” – OB “Rigorous”	2.81	13.843	.882	1.078	4.55	3.19	245	.002
Pair 3	TB “Regular” – TB “Rigorous”	2.36	13.099	.835	.717	4.01	2.82	245	.005

5.3.4 Clustering Analysis

A hierarchical clustering approach was used to identify homogenous groups of cases - the so-called “clusters”. Ward’s clustering method was used to apply a maximum distance measure to aggregate cases that reveal similar patterns and levels of response. In doing so, patterns were extracted between organizations and unique features of each pattern were identified. These are expected to reflect the Big Data maturity level of organizations based on the degree to which they perform Big Data behaviours regularly and rigorously.

Graphical representation of the clustering analysis is provided via a dendrogram in Figure 5-15 for the regularity scale and at Figure 5-16 for the rigor scale. The dendrogram is a cluster tree which illustrates arrangement of clusters produced by hierarchical clustering. Individual cases are arranged in groups along the horizontal axis of the dendrogram based on similarity between them. Similar groups are clustered together until all groups are organized as a tree. The dendrogram vertical axis or height shows distance between cases. Similar cases have the same height whilst the greater the difference in height the more dissimilarity. The cluster membership list for the 246 cases on regularity and rigor scales is given in Appendix D.

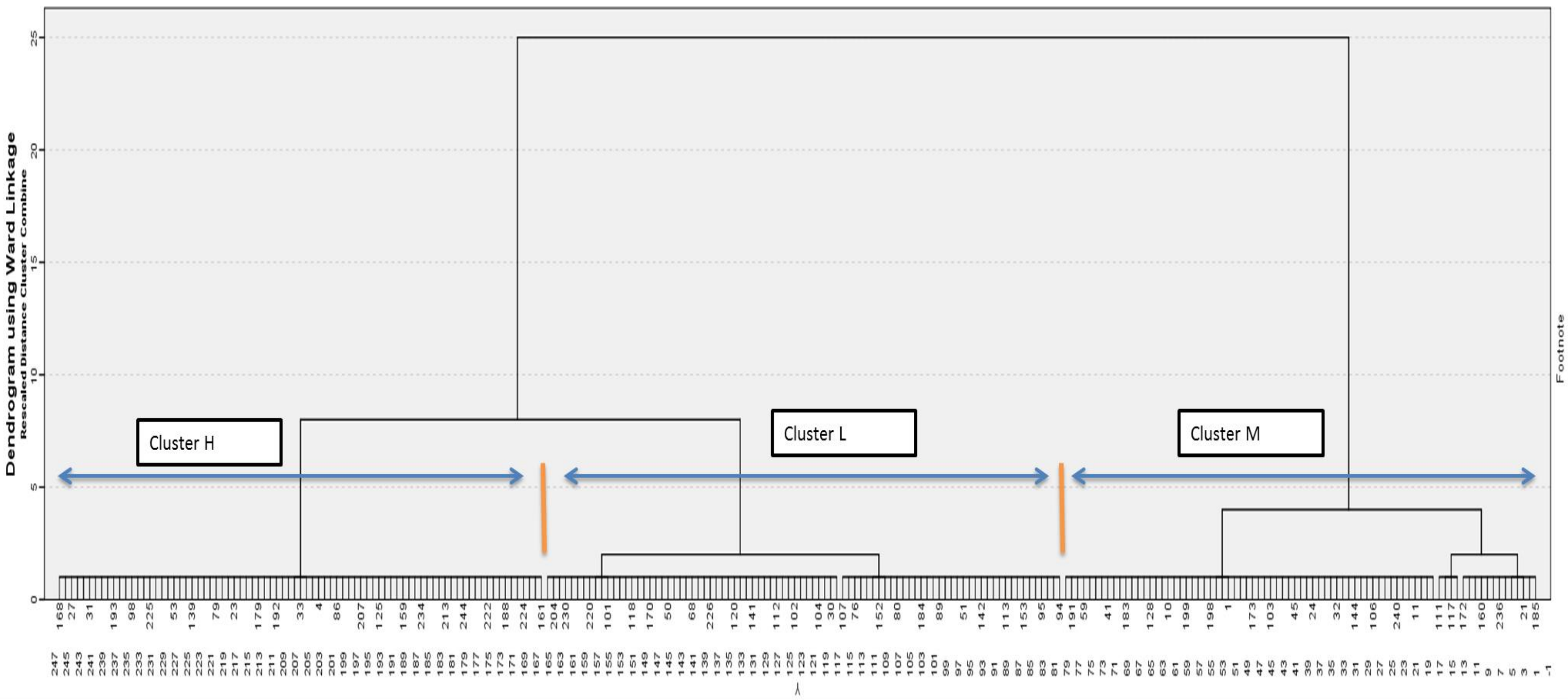

Figure 5-15. Clustering Cases - Regularly

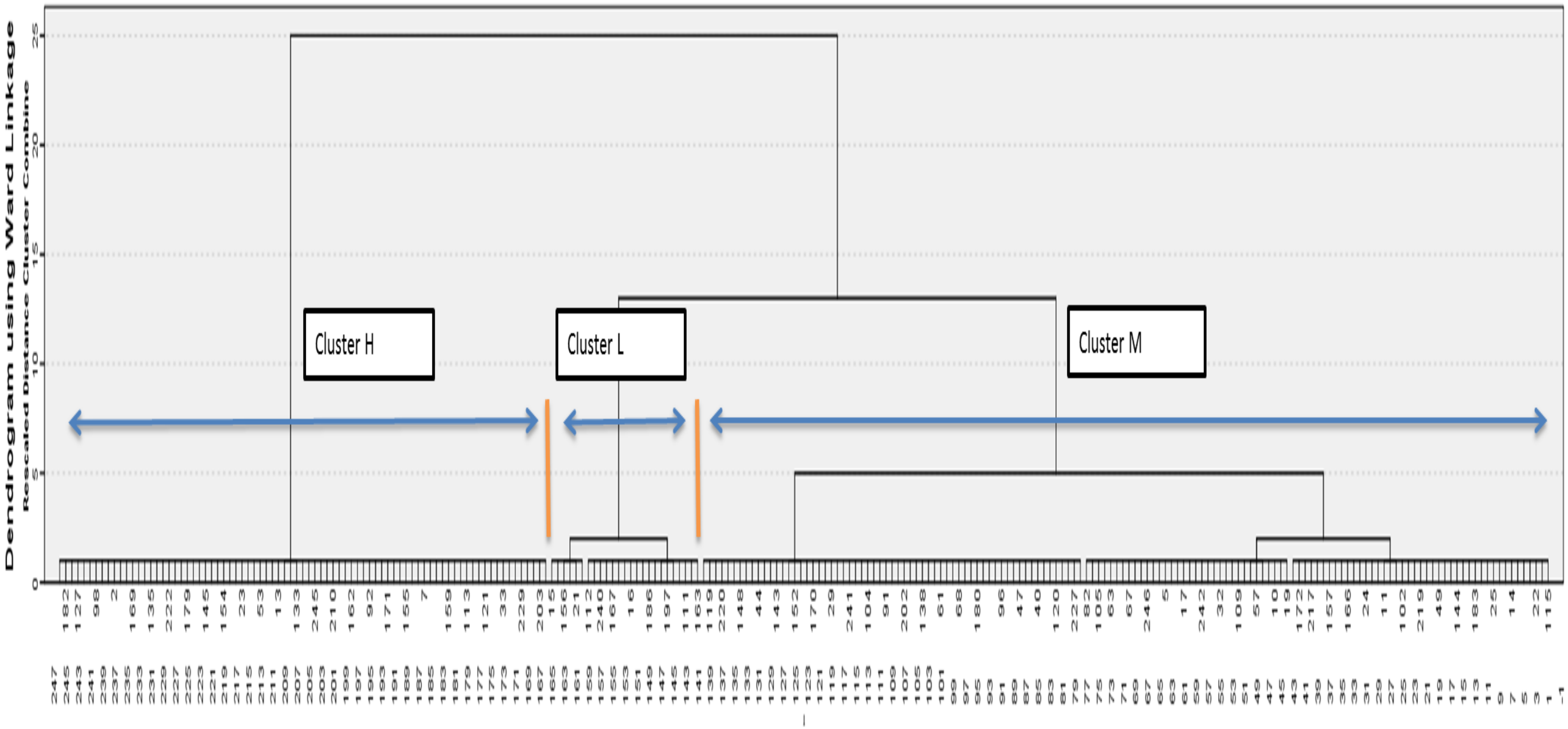

Figure 5-16. Clustering Cases –Rigorously

For clarification, orange lines are drawn on the dendrogram (see Figure 5-15 and Figure 5-16) between different clusters using the cluster membership list given in Appendix D. The average score for each cluster is given in Table 5-8 and Table 5-9 and further visualized in Figure 5-17. Results imply good discrimination between the three types of organizations represented by the three clusters: Cluster H (is a cluster with a high level of rigor or regularity), Cluster M (is a cluster with a medium level of rigor or regularity) and Cluster L (is a cluster with a low level of rigor or regularity). The discrimination is based on the average values of how regularly and rigorously various BD behaviours are performed within each cluster.

Table 5-8. Regularly: Average of clusters

REGULAR	L	M	H
Count	79	86	81
MEAN	53.91	74.42	93.04
STD. DEV.	11.40	5.90	4.61

Table 5-9. Rigorously: Average of clusters

RIGOR	L	M	H
Count	25	140	81
MEAN	32.07	66.60	91.76
STD. DEV.	11.25	9.33	5.46

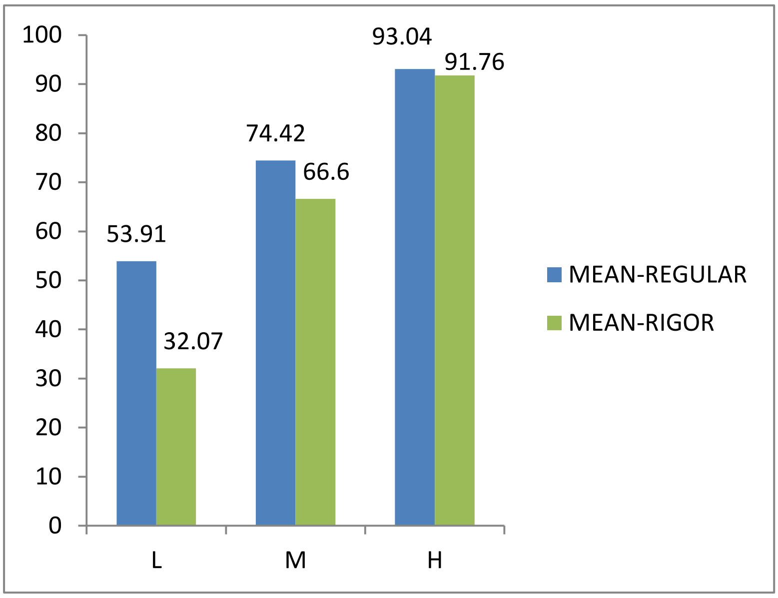

Figure 5-17 Average of Clusters

More organizations fit in Cluster M compared to Cluster L and H (Figure 5-18). This is more pronounced for the rigor scale, where 140 organizations fit in Cluster M compared to 25 in Cluster L and 81 in Cluster H. Detailed results of the clustering analysis are available in Appendix E and Appendix F.

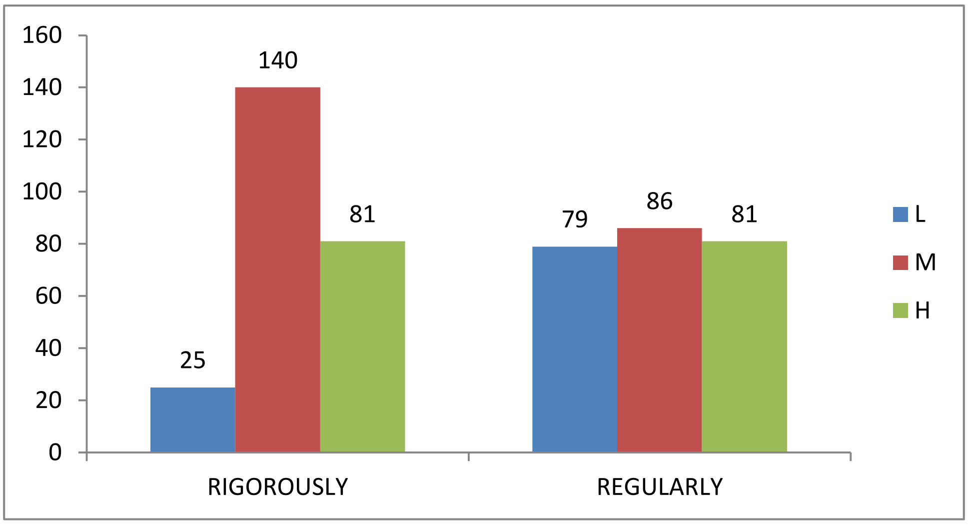

Figure 5-18. Count of Cases in Clusters

5.3.4.1 Clustering Analysis by Area

The average values are calculated per cluster for PB, OB and TB's. Results are shown in Table 5-10 and Table 5-11 on the regularity and rigor scales respectively. Results are also depicted in Figure 5-19 and Figure 5-20 on both regularity and rigor scales.

Table 5-10. Regularly: Average of Clusters per Area

Regularity	People		Organization		Technology	
	Average	StdDev	Average	StdDev	Average	StdDev
H	90.91	7.96	93.73	4.54	93.11	4.82
M	72.66	10.51	75.25	7.55	74.36	6.83
L	55.18	12.96	54.65	12.26	52.78	13.15

Table 5-11. Rigorously: Average of Clusters per Area

Rigor	People		Organization		Technology	
	Average	StdDev	Average	StdDev	Average	StdDev
H	89.47	8.73	92.08	6.04	92.13	5.59
M	65.74	12.86	67.05	11.22	66.51	9.77
L	35.29	16.59	33.71	13.78	30.08	12.40

Cluster H performs all activities across PB, OB, and TB with a high degree of regularity and rigor (regularity is between 90.91% and 93.73% whilst rigor is between 89.47% and 92.13%). Cluster L performs all activities across PB, OB, and TB with a relatively low degree of regularity and rigor (regularity is between 52.78% and 55.18% whilst rigor is between 30.08% and 35.29%).

Cluster M performs all activities across PB, OB, and TB with a medium degree of regularity and rigor (regularity is between 72.66% and 75.25% whilst rigor is between 65.74% and 67.05%). Detailed results of the clustering analysis are available in Appendix G and Appendix H.

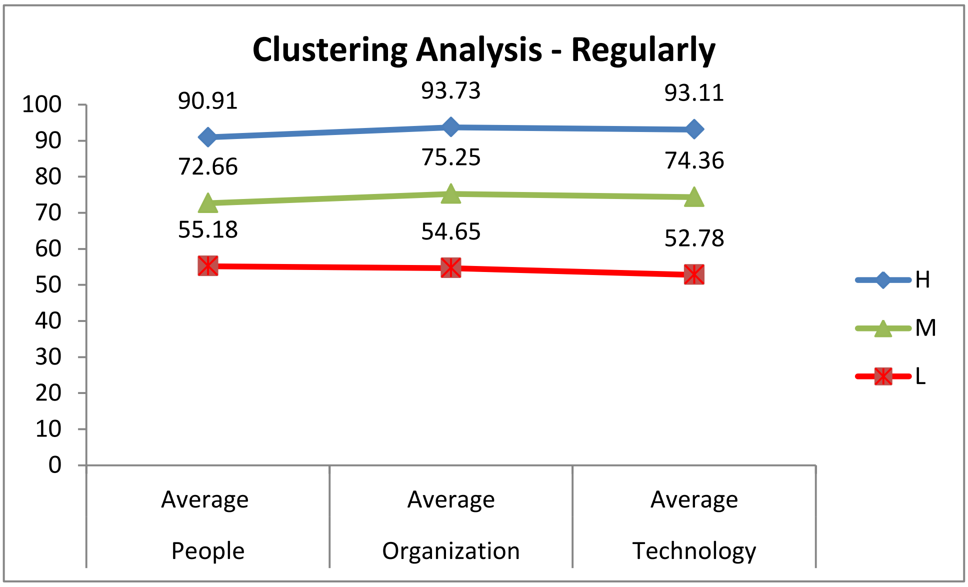

Figure 5-19. Regularly: Clustering Analysis by Area

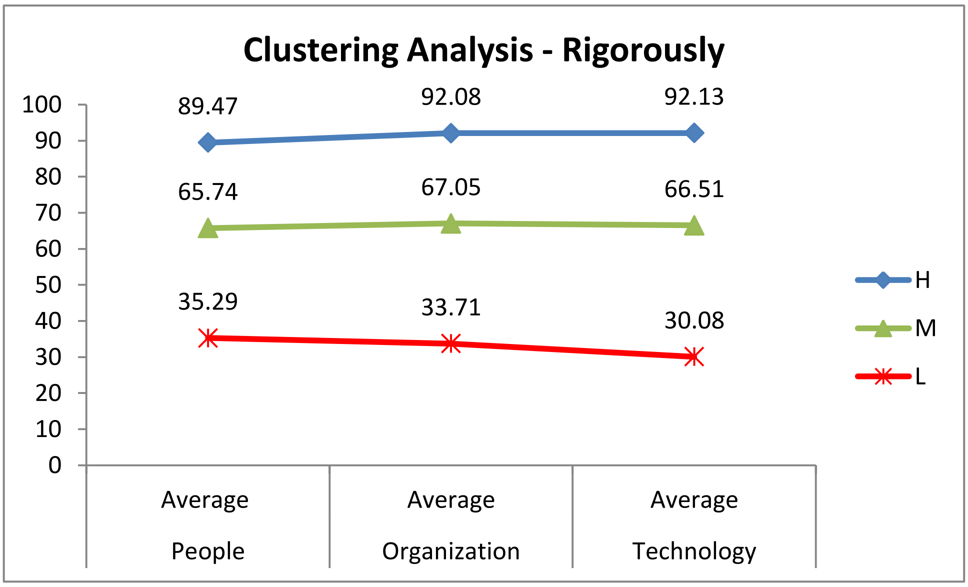

Figure 5-20. Rigorously: Clustering Analysis by Area

5.3.4.2 Organizational Attributes and Clusters

A key finding from clustering analysis is there are 3 levels of organizational maturity with respect to Big Data. These levels are named as Cluster H, Cluster M, and Cluster L. These clusters represent organizations with high, medium, and low scores on the regularity and rigor scales. Further analysis is conducted below to explore the relationship between these different levels of maturity with respect to Big Data and organization profiles.

5.3.4.2.1 Headquarter Location

As shown in Figure 5-21, most surveyed organizations that responded are headquartered in North America (United States and Canada). With respect to regularity, most organizations are in Cluster M, whilst most organizations headquartered in the Middle East are in Cluster L. Three of four European organizations are in Cluster M and one in Cluster H. Four out of six organizations headquartered in Latin America and the Caribbean are in Cluster H and two in Cluster M. Finally, both African headquartered organizations are in Cluster L.

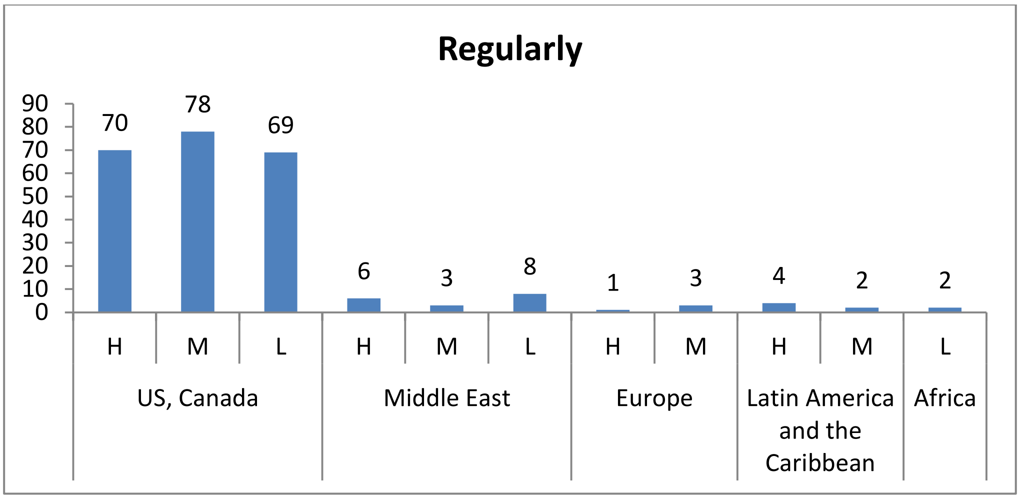

Figure 5-21. Regularly: Headquarter Location vs. Level

With respect to rigor (see Figure 5-22), North American organizations are mostly in Cluster M. However, the difference between clusters in terms of frequency is more pronounced (i.e. the difference between Cluster M and L is 98). With respect to regularity, the difference between the M and L clusters is only 9. For the Middle East, most organizations are in Cluster M. European organizations are equally distributed between Cluster M and H. With Latin American and Caribbean organizations, the maximum number of responses is in Cluster H - similar to the regularity results. More African organizations are clustered in Cluster M. This is in contrast to the regularity scale, where most African organizations are in cluster L.

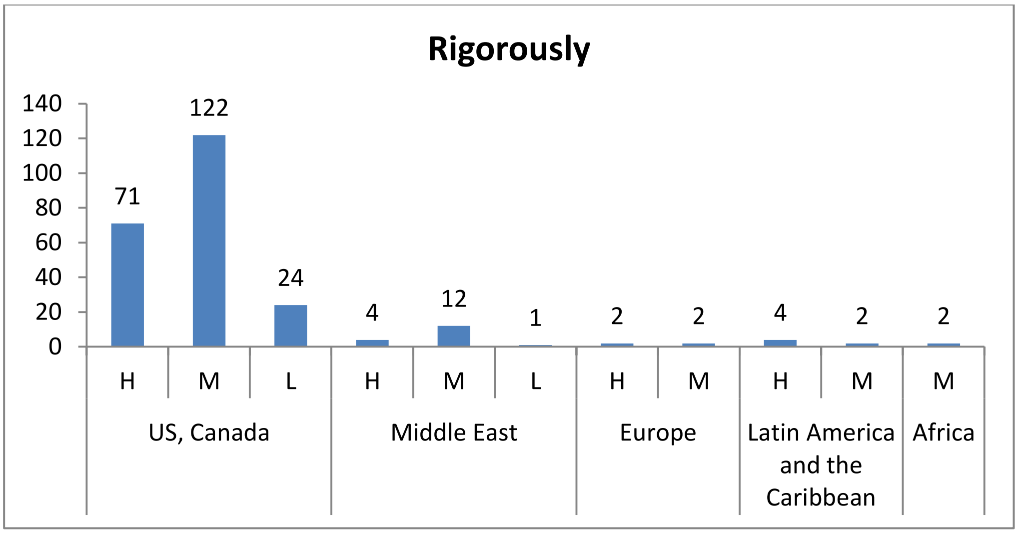

Figure 5-22. Rigorously: Headquarter location vs. Level

5.3.4.2.2 Size

The largest organizations surveyed tend to be in Cluster H for both regularity and rigor (Figure 5-23 and Figure 5-24). Organizations of other sizes tend to be in Cluster M for rigor with pronounced frequency differences between Cluster H, M and L. One obvious observation here is that organizations sized 10-50 employees, 50 – 250 employees, and 1000 – 2500 employees appear in Cluster L in the regularity result. To evaluate the difference statistically, a Paired Sample t–test was used to determine whether mean value difference between large and smaller organizations is zero in regularity. Results show that $t > 1.54$ for the first case (1000 – 2500 employees) and $t > 2.4$ for the remaining cases. The significance level is less than .05 for all cases except the first case where the significance level is .129 as seen in Table 5-12. This implies a systematic difference between large and other sized organizations and that difference is significant, except for the first case. The correlation with organizations less than 10 employees cannot be computed because there is only one case exist in this category.

The same test is performed to determine whether the mean value difference between large and smaller organizations is zero on the rigor scale. Results indicate that t>2.6 for all cases as seen in Table 5-13. The significance level is less than .05 for all cases. This indicates a systematic, significant difference between large and small organizations.

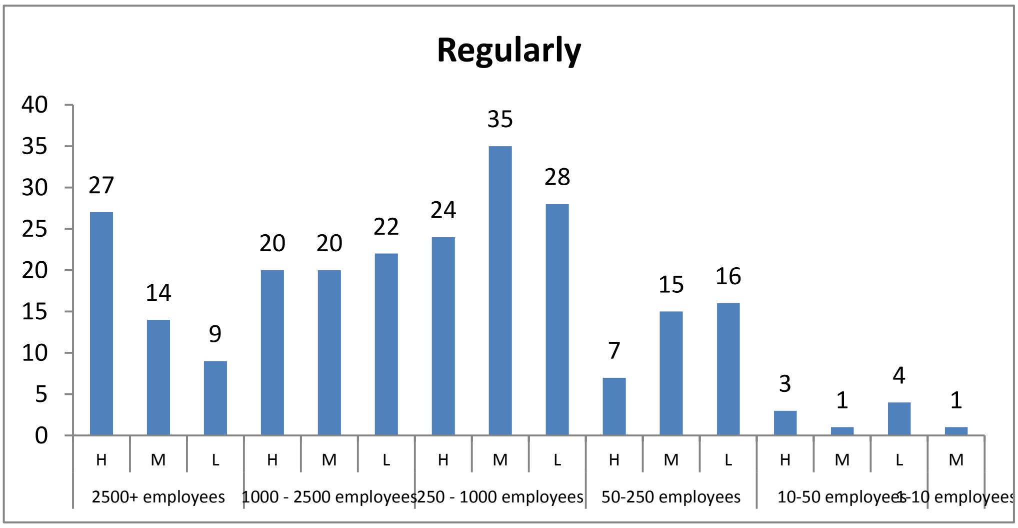

Figure 5-23. Regularly: Size vs. Level

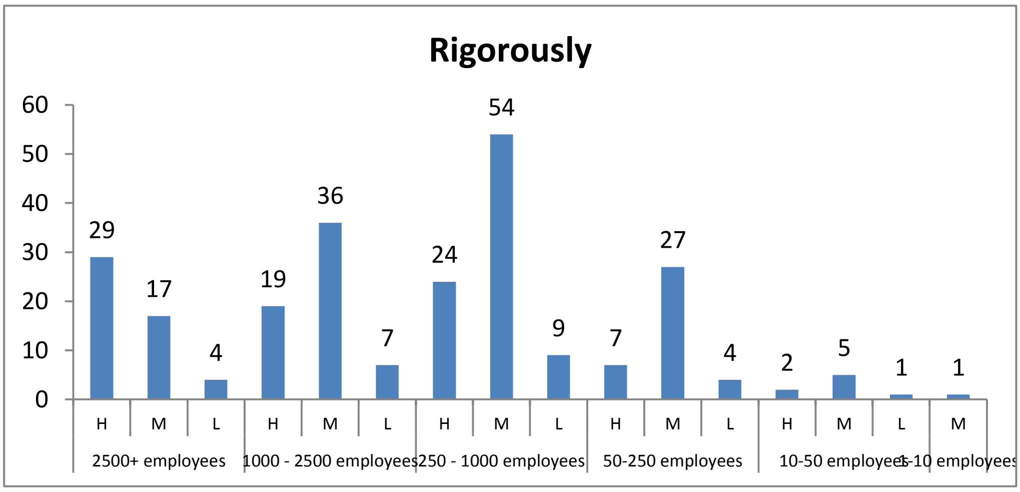

Figure 5-24. Rigorously: Size vs. Level

Table 5-12. Regularly: Paired Sample t –Test (Size)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Dev.	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Regularity2500 - Regularity1000	5.72	26.188	3.703	-1.726	13.159	1.544	49	.129
Pair 2	Regularity2500 - Regularity250	10.22	25.725	3.638	2.912	17.534	2.810	49	.007
Pair 3	Regularity2500 - Regularity50	11.14	25.474	4.132	2.766	19.512	2.696	37	.011
Pair 4	Regularity2500 - Regularity10	17.52	20.600	7.283	.297	34.741	2.405	7	.047

Table 5-13. Rigorously: Paired Sample t –Test (Size)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Rigor2500 - Rigor1000	11.523	30.633	4.332	2.817	20.229	2.660	49	.011
Pair 2	Rigor2500 - Rigor250	14.585	29.343	4.150	6.246	22.925	3.515	49	.001
Pair 3	Rigor2500 - Rigor50	13.584	21.449	3.479	6.534	20.634	3.904	37	.000
Pair 4	Rigor2500 - Rigor10	17.214	17.217	6.087	2.820	31.607	2.828	7	.025

5.3.4.2.3 Year Founded

Surveyed organizations founded between 1990 and 2005 represent exactly 50% of respondents. Most are in Cluster M on both regularity and rigor scales (see Figure 5-25 and Figure 5-26). Organizations founded before 1990 and after 2005 appear mostly in Cluster M on the rigor scale, but tend to be in Cluster L on the regularity scale. Moreover, these organizations perform Big Data behaviours less regularly than those founded between 1990 and 2005.

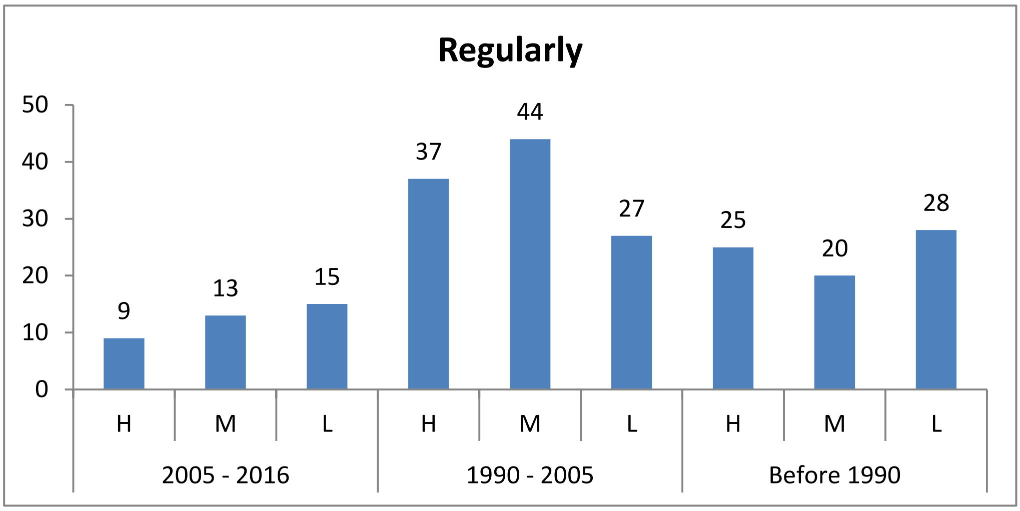

Figure 5-25. Regularly: Year Founded vs. Level

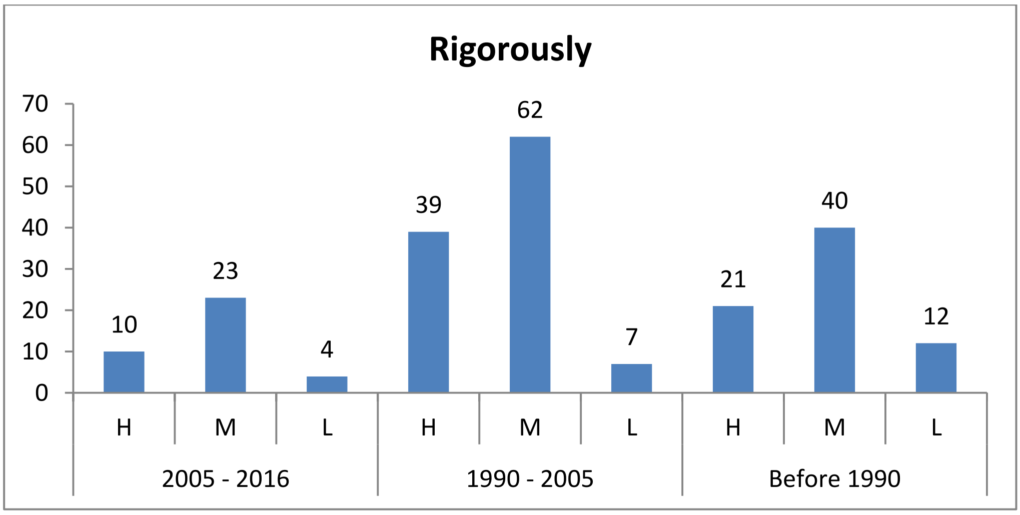

Figure 5-26. Rigorously: Year Founded vs. Level

5.3.4.2.4 Industry

The technology sector represents 33% of all survey respondents. Most are in Cluster M on both regularity and rigor scales (Figure 5-27 and Figure 5-28). This is also the case for all other sectors for the rigor scale (except for Consumer Goods/Retail, which is mostly in Cluster H). However, results are different for the regularity scale. There are small differences between the categories. For example, Manufacturing, Finance and Consumer Goods/Retail are clustered mostly in H; Healthcare, Energy, Utilities and Government are clustered mostly in L.

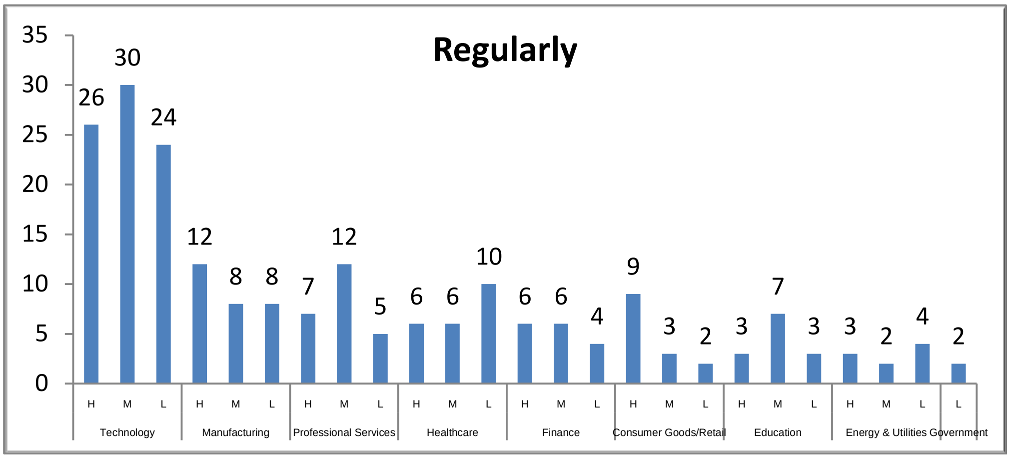

Figure 5-27. Regularly: Industry vs Level

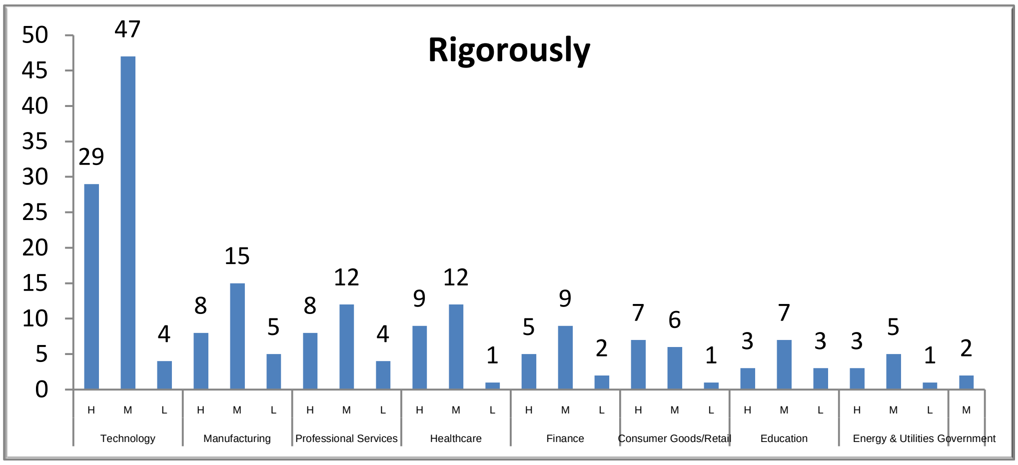

Figure 5-28. Rigorously: Industry vs. Level

5.3.4.2.5 Online Business

The percentage of surveyed organizations performing online business is 95%. Most are in Cluster M, followed by H (see Figure 5-29 and Figure 5-30). Organisations that do not conduct online business are in Cluster L on the regularity scale and M on rigor scale.

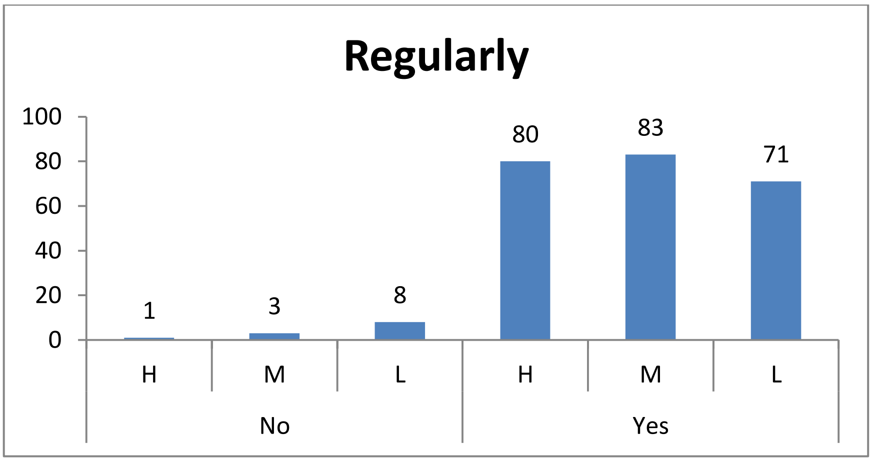

Figure 5-29. Regularly: Online vs. Level

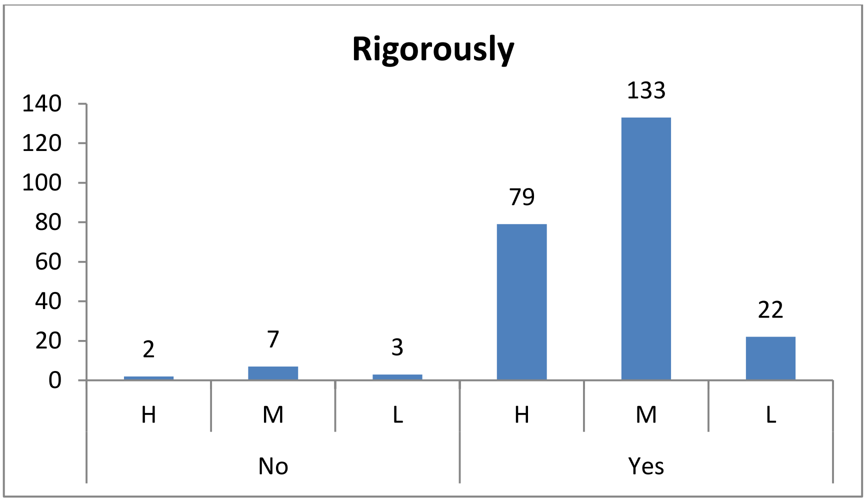

Figure 5-30. Rigorously: Online vs. Level

5.3.5 Summary of Part 2

Based on the results provided above, organizations in North America are no different from organizations headquartered in other regions of the world. It is noted there is significant differences between large and small organizations. To assess the difference statistically, a paired samples t–test was used to determine whether mean value difference between large and smaller organizations is zero in regularity. The test produces $t > 1.54$ for the first case (2500 – 1000 employees) and $t > 2.4$ for the remaining cases. The significance level is less than .05 for all cases except pair-1 (2500 – 1000 employees) where the significance level is .129, as shown in Table 5-12. This implies systematic, significant differences between large and small organizations.

No pattern is detected for the relationship between the age of an organization and Big Data maturity level. It is observed that there is some relation between organization industry and maturity level. Organizations in the Consumer Goods/Retail industry and those in Manufacturing and Finance tend to be in Cluster H on regularity scale. This implies that these types of organizations are more mature with respect to Big Data than other industry.

5.4 Application of Clustering Analysis on BDMM

Using clustering analysis, three clusters of organizations (H, M, and L) were identified based on the average values of regularity and rigor with which organizations perform BD activities within each of the clusters. The average survey responses across PB, OB & TB are seen in Figure 5-31 and Figure 5-32. It is therefore argued that the three clusters are related to BDMM levels of Data Processing (DP), Business Intelligence (BI), and Big Data (BD). These maturity levels associated with the proposed BDMM were discussed in Section 3.4.

The highest pattern of clustering analysis (Cluster H) implies that organizations in this cluster have achieved the BD maturity level (blue line). Organizations in H perform all behaviours with a significant degree of regularity and rigor (over 90% and 89% respectively). The middle pattern of the clustering analysis (Cluster M) implies the existence of BI maturity level (red line). Organizations in M perform all behaviours with average regularity and rigor (74% and 66% respectively). The lowest pattern of clustering analysis (Cluster L) implies the existence of DP level (green line). Organizations in L all behaviours with low regularity and rigor (<56% and <36% respectively).

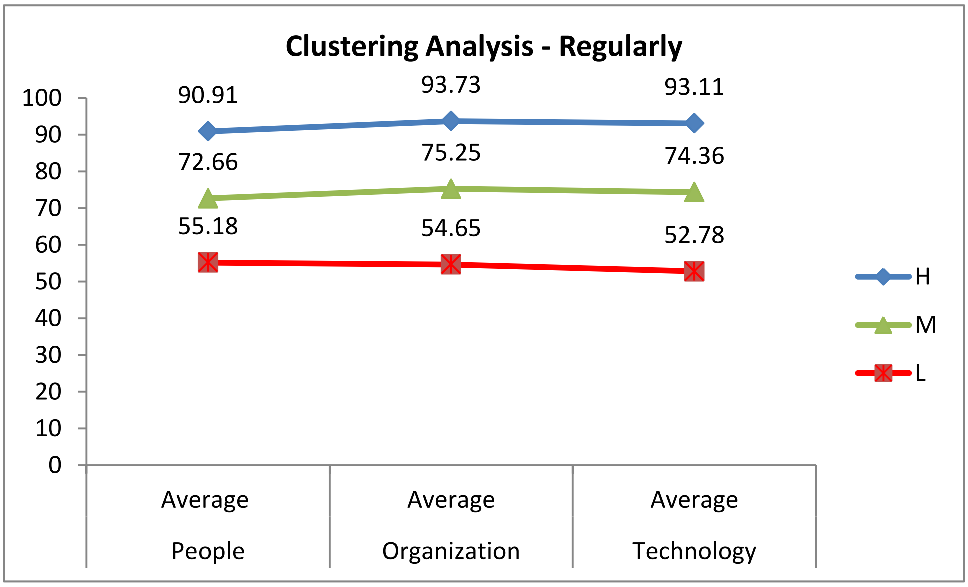

Figure 5-31. Regularly: Clustering Analysis by Area

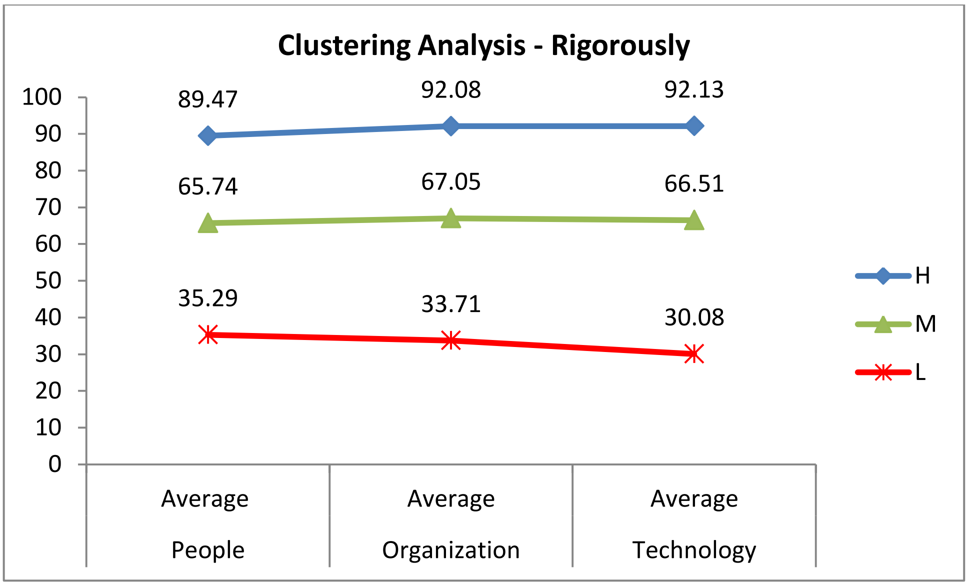

Figure 5-32. Rigorously: Clustering Analysis by Area

5.5 Implementation of BDMM

In order for the BDMM (see Figure 5-33) to be useful as a benchmarking tool, the results of clustering analysis are used to calculate the boundaries of three clusters: H, M, and L. This enables the assigning of organizations into one of the three BDMM levels: DP, BI or, BD of the (see Section 3.4).

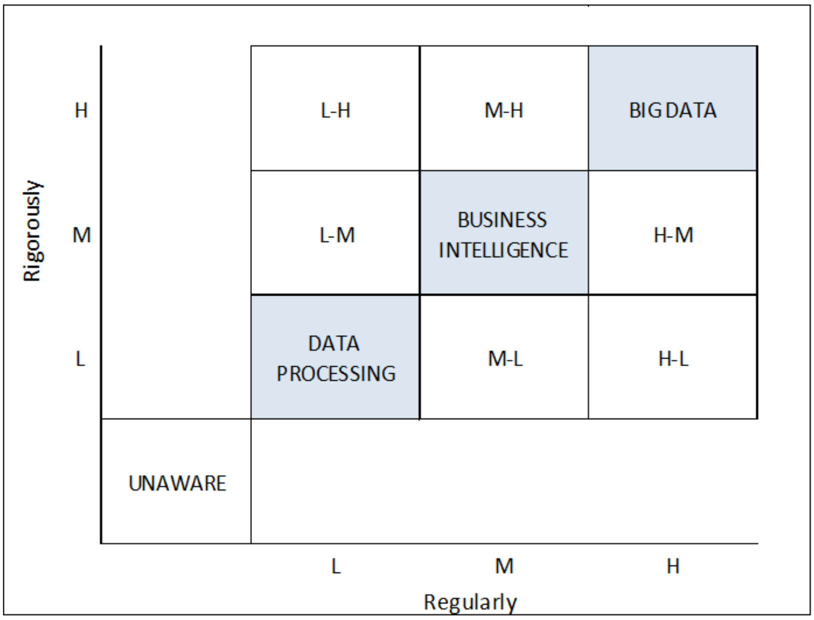

Figure 5-33. Big Data Maturity Model (BDMM)

5.6 Computing the Boundaries of Maturity Levels

To calculate the upper (max) and lower (min) boundaries for each level, the average values and standard deviation of Clusters H, M, and L are used. One standard deviation is added or subtracted to the mean for the upper (max) and lower (min) boundaries. These calculations for the regularity scale are shown in Table 5-14 and Figure 5-34. The calculations for the rigor scale are shown in Table 5-15 and Figure 5-35.

Table 5-14. Regularly: Boundaries of Maturity Levels

REGULAR	L	M	H
MEAN	54	74	93
STD. DEV.	11	6	5
MAX	65	80	98
MIN	43	69	88

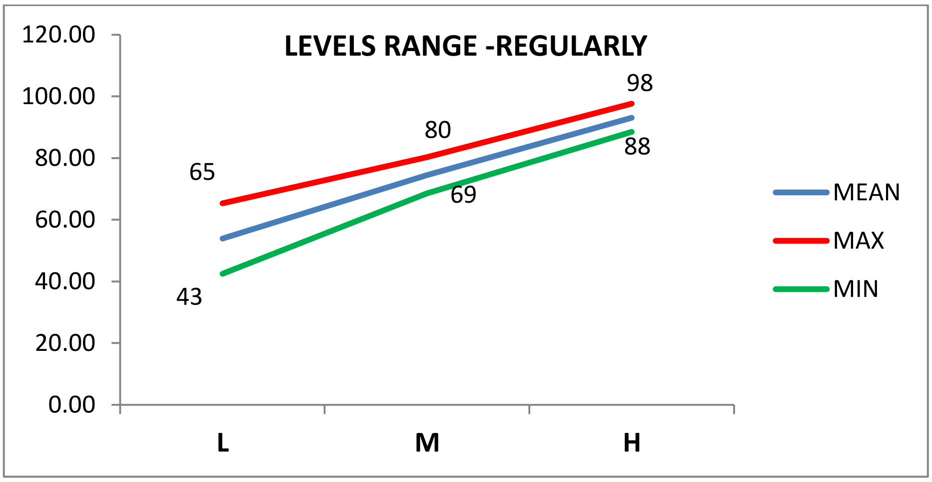

Figure 5-34. Regularly: Boundaries of Maturity Levels

Table 5-15. Rigorously: Boundaries of Maturity Levels

RIGOROUS	L	M	H
MEAN	32	67	92
STD. DEV.	11	9	5
MAX	43	76	97
MIN	21	57	86

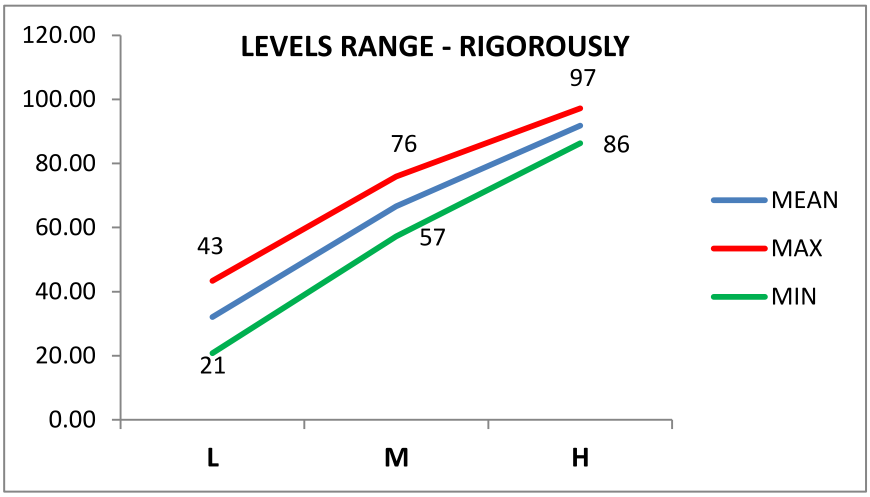

Figure 5-35 Rigorously: Boundaries of Maturity Levels

There is a discontinuity between levels (i.e using the mean and standard deviation approach) as seen in Table 5-14 and Table 5-15. In other words, there are gaps between upper (max) and lower (min) boundaries of levels H, M and L. This indicates that some organizations may fall between two levels. Figure 5-36 depicts these gaps between levels. A simple solution designed to eliminate these gaps is discussed in the next section.

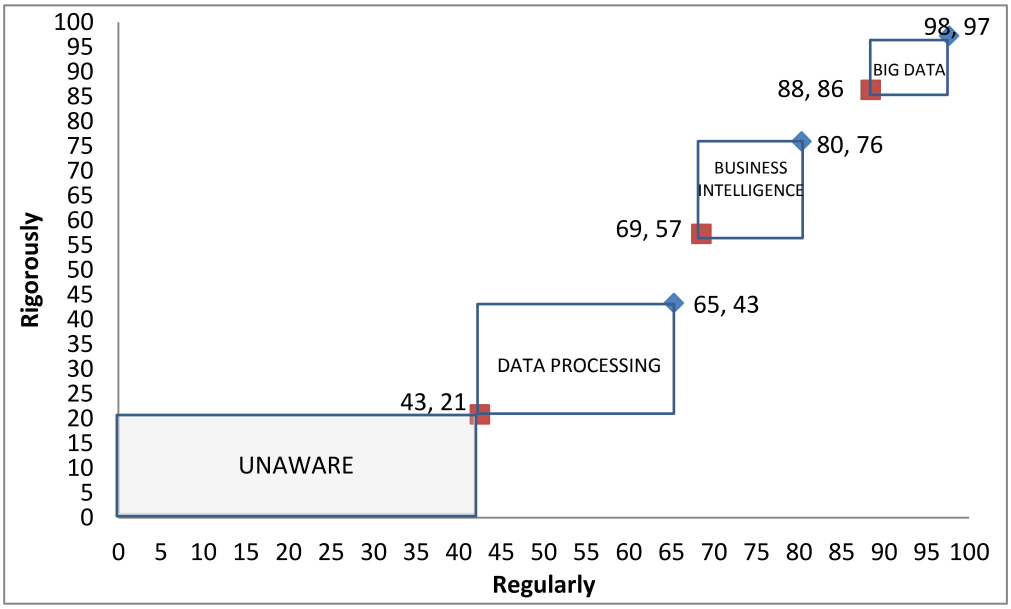

Figure 5-36. Gaps between Levels

5.7 Revised Boundaries of Maturity Levels

To address the issue of gaps between levels, the upper limit of one level has been extended to the next higher level. This is reflected in the updated upper limits provided in Table 5-16 for the regularity scale. The same process is repeated for the rigor scale, as shown in Table 5-17. This approach removes the discontinuity without affecting the sequence of the maturity levels (see Figure 5-37). For example, assume organization X falls between L and M i.e. exhibits higher values than L but lower values than M. In the updated levels, the organization X will join L, because it qualifies to join the lower level but does not qualify to join the higher level.

Table 5-16. Regularly: Revised Boundaries of Maturity Levels

Levels	Regularly		Regularly (Updated)	
	Lower Limit	Upper Limit	Updated Lower Limit	Updated Upper Limit
L	43	65	43	69
M	69	80	69	88
H	88	98	88	100

Table 5-17. Rigorously: Revised Boundaries of Maturity Levels

Levels	Rigorously		Rigorously (Updated)	
	Lower Limit	Upper Limit	Updated Lower Limit	Updated Upper Limit
L	21	43	21	57
M	57	76	57	86
H	86	97	86	100

5.8 The Big Data Maturity Model as a Benchmark

In order to use the BDMM as a benchmarking tool, the updated boundary levels were provided in Figure 5-37.

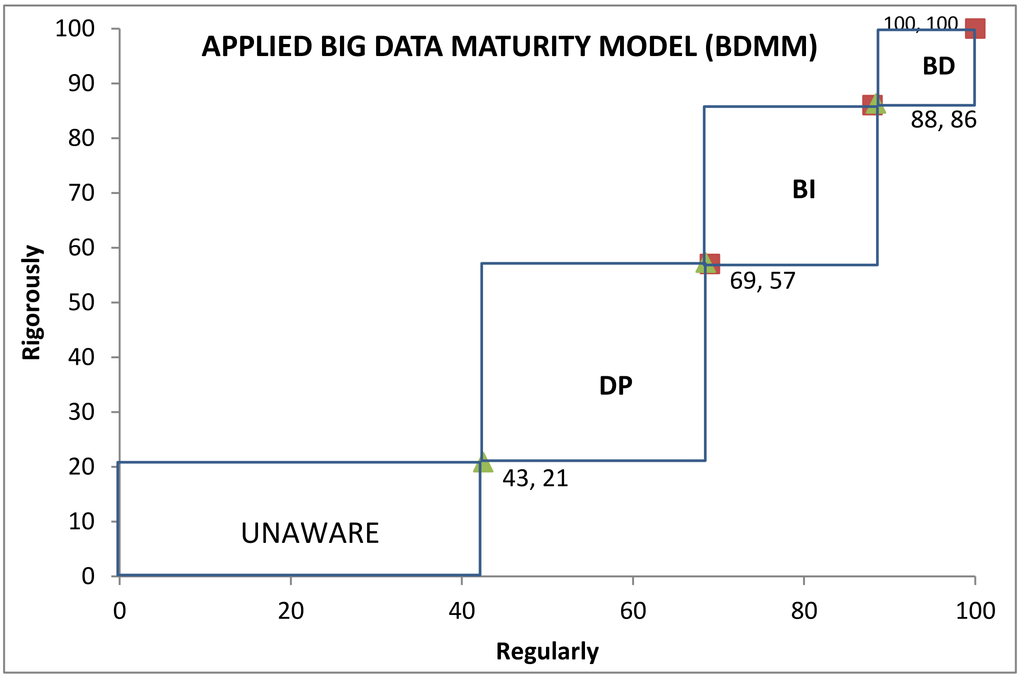

Figure 5-37. Applied Big Data Maturity Model

It is therefore argued that the BDMM can be used as a benchmarking tool to assess an organisation's Big Data maturity. It can also help organisations identify deficiencies in behaviours or capabilities that need to be rectified to achieve a higher level of maturity with respect to Big Data

5.9 Summary of Chapter

This chapter has provided an overview of the results obtained with the help of the survey. It provides descriptive statistics with discussion about the general demographics questions included in part one of the survey. Then the chapter provides descriptive statistics of the 45 questions included in part two of the survey with explanations and interpretations of these descriptive. For each question in part two of the survey, results were obtained using two scales to record the regularity and rigor of behaviours as perceived by the survey participants. In line with the exploratory nature of this research, we have utilized clustering analysis to explore the hidden patterns or clusters in the data collected. Results imply the existence of three categories of different levels of regularity and rigor, indicating three levels of maturity with respect to Big Data. Clusters were then linked to the BDMM proposed in section 3.4. Cluster boundaries were adjusted to eliminate the gaps among the clusters so that the model can be used as a benchmarking tool.

CHAPTER 6

DISCUSSION

This chapter provides a summary of the study; discusses the empirical results in the light of the main objectives of the study, literature review, and conceptual framework; and provides analysis and interpretation of the empirical results. Furthermore, the chapter discusses implications of the results, highlighting the main contributions and some limitations. The chapter concludes with directions for future research in the area of Big Data capabilities.

6.1 Overview of Study

The study sought to address the problem related to the lack of a practical maturity model for the era of Big Data. It was argued that most existing organizational assessment tools and maturity models in the field of data analytics have been created with Business Intelligence and Data Warehousing in mind. Several existing maturity models were examined to identify their limitations in relation to Big Data. Thus, it was argued that there was a need for a new Big Data Maturity Model (BDMM).

Given this need, a framework for key Big Data Capabilities (BDC) was developed based on the Socio-Technical Perspective (STP) and the Resource-Based View (RBV) of the firm. This BDC framework consists of three main capability areas: People, Organization, and Technology. People capability refers to organization capacity required to acquire and develop people skills, knowledge, support, etc., necessary to build, use, and operate BD technology in an organization. Organization capability refers to organization capacity required to develop and promote organization strategy, process, structure, and culture in relation to BD. Technology capability refers to organization capacity required to deploy hardware and software that can support BD

Each of these three broad areas of BD capabilities is populated by specific behaviours or processes reflecting these capability areas. These behaviours are collectively referred to as People Behaviours (PB), Organizational Behaviours (OB), and Technology Behaviours (TB).

Based on the BDC framework, a new BDMM was developed. It consists of four levels derived from chronological progress of data processing; Unaware, Data Processing (DP), Business Intelligence (BI), and Big Data (BD). Thus, each level of the BDMM represents a historical period in data processing.

6.2 Discussion and Implications of Results

To assess the capabilities of organizations in relation to BD, a set of behaviours has been developed. These behaviours represent tangible manifestations of organizational BD capabilities. The behaviours have been developed based on the BDC framework. A total of 45 validated behaviours are used in the study: 6 for PB, 16 for OB, and 23 for TB.

Organizations perform behaviours with a certain level of regularity and rigor - depending on the level of organizational maturity in relation to Big Data. Mature organizations perform all three behaviours with a high level of regularity and rigor. Low maturity organisations perform these behaviours with a low level of regularity and rigor (i.e. in an ad-hoc or informal manner). Therefore, the level of maturity of an organization can be measured by assessing how regularly and rigorously organizations perform behaviours across the three areas of Big Data capabilities. The sections below discuss the results obtained with the help of this scale.

6.2.1 Relationship between Areas: PB, OB and TB

Results show that PB, OB, and TB are highly correlated. The correlations were formally explored using the Bivariate Pearson Inter – item correlation test. The evidence also points to no systematic differences between the three areas of behaviour among the companies in the dataset.

6.2.1.1 Implication 1: Relationship between Areas: PB, OB and TB

The strength of the statistical relationship or correlation between the three areas of capabilities implies that the three areas of BDC reflect the same concept – organizational maturity with respect to Big Data. It also implies a direct relationship between organization maturity and the three areas. So, it is necessary that all three areas move to a higher level of maturity in order for an organization to also move to a higher level of Big Data maturity in the proposed BDMM.

6.2.2 Regularity vs. Rigor

Results indicate that organizations systematically perform behaviours more regularly than rigorously. The significant difference is confirmed via a Paired Sample t – Test. This difference implies organizations usually prioritize regularity over rigor. Moreover, results show a correlation between rigor and regularity in three areas of PB, OB and TB.

6.2.2.1 Implication 2: Regularity vs. Rigor

Despite the need for a balance between regularity and rigor to achieve consistent Big Data maturity progression, the difference between regularity and rigor implies organizations are driven by achievement and implementation of activities rather than the rigor of procedures used. This can be due to the fact that organizations are motivated by achievement (regularity) rather than following formal processes (rigor). This result aligns with David McClelland theory, mentioned in his 1961 book, "The Achieving Society". He identified three primary motivators people require, with one of those motivators being the need for achievement.

6.2.2.2 Implication 3: Regularity vs. Rigor

The correlation between regularity and rigor indicates that there is a relationship between regularity and rigor. This correlation implies organizations performing behaviours more rigorously are committed to perform behaviours more regularly. However, this relationship does not necessary imply causality. Perhaps both scales are an indicator of another, broader concept, such as organizational commitment to Big Data.

6.2.3 Clustering Analysis

The clustering analysis of survey responses shows three distinct clusters within the sample based on the average value for regularity and rigor. Cluster with low average value is called Cluster L; cluster with medium average value is called Cluster M; and cluster with high average value is referred to as Cluster H.

Relationship between clustering levels and some attributes of organizations is explored for five organisational attributes: headquarter location, size, year founded, industry sector, and whether the organization conducts business online. It is observed that organizations are distributed nearly equally among the three levels on regularity scale (with Cluster M being somewhat more numerous). Distribution on the rigor scale shows that the majority fit in Cluster M, followed by H. Only a small percentage of organizations is in L. It is observed that large organizations, specifically in the Consumer

Goods/Retail sector, tend to be in Cluster H on both regularity and rigor scales.

6.2.3.1 Implication 4: Clustering Analysis

One finding of this study is that large organizations tend to perform behaviours more regularly and rigorously compared to smaller organizations. This implies that large organizations are more likely to be at the Big Data maturity level. Another finding implies organizations working in Consumer Goods/Retail sector tend to perform behaviours more regularly and rigorously compared to organizations working in other industries. This also implies that Consumer Goods/Retail sector presents organizations with better opportunities for developing Big Data capabilities.

6.3 Big Data Maturity Model

Results of the clustering analysis correspond to the proposed maturity levels derived from the three eras of computing and integrated into the proposed BDMM (see Figure 5-33). The boundaries of clusters that are believed to represent various levels of maturity were slightly modified in order for the BDMM to be usable as a benchmarking tools. These modifications are reflected in the updated BDMM (see Figure 5-37).

The resulting clusters are linked to the theoretical level of maturity as follows. Cluster H matches the Big Data maturity level. Organizations at this level exhibit a high level of regularity and rigor with respect to Big Data behaviours in all three areas: PB, OB, and TB. Cluster M matches the BI level.

Organizations in this cluster tend to have a medium level of regularity and rigor with respect to BD with respect to PB, OB, and TB. Cluster L matches the DP level. Organizations at this level show a low level of regularity and rigor with respect to BD behaviours in all three areas: PB, OB, and TB. Another level called Unaware is added to reflect organizations outside of any particular cluster. Unaware represents organizations below the lower boundaries for regularity and rigor in Cluster L. This means that these organizations have no systematic capabilities or competencies in Data Analytics. There may be individual efforts among these organizations to perform behaviours regularly or rigorously, but these efforts are unplanned and not very formal..

6.3.1.1 Implication 5: Relation between maturity levels and scales

The BD maturity level (see Figure 5-37) covers 12% (88% – 100%) on regularity scale and 14% (86% – 100%) on rigor scale. The BI maturity level covers 19% (69% – 88%) on regularity scale and 29% (57% – 86%) on rigor scale. Finally, DP maturity level covers 26% (43% – 69%) on regularity scale and 36% (21% – 57%) on rigor scale. These ranges show that the spread in regularity and rigor of behaviours tends to decrease when the maturity level increases (i.e. there is an inverse relationship). The gap between organizations at extreme sides of DP maturity level is relatively big. This means that there are great differences in the regularity and rigor with which organizations perform behaviours in this maturity level. Thus, efforts to upgrade organizations from DP to BI level and from BI and BD level may vary.

6.3.1.2 Implication 6: Distribution of Organizations on BDMM

Based on one scale (i.e. regularity or rigor), organizations canfit within one of the four BDMM levels: Unaware, DP, BI, or BD - depending on behaviour performance i. However, it is observed that some organizations can also fit into two maturity levels if the two scales are used separately to classify an organization into a level of maturity. For example, if an organization exhibits regularity of behaviours at the BI level but rigor at the DP level – it will be hard to place the organization into one of these two levels (see Figure 6-1).

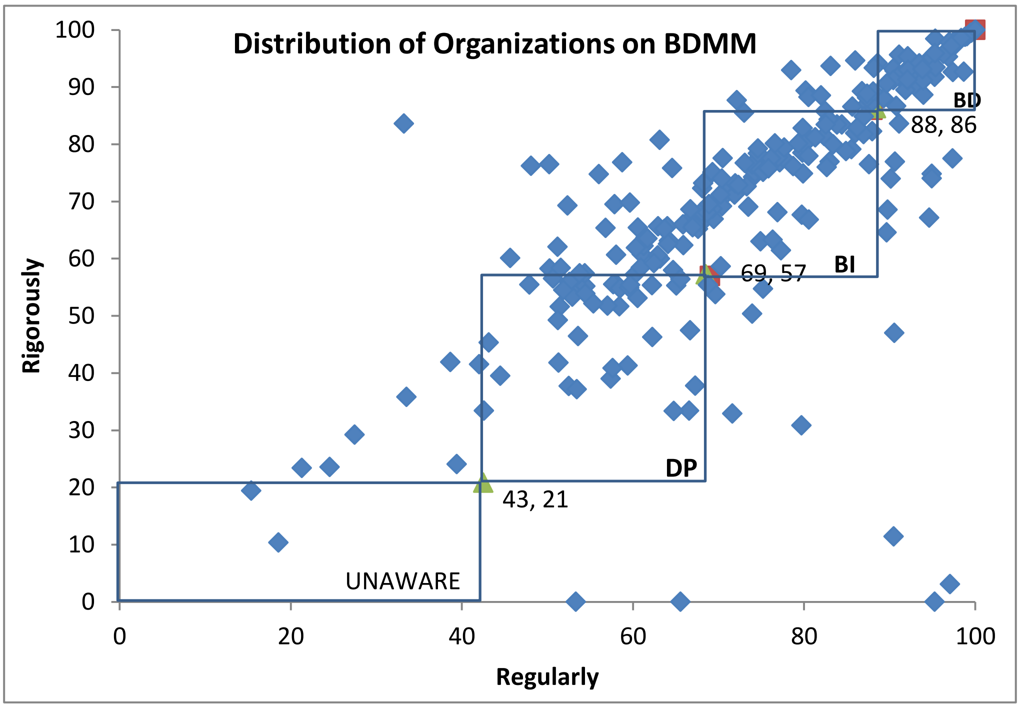

Figure 6-1. Distribution of Organizations on BDMM

6.4 Contributions

The first contribution of this study is a simple and different theoretical stance based on a historical perspective. This perspective is used to differentiate between various maturity levels with respect to Data Analytics among organizations. This study identifies four computing eras: primary DP era before 1960; advanced DP era between 1960 and 1990; BI era between 1990 and 2005; and BD era after 2005. Each era is largely defined by unique technologies and practices related to data processing. This historical approach helps differentiate between organizations at various levels of data processing maturity in a simple and unambiguous fashion.

The second contribution is a simple and practical conceptualization of the Big Data concept. Some common definitions for Big Data are discussed first. These common definitions were proposed by such organizations as Gartner Inc., IBM, SAS, Oracle, and Microsoft. One common definition of Big Data is in terms of the 3 V's of data: volume, variety and velocity (Ward & Barker, 2013). This study also chooses to adopt the 3 V's and links them to well-known Big Data technologies (e.g. MapReduce, NoSQL, In-Memory Analytics, etc.) and related managerial practices. It is concluded that Big Data often refers not only to data itself, but to the tools and practices for analysing, processing, and managing massive, complex, and rapidly evolving data sets.

The third contribution of this study is a clear articulation of the differences between BI and BD. This study argues that BD is similar to the concept of BI, yet it is different on a number of important dimensions. Both BI and BD are used to denote usage of data management technologies and computer-based analytical tools for discovering actionable business knowledge and facilitating decision making based on organizational data. Yet this study highlights that both BD and BI have three major differences related to the type of data used. These differences are related to the aforementioned 3 V's (see Table 6-1).

Table 6-1. BI Vs. Big Data

Characteristics	Big Data	Business Intelligence
Volume	Infinite	Finite
Velocity	Real time	Offline
Variety	Unstructured	Structured

The fourth contribution of this study is the identification of barriers and challenges that prevent organizations from fully exploiting BD opportunities. These barriers include outdated IT infrastructure, the inherent complexity and messiness of BD, lack of data science skills within organizations, privacy concerns and organizational cultures that are not conducive to data-driven operations or data-driven decision making. This study proposes some tactics for addressing these barriers and challenges related to BD, which involve changes to technology infrastructure through the utilization of commodity hardware and specialized BD software, focus on privacy, promotion of BD and analytic skills through collaboration with educational institutions, and creation of a clear organizational BD vision. The fifth contribution is a review of 20 BI/DW maturity models spanning 2001 to 2013. This review shows that most BI maturity models are mostly focused on IT components, such as IT infrastructure, software applications, and databases (Lahrmann et al., 2010). They are less focused on components related to people and organizations. Examples of important and overlooked organizational components related to data analytics are organizational strategies and organizational culture. Examples of overlooked people components include user skills and expertise as well as top management support for Big Data analytics.

The sixth contribution is the development of a framework for BDC. Consistent with literature review and STP and RBV, the BDC framework consists of three different areas of capabilities: technology, people and organization (see Figure 6-2).

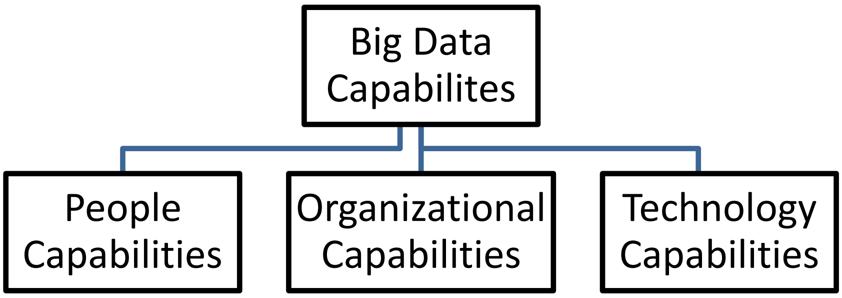

Figure 6-2: Big Data Capability (BDC) Model

People capabilities are related to organizational capacity required to acquire and develop people skills, knowledge, support, etc. needed to build, use, and operate Big Data technologies and processes. Organizational capabilities are related to organizational capacity required to develop and promote organization strategy, processes, structures, and values for adapting and using Big Data. Technology capabilities refer to organizational capacity to deploy hardware and software solutions supporting Big Data analytics. The people, organizational and technology behaviours are shown to be of similar importance for assessing organizations maturity in relation to BD.

The seventh and main contribution is the development of a BDMM. The proposed model consists of four levels: Unaware, DP, BI, and BD. The model is populated by specific behaviours related to the main areas of capabilities mentioned earlier: people, organization, and technology. These behaviours are identified using the BDC framework. It is assumed that each behaviour is performed with a certain level of regularity and rigor. As organizations perform behaviours more regularly and rigorously, maturity increases. Organizations perform the behaviour more regularly if it follows a strict schedule. It does not matter how frequently as long as it adheres to schedule. Organizations perform the behaviour rigorously if it follows formal processes or procedures rather than performing the activity in an ad hoc manner (Piccoli, et al., 2011).

A total of 45 behaviours were included in the BDMM. Boundaries for different levels were identified so that the model can not only guide organizations on which behaviours need to be emphasized for improving maturity, but also benchmark organizations in relation to the maturity levels and other organizations within these maturity levels.

Overall, the BDMM, when used as a benchmarking tool, can help organizations accomplish the following:

- Assesses current organizational capabilities in relation to Big Data

- Standardize schedule and process consistency among various Big Data processes
- Help managers identify specific actions or behaviours for improving Big Data capabilities for the purpose of gaining competitive advantage via Big Data use

6.5 Limitations

Although this study has achieved its main objectives, the results of this study should be viewed in the light of several limitations. First, this study includes participants of certain positions or titles, such as the CIO, CDO, CTO, and DDS. Therefore, the results may not be generalizable to other organizational roles. Second, it is close to impossible to explore all relevant Big Data behaviours. Moreover, the BDMM and the corresponding survey need to be concise to be practical. Thus, some important Big Data behaviours might have been omitted. Finally, the survey includes only five general attributes of organizations: headquarter location, size, year founded, industry sector, and whether the organization conducts business online. Perhaps, other attributes need to be included for a better understanding of how organizational characteristics influence Big Data capabilities.

6.6 Suggestions for Future Research

Because of the exploratory nature of this study, there are numerous opportunities to extent this research via formal statistical tests so that a better understanding of Big Data maturity among organizations is gained. For example, additional questions that can be potentially explored in relation to the BDC and BDMM frameworks proposed in this study are as follows:

- Is there a causal relationship between regularity and rigor?
    - Is there a causal relationship between individual Big Data capabilities as well as the three broader areas of these capabilities (PB, OB, and TB)?

The study could also be extended in several dimensions. For example, the following research questions can be explored based on the exploratory results obtained in this study:

1. Are there systematic differences between countries and industries with respect to Big Data capabilities?
2. What is the missing attribute that would explain why some firms fall into one level of maturity and not the other?
3. Why the behaviours are so tightly clustered on both scales?
4. What would be the results if the survey includes different job titles?

CHAPTER 7

CONCLUSION

This chapter has three main aims. First, the chapter summarizes the main goal of this study - which is to develop a maturity model suitable for the era of Big Data. Second, it provides an overview of the chapters of this dissertation to show how various activities and accomplished objectives in each of the chapters contributed to the attainment of the main goal of the study. Finally, this chapter ends with a set of board directions and recommendations for further research in this area.

The main goal of this study is to develop a Maturity Model that can be used for evaluating organizational maturity in relation to Big Data. The model is meant to be an improvement over existing maturity models created, either explicitly or implicitly, for the previous computing eras of basic Data Processing (DP) and Business Intelligence (BI). With this main goal in mind, a number of smaller objectives have been met in several research phases comprising this research projects. These objectives are discussed in more detail below.

Chapter 2 discusses, in a chronological order, the developments in computing and data processing technologies during the 20th and 21st centuries. Four specific phases in development have been discussed together with specific technologies and processes driving these phases: (1) Primary Data Processing (before 1960's); (2) Advanced Data Processing (1960's to 1990's); (3) Business Intelligence (from 1990's to approximately 2005); and (4) the Big Data era (from 2005 onwards). This order was used as the basis for describing levels of the Big Data Maturity Model (BDMM) developed in this study. In addition to that, several definitions of Big Data were discussed. It was concluded that the original conceptualization of Big Data by Gartner in terms of the so-called "3V's" (volume, variety, and velocity) is the most fundamental one and can serve as a conceptual foundation for the BDMM. Also, with this definition of Big Data, the difference between Business Intelligence and the Big Data movement was articulated. Also, the articulation of this distinction between BI and Big Data made it clear that there was a need to develop a maturity model departing from the standards and expectations of the BI era and suitable for the new era of Big Data. This model should not only serve as a tool for assessing organizational maturity in relation to Big Data, but also as a road map for identifying a clear path of improvements for organizations wishing to develop capabilities in relation to Big Data.

Chapter 3 discusses the key theoretical perspectives employed by this study: the Socio-Technical Perspective (STP) and the Resource Based View (RBV) of the firm. Under STP, an organization is viewed as a combination of social and technical systems. The main reason for selecting this meta-theory is that it allows a more holistic approach to implementing Big Data initiatives in organizations. Consistent with this theoretical lens, this study focuses on people, technology, and organizational capabilities in relation to Big Data analytics. RBV explains how internal resources, including tangible and intangible ones, impact organizational performance. A fundamental proposition of RBV is that organizations should have resources that are heterogeneous and immobile, so that they can outperform their competitors and achieve Sustained Competitive Advantage (SCA).

Using STP and RBV, this study proposes a Big Data Capabilities (BDC) framework populated with key resources and capabilities needed for creating a sustained competitive advantage for an organization via Big Data analytics. The design of the framework fills in the gaps identified in relation to the previously developed BI maturity models. This BDC framework consists of three areas of capabilities and behaviours reflecting these capabilities: people, organization, and technology. Each capability area of BDC consists of several sub-dimensions representing various resources (e.g. hardware, software, networking for the technology dimension).

The proposed BDMM extends BDC, so that organizational maturity in relation to Big Data can be classified into a distinct level and measured numerically. Big Data capabilities are represented by various behaviours within each of the three areas indicative of these capabilities. Capabilities refer to organizations capacity to perform certain actions. Behaviours are the tangible representations of these capacities (Piccoli, et al., 2011). Such behaviours are measured at two levels: the process level and the organizational level. Each of these two levels are operationalized by a scale measuring regularity (process level) and rigor (organizational level) of a relevant behaviour. As organizations perform Big Data behaviours more regularly and rigorously, Big Data maturity increases. Organizations perform the behaviours regularly if a strict schedule is followed. It does not matter how frequently as long a schedule is adhered to. Organizations perform the behaviours rigorously if formal processes or procedures are followed. The proposed BDMM consists of four levels of maturity derived from the chronological progress of Data Processing. These four levels are as follows: Unaware, Dada Processing (DP), Business Intelligence (BI), and Big Data (BD). So each level of the BDMM corresponds to one of these historical periods. The proposed BDMM assumes organization capability improves over time through continuous improvement, which is measured through the regularity and rigor levels of relevant behaviours.

Chapter 4 describes how the data was collected to provide a preliminary, exploratory validation of the BDMM developed in earlier chapters. A survey is designed based on the developed BDC framework to assess organization behaviours in the following areas: People Behaviours (PB), Organizational Behaviors (OB), and Technology Behaviors (TB). Organizations are expected to perform these behaviours with certain levels of regularity and rigor, depending on their level of Big Data maturity. The organizational maturity level is measured by confirming how regularly and rigorously the firm performs a predefined set of behaviours. A continuous scale ranging from 0 to

100 is used with the following anchors: “not at all”, “somewhat”, and “extremely”.

The survey used for collecting the data was built and distributed using a popular online platform for surveys. We believe the platform allowed us to implement the survey in a way that is simple, secure, and protective of identity of survey respondents. Overall, the data collected does not contain personal information that can be used to identify survey respondents and their organizations. A clearance from the university’s Institutional Review Board (IRB) and the Human Subjects Committee was received to use the survey for data collection. The survey consists of 45 validated behaviours broken down into category behaviours: PB (6 items), OB (16 items), and TP (23 items). In addition to that, 6 items measuring the general demographics of survey respondents are included in Part 1 of the survey. This includes as question with disqualifying logic used to exclude participants whose job title does not match any of the following titles: CIO, CTO, CDO, and DDS. The selection of participants was done in line with the research objective and to ensure that each respondent had knowledge, experience, and awareness in Big Data analytics and how Big Data was used in his or her organization.

A well-known third-party research panel provider was hired to distribute the survey online among the target population and collect the data. A total of 528 responses were received. The number of completed survey responses that meet the accepted selection criterion (title of participant) is 246. A small number of completed forms had missing values. The missing values were replaced with averages.

Chapter 5 outlines the results of data analysis. Part 1 of the survey consists of 5 questions designed to collect details about the survey respondents and their organizations The results indicate that 88% of organizations surveyed are headquartered in North America. Small organizations (<50 employees) and very small organizations (<10 employees) combined make up less than 4% of surveyed organizations. Organizations founded after 1990 represent 67% of the sample. Organizations in the technology sector represent 33%. Finally, majority or 95% of organizations surveyed conduct online business. Information derived from the analysis of basic respondent characteristics may suggest some limitations in the sample with respect to representativeness, but this has to be explored in more detail by further research.

Part 2 of the survey consists of 45 questions organized into behavioural areas of PB, OB, and TB. For each question, participants are asked to report the degree of regularity and rigor with which their organization performs the desired behaviour. The differences between PB, OB, and TB are found small on both scales (regularity and rigor). The Bivariate Pearson Inter–item correlation test is performed to formally confirm that there are no systematic differences between individual items within the PB, OB, and TB capability areas. Results also show a high degree of correlation between all three areas. All the correlations are statistically significant. Results also indicate a statistically significant difference in mean values between regularity and rigor.

Thus, it is inferred that organizations prioritize regularity over rigor. This could be explained by the fact that organizations are generally driven by actions and accomplishments rather than how rigorously these actions were performed.

A hierarchical clustering approach is used to identify homogenous groups or "clusters" of organizations within the sample. The clusters are extracted by matching unique features of each organization with the features of each cluster. These clustering results are parallel to the BDMM levels based on historical periods in data processing. Empirically, these levels are based on the degree with which organizations perform Big Data behaviours regularly and rigorously. Results imply good discrimination between the three types of organizations represented by three clusters: Cluster H (high level of rigor or regularity), Cluster M (medium level of rigor or regularity) and Cluster L (low level of rigor or regularity). The discrimination is based on the average values of regularity and rigor with which various BD behaviours are performed within each cluster. Furthermore, the following findings and observations are made based on the calculated levels of maturity:

- Organisation location does impact Big Data maturity. For example, there are no differences between organizations headquartered in North America or elsewhere.
- Organizational size seems to have impact on Big Data Maturity. A Paired Sample t–test is used to determine whether mean value difference between large and smaller organizations is zero in

regularity. The results imply a systematic difference between large (> 2500) and other sized organizations and that the difference is statistically significant. Large organizations are mostly in Cluster H on both regularity and rigor scales.

- No pattern is detected for the relationship between the age of an organization and the Big Data maturity level.
- Organizations in the Consumer Goods/Retail sector tend to be in Cluster H. Also, organizations in Manufacturing and Finance tend to be in Cluster H, but only on the regularity scale. This implies that organizations in Manufacturing and Finance sectors are more mature with respect to scheduling Big Data activities but exhibit less rigor in performing these activities. No clear patterns were identified for organizations in other industries.

Based on the results of clustering, it is argued that the three clusters (L, M, and H) are related to BDMM levels of DP, BI, and BD discussed previously. The highest pattern of clustering analysis (Cluster H) implies that organizations have achieved the BD maturity level. Organizations in H perform all behaviours with a significant degree of regularity and rigor (over 90% and 89% respectively). The middle pattern of clustering analysis (Cluster M) implies the existence of BI maturity level. Organizations in M perform all behaviours with average regularity and rigor (74% and 66% respectively). The lowest pattern of clustering analysis (Cluster L) implies the existence of DP level. Organizations in L perform all behaviours with low regularity and rigor

(<56% and <36% respectively). The boundaries of maturity levels are calculated so that the BDMM can be used as a benchmarking tool to assess organizational capabilities in relation to Big Data. This may help managers identify missing Big Data capabilities and improve on them for the purpose of gaining competitive advantage via Big Data. Specifically, the BDMM can help organizations improve schedule and process consistency within organization with respect to Big Data activities.

7.1 Further Research

The exploration of Big Data capabilities performed as a part of this study has identified a wide range of further research opportunities for researchers. For example, limiting the survey participants to certain positions or titles (CIO, CDO, CTO and DDS) can potentially decrease the understanding of full range of perceptions in relation to Big Data activities within various organizations. Future research should consider involving a wider range of positions or titles to get a better view of Big Data activities and maturity within organizations. A socio-technical view probably requires a wider diversity of opinions from people in non-technical roles. This wider perspective many also provide valuable insights on how these different positions and titles in organizations influence Big Data maturity. For parsimony and practicality reasons, the survey included 45 questions only. Conducting a more comprehensive study that explores a wider range of Big Data behaviours can potentially provide valuable information for improving the BDC framework. For the same reasons, the survey included only five general attributes of organizations:

headquarter location, size, year founded, industry sector, and whether the organization conducts business online. Perhaps extending this list to include other attributes of respondents and organizations can provide additional insights on how these individual and organizational characteristics influence Big Data behaviours. Also a qualitative study with in-depth interviews and observations could be synergistic with this work and provide additional valuable insights

There are several other potentially fruitful opportunities to extent this research by adopting a explanatory approach. There may be a need to determine if there are a causal relationship between regularity and rigor. Moreover, the relationships among individual Big Data behaviours can be investigated as well. This can provide a more granular understanding of how organizations develop and practice various Big Data activities. Furthermore, systematic differences between countries and sectors can be investigated further. Another area explore would be the clusters of behaviours within each of the three areas of Big Data capabilities: PB, OB, and TB.

In conclusion, this research set out to develop a maturity model that that would take account of the new developments in Big Data analytics. This goal has been successfully achieved through the use of Big Data Capability (BDC) framework which has been formulated in this study. The BDC framework allows to describe the Big Data Capabilities in three areas: people, organization and technology. The developed model can be used as a benchmarking tool to assess organizational capabilities in relation to Big Data.

At the strategic level, the BDMM offers valuable guidance to managers when it comes to identifying gaps in Big Data capabilities within their organizations and formulating a clear roadmap for improvements in these capabilities. All this can help organizations achieve Sustainable Competitive Advantage (SCA) in the era of Big Data.

REFERENCES/BIBLIOGRAPHY

Ahuja, S. P., & Moore, B. (2013). State of big data analysis in the cloud. Network and Communication Technologies, 2(1), 62.

Alter, S. (2013). Work system theory: overview of core concepts, extensions, and challenges for the future. Journal of the Association for Information Systems, 14(2), 72.

Barney, J. (1991). Firm resources and sustained competitive advantage. Journal of management, 17(1), 99-120.

Barney, J. B., & Arikan, A. M. (2001). The resource-based view: Origins and implications. Handbook of strategic management, 124-188.

Bhargava, H., & Power, D. (2001). Decision support systems and web technologies: a status report. AMCIS 2001 Proceedings, 46.

Blanchard, B. S., & Fabrycky, W. J. (2006). Systems Engineering and Analysis, International Series in Industrial and Systems Engineering.

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: a socio-technical perspective, part II: the application of socio-technical theory. MIS quarterly, 11-28.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of ‘big data’. McKinsey Quarterly, 4(1), 24-35.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188.

Collett, S. (2011). Why big data is a big deal. Computerworld, 45(20).

Cortada, J. W. (2000). Before the computer: IBM, NCR, Burroughs, and Remington Rand and the industry they created, 1865-1956. Princeton University Press.

Courtney, M. (2012). The larging-up of big data. Engineering & Technology, 7(8), 72-75.

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. Foreign Aff., 92, 28.

Davenport, T. H., Barth, P., & Bean, R. (2012). How big data is different. MIT Sloan Management Review, 54(1), 43.

Davenport, T. H., & Dyché, J. (2013). Big data in big companies. International Institute for Analytics, 3.

Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business Press.

Davenport, T. H., Harris, J. G., De Long, D. W., & Jacobson, A. L. (2001). Data to knowledge to results: building an analytic capability. California Management Review, 43(2), 117-138.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

Dijcks, J. P. (2012). Oracle: Big data for the enterprise. Oracle white paper, 16.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. Communications of the ACM, 54(4), 86-96.

Douglas, M. (2013). Big Data Raises Big Questions. Government Technology, 26(4), 12-16.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5.

Ferguson, M. (2012). Architecting a big data platform for analytics. A Whitepaper prepared for IBM, 30.

Fujitsu. (2014). Infrastructure Products from Fujitsu – Ideal for Big Data [White paper]. Retrieved from https://sp.ts.fujitsu.com/dmsp/Publications/public/wp-bigdata-ideal-products.pdf

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

Glossary, G. I. (2013). Big Data defintion. Retrieved from http://www.gartner.com/it-glossary/big-data

Gobble, M. A.(2013) Big Data: The Next Big Thing in Innovation. Research and Technology Management, 56(1), 64-66.

Gorry, G. A., & Morton, M. S. S. (1989). A framework for management information systems. MIT Sloan Management Review, 30(3), 49.

Guide, P. (2013). Getting Started with Big Data. Retrieved from https://www.intel.co.jp/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf

Henschen, D. (2015). top big data analytics platforms. InformationWeek.

Hoffman, S., & Podgurski, A. (2013). Big bad data: law, public health, and biomedical databases. The Journal of Law, Medicine & Ethics, 41(1_suppl), 56-60.

Hortonworks. (2014). Hortonworks Data Platform : Cluster Planning Guide. [White paper]. Retrieved from https://www.puurdata.nl/wp-content/uploads/bk_cluster-planning-guide-20140502.pdf

Hribar Rajterič, I. (2010). Overview of business intelligence maturity models. Management: Journal of Contemporary Management Issues, 15(1), 47-67.

IBM. (2013a). IBM Selected by CenterPoint Energy to Enhance Customer Engagement with New Service Platform [Press release]. Retrieved from https://www-03.ibm.com/press/us/en/pressrelease/40621.wss

IBM. (2013b). Dublin City Council: Leveraging the leading edge of IBM Smarter Cities research to reduce traffic congestion [White paper]. Retrieved from https://www.ibm.com/midmarket/ie/en/att/pdf/IBM_DCC_130715.pdf

IBM. (2014). Automotive manufacturer increases productivity for cylinder-head production by 25 percent [White paper]. Retrieved from http://presidionwp.s3-eu-west-1.amazonaws.com/wp-content/uploads/2015/08/Daimler_Group.pdf

Jeon, Y. H. (2012). Impact of Big Data: Networking Considerations and Case Study. International Journal of Computer Science and Network Security, 12(12), 30-34.

Johnson, J. E. (2012). Big data+ Big analytics= Big opportunity: Big data is dominating the strategy discussion for many financial executives. As these market dynamics continue to evolve, expectations will continue to shift about what should be disclosed, when and to whom. Financial Executive, 28(6), 50-54.

Keen, P. G. (1981). Information systems and organizational change. Communications of the ACM, 24(1), 24-33.

Lahrmann, G., Marx, F., Winter, R., & Wortmann, F. (2010). Business Intelligence Maturity Models: An Overview. In Proceedings of the VII Conference of the Italian Chapter of AIS, Naples, Italy.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT sloan management review, 52(2), 21.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. Journal of Broadcasting & Electronic Media, 57(1), 34-52.

Loebbecke, C., Bienert, J., & Sunyaev, A. (2013). A parallel platform for big data analytics: A design science approach. International Journal of Science, Engineering and Computer Technology, 3(5), 152.

Lucas, H. C. (1971). Computer-based information systems in organizations. Sra.

Madden, S. (2012). From databases to big data. IEEE Internet Computing, 16(3), 4-6.

Marshall, C. (2012). Big Data, the crowd and me. Information Services and Use, 32(3-4), 215-226.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Mintzberg, H., & Quinn, J. B. (1992). ’The Strategy Process: Concepts, Contexts. Cases.

Needle, D. (2010). Business in context: An introduction to business and its environment. Cengage Learning EMEA.

O'reilly, T. (2009). What is web 2.0. " O'Reilly Media, Inc.".

Pearson, T., & Wegener, R. (2013). Big data: the organizational challenge. Bain Co.

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. Journal of Asynchronous Learning Networks, 16(3), 9-20.

Piccoli, G., & Ives, B. (2005). IT-dependent strategic initiatives and sustained competitive advantage: a review and synthesis of the literature. MIS quarterly, 29(4), 747-776.

Piccoli, G., Carroll, B., & Hall, L. (2011). Network exploitation capability: Mapping the electronic maturity of hospitality enterprises [Electronic article]. Cornell Hospitality Report, 11(18), 6-14.

Porter, M. E. (1980). Competitive strategy: Techniques for analyzing industries and competitors (Vol. 267). New York: free press.

Powell, R. (2001). DM Review: A 10 Year Journey. DM Review, February. Retrieved from www.dmreview.com

Power, D. J. (2007). A brief history of decision support systems. Retrieved from http://DSSResources.COM/history/dsshistory.html

Ravasi, D., & Schultz, M. (2006). Responding to organizational identity threats: Exploring the role of organizational culture. Academy of management journal, 49(3), 433-458.

Ray, S. S., Bandyopadhyay, S., & Pal, S. K. (2009). Combining multisource information through functional-annotation-based weighting: gene function prediction in yeast. IEEE Transactions on Biomedical Engineering, 56(2), 229-236.

Rijmenam, M. (2014). Think bigger: Developing a successful big data strategy for your business. Amacom.

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

SAS Inc. (2015). Big Data: What it is & why it matters.

Schouten, P. (2013). Big data in health care: solving provider revenue leakage with advanced analytics. Healthcare Financial Management, 67(2), 40-43.

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2011). Business Intelligence Maturity Models Toward New Integrated Model [White paper]. Retrieved from http://www.nauss.edu.sa/acit/PDFs/f2873.pdf

Simon, P. (2013). Too big to ignore: the business case for big data (Vol. 72). John Wiley & Sons.

Sint, R., Schaffert, S., Stroka, S., & Ferstl, R. (2009). Combining unstructured, fully structured and semi-structured information in semantic wikis. Fourth

Workshop on Semantic Wikis–The Semantic Wiki Web 6 th European Semantic Web Conference Hersonissos, (p. p.73).

Tan, C. S., Sim, Y. W., & Yeoh, W. (2011). A maturity model of enterprise business intelligence. Communications of the IBIMA.

TCS. (2013). Cultural challenges impact success of Big Data initiatives [Press release]. Retrieved from http://www.tcs.com/news_events/press_releases/Pages/Cultural-challenges-success-Big-Data-initiatives.aspx

Thamir, A., & Theodoulidis, B. (2013). Business Intelligence Maturity Models: Information Management Perspective. In Information and Software Technologies (pp. 198–221). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-41947-8_18

Treiblmaier, H., & Filzmoser , P. (2011). Benefits from Using Continuous Rating Scales in Online Survey Research. Thirty Second International Conference on Information Systems, Shanghai 2011, (pp. 1 - 14)

Trelles, O., Prins, P., Snir, M., & Jansen, R. C. (2011). Big data, but are we ready? Nature Reviews Genetics, 12(3), 224–224.

Vance, A. (2009). Hadoop, a free software program, finds uses beyond search. New York Times. March, 16.

Vasudeva, A. (2012). NextGen Infrastructure for Big Data. Retrieved from http://imexresearch.com/big_data_infrastructure.pdf

Vorhies, B. (2013). The Big Deal About Big Data: What's Inside – Structured, Unstructured, and Semi-Structured Data. Retrieved from http://data-magnum.com/the-big-deal-about-big-data-whats-inside-structured-unstructured-and-semi-structured-data/

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definition. arXiv Preprint arXiv:1309.5821.

Wernerfelt, B. (1984). A resource-based view of the firm. Strategic management journal, 5(2), 171-180.

APPENDICES

Appendix A. All Behaviours

Q.	Category	Code	Behaviour
1	People	PB1	1) We appoint specific individuals in our organization to carry out data analytics initiatives in their respective organizational units.
2	People	PB2	2) We train employees on how to use data analytics.
3	People	PB3	3) We encourage employees to use data analytics in their work.
4	People	PB4	4) We have a Chief Information Officer (CIO), Chief Data Analytics Officer (CDO), or a Chief Data Scientist (CDS) who is involved in developing organization strategies.
5	People	PB5	5) Our employees collaborate with universities to remain well-informed about developments in data analytics and influence data analytics curriculum.
6	People	PB6	6) Our employees participate in specialized groups, organizations, or conferences devoted to disseminating knowledge on data analytics.
7	Organizational	OB1	7) We use data analytics for organizational decision making.
8	Organizational	OB2	8) We use data analytics in several departments or organizational units.
9	Organizational	OB3	9) We integrate data across multiple departments and make the data available to our management.
10	Organizational	OB4	10) We perform data analytics across all organizational sub-units or sub-systems in an integrated fashion.
11	Organizational	OB5	11) We develop analytical models for improving organizational performance, managing risks, and finding new opportunities.
12	Organizational	OB6	12) We utilize data analytics to predict future performance of our organization by developing

			various future scenarios that can be changed or customized using model parameters.
13	Organizational	OB7	13) We develop decision making processes and policies based on data analytics.
14	Organizational	OB8	14) Data analytics is performed in our organization in a centralized fashion (i.e. there is a formal or ad-hoc unit responsible for data analytics across the entire organization)
15	Organizational	OB9	15) We integrate data analytics in our organizational processes.
16	Organizational	OB10	16) We formulate and implement data analytics goals and objectives for our organization.
17	Organizational	OB11	17) We use a formal approach to establishing and reviewing data analytics processes and policies.
18	Organizational	OB12	18) We use data analytics for improving organization performance.
19	Organizational	OB13	19) We integrate data analytics in our organizational governance structure and governance processes.
20	Organizational	OB14	20) We use data analytics to automate decision-making processes.
21	Organizational	OB15	21) We use data analytics for competing with other organizations.
22	Organizational	OB16	22) We use data analytics to create and exploit new opportunities.
23	Technological	TB1	23) We use traditional spreadsheet tools (e.g. Excel) for data management, analysis, and reporting.
24	Technological	TB2	24) We analyse structured data.
25	Technological	TB3	25) We analyse archival data (i.e. data at rest).
26	Technological	TB4	26) We use data analytics to analyse past organizational problems and to get a historical overview of our organizational performance.
27	Technological	TB5	27) We employ desktop computers for data management and storage.
28	Technological	TB6	28) We utilize Business Intelligence (BI) tools (e.g. SAP BI, IBM BI, Microsoft SQL Server - BI, etc.) to manage and analyse data.
29	Technological	TB7	29) We analyse semi-structured data.
30	Technological	TB8	30) We analyse real time or nearly real time data.
31	Technological	TB9	31) We perform predictive modelling on our data.
32	Technological	TB10	32) We perform data mining on our data.
33	Technological	TB11	33) We employ servers/mainframes for data

			storage and management.
34	Technological	TB12	34) We employ Business Intelligence (BI) tools to generate reports.
35	Technological	TB13	35) We use a data center as our main hardware/software platform for Business Intelligence (BI).
36	Technological	TB14	36) We utilize Hadoop/MapReduce and NoSQL databases (e.g. MongoDB, Apache Cassandra, etc.) to manage and analyse a variety of data
37	Technological	TB15	37) We analyse various types of data (e.g. transactional data, clickstream data, sensor data, videos, pictures, social media, text data, etc.)
38	Technological	TB16	38) We utilize in-memory and streaming database technology (e.g. SAP HANA, IBM Streaming Analytics, etc.) to analyse real time data.
39	Technological	TB17	39) We employ cloud solutions to store and manage a virtually unlimited volume of data.
40	Technological	TB18	40) We use scalable commodity servers to store and manage data.
41	Technological	TB19	41) We employ specialized reporting and data visualization tools that can be integrated with Hadoop or NoSQL platforms.
42	Technological	TB20	42) We perform traditional analysis techniques on our data, such as descriptive statistics, segmentation, sensitivity analysis, etc.
43	Technological	TB21	43) We perform advanced analysis techniques on our data, such as simulations, optimization, and predictive modelling.
44	Technological	TB22	44) We develop interactive or customizable reports for data presentation.
45	Technological	TB23	45) We utilize innovative approaches and new data sources in our data analysis (e.g. real-time analytics, video and voice analytics, image analytics, streaming analytics, geospatial analytics, sensor data analytics, text mining, machine learning, etc.)

Appendix B. Informed consent statements

Dear Participant,

This study explores the use of Big Data Analytics in organizations. This research can potentially benefit various organizations by identifying the factors that facilitate or hinder adoption and use of Big Data.

We would be very grateful if you could participate in this survey. Completing this survey should take approximately less than 20 minutes of your time.

Your participation in this survey is absolutely voluntary. You are free to choose not to participate in the survey or discontinue the survey at any time. Your responses will be anonymous. The information that you provide through this survey will not be linked to your identity or identify of your organization. Because this survey is anonymous, the researchers reserve the right for a variety of ethical uses of this data at their own discretion. The use may involve sharing this data with other researchers or the public.

The study is conducted by Abdulkhaliq Alharthi (Abu Dhabi University), Dr. Vlad Krotov (Murray State University), and Dr. Gabriele Piccoli (Louisiana State University).

Please feel free to contact Abdulkhaliq Alharthi at alharthi2000@yahoo.com if you have any questions or comments about this survey.

Thank you!

Appendix C. Instruction to Participate in Survey

PLEASE READ THE INSTRUCTION BELOW BEFORE PROCEEDING TO THE SECOND PART OF THE SURVEY

Please tell us how **Regularly** (this is related to the **schedule**) and how **Rigorously** (this is related to whether you follow **formal processes**) your organization performs each of the activities listed in the following categories:

1- Individual behaviors
2- Organizational behaviors
3- Technological behaviors

Please chose Regularity and Rigor on the percentage scale using the two slider bars below for each of the activities listed in these categories.

Extremely Regular means you follow a strict **schedule**. It does not matter how frequently as long as you adhere to the schedule. More than 90% can be considered as "Extremely regularly". If you don't perform a particular activity at all, then your response should be 0% ("Not at all"). If you perform a an activity occasionally (once in a while, but there is no fixed schedule), then your response should be closer to 50% ("Somewhat regularly").

Extremely Rigorously means you follow formal ***processes*** or ***procedures*** rather than performing the activity as you see fit. For example, if you have a clear set of policies and procedures that everybody follows in relation to a particular activity, then you should select 90% or above (corresponds to "Extremely rigorously)". If you have some policies and procedures but they don't exhaustively describe an activity or these policies and procedures are rarely followed, then your answer should be closer to 50% ("Somewhat rigorously"). If an activity is performed on a completely "ad hoc" basis and formal policies and procedures (if any) in relation to this activity tend to be ignored, then your answer should be closer to 0% ("Not at all")

Please note that even if you intent to select 0% ("Not at all") for a particular answer, you still need to use a slider to select 0%. Otherwise, the system will tell you that you haven't answered this question. This is done to make sure that none of the questions are omitted and the level of 0% is not selected by accident.

Appendix D. Cluster Membership

Regularly		Rigorously		Regularly		Rigorously	
Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership
1	1	1	1	42	2	42	2
2	2	2	2	43	3	43	1
3	1	3	1	45	1	45	1
4	2	4	2	46	3	46	1
5	3	5	1	47	1	47	1
6	2	6	2	48	1	48	1
7	2	7	2	49	1	49	1
8	1	8	1	50	3	50	1
9	3	9	1	51	3	51	2
10	1	10	1	52	2	52	2
11	1	11	1	53	2	53	2
12	2	12	2	54	3	54	1
13	2	13	2	55	2	55	2
14	3	14	1	56	3	56	1
15	2	15	3	57	1	57	1
16	1	16	3	58	3	58	1
17	1	17	1	59	1	59	1
18	2	18	2	60	3	60	2
19	1	19	1	61	2	61	1
20	3	20	1	62	2	62	3
21	1	21	3	63	1	63	1
22	3	22	1	64	3	64	2
23	2	23	2	65	1	65	3
24	1	24	1	66	2	66	2
25	3	25	1	67	1	67	1
26	2	26	2	68	3	68	1
27	2	27	2	69	1	69	1
28	1	28	1	70	3	70	1
29	3	29	1	71	3	71	1
30	3	30	3	72	1	72	1
31	2	31	2	73	1	73	1
32	1	32	1	74	3	74	1
33	2	33	2	75	1	75	1
34	1	34	1	76	3	76	1
35	2	35	2	77	1	77	3
36	1	36	1	78	1	78	1
37	3	37	1	79	2	79	1
38	3	38	1	80	3	80	2
39	2	39	1	81	2	81	2
40	3	40	1	82	1	82	1
41	1	41	1	83	2	83	2

Continue: Cluster Membership

Regularly		Rigorously		Regularly		Rigorously	
Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership
84	3	84	2	126	2	126	2
85	3	85	1	127	2	127	2
86	2	86	2	128	1	128	1
87	1	87	1	129	3	129	1
88	1	88	1	130	3	130	1
89	3	89	1	131	3	131	1
90	2	90	2	132	2	132	2
91	3	91	1	133	2	133	2
92	3	92	2	134	3	134	1
93	1	93	1	135	2	135	2
94	3	94	1	136	1	136	1
95	3	95	1	137	1	137	3
96	2	96	1	138	3	138	1
97	1	97	1	139	2	139	2
98	2	98	2	140	2	140	2
99	2	99	2	141	3	141	1
100	2	100	1	142	3	142	2
101	3	101	1	143	3	143	1
102	3	102	1	144	1	144	1
103	1	103	3	145	2	145	2
104	3	104	1	146	3	146	1
105	1	105	1	147	3	147	1
106	1	106	3	148	3	148	1
107	3	107	1	149	2	149	2
108	1	108	1	150	1	150	1
109	1	109	1	151	2	151	2
110	1	110	1	152	3	152	1
111	1	111	3	153	3	153	2
112	3	112	3	154	2	154	2
113	3	113	2	155	2	155	2
114	3	114	2	156	1	156	3
115	1	115	1	157	3	157	1
116	2	116	2	158	3	158	1
117	1	117	1	159	2	159	2
118	3	118	2	160	1	160	2
119	3	119	1	161	2	161	2
120	3	120	1	162	2	162	2
121	2	121	2	163	1	163	3
122	2	122	2	164	3	164	2
123	2	123	1	165	3	165	1
124	1	124	1	166	3	166	1
125	2	125	2	167	1	167	3

Continue: Cluster Membership

Regularly		Rigorously		Regularly		Rigorously	
Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership	Case	Cluster Membership
168	2	168	2	210	3	210	2
169	2	169	2	211	2	211	2
170	3	170	1	212	3	212	1
171	2	171	2	213	2	213	2
172	1	172	1	214	2	214	1
173	1	173	1	215	3	215	1
174	2	174	2	216	2	216	2
175	3	175	3	217	3	217	1
176	2	176	2	218	2	218	1
177	1	177	1	219	1	219	1
178	2	178	1	220	3	220	1
179	2	179	2	221	3	221	3
180	3	180	1	222	2	222	2
181	2	181	2	223	3	223	1
182	2	182	2	224	2	224	2
183	1	183	1	225	2	225	2
184	3	184	1	226	3	226	1
185	1	185	3	227	2	227	1
186	1	186	3	228	1	228	1
187	1	187	3	229	2	229	2
188	2	188	2	230	3	230	1
189	1	189	3	231	1	231	3
190	3	190	1	232	1	232	1
191	1	191	1	233	1	233	1
192	2	192	2	234	2	234	1
193	2	193	2	235	1	235	1
194	3	194	1	236	1	236	3
195	2	195	1	237	3	237	1
196	1	196	1	238	2	238	2
197	1	197	3	239	1	239	1
198	1	198	1	240	1	240	3
199	1	199	1	241	3	241	1
200	3	200	1	242	1	242	1
201	3	201	1	243	3	243	1
202	2	202	1	244	2	244	2
203	3	203	2	245	2	245	2
204	3	204	1	246	1	246	1
205	1	205	1				
206	3	206	1				
207	2	207	2				
208	2	208	2				
209	3	209	1				

Appendix E. Clustering Analysis – Regularly

Cluster - H				Cluster - M				Cluster - L			
Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.
1	95.36	56	95.38	1	69.62	56	79.31	1	58.18	56	62.93
2	92.44	57	90.13	2	64.58	57	83.11	2	53.82	57	66.69
3	84.42	58	96.31	3	67.62	58	79.73	3	61.04	58	51.87
4	87.36	59	96.98	4	79.89	59	71.89	4	54.44	59	18.56
5	87.00	60	100.00	5	68.78	60	86.76	5	64.04	60	38.64
6	94.89	61	87.38	6	68.02	61	71.96	6	52.49	61	15.38
7	90.49	62	93.91	7	77.09	62	68.69	7	51.44	62	27.47
8	85.98	63	98.82	8	66.56	63	76.16	8	64.71	63	60.22
9	95.27	64	90.56	9	69.44	64	71.62	9	24.56	64	50.67
10	91.96	65	87.58	10	70.51	65	72.33	10	64.20	65	39.42
11	100.00	66	90.44	11	73.49	66	76.60	11	58.42	66	58.30
12	100.00	67	92.16	12	75.20	67	76.91	12	57.71	67	60.91
13	93.24	68	98.36	13	77.16	68	83.40	13	56.80	68	59.58
14	97.40	69	86.22	14	64.74	69	70.22	14	57.02	69	51.18
15	90.64	70	85.56	15	74.07	70	79.90	15	52.91	70	51.56
16	91.84	71	95.31	16	78.49	71	82.00	16	66.62	71	59.40
17	94.89	72	97.38	17	70.78	72	59.64	17	65.58	72	58.04
18	96.53	73	89.64	18	74.91	73	73.96	18	52.58	73	61.71
19	96.84	74	85.67	19	70.29	74	81.31	19	60.58	74	54.62
20	82.98	75	97.20	20	82.49	75	82.53	20	65.07	75	43.13
21	97.04	76	94.91	21	73.00	76	70.38	21	54.38	76	63.69
22	100.00	77	92.09	22	70.52	77	74.89	22	60.53	77	57.64
23	94.64	78	89.67	23	68.11	78	80.56	23	42.02	78	59.69
24	99.89	79	88.07	24	74.29	79	75.20	24	58.49	79	62.22
25	90.60	80	87.36	25	80.60	80	79.71	25	65.84	Total Mean	53.9089
26	90.44	81	93.84	26	76.27	81	74.60	26	47.91	Std. Dev.	11.39668
27	88.62	Total Mean	93.0384	27	80.18	82	68.89	27	67.13		
28	94.91	Std. Dev.	4.61238	28	83.87	83	62.44	28	62.91		
29	99.56			29	65.89	84	75.87	29	53.34		
30	99.04			30	78.71	85	84.89	30	45.64		
31	95.27			31	68.22	86	73.09	31	53.58		
32	97.53			32	79.84	Total Mean	74.4214	32	58.76		
33	92.13			33	79.82	Std. Dev.	5.89536	33	61.49		
34	93.96			34	77.56			34	66.98		
35	89.76			35	73.93			35	65.56		
36	92.51			36	64.09			36	64.78		
37	87.93			37	69.31			37	62.27		
38	100.00			38	82.67			38	57.42		
39	100.00			39	67.27			39	57.87		
40	91.11			40	80.51			40	54.47		
41	99.44			41	82.36			41	61.27		
42	97.93			42	72.16			42	42.58		
43	92.11			43	80.00			43	50.24		
44	91.04			44	73.31			44	53.31		
45	88.16			45	74.60			45	53.78		
46	87.98			46	68.04			46	55.40		
47	98.67			47	77.31			47	59.41		
48	86.00			48	66.71			48	33.53		
49	89.33			49	70.49			49	60.36		
50	93.47			50	69.31			50	51.22		
51	90.78			51	82.89			51	21.31		
52	100.00			52	76.96			52	33.24		
53	97.71			53	77.67			53	44.51		
54	86.98			54	75.64			54	53.44		
55	91.11			55	65.87			55	48.09		

Appendix F. Clustering Analysis – Rigorously

Cluster - H				Cluster - M						Cluster - L	
Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.	Case	Avg.
1	98.44	56	89.29	1	55.38	56	76.18	111	58.44	1	11.40
2	91.80	57	100.00	2	54.89	57	73.22	112	71.18	2	37.71
3	83.47	58	97.44	3	53.78	58	65.49	113	77.36	3	23.58
4	87.04	59	81.61	4	61.02	59	78.23	114	76.51	4	33.44
5	84.87	60	83.60	5	75.82	60	77.64	115	69.78	5	3.07
6	95.29	61	93.47	6	55.51	61	74.84	116	54.89	6	41.56
7	94.67	62	95.82	7	65.64	62	56.40	117	50.40	7	45.75
8	91.78	63	97.89	8	65.22	63	71.37	118	81.22	8	33.33
9	90.98	64	100.00	9	51.62	64	76.44	119	73.98	9	39.02
10	100.00	65	88.13	10	57.98	65	62.75	120	79.13	10	33.42
11	99.96	66	91.18	11	74.91	66	75.11	121	75.58	11	38.53
12	89.73	67	98.78	12	55.42	67	46.29	122	66.82	12	35.84
13	92.60	68	88.56	13	63.16	68	76.00	123	77.53	13	23.42
14	89.51	69	91.22	14	66.24	69	69.53	124	62.07	14	42.33
15	92.96	70	93.02	15	51.67	70	53.89	125	75.16	15	37.18
16	95.02	71	83.44	16	78.69	71	62.22	126	79.24	16	33.81
17	96.04	72	98.58	17	55.51	72	58.29	127	69.53	17	13.10
18	95.31	73	82.67	18	65.36	73	71.37	128	74.04	18	41.93
19	85.78	74	95.42	19	52.18	74	80.00	129	58.38	19	19.44
20	85.53	75	90.62	20	66.96	75	72.62	130	59.22	20	29.22
21	98.04	76	86.56	21	70.71	76	68.58	131	60.64	21	24.07
22	89.40	77	96.57	22	76.96	77	57.44	132	63.51	22	30.87
23	99.84	78	90.84	23	69.22	78	52.16	133	64.71	23	42.72
24	92.87	79	89.20	24	53.33	79	78.40	134	76.53	24	45.33
25	83.42	80	88.80	25	54.73	80	68.24	135	75.78	25	41.57
26	93.24	81	93.11	26	76.82	81	61.47	136	65.18	Total Mean	32.0658
27	94.20	Total Mean	91.756	27	66.76	82	68.62	137	78.80	Std. Dev.	11.2512
28	82.84	Std. Dev.	5.4557	28	62.41	83	71.93	138	55.42		
29	99.49			29	71.37	84	77.56	139	76.76		
30	98.91			30	55.91	85	68.76	140	55.33		
31	88.24			31	65.42	86	77.73	Total Mean	66.5998		
32	81.44			32	74.29	87	61.91	Std. Dev.	9.32823		
33	97.98			33	71.96	88	79.26				
34	87.69			34	63.02	89	76.29				
35	91.24			35	55.31	90	74.45				
36	88.64			36	58.67	91	49.74				
37	93.76			37	57.44	92	80.16				
38	87.44			38	76.93	93	67.80				
39	100.00			39	53.86	94	71.18				
40	96.06			40	54.82	95	73.16				
41	95.64			41	69.20	96	68.17				
42	99.44			42	65.98	97	77.53				
43	97.87			43	72.29	98	76.13				
44	95.38			44	75.07	99	65.79				
45	84.33			45	55.44	100	47.49				
46	91.89			46	65.87	101	73.98				
47	93.36			47	80.81	102	72.84				
48	82.27			48	60.53	103	54.51				
49	93.67			49	64.23	104	80.16				
50	92.69			50	60.09	105	68.09				
51	81.96			51	67.16	106	57.22				
52	88.02			52	46.47	107	80.00				
53	83.62			53	62.36	108	47.02				
54	94.11			54	76.87	109	56.51				
55	86.69			55	63.69	110	69.72				

Appendix G. Clustering Analysis by Area: Regularly

Cluster – H (Regularly)

Case	PB	OB_	TB	Case	PB	OB_	TB
1	92.00	96.44	95.48	53	97.17	97.44	98.04
2	89.67	93.63	92.35	54	79.67	88.63	87.74
3	76.67	86.38	85.09	55	89.83	89.88	92.30
4	70.17	94.31	87.00	56	97.50	96.31	94.17
5	84.50	83.88	89.83	57	93.00	89.56	89.78
6	91.50	96.25	94.83	58	92.50	96.38	97.26
7	89.00	90.38	90.96	59	90.00	97.81	98.22
8	97.00	89.63	80.57	60	100.00	100.00	100.00
9	96.00	94.13	95.87	61	87.50	93.63	83.00
10	91.83	91.69	92.17	62	89.33	95.44	94.04
11	100.00	100.00	100.00	63	99.33	98.88	98.65
12	100.00	100.00	100.00	64	84.33	94.31	89.57
13	92.00	94.50	92.70	65	82.17	87.44	89.09
14	96.50	97.94	97.26	66	89.83	90.44	90.61
15	92.83	92.63	88.70	67	98.83	91.56	90.83
16	92.17	92.69	91.17	68	97.00	98.81	98.39
17	92.67	94.56	95.70	69	81.33	88.50	85.91
18	95.50	96.44	96.87	70	76.50	85.38	88.04
19	98.33	98.50	95.30	71	97.83	95.69	94.39
20	76.00	88.63	80.87	72	97.33	97.75	97.13
21	95.50	97.06	97.43	73	63.00	88.94	97.09
22	100.00	100.00	100.00	74	75.83	84.38	89.13
23	94.50	95.44	94.13	75	95.67	97.56	97.35
24	100.00	100.00	99.78	76	93.33	95.63	94.83
25	93.50	89.19	90.83	77	85.50	93.06	93.13
26	88.50	90.44	90.96	78	89.17	93.50	87.13
27	93.17	89.31	86.96	79	84.83	90.13	87.48
28	95.83	96.69	93.43	80	84.67	86.94	88.35
29	100.00	100.00	99.13	81	93.67	96.75	91.87
30	99.00	99.00	99.09	Mean	90.913	93.727	93.113
31	100.00	95.25	94.04	Std. Dev.	7.9637	4.5441	4.8178
32	96.83	97.44	97.78				
33	95.17	92.19	91.30				
34	92.17	93.56	94.70				
35	87.17	85.81	93.17				
36	92.83	91.50	93.13				
37	80.67	88.31	89.57				
38	100.00	100.00	100.00				
39	100.00	100.00	100.00				
40	90.00	92.56	90.39				
41	100.00	98.44	100.00				
42	99.00	98.56	97.22				
43	83.50	97.06	90.91				
44	87.67	90.69	92.17				
45	78.67	90.06	89.30				
46	89.50	92.81	84.22				
47	97.50	98.69	98.96				
48	80.33	85.63	87.74				
49	81.83	85.31	94.09				
50	77.67	98.06	94.39				
51	92.50	89.63	91.13				
52	100.00	100.00	100.00				

Cluster – M (Regularly)

Case	PB	OB	TB	Case	PB	OB	TB
1	67.17	70.69	69.52	57	87.67	83.38	81.74
2	64.50	64.50	64.65	58	86.50	79.06	78.43
3	67.00	76.69	61.48	59	62.67	71.63	74.48
4	78.17	79.06	80.91	60	90.67	86.13	86.17
5	53.33	74.69	68.70	61	66.83	66.06	77.39
6	72.17	70.50	65.22	62	72.00	69.94	66.96
7	68.50	76.13	80.00	63	79.00	78.63	73.70
8	68.33	63.75	68.04	64	51.00	82.63	69.35
9	83.67	72.00	63.96	65	59.33	74.13	74.48
10	73.17	72.38	68.52	66	78.00	77.50	75.61
11	71.50	73.19	74.22	67	73.00	77.50	77.52
12	76.67	76.00	74.26	68	86.67	83.94	82.17
13	83.50	75.19	76.87	69	78.33	67.19	70.22
14	49.00	50.31	78.89	70	77.37	80.20	80.52
15	80.33	67.69	76.87	71	81.83	82.94	81.39
16	64.33	78.94	81.87	72	58.17	60.44	59.48
17	71.17	69.94	71.26	73	60.50	77.88	74.74
18	77.50	81.38	69.74	74	78.50	81.63	81.83
19	65.50	69.88	71.83	75	82.67	83.81	81.61
20	83.67	81.81	82.65	76	81.67	77.25	62.65
21	74.50	72.88	72.70	77	75.83	77.75	72.65
22	62.88	70.25	72.70	78	77.00	83.94	79.13
23	60.33	71.63	67.70	79	56.67	59.00	91.30
24	74.83	74.25	74.17	80	76.67	85.75	76.30
25	79.33	83.81	78.70	81	74.00	72.94	75.91
26	81.83	80.27	72.25	82	57.33	69.06	71.78
27	83.50	79.88	79.52	83	61.17	59.81	64.61
28	74.83	85.75	84.91	84	86.83	83.63	67.61
29	80.00	66.50	61.78	85	91.00	81.88	85.39
30	85.67	81.06	75.26	86	69.33	60.06	83.13
31	64.50	69.63	68.22	Mean	72.664	75.251	74.363
32	77.83	77.31	82.13	Std. Dev.	10.511	7.554	6.831
33	93.50	77.56	77.83				
34	85.83	78.63	74.65				
35	74.00	74.44	73.57				
36	61.17	65.25	64.04				
37	70.83	72.19	66.91				
38	62.50	91.88	81.52				
39	66.17	68.31	66.83				
40	78.67	82.50	79.61				
41	82.83	84.44	80.78				
42	74.83	71.88	71.65				
43	87.50	80.00	78.04				
44	58.50	78.06	73.87				
45	71.33	80.13	71.61				
46	50.50	60.06	78.17				
47	72.83	77.75	78.17				
48	67.50	69.31	64.70				
49	50.50	72.44	74.35				
50	72.17	71.94	66.74				
51	80.83	83.13	83.26				
52	77.17	75.63	77.83				
53	78.83	76.94	77.87				
54	54.67	77.38	79.91				
55	59.40	88.75	56.48				
56	79.67	79.50	79.09				

Cluster – L (Regularly)

Case	PB	OB_	TB	Case	PB	OB_	TB
1	55.83	58.75	58.39	57	65.17	65.44	67.96
2	47.33	55.75	54.17	58	54.17	54.75	49.26
3	57.00	56.06	65.57	59	0.00	0.00	36.30
4	57.67	50.06	56.65	60	28.17	38.81	41.26
5	65.67	65.56	62.57	61	22.33	16.63	12.70
6	59.33	51.56	51.35	62	55.83	49.88	4.48
7	48.33	52.44	51.57	63	56.67	53.63	65.74
8	68.00	63.38	64.78	64	54.33	50.31	49.96
9	17.67	21.75	28.30	65	46.33	51.81	29.00
10	77.17	64.63	60.52	66	45.74	44.00	63.50
11	57.67	60.31	57.30	67	56.83	66.75	57.91
12	50.50	58.44	59.09	68	71.50	56.31	58.74
13	68.00	56.06	54.39	69	56.67	52.56	48.78
14	60.00	61.00	53.48	70	59.33	54.63	47.39
15	55.83	54.13	51.30	71	70.33	63.94	53.39
16	78.50	66.44	63.65	72	61.67	58.25	56.96
17	52.67	69.81	66.00	73	67.67	58.06	62.70
18	51.17	51.63	53.61	74	35.33	42.25	56.87
19	67.83	59.88	59.17	75	41.00	44.44	42.78
20	62.00	59.94	69.43	76	63.67	59.81	66.39
21	47.17	51.56	58.22	77	62.67	62.81	52.74
22	60.33	57.94	62.39	78	51.67	55.81	64.48
23	41.67	38.50	44.57	79	58.67	60.50	64.35
24	52.83	59.56	59.22	Mean	55.175	54.653	52.782
25	74.67	78.88	54.48	Std. Dev.	12.963	12.264	13.147
26	45.50	46.56	49.48				
27	65.17	65.75	68.61				
28	63.67	63.50	62.30				
29	53.33	45.19	59.02				
30	47.67	46.69	44.39				
31	49.67	54.25	54.13				
32	67.33	62.81	53.70				
33	60.50	65.88	58.70				
34	69.67	64.38	68.09				
35	51.67	71.75	64.87				
36	54.83	62.25	69.13				
37	66.67	59.50	63.04				
38	64.67	60.63	53.30				
39	59.00	59.06	56.74				
40	56.17	52.94	55.09				
41	55.50	62.38	62.00				
42	56.00	58.50	28.00				
43	53.17	62.75	40.78				
44	70.00	56.63	46.65				
45	54.33	53.38	53.91				
46	59.17	49.63	58.43				
47	58.64	60.33	56.35				
48	43.33	37.94	27.91				
49	65.83	61.75	57.96				
50	52.00	49.94	51.91				
51	27.17	19.81	20.83				
52	50.67	39.44	24.39				
53	43.00	56.31	36.70				
54	53.83	61.81	47.52				
55	54.67	57.56	39.78				
56	65.50	62.94	62.26				

Appendix H. Clustering Analysis by Area – Rigorously

Cluster – H (Rigorously)

Case	PB	OB	TB	Case	PB	OB	TB
1	98.33	98.56	98.39	51	73.33	80.69	85.09
2	88.83	92.56	92.04	52	83.83	85.19	91.09
3	84.67	84.88	82.17	53	78.83	82.31	85.78
4	74.67	95.56	84.35	54	83.17	98.19	94.13
5	81.33	81.50	88.13	55	87.17	87.75	85.83
6	95.00	96.38	94.61	56	90.83	88.50	89.43
7	98.83	93.00	94.74	57	100.00	100.00	100.00
8	89.50	91.44	92.61	58	97.17	97.56	97.43
9	91.00	91.25	90.78	59	72.50	81.85	83.83
10	100.00	100.00	100.00	60	83.00	81.88	84.96
11	100.00	100.00	99.91	61	95.33	94.38	92.35
12	88.83	90.50	89.43	62	93.50	95.25	96.83
13	94.17	93.25	91.74	63	95.17	98.19	98.39
14	93.50	89.44	88.52	64	100.00	100.00	100.00
15	84.33	90.25	97.09	65	88.00	95.38	83.13
16	92.17	95.50	95.43	66	85.17	92.06	92.13
17	96.67	96.06	95.87	67	99.33	98.88	98.57
18	96.67	95.94	94.52	68	87.17	87.75	89.48
19	86.00	86.69	85.09	69	91.33	89.69	92.26
20	73.33	84.88	89.17	70	98.00	91.44	92.83
21	85.28	100.00	100.00	71	89.33	84.00	81.52
22	92.50	90.69	87.70	72	98.33	98.75	98.52
23	100.00	100.00	99.70	73	83.33	84.00	81.57
24	92.33	92.13	93.52	74	100.00	97.62	92.70
25	59.83	88.19	86.26	75	63.33	91.81	96.91
26	94.17	93.50	92.83	76	87.50	72.13	96.35
27	93.17	95.94	93.26	77	98.33	96.56	96.12
28	80.00	81.25	84.70	78	87.50	91.31	91.37
29	100.00	100.00	99.00	79	83.50	92.31	88.52
30	99.67	98.81	98.78	80	87.83	88.75	89.09
31	91.50	89.13	86.78	81	87.50	95.13	93.17
32	83.00	85.75	78.04	Total Mean	89.472	92.075	92.131
33	97.17	97.94	98.22	Std. Dev.	8.7295	6.0385	5.5916
34	88.83	85.44	88.96				
35	92.00	91.25	91.04				
36	81.33	85.56	92.70				
37	93.17	93.81	93.87				
38	86.00	88.88	86.83				
39	100.00	100.00	100.00				
40	70.46	100.00	100.00				
41	96.67	94.63	96.09				
42	100.00	98.44	100.00				
43	99.33	99.06	96.65				
44	87.17	98.50	95.35				
45	80.33	85.94	84.26				
46	88.33	92.06	92.70				
47	86.50	94.25	94.52				
48	82.00	84.63	80.70				
49	98.17	93.00	92.96				
50	91.17	92.44	93.26				

Cluster – M (Rigorously)

Case	PB	OB	TB	Case	PB	OB	TB	Case	PB	OB	TB	Case	PB	OB	TB
1	51.83	53.63	57.52	40	49.17	54.00	56.87	79	78.83	85.31	73.48	118	79.67	80.69	82.00
2	51.50	56.25	54.83	41	62.67	67.00	72.43	80	74.33	68.44	66.52	119	85.00	81.63	65.78
3	53.33	52.81	54.57	42	73.17	78.56	55.35	81	59.67	61.25	62.09	120	68.33	80.31	81.13
4	58.01	58.63	63.48	43	71.17	67.31	76.04	82	75.00	66.13	68.70	121	77.00	77.75	73.70
5	78.50	75.00	75.70	44	77.00	74.69	74.83	83	72.01	70.38	72.99	122	64.00	70.19	65.22
6	60.51	48.13	59.35	45	61.50	56.56	53.09	84	66.17	79.00	79.52	123	75.67	77.00	78.39
7	64.50	67.44	64.70	46	57.50	67.13	67.17	85	71.17	70.38	67.00	124	57.50	63.56	62.22
8	57.00	75.19	60.43	47	79.50	85.00	78.23	86	77.83	76.19	78.78	125	90.17	88.88	61.70
9	48.67	52.44	51.83	48	58.17	59.69	61.74	87	70.67	66.19	56.65	126	72.50	78.81	81.30
10	45.50	43.75	71.13	49	44.19	70.20	65.30	88	82.00	78.75	78.90	127	57.17	70.63	72.00
11	67.17	77.81	74.91	50	51.00	60.69	62.04	89	56.17	78.13	80.26	128	44.83	69.94	84.52
12	50.00	55.63	56.70	51	71.33	62.06	69.61	90	67.06	82.99	70.43	129	62.50	63.13	54.00
13	73.50	64.81	59.30	52	45.67	46.06	46.96	91	49.67	52.14	48.09	130	58.67	57.00	60.91
14	68.50	69.31	63.52	53	69.50	66.13	57.87	92	79.83	80.00	80.35	131	71.83	61.13	57.39
15	50.00	48.81	54.09	54	77.17	78.69	75.52	93	75.83	66.88	66.35	132	75.33	62.63	61.04
16	73.33	79.25	79.70	55	67.00	67.13	60.43	94	63.00	68.75	75.00	133	54.17	64.00	67.96
17	55.83	51.06	58.52	56	85.33	76.94	73.26	95	67.50	69.94	76.87	134	68.67	75.94	79.00
18	80.67	60.44	64.78	57	70.83	75.63	72.17	96	75.41	69.75	65.18	135	78.67	85.13	68.52
19	61.68	54.06	48.39	58	67.83	66.50	64.17	97	73.83	76.94	78.91	136	68.67	61.75	66.65
20	72.17	72.38	61.83	59	87.00	76.14	77.39	98	100.00	99.81	53.42	137	81.67	80.94	76.57
21	71.67	74.31	67.96	60	86.00	78.69	74.74	99	77.83	67.98	61.13	138	48.83	50.69	60.43
22	81.33	79.31	74.17	61	72.50	74.50	75.70	100	35.83	48.69	49.70	139	76.17	66.13	84.30
23	66.67	70.45	69.03	62	45.50	62.38	55.09	101	67.50	75.69	74.48	140	56.67	54.63	55.48
24	55.83	55.31	51.30	63	70.46	71.90	71.24	102	61.50	74.50	74.65	Mean	65.73	67.05	66.51
25	60.83	53.06	54.30	64	73.50	76.31	77.30	103	66.17	56.94	49.78	Std. Dev.	12.86	11.21	9.765
26	87.50	75.06	75.26	65	61.00	60.31	64.91	104	86.00	82.81	76.78				
27	79.33	67.94	62.65	66	80.00	81.44	69.43	105	59.33	66.50	71.48				
28	51.00	49.69	74.23	67	49.83	46.25	45.39	106	65.83	50.38	59.74				
29	70.46	71.90	71.24	68	37.50	85.63	79.35	107	75.67	82.75	79.22				
30	61.33	55.88	54.52	69	75.83	71.06	66.83	108	33.17	37.81	57.04				
31	69.17	63.25	65.96	70	51.50	54.69	53.96	109	58.50	62.25	52.00				
32	79.83	70.19	75.70	71	56.50	63.31	62.96	110	48.30	66.76	77.36				
33	66.67	72.13	73.22	72	55.00	68.50	52.04	111	60.83	64.13	53.87				
34	78.50	68.38	55.26	73	70.46	71.90	71.24	112	76.17	71.06	69.96				
35	38.00	40.06	70.43	74	72.50	79.38	82.39	113	78.62	73.91	79.43				
36	61.67	60.94	56.30	75	75.83	66.69	75.91	114	71.67	74.81	78.96				
37	65.00	53.25	58.39	76	49.00	69.69	72.91	115	72.83	67.81	70.35				
38	82.83	72.81	78.26	77	57.67	56.88	57.78	116	43.00	57.31	56.30				
39	54.91	52.38	54.61	78	47.83	45.81	57.70	117	34.83	50.88	54.13				

Cluster – L (Rigorously)

Case	PB	OB	TB
1	11.83	13.19	10.04
2	34.17	33.06	41.87
3	17.00	20.25	27.61
4	31.67	36.25	31.96
5	10.83	1.88	1.87
6	42.50	38.63	43.35
7	30.50	49.50	47.13
8	45.00	27.44	34.39
9	39.33	34.38	42.17
10	53.50	38.44	24.70
11	67.00	36.00	32.87
12	46.00	41.06	29.57
13	31.17	21.56	22.70
14	27.83	45.25	44.08
15	31.83	38.56	37.61
16	26.67	37.06	33.41
17	0.00	0.00	25.62
18	30.17	41.88	45.04
19	22.33	23.06	16.17
20	52.00	52.56	7.04
21	23.67	33.88	17.35
22	36.83	39.31	23.43
23	62.91	48.11	33.70
24	51.50	49.06	41.13
25	56.08	42.38	37.22
Mean	35.293	33.709	30.080
Std. Dev.	16.592	13.775	12.404