

The Role of Big Data in Smart City Planning

Weiwei Jiao

A dissertation submitted to Auckland University
of Technology for the degree of Master of
Business (MBus)

2018

Department of Business Information Systems
Faculty of Business, Economics and Law

Abstract

With the development of urbanization, the rapid growth of urban population density has brought a raise in different types of services and higher demand for up-to-date infrastructures. In order to provide high-quality services and advanced infrastructure to citizens, it is important to understand what their requirements and demands are. The data scattered throughout every corner of the city provides effective information for smart city initiatives. This data comes from sensors, smart mobile phones, and social media, which reflect the big data era. It is essential to investigate big data effects in the smart city due to the increase in data resources and the number of residents. This dissertation undertakes a systematic literature review of 27 academic journals and conference papers which span from 2013 to 2017 in order to study the role of big data in smart city development. It aims to provide a systematic synthesis of the literature to reveal how big data technology supports the evolution of a city to become smart, the benefits of big data applications in the smart city, and the potential negative aspects of big data currently facing the smart city. This dissertation explains how data is collected in an urban environment. It then illustrates two computing infrastructures: cloud- computing and fog computing; three analytical methods: descriptive analytics, predictive analytics, and prescriptive analytics; and two data processing platforms: Hadoop and Spark. Several big data applications are also presented: smart transportation, smart energy, smart security, smart environment, smart healthcare, and smart education. The challenges of big data in the smart city are discussed by focusing on the inappropriate use on big data, big data's defects, and the growing demand for resources.

Keywords: big data, smart city, big data technology, big data applications, urban development, citizens' lives

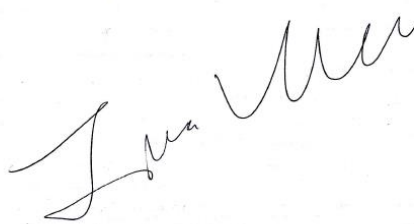
Table of Contents

<i>Attestation of Authorship</i>	3
<i>Acknowledgements</i>	4
<i>Introduction and background</i>	5
<i>Methodology</i>	9
<i>Findings</i>	12
Theme: Big data and its underlying process in the smart city	12
Urban big data collection	13
The computing infrastructure for big data management	14
Analytical approaches and techniques in big data analytics	16
Data processing platforms	18
Theme: Benefits of big data applications in smart city initiatives	19
Smart transportation	20
Smart energy	22
Smart security	23
Smart environment	25
Smart healthcare	26
Smart education	28
Theme: The challenges that big data encounters in the smart city	29
The risks of improper uses of big data	29
The limitations of data and information sharing	31
Continued growth in demand for resources	32
<i>Discussion and Conclusion</i>	35
<i>Reference lists</i>	38
<i>Appendix</i>	40
Table 1: List of papers used in the data analysis	40
Table 2: Themes, sub-themes and open codes for the findings	44
Table 3: Finding Summary	46

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signed:

A handwritten signature in black ink, appearing to be 'J. van der...' followed by several loops and a final flourish.

Date: 31st July 2018

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Angsana A. Techatassanasoontorn for her guidance, encouragement and patience. She directed and inspired me from preliminary ideas to how to better conduct this research. When I was not confident with my linguistic skills, she always motivates me and encourage me to keep going. When I had problems with the progress, she always provides useful suggestions with professional knowledge and warmth. Also, I would like to thank my proof reader for all the efforts in the grammar modification of this dissertation. Finally, I am grateful to my family and friend for their support and trust during this period.

Introduction and background

During the last five years, local governments have increasingly been using new technologies to implement various projects in order to build smart city. The term 'smart city' is associated with many areas of urban life: environment, business, industry, agriculture, transportation, and education. A smart city has been defined in different ways in the literature. Giffinger et al. (2007) synthesize the proposed six features of smart city into the following: smart economy, smart people, smart governance, smart mobility, smart environment, and smart living, and conclude that "a smart city is a city well performing in a forward-looking way in these six characteristics, built on the smart combination of endowments and activities of self-decisive, independent and aware citizens" (p. 11). Caragliu et al. (2011) defines when city management integrates traditional construction and information and communications technology (ICT) solutions to support human activities with a high quality of life and sustainable economic growth, then the city can be smart. Ahlgren et al. (2016) explain that smart cities are the vision of urban development, achieved by integrating various advanced ICT solutions to address a city's assets with the goals of environmental sustainability, an enhanced living benchmark, and an improved city economy. The goal of a smart city is to facilitate the development of city infrastructure and increase the contributions of government to society by utilizing advanced ICT. Therefore, a smart city not only involves technology, but also the environment and culture, which play significant roles in smart city initiatives (Dameri, 2017). Smart city projects should concern three aspects to enhance the smartness of a city: effectiveness, environmental consideration, and innovation. Effectiveness refers to the ability of a city to provide services to public or private subjects. Environmental consideration refers to the impact of environmental quality on urban areas. Innovation means the process of newly emerged technologies that are used in smart city development (Dameri, 2017). The creation of public values is identified as a complex performance which involves all the stakeholders in smart city planning, like residents, firms, local authorities, and governmental organizations (Dameri, & Rosenthal-Sabroux, 2016). A smart city brings benefits to several stakeholders in city planning. For example, it provides citizens with a better standard of life in respect to the environment (pollution), public services, health and well-being, and economic and work opportunities, and it helps public administrations

to offer public services with higher quality (and widespread e-services) to residents, such as public transportation. It also provides businesses with more opportunities to widen their business scope and a more creative and economic condition (Dameri, 2017; Jin et al., 2014).

Big data plays a significant role in a smart city environment. The prevalence of big data in the information system (IS) research field has increased over the last decade. Mauro, Greco and Grimaldi (2015) maintain about the main concept of big data can be related to three dimensions: information, technology, and its impact on society, through which it connects closely to the goal of a smart city. Therefore, big data “represents the Information assets characterized by such a High Volume, Velocity, and Variety to require specific Technology and Analytical Methods for its transformation into Value” (Mauro et al., 2015, p. 103). Chen, Mao, and Liu (2014) refer to several technologies related to big data, encompassing cloud computing and the Internet of Things (IoT). The roles of cloud computing and IoT technology are essential in integrating, mining, and analyzing urban big data. The goal of cloud computing is to store and process big data with its powerful computing capacity. The environment of IoT covers a huge number of sensors and various devices that are used to collect big data, and this involves network connectivity (Rose, Eldridge & Chapin, 2015). Jin et al. (2014) argue that big data will become outdated if city management handles it using traditional manual and semi-automated analytical techniques, such as spreadsheets. They encourage the use of IoT systems that incorporate sensors, data storage devices, and a computer terminal to deal with big data in city operations. Similarly, Rathore et al. (2016) indicate that it is difficult for cities to become smarter without using an IoT based system to work on the overwhelming amount of data. They suggest that a smart city system could establish sensors in different places to collect data and use a Hadoop ecosystem to analyze data and present results. Ahlgren et al. (2016) also emphasize that IoT systems have leverage on data volumes and smart city development, allowing a city to achieve benefits for the public such as decreased expenditure, increased productivity, and improved quality of life.

In addition to collecting and storing data based on cloud computing and IoT, there is a necessary technical process in analyzing and extracting information from big data – big data

analytics (Gandomi & Haider, 2015). Big data analytics have proven performance in acquiring useful information from the data produced by platforms (Ahmed et al., 2017). Also, big data analytics is an advanced method to perform real-time management (Babar & Arif, 2017). For example, a traffic control department may use big data applications to capture real-time, up-to-date information on the highway to forecast traffic or avoid incidents.

Big data can bring positive changes and ideas to the operating method of city management and transform and upgrade traditional industries to facilitate the intelligence of a city (Pan et al., 2016). Hence, big data should be explored and extracted appropriately so it can work for building a smarter city. Hashem et al. (2016) indicate that big data needs the right applications to assist with decision-making in the construction of smart city services and resources. Big data analytics applications have been utilized in some fields of smart city solutions like smart grid, smart education, smart transportation, and smart healthcare (Ahmed et al., 2017; Hashem et al., 2016). For example, healthcare specialists can use such applications to analyze a large amount of personal data for their patients and detect illness from the history of medical cases to improve the accuracy of diagnosis (Ahmed et al., 2017). Also, big data applications can reduce traffic accidents and traffic congestion, and indirectly optimize freight logistics (Ahmed et al., 2017). On the other hand, big data can also have negative implications because of its characters of volume, variety, velocity and veracity, such as data reduction, data structuring and data visualization (Chauhan, Agarwal & Kar, 2016). Also, human behavior is the important mediator between big data and smart applications (Chauhan, Agarwal & Kar, 2016), which means that all data generating systems need labor to operate and update. If people do not accept big data technology, there will be no space for big data to benefit the city development.

To conclude, it is obvious that big data is becoming essential to the implementation of a smart city. There are also potential risks in using big data in smart city initiatives. The purpose of this study is to review and synthesize the relevant literature with the aim of investigating the role of big data in a smart city based on the research questions: 1) how does big data support smart city initiatives? 2) what are the benefits of big data to smart city initiatives? 3) what are the challenges of big data in smart city initiatives? The paper is organized as follows. Section 2

discusses the methodology used in this study, which is a systematic literature review, and the research process associated with the methodology. Section 3 focuses on the findings retrieved from the identified literature. Section 4 discusses the findings, illustrates the knowledge we already learned, points out implications of future areas of research, and presents limitations of this study.

Methodology

This study used a systematic literature review as the methodology and followed the framework suggested by Templier and Paré (2015). This involved formulating the problem, searching the literature, screening for inclusion, assessing quality, extracting data, and analyzing and synthesizing the data. The first part of this research was to formulate the problems, which were as follows: to understand how big data supports smart city initiatives, to determine the different benefits of big data to smart city operations, and to identify the challenges of big data in smart city initiatives. Then, the second step was to undertake a search of the literature and identify the relevant academic articles. The AUT Library Homepage from the Student Digital Workplace was largely used to conduct the literature search as it provided easy access and quick links to other scholarly databases such as ACM Digital Library, JSTOR, ProQuest Computing, IEEE Xplore, and SpringerLink. For the preliminary scoping search, the initial search terms used were: 'big data and smart city initiatives', 'big data and city intelligence', 'big data and smart city planning', and 'big data and smart city benefits', which resulted in 15735 to 16526 journal articles and conference papers in the AUT Library databases. After browsing the titles and abstracts of the articles, it became clear that the search terms need to be progressively more specific. Hence, additional search terms were identified: 'big data and sustainable cities', 'big data and sustainable urban planning', 'urban big data and the smart city', 'smart city security issues', 'smart city and citizen privacy', and 'big data challenges and citizens'. This resulted in 7438 – 9630 journal articles and conference papers in the AUT Library databases. By reviewing the titles of these articles, the relevant articles were selected for additional screening of the abstract, introduction and conclusion sections. If the topic discussed in the article is related to this research, the article will be putting into a candidate pool. After identifying 40 candidate articles as the initial pool, forward and backward searches were conducted for further relevant papers (Bandara et al., 2015). In forward searching, Google Scholar was used to find articles that cited the paper that had been initially identified. For backward searching, the citations in the identified papers were examined to find relevant empirical papers. At the end of the literature search, 51 primary articles including academic journals and conference papers were found for inclusion.

The third step was to select relevant articles to be included in this study. After browsing through the primary articles, irrelevant papers were removed, which included those articles that had duplicate findings, conceptual papers, and those that did not focus on the association between big data and the smart city, instead referring to big data or the smart city as a single topic. In total, 29 papers remained on the list. After screening studies for inclusion, the next step was to read the full text to examine the quality of the relevant papers. During this process, notes and comments were made beside each primary article selected in the third step. After reading all the papers, their quality was assessed by comparing the notes made on each paper. After eliminating irrelevant papers and assessing the quality of papers, 27 articles were chosen for further coding and analyzing. The span of these articles was 5 years from 2013 to 2017. The list of articles is outlined in Table 1 in the appendix.

The next step was to extract applicable information from the 27 articles that could answer the research questions. A qualitative tool – NVivo – was used for data extracting. An inductive method was used, and a three-layer coding framework was created at this stage. First, open codes were created as phrases with supporting evidence based on the sentences and paragraphs from the articles. At the end of open-coding, 54 initial open codes were identified as the nodes in the NVivo interface. Some examples of open codes were ‘clustering techniques’, ‘cloud computing’, ‘save energy’ and ‘people’s privacy’. Second, the sub-themes were derived from merging conceptually linked open codes together. For example, open codes such as ‘smart parking’, ‘handle traffic issues’ and ‘improve bus service’, described the operations and benefits of smart city applications that involve with the field of transportation in urban services in relation to big data. Hence, these codes were categorized into the sub-theme ‘smart transportation’. Third, in the last layer, three main themes were built on the interrelationship between sub-themes and open codes, namely, ‘big data and its underlying process in the smart city’, ‘benefits of big data applications in smart city initiatives’, and ‘the challenges that big data encountering in the smart city’. This process required the integration and refinement of the identified sub-themes and open codes by repeatedly reading the selected articles. The analysis process will summarize the connections between each sub-theme and its related main theme, then interpret all sub-themes one by one in a top-down order. For example, the sub-theme

‘urban big data collection’ describes about how big data is being collected at the beginning of the whole process, thus, it will be discussed under the concept of the main theme – ‘big data and its underlying process in the smart city’. The full list of coding results is presented in Table 2 in the appendix.

Findings

This section presents the synthesized knowledge as findings based on the thematic analysis of the selected articles. The section illustrates how big data is being collected all around the smart city environment and shows the analytical methods and techniques, computing infrastructure and platforms, that are related to big data processing and management. The applications of big data that are beneficial to smart city development are also introduced in this section.

Furthermore, the challenges confronted by big data in smart city initiatives are stated. The findings are presented in accordance with the sequence of identified themes and sub-themes. Under the first theme – ‘Big data and its underlying process in the smart city’ – the sub-themes are urban big data collection, the computing infrastructure for big data management, analytical approaches and techniques in big data analytics, and data processing platforms. Under the second theme – ‘Benefits of big data applications in smart city initiatives’ – the sub-themes are smart transportation, smart energy, smart security, smart environment, smart healthcare, and smart education. Under the third theme – ‘The challenges that big data encounter in the smart city’ – the sub-themes are the risks of improper uses of big data, the limitations of data and information sharing, and continued growth in demand for resources. The summary of findings is presented in Table 3 in the appendix.

Theme: Big data and its underlying process in the smart city

In this main theme, the sub-themes show the various forms and sources of big data and the collection methods. The sub-themes then present two computing infrastructures for data collection and storage: cloud-computing and fog computing. To analyze the collected data, three analytical types are introduced: descriptive analytics, predictive analytics, and prescriptive analytics, with some analytical techniques discussed simultaneously. The two popular platforms of Hadoop and Spark, which process and analyze big data to build the smart city, are showed subsequently.

Urban big data collection

Big data is the overwhelming volume of data that is generated from heterogeneous sources everywhere in the city. Hashem et al. (2015) describe data sources as inclusive of sensing devices, internet data, transactional information, social media, and machine-generated data. Crittenden (2017) categorizes data into three types: passive data (data collected by fixed sensors), transactional data (data created by an individual or organization and can be tracked but its main purpose is not to provide information), and explicit data (data contributed actively and purposely such as data from mobile phones that is entered by users through the Internet). Mobile phone data provides the most information and is important to experts studying and measuring citizens' mobility and sociality at the urban level, the effect of events in cities, and the revival and improvement city infrastructure for citizens (Bergamaschi, et al., 2016). The most common data source in the emerging smart city is sensors that are deployed at different locations. Along with the infusion of IoT solutions, many urban sensors can be wirelessly connected to the operational terminal, which means that the authorities are able to install large-scale sensors at every corner of the city and are no longer limited to wire connections (Thakuriah, Tilahun & Zellner, 2017). For example, sensors are placed beside dams, rivers, and other open places to provide climate and water information, such as temperature, moistness, atmospheric pressure, rainfall, and water levels, and to collect weather-related data to improve the efficiency of the smart city (Rathore et al., 2016). Therefore, urban weather stations can detect environmental disasters in advance and take precautions to mitigate damage to cities and citizens. Another example of sensors embedded in urban lives is the use of wearable sensors such as sports watches to report on people's health conditions (Ang & Seng, 2016). Many up-to-date sports watches are able to wirelessly connect to a mobile phone to transmit data relating to heart rate, body temperature, and blood pressure, which can evolve into a solution for those requiring assisted living (Ang & Seng, 2016). Sensors not only present in the form of a machine, but also as human dynamics expressed on social media. Social media such as Facebook allow users to share geographic information, as people normally publish their posts or comments on events that take place in a real-time location (Ang & Seng, 2016). Once people post or comment about an event from the same location, the geolocation data can be

archived for event examination. If the authorities need to review or investigate the details of an event, they can find new clues from the information posted by individuals on social media and track them based on the location and time of the posts. Geographic information discovery can also be utilized to track criminal behavior, which may be useful for police who need to predict where, when, and why specific crimes are likely to happen (Bello-Orgaz, Jung & Camacho, 2016). In addition, scientists can collect knowledge from human-generated data in social media to identify people's perceptions of environmental changes and link these to machine-generated data to analyze and detect potential disaster (Ang & Seng, 2016). The data from a combination of different data sources also provides meaningful information. For example, the GPS data from a person's personal devices that tracks their movement around urban areas and the data from air quality sensors that is mapped on publicly accessible online platforms can support the measurement of environmental risks (Crittenden, 2017). Also, in the transportation domain, data is collected from sensors, cameras, radar, and GPS tools that are placed on streets and in vehicles and are used to determine traffic conditions (Alshawish, Alfagih, & Musbah, 2016).

All the data from multiple sources is collected and transmitted via the IoT within the smart city environment (Li, Cao & Yao, 2015). The IoT is the network that empowers seamless data collection and exchange among different objects (such as mobiles, tablets, and digital cameras) embedded with software, electric equipment, and sensors, and it provides a platform using the existing network to support remote sensing and control of those objects to enhance productivity, financial efficiency, and precision within the smart city environment. Therefore, IoT enables the interconnection of heterogeneous devices with each other through the network, which generates a large amount of data, thus creating innovations and new services for building smart cities (Hashem, et al., 2015; Kumar & Prakash, 2016; Rathore et al., 2016).

The computing infrastructure for big data management

The ICT infrastructure provides strong supports to the authorities in smart urban operations. Cloud-computing is now becoming a vital solution in its support of smart city big data collection, storage, and analysis because of its commercial services, low-cost, and geographical

distribution. Cloud-computing describes multiple kinds of computing models that enable many computers or clusters to connect and share capabilities delivered through a real-time communications network (Bibri, 2017; Hashem et al., 2016). Cloud-computing encompasses three types of services: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). These services connect city computing associates with diverse smart city applications and involve an arrangement of powerful machines distributed throughout the environment to approach the demand of different urban domains in respect of the use of big data analytics instruments, approaches, techniques, and technologies in the context of the smart city (Bibri, 2017). Cloud-computing is an efficient solution that provides technological means to facilitate data process in both the general public and specific urban domains. It meets the large-scale demand for computing resources by rapidly processing huge amounts of data with its cloud-based virtual infrastructure which integrates monitoring devices, visualization platforms, storage devices, analytics tools, and end-user delivery (Bibri, 2017; Hashem et al., 2016).

Fog computing can be regarded as an alternative computing architecture that extends cloud-computing from the central to the edge operation based on the IoT and big data analytics. The key idea of fog computing infrastructure is to use one or more cooperating near-user edge devices among the city's network to process the data instead of deploying servers in a cloud data center. This facilitates the performance of computation, storage, communication, and network services between edge devices and cloud computing data centers (Bibri, 2017; Tang et al., 2017). Fog computing is designed to fit well with big data applications based on its diverse characteristics that include a low latent period, location knowledge, mobility, and broad geographical distribution. It distributes computing tasks all through the IoT thus decreasing the network communication overhead and accomplishing high-performance computing competence (Tang et al., 2017). Tang et al. (2017) introduced a four-layer hierarchical fog computing architecture which from bottom to the top includes sensory nodes as numerous sensors, edge computing nodes, intermediate computing nodes, and a data center that supports data collection, storage, and analysis, and provides quicker response and more efficient management in smart cities.

Analytical approaches and techniques in big data analytics

Big data analytics encompass an accumulation of complex specified software applications and database systems that are kept operating by up-to-date data analysis software in the context of the smart city. These software can transform abundant urban data into valuable information needed to make accurate decisions based on insights from connections to various urban areas on issues such as transportation, environment, and energy (Bibri, 2017). This theme comprehends the different kinds of analytics and analytical techniques in big data analytics.

There are three main types of big data analytics in the smart city environment: descriptive analytics, predictive analytics, and prescriptive analytics (Watson, 2014). Descriptive analytics is used to describe what has/is occurred/occurring from the data that has already been collected in the city. It also provides an evaluation of why the events happened based on reports, dashboards, and real-time investigation, using, for example, a data-mining technique which primarily is used to extract knowledge from the data pool (Alshawish, Alfagih & Musbah, 2016). Data mining has become a meaningful way of investigating and exploring the rules of natural and societal changes surrounding people's lives, performance, and preferences, and the trends of social development and public opinion (Li, Cao & Yao, 2015). Predictive analytics predicts what will happen in the future. For example, it can predict the electrical demand for the coming year based on electricity consumption data over the past few years. Based on vehicular traffic-generated data, governments can predict which roads need to be expanded and where more parking areas are needed (Rathore et al., 2016). Predictive analytics can also predict potential crime risks and identify high frequency criminal locations and timelines for better preparation and victim protection (Rathore et al., 2016; Broeders et al., 2017). Predictive analytics integrates past and present data for further data analysis. Typical analytical techniques are implemented in predictive analytics such as regression and A/B testing (A and B refer to two variants). Regression identifies how the value of dependent variables will change along with the change in value of one or more independent variables. A/B testing utilizes different variants of a set to test data and to determine its effects on the situation. After determining the most effective variant, the expert can make better predictions based on how modifications on the variant can

be made to facilitate a circumstance (Alshawish, Alfagih & Musbah, 2016). Prescriptive analytics is often used to identify what is the best decision as it can suggest the optimal solution under a given condition, such as using navigation software on a mobile (Watson, 2014). The evolution of the analysis of big data is specifically moving from descriptive to predictive to prescriptive analytics (Watson, 2014), which will eventually identify the entirety of big data analytics progression.

Big data analytics can be implemented both offline and online. In the situation of a smart city, some domains like education, market economic and healthcare can be undertaken historical or offline analysis. However, making decisions within a specific timeline in order to provide high quality and more efficient services to citizens is encouraged by city builders (Al Nuaimi et al., 2015). Hence, real-time big data analytics acts as an increasingly essential mechanism in smart city operations. Many city authorities now use real-time analytics to manage how a city operates and is controlled (Kitchin, 2014). For example, the Office of Policy and Strategic Planning for New York city has created a one-stop data analytics center. The mission of this center is to draw together terabytes of data stream from a variety of agencies, which is then analyzed and managed by several groups of analysts with the aim of ensuring the city operates more efficiently and effectively (Kitchin, 2014). Similarly, the Centro De Operacoes Prefeitura Do Rio in Rio de Janeiro has also built functional analytics center deploying a virtual operations platform that provides access for city officials to extract real-time information (Kitchin, 2014). For instance, the platform allows police to upload situations and check real-time information updated by the data center when they out on tasks. In addition to platform work among official people, London has implemented a 'city dashboard' that enables citizens to acquire information on the weather, real-time transport schedules, water demand, and electricity usage and provides visualizations for citizens to be involved in joint development and monitoring of the city (Kitchin, 2014). Real-time analytics is essential and actively used in many smart city domains. In areas that are surrounded by unstable air pollution with harmful gases that are dangerous to people's health, real-time analytics can collect data on metals, carbon monoxide, sulphur dioxide, ozone, and noise. Harmful gases in the atmosphere can then be analyzed and alerts provided when any of the gases surpass a specific threshold, providing pre-warning to citizens

to avoid outside physical activities or to take protective measures when necessary. Besides outdoor environmental data, sensors that monitor the smoke and temperature within a building can detect a fire in real-time (Rathore et al., 2016). Real-time pollution monitoring can potentially aid people in maintaining their health. In the traffic domain, transportation authorities can gain real-time information on road incidents or growing traffic densities and provide updates for concerned citizens. Real-time traffic information allows citizens to choose alternative routes or to follow suggestions from online navigation devices to get to their destination, as well as saving fuel consumption and time, and reducing air pollution resulting from heavy traffic (Alshawish, Alfagih & Musbah, 2016; Rathore et al., 2016). Authorities can also guide citizens to find a vacant parking lot within a city with real-time monitor from all parking lots in the city (Rathore et al., 2017).

Data processing platforms

The computing operated models discussed above provide basic technological support for big data platforms such as Hadoop and Spark. Hadoop is an open source architecture and can be used for free for different commercial usages. It works by processing data in parallel through computing nodes and can handle large amounts of data in the shortest possible time (Alshawish, Alfagih & Musbah, 2016). Hadoop allows the distribution and processing of the load among the cluster nodes, thus improving processing capability. When dealing with unstructured data, it is also possible to add or remove nodes in the cluster in light of the requirements, and to make homogenous clusters with a different gathering of machines rather than the expensive choice of using one supercomputer (Bibri, 2017). Hadoop implements two main functions: the HDFS (Hadoop Distributed File System) and the MapReduce model. HDFS splits the large data files into various fixed-size blocks and stores these blocks in a large cluster. Each block is replicated on multiple nodes and each node, processing its stored blocks, can be managed in parallel (Alshawish, Alfagih & Musbah, 2016; Bibri, 2017; Rathore et al., 2017). HDFS is file based and can manage and store any type of data once the data is put in a file (Watson, 2014). MapReduce contains six phases: input, spitting, mapping, shuffling, reducing, and output. A MapReduce job first divides the input data files in HDFS into blocks and distributes them across

nodes. These blocks are then transformed into key-value pairs in the mapping process. The shuffling process further sorts the key-value pairs and produces the results. The MapReduce system then transfers the shuffle outputs to the reducing process. Lastly, the system merges and processes input in the reducing process, after which the final output is produced (Alshawish, Alfagih & Musbah, 2016; Bibri, 2017; Watson, 2014). This output can be delivered to a data warehouse where it might be combined with other data for further data analysis to facilitate the performance of big data applications (Watson, 2014).

While Hadoop processes data in parallel, Spark is also an open-source computing framework which is more efficient in real-time data processing. The Spark architecture enables the reuse of a working set of data over various parallel tasks which means it can bolster the processing operations while maintaining the versatility and fault tolerance of MapReduce. While the Hadoop MapReduce system is only suitable for batch processing, Spark's in-memory primitives provide performance up to 100 times quicker for specific applications by permitting user programs to load data into a clusters memory and query it repeatedly (Bello-Orgaz, Jung, & Camacho, 2016; Bibri, 2017; Rathore et al., 2017).

Theme: Benefits of big data applications in smart city initiatives

There is a diversity of urban data, for example, traffic flow data, energy data, environmental data, and human mobility data. Each kind of data can provide useful information for big data application in different urban domains. There are several common smart city applications under this theme. Based on the techniques and technologies described in the last theme, city planners can facilitate the performance and interaction of urban services as well as improve the life quality of citizens by implementing different kinds of big data applications. The sub-themes delve into these successful big data applications that transform a city into a smart city in terms of different urban domains and indicate a variety of advantages of smart city services that benefit citizens.

Smart transportation

Big data can offer values for more efficient decision making to address traffic issues. By investigating real-time big data, traffic authorities can monitor traffic performance and patterns and adjust traffic control, thus providing citizens with alternative routes to reduce city road congestion as well as to minimize air pollution from cars' exhausts (Al Nuaimi et al., 2015; Kumar & Prakash, 2016). They can also provide warning signs on roads with a high-frequency of incidents and predict traffic conditions when opening new roads to alleviate traffic problems. They can also investigate future trends by analyzing the historical data of accidents that involved different factors such as driver speed (Al Nuaimi et al., 2015; Hashem et al., 2016; Shukla, Balachandran & Sumitha, 2016). A smart transportation system can additionally reduce supply chain wastage by strengthening delivery capacity and optimizing freight movements (Al Nuaimi et al., 2015; Hashem et al., 2016). Traffic big data can be collected through various sensors, such as smart traffic lights and on-vehicle communication devices such as mobile phones (Al Nuaimi et al., 2015). For example, the city of Stockholm in Sweden has placed sensor-like radio frequency identification, laser scanning, and automatic photography on the downtown roads, which helps transport authorities detect the inflow and outflow of cars at rush hours in order to levy road tax and control bad traffic and the environmental conditions (Wu et al., 2018). The government of Los Angeles was the first city in the world to achieve 4,500 smart traffic lights that work simultaneously, which means that people can save 12% of driving time when they drive on the main roads in Los Angeles (Wu et al., 2018). The traffic light sensors send real-time data for processing and the lights change based on current traffic flow (Ismail, 2018). In addition, in Rathore et al.'s (2017) Intelligent Smart Transportation model, the data is gathered by road sensors and the vehicular network. Road sensors are installed at each junction and the vehicular network data is based on sensors affixed to moving vehicles, providing mobility information, speed, and location. The vehicular network data can also be analyzed against historical data and patterns to predict the risk of accidents as well as the potential cost or differential rates that might be applied to make amends for expanded financial risk (Elmaghraby & Losavio, 2014).

Many cities have successfully initiated smart transportation systems by taking advantage of big data. For example, Singapore has implemented a CtiyMIND platform to create a comprehensive and reliable city vision by using advanced analytics to connect various interfaces and outputs (from video and facial recognition cameras to mobile phones) with network intelligence, power sensors, and big data technologies (Kumar & Prakash, 2016). Zhejiang in China has deployed 1000 digital surveillance checkpoint systems that can continuously capture image and video data such as time, location, and vehicle information, which is stored in a centralized data centre for future retrieval (Kumar & Prakash, 2016). This therefore builds direct or indirect communication bridges between transport authorities and citizens and enables traffic police to easily obtain particular car characteristics and plate numbers through the big data network.

Moreover, bus services are being improved through the implementation of big data technology as part of the smart city revolution. Ordinarily, buses deliver passengers between bus stops throughout the city according to a particular timetable. However, real bus timetables do not usually align with specified timetables, which result in passengers missing their bus or extending their waiting time. Hence, in smart cities, authorities have installed numerous sensors and cameras in the buses and bus stops and use big data analytics to calculate the time variation between buses at each stop joined together with the quantity of delayed passengers. Feedback function is then used to create a controlled solution that measures the deviation of parameters from the expected result, and then interacts with the buses. With the analysis results, decisions can be made on whether to assemble more labor or buses, or to take appropriate actions to correct any inconvenience and achieve a better and higher benchmark of service (Ismail, 2016).

IBM scientists have estimated that more than 30% of traffic in 20 international cities is because of drivers searching for parking (Crittenden, 2017). Car park authorities can implement smart parking to alert citizens to the nearest free parking spaces or more suitable parking locations – information which helps citizens in their daily lives in smart cities. Smart parking is achieved by tracking vehicles entering and leaving different parking areas; thus, a smart parking lot can be designed by considering the quantity of vehicles within a zone or developed where there are

more vehicle flows (Rathore, 2016). This system uses big data analytics to reduce fuel consumption (Kumar & Prakash, 2016) and saves time, allowing people to spend more time doing other activities rather than parking (Rathore, 2016).

Smart energy

The concept of a smart energy system includes a smart grid and a smart metering system which is applied by connecting together the traditional power grid, renewable energy, and ICTs. The large scale of data is collected from the smart grid devices involving smart meters and sensors that produce data on energy consumption every 15 or 30 minutes. This data is sent to the cloud where there is a big data computing platform for data analytics. Data from the smart power grid facility is processed, controlled, and transformed into valuable foresights in the cloud, enabling real-time management. Energy companies can receive outage reports through the smart grid system instead of inspecting it themselves. Accordingly, in the smart energy infrastructure, big data analytics operates in three main areas: 1) practical analytics such as real-time visualization and grid simulation; 2) grid operation analytics such as renewable energy storage, power quality, power outage recovery, and grid optimization; 3) consumer analytics such as consumer behavior and the pricing strategy of usage time (Alshawish, Alfagih & Musbah, 2016).

The smart energy system allows inspection, analysis, control, and real-time communication inside the energy supply chain to help enhance the effectiveness of energy management, reduce energy waste, and increase transparency and dependability (Kumar & Prakash, 2016). The predictions can be performed in a near real-time manner from current energy production to consumption by the effective analysis on energy relevant data. This allows decisions to be made related to the level of the electricity supply based on the actual needs of citizens and the conditions of all influences (Al Nuaimi et al., 2015). Analysts can identify energy usage patterns by analyzing utilities data collected regularly over the years. Recognition of seasonality patterns helps forecast when and where regional demand will be high or low. Thus, energy companies can take actions in advance to achieve intelligent services (Kumar & Prakash, 2016). Smart

energy systems use dynamic pricing structures for electrical usage to flatten highest consumptions by charging high fees during peak hours and reducing costs at other times, which helps balance power usage and avoid power waste to enhance energy resource efficiency (Al Nuaimi et al., 2015). For example, increasing the prices of energy usage before 7 pm encourages people to use laundry or dishwashing machines after that time to reduce the power supply pressure and avoid power outage. Citizens can also use their smart phones to get living energy prices as well as adjust their usage based on their need and affordable prices, thereby reducing the pressure on energy costs (Bibri, 2017). Therefore, using particular pricing models based on the data of supply, demand, and production and consumption is an effective solution to enhance energy resource efficiency as a strategic goal (Al Nuaimi et al., 2015).

Moreno et al. (2017) introduced a smart building energy system to achieve energy efficiency that estimates indoor localization and predicts thermal comfort characterizations and energy consumption. This system collects data about people's activities and measures their indoor localization in the building based on a single active Radio-Frequency Identification (RFID - a kind of identification and traceability sensor) system (Bibri, 2017). It is important to gather accurate information on location for optimum energy efficient services. Once the energy consumption information is gathered, this system follows an optimization mechanism that is able to maximize comfort and simultaneously minimize energy waste. The benefits of the smart energy system within buildings is the ability to control the environment, including temperature and ventilation, observe system execution, monitor levels of carbon emissions, and manage lights according to occupants' activities, thereby minimizing energy waste (Bibri, 2017).

Smart security

Big data technology is increasingly achieving many benefits in relation to the public safety area and providing citizens with high quality services. Smart safety systems combine big data technology with traditional public safety to provide automatic sensing of incidents to achieve a more proactive, predictive understanding of situations instead of just reporting what happened after the incident (Alshawish, Alfagih & Musbah, 2016). Big data analytics can be used to reveal

crimes and help rebuild past events for police investigations. Unexpected connections play a huge role in various criminal cases, particularly in cases that happen in a rich data context and have specific repeated patterns. Pattern recognition is an essential requirement of big data analytics in the urban safety domain. For example, cases such as financial fraud show repeated patterns and diverse case materials can be used for profile construction (Broeders et al., 2017). In the security area, there are high expectations on predicting which location has a high crime risk and who is likely to be the criminal by analyzing crimes trends, victims' reports, and characteristics of likely criminals (Crittenden, 2017). Predictions can facilitate the likelihood of arresting offenders by providing more insight into their behavior. This information allows urban police to take preventive measures and warn individuals and organizations about potential risks (Broeders et al., 2017).

Moreover, due to the large scale and the diversity of the database, big data can perform more accurate risk analysis for public safety. Big data analytics can be used in regulatory, public security, and elementary investigations as part of the rules of safety policies. It can reveal unexpected relationships that can be used for risk profiles, resulting in the correct use of significant governmental resources, such as ambulances and police surveillance, and simultaneously achieve more effective inspections and investigations. Many cities use big data to achieve smart applications for their citizens' safety. Singapore is a notable smart city that has implemented a smart safety program called Risk Assessment and Horizon Scanning, under the control of the National Security Coordination Center, in order to achieve national security and protect public safety. This program operates by collecting and analyzing big data and proactively managing national threats like terrorist attacks, infectious diseases, and financial crises (Al Nuaimi et al., 2015). In South Korea, the relevant industry authorities, including the Ministry of Food, Agriculture, Forestry, and Fisheries and the Ministry of Public Administration and Security (MOPAS) have launched the Prevention of Foot and Mouth Disease Syndrome System to protect the health and safety of local citizens. The system collects relevant big data on overseas animal diseases, customs/immigration records, farm surveys, livestock migration, and livestock farmers to identify virus risks, predict the potential causes of disease and disaster, and warn citizens to pay attention.

Smart environment

Big data has enormous opportunities to protect the environment and thus enhance the quality of citizens' lives by monitoring air and water quality and reducing air pollution. The smart environment system offers real-time and fine-grained air quality information to better inform people about potential pollution hazards and assist their daily decisions (Ang & Seng, 2016).

The smart environment system collects environment related data by creating monitoring stations in the city and deploying sensors in cars. The data is then analyzed both in terms of the current and future air quality by using floating car data as the main source of car emissions, derived from personal GPS, Wi-Fi signals, smart phones, and loop detectors, which detect car movements at a certain time. Widespread sensors are used to identify air and water pollutants, which can lead to the elimination of multiple pollutants that are harmful to public health (Bibri, 2017). For example, there are only 22 stations in Beijing over an area of 50 square kilometers which means there are still many areas that cannot obtain air quality information through monitoring stations. For these areas, the big data approach combines direct air quality information data obtained from regions with monitoring stations with various other indirect data sources (such as social network data, real-time traffic data, historical time-series data, and urban layout) to deduce air quality information for the whole urban area (Ang & Seng, 2016).

Also, the smart environment system can provide weather information including temperature, humidity, rain level, and air pressure, that helps improve the country's agricultural land use by providing more accurate demand forecasts (Al Nuaimi et al., 2015; Ang & Seng, 2016). Through this system, future environmental changes can be predicted based on different requirements on geographic locations. Simultaneously, natural disasters like floods and earthquakes can be predicted to save many lives and resources (Bibri, 2017). For example, Japan has implemented a Disaster Behavior Analysis and Probabilistic Reasoning System (DBAPRS) for disaster management. This system obtained historical data from mobile sensors relating to the GPS records of the movements of 1.6 million people across Japan between 2010 and 2011 in order to analyze and simulate people's evacuation behaviors during the Great East Japan Earthquake and the Fukushima nuclear accident. The geographic location data included longitude, latitude,

and time period for each person. Based on the analysis results, national geographic authorities are able to predict population mobility and evacuation in many cities affected by disasters across Japan and use this information for the development of future disaster management strategies (Ang & Seng, 2016).

Many cities now use big data to reduce waste and improve waste management. Waste is a central aspect of society and the methods of eliminating relate to economic development, consumer habits, and environmental improvement. Big data can facilitate waste management and enable governments to better comprehend real-time social performance and enhance citizens' lives. Sensors are installed inside rubbish bins and provide data. Cities can use cloud computing with big data analytics to predict if bins are full enough for collection and to plan better rubbish collection routes in order to enhance collection efficiency (Crittenden, 2017).

Smart healthcare

Healthcare improvements can be achieved through improvements in preventive health treatment, diagnostic methods, cure tools, medical record management, and patient care (Al Nuaimi et al., 2015). The rapid growth of the world's population has meant rapid improvements in treatment patterns and many decisions behind these transforms are data-driven. In the past decade, healthcare suppliers have gathered, generated and controlled heterogeneous and e-health data that has aided the development of personalized medicine, disease prevention, and effective healthcare institutions. This data can also be used by insurance companies and some government agencies (Al Nuaimi et al., 2015; Bergamaschi et al., 2016). This enormous data availability guarantees unlimited possibilities for enhancing the sustainability and quality of healthcare systems (Bergamaschi et al., 2016). In South Korea, the Ministry of Health and Welfare has established a comprehensive Social Welfare Management Network to manage big data. It analyzes 385 different types of public data in 35 institutions, and comprehensively manages benefits and services offered by the central government and local governments and provides recipients with assistance. The appropriate investigation of huge healthcare data can help predict plagues, cures, and illness, enhance personal life satisfaction, and lessen the

number of preventable deaths. Doctors can enrich the sum of the information accumulated for specific patients' health problems and maintain a continuous record through smart healthcare devices related to home or clinics. This allows doctors to gain real-time data such as blood pressure, blood sugar, and sleep patterns through daily or on-demand monitoring, which helps them understand a patient's physical condition and mark potential health problems, thus providing a more comprehensive medical record. Furthermore, the analysis of a large amount of medicinal services data can empower doctors to identify the warning signs of serious disease at the beginning of treatment, that thus save many lives (Al Nuaimi et al., 2015; Hashem et al., 2016). Many developed countries have implemented smart healthcare systems to extract value from big data.

A typical example is Italy which has a large aging population. The effective management of patients' chronic diseases is essential in order to avoid complications and disability, as well as the sustainable development of the national economy. Organization of the medical system in Italy is hierarchical and decentralized. The local government is in charge of ensuring the overall goals and basic principles of the medical system. There are 21 local governments responsible for providing independent medical services through the local medical institution network. Due to this decentralized and independent system, medical data management systems cannot access and operate across different regions. With this background, the government has initiated a large-scale data analysis platform with the aim of unifying the analytical methods used in the administrative e-health management of records of regional units. This platform supports analysis including the entire processing cycle of big data, from distributed data collection and storage to each system and parallel implementation of analytics and results output. The data encompass four years concentrated records relating to 7 million citizens. This platform supports hiding the extreme diversity of regional information systems by supporting a common model of extraction and remapping of administrative records through documenting all interactions between each citizen and the public health system. Authorities can compare different regional public health systems through available big data to improve the quality of the prediction algorithm, thereby forecasting certain chronic diseases (Bergamaschi et al., 2016).

Smart education

Big data in the field of education is mainly targeted at relevant persons (e.g., students, teachers, parents, managers, and other staffs), infrastructure (e.g. schools, classrooms, computer facilities, educational sites, libraries), and other relevant features (e.g. exams, grades, school events, announcements, assessments, etc.). Big data allows analysts to analyze and extract useful information relating to educational trends and patterns, and to observe the shortcomings of the current education model, so as to improve and strengthen the available resources for education. Moreover, big data combined with ICT as a solution has laid the foundation for the intelligent development of education and will help build a knowledge-based society that will improve a country's competitiveness. Smart education not only improves the efficiency, effectiveness, and productivity of the education process (Ismail, 2016), but also strengthens educational control and assessment, thus providing better support for citizens' life-long learning (Al Nuaimi et al., 2015). Smart education applications attract people to actively engage in a solid learning environment so that they can adapt to the rapid changes in society. In addition, correctly generating and processing the required information based on real-time big data collected in the field has a positive impact on the level of knowledge, teaching approaches, and learning tools, and thereby the transference or acquisition of knowledge (Al Nuaimi et al., 2015). For example, educational institutions can achieve personalized teaching programs. Big data can mine learning information and generate a unique data trajectory for each student to gain knowledge of the student's performance and learning methods. By focusing on data analysis, teachers can not only understand a student's performance within a fixed time period, but they can also analyze and understand the student's existing knowledge in a more comprehensive and meticulous way in order to develop a personalized teaching program for the student (Al Nuaimi et al., 2015; Ismail, 2016). Online tools can support teachers to assess broader student learning abilities, such as the time they put into reading, the location from which they obtain electronic resources, and the speed with which they gain key knowledge. Smart education can also be applied to the academic research domain. For example, biologists can observe and analyze the records of both texts and digital pictures throughout the whole life cycle of animals by using big data analytics, which improves their research efficiency and new discoveries. Big

data analytics also applies to many other fields of science and research, such as astronomy, medical experimentation, manufacturing, environmental research, and economic and financial analysis (Ismail, 2016).

Theme: The challenges that big data encounters in the smart city

The rapid development of big data in recent years has brought tremendous benefits to the construction of smart cities and citizens' daily lives, as discussed in the previous theme. However, big data can also be seen as a double-edged sword, because if it is used improperly, it will restrict the progress of our society. Using this theme, the risks, negative aspects, and challenges of using big data to create smart cities will be discussed. These difficulties include the improper use of big data that leads to breaches in information security and personal privacy, and wrong judgement in issues, the flaws of big data itself, such as data quality issues, are hard to modify and various information is difficult to unify, additionally, the support from talented human resources and huge financial costs will be heavily burden with the continued mining of big data in the future.

The risks of improper uses of big data

The first risk discussed here is information security, which includes national confidential information security and personal information security. The smart city has established a public management and service platform that benefits multiple participants as well as achieving multi-level resource sharing. Any leak of reliable information may be used by hackers to invade the system and steal confidential data by programming malicious codes. The scale of a big data breach could be very large, leading to the collapse of the entire system. This could cause serious reputational damage and may have legally associated consequences (Kumar & Prakash, 2016; Wu, 2018). For example, ransomware temporarily suspended ticketing works in the San Francisco Municipal Transportation System (MUNI) in November 2016, resulting in out of service transportation and affecting citizens' regular travel (Crittenden, 2017). In addition, smart applications that operate across different objects also lack high security because data is

generated and delivered to various types of networks, some of which may be insecure. Most of the big data technologies lack of adequate security, including Hadoop (Al Nuaimi et al., 2015). From the point of view of personal privacy, the terminal sensor acquires a large amount of information related to citizens' transportation, consumption, medical care, and finance through scanning, photographing, positioning, and tracking (Wu, 2018). Sensor systems can be government or privately owned (Thakuriah, Tilahun & Zellner, 2017). Although this kind of personal and municipal data is very important for the development of smart cities, without reasonable restrictions, it is difficult to ensure personal privacy and simultaneously the security of personal information is threatened (Wu, 2018). Martinez-Balleste, Pérez-Martínez, and Solanas (2013) categorized citizens' privacy into five aspects: identity, query, location, footprint, and owner. When this information is provided to public service providers, other third parties can use it to associate users and their personal activities, profile users, and harvest information about their habits. In this regard, Elmaghraby and Losavio (2014) emphasize that location data is a key security issue. Individuals who using an intelligent transportation system that include GPS data, travel data, vehicle status, and contact numbers can be potential targets (Ismail, 2016). This system can track the destination and departure location during the time of use and store the actual route. It is also possible to access contact lists and messages, as well as content that may be kept secret for personal, commercial, or other reasons. For example, many people save home addresses on navigation software, allowing a thief to use this information in order to rob a house once the owner has left.

Nowadays, more and more national government departments have started to use the huge amount information available relating to the privacy of their citizens. More and more data has become available for surveillance, investigation, and prosecution, providing police with information on a large number of case investigations. In order to conduct more efficient and accurate case investigations, government agencies have begun to collect more and more private data and analyzing it, even to the extent of purchasing a large number of private databases. For example, the Dutch tax authorities are using private data, such as parking and transportation information related to private organizations (Broeders et al., 2017). As more and more diverse databases are used, personal data of people with no historical criminal record is

increasingly being collected and analyzed, which impairs their personal privacy. However, personal privacy rights can only be legally triggered by the principle of personal attack, which is not common in the case of big data. Thus, the huge scale of collection, storage, and analysis of big data by government agencies, as well as forced real-name systems on the Internet, results in threats to citizens' privacy and freedom of expression. This undermines civil liberties and could lead people to change their behavior and limit their freedom because they know they are being monitored. This also creates a disruptive influence on those who play a key role in the democratic movement, such as reporters, writers, and lawyers (Broeders et al., 2017).

The final risk discussed under this sub-theme is the negative consequences of the predictive analysis of big data. Big data analytics is based on historical data and existing data models for forecasting, but it can only provide a partial prediction of the future alongside probability maps, which means that those who deal with the results of analysis should treat it as a possibility rather than a straightforward result. Big data is characterized by "locality, selectivity, and representativeness," (Broeders et al., 2017) and can be distorted by the requirements used to capture the data. If not corrected, bias associated with each data set may turn into discrimination and unfair treatment of specific groups in society. When using data on a large scale, the results of big data analysis may complement each other, thereby amplifying social and economic inequalities (Broeders et al., 2017).

The limitations of data and information sharing

Data is facing some challenges in terms of the progress of its contribution to the smart city. The availability of data is often constrained because data is randomly collect among the entire urban environment. Sometimes data does not exist at all, sometimes there are problems with retention period, and sometimes different data platforms cannot interoperate. The quality of the data is also not guaranteed because the data may be outdated, damaged, biased, or even manipulated. As data becomes larger and larger, the amount of false data grows, meaning that data is uncertain (Batty, 2013). There are many new data formats, many of which are unstructured, and transforming them into structured formats for analytics requires management

and classification using advanced database systems (Al Nuaimi et al., 2015). For example, data from social media is in an unstructured form and needs to be pre-processed to make sense for customers and big data analytics tools because the integrity and quality of input data determines whether the output information is valuable. The structured feature of big data includes many different attributes that make it difficult to match the corresponding smart applications and are not easily accessible. These deficiencies can affect big data analysis and lead to erroneous predictions. Even if there are a variety of advanced tools to quickly analyze data, if the information used for decision making is inaccurate, it will waste time and resources and increase costs, which will reduce the efficiency of building smart cities (Broeders et al., 2017; Ismail, 2016; Shukla, Balachandran & Sumitha, 2016).

Sharing data and information between different public organizations and city departments is also a challenge. Each city agency and department usually have confidential information that belongs only to themselves or to customers, and most people are reluctant to share this information with proprietary data. Even sharing datasets inside an enterprise is difficult because each part might have its own independent database. The constraints of privacy terms and conditions make some data difficult to share between different objects. The challenge here is the conflict between privacy and the universality of big data.

Moreover, due to the diversity, heterogeneity, and differences in the basic structure and pattern of data sources in each department and organization, it is difficult to quickly create a unified mechanism for data to ensure that extracting new knowledge under the circumstances does not affect the data real-time feature (Al Nuaimi et al., 2015; Crittenden, 2017).

Continued growth in demand for resources

With the continuous emergence of new technologies, the human and financial resources that support this development are under tremendous pressure. From the perspective of human resources, the requirements for professionally skilled technicians are steadily increasing.

Hence, lack of advanced data analysis skills can be an obstacle to the effective use of big data

for urban management. The skills to manage and analyze large datasets and develop insights for effective decision making or operational improvement are currently scarce, especially in the public sector (Kumar & Prakash, 2016). A complete analysis team requires data scientists, developers, and analysts with professional domain knowledge (Balachandran & Sumitha, 2016). Analysts are generally classified into two categories: One is business intelligence analysts who are part of the business intelligence or analytics department and work throughout the organization. The other is business analysts who work in business units such as marketing and execute analytical work there. Data scientists are well-trained, highly skilled, and experienced professionals who discover new insights into big data. Each team needs different skills when dealing with big data, including a mixed knowledge of business, data, and analytics expertise, which is difficult to organize (Watson, 2014).

The cost of the resources needed to create infrastructure in a smart city construction cannot be underestimated (Crittenden, 2017), and it is a cost that cannot be predicted for the future. In order to develop a smart city, governments are installing an increasing number of sensors to monitor and record information, as well as using a large number of new technologies in order to promote citizens' lives. All of those implementations require numerous natural and human resources which will become more and more expensive. In addition, if a new smart application project is not executed properly from the beginning, it can lead to many problems, such as very high expenses, and might have negative impact on the city. For instance, the cost of testing smart traffic lights and signal systems is very high. These tests not just generate high expenses in terms of resources, but also cause consumption issues when the machine is deployed and tested. Moreover, there will be a need to replace expensive hardware and software to further develop and test smart city infrastructure and applications in the future (Al Nuaimi et al., 2015). The speed of data growth is also much faster than the speed of storage technology, which will lead to higher and higher storage costs. The rapid growth of data and high cost storage are becoming important factors that constrain urban security and the development of other systems. In China, most cities choose to shorten data retention time and reduce data storage quality to reduce costs. However, the disappearance of a large amount of valuable historical data also affects the available value of the analysis results (Li, Cao & Yao, 2015). At the same time, the

continuous development of smart cities requires more advanced technologies to assist this development, while maintaining and accelerating the pace of technology is very difficult and costly (Kumar & Prakash, 2016).

Discussion and Conclusion

The presented findings of this dissertation describe how big data is collected, processed, and applied, to support smart city initiatives, as well as illustrating several benefits of big data for smart city development. This dissertation also explains some of the challenges of dealing with big data and the insufficient aspects of big data in smart city planning. Data analytics involves actively collecting data across all corners of the city and analyzing the given databases based on the efficient development of big data technologies. A successful implementation of big data projects to support smart cities requires appropriate ICT infrastructure. While big data supports smart cities to improve living standards and service efficiency to meet the demands of citizens, ICT infrastructure provides a useful and unique solution to support application development and service operation. Sensors and detectors installed in different geographical locations are continuously monitoring and collecting data through IoT and interact directly with the social environment. Data from citizens' devices also provides a steady stream of information to relevant authorities and helps urban planners improve the quality of urban services and methods of interaction. After the data is collected, it needs to be managed and analyzed, which requires advanced computing infrastructure. Therefore, it is an important task for urban planners to encourage the continuous development of new ICTs to cater to the rapid development of cities in the future.

More and more big data applications can help city planners to more effectively use resources. A monitoring system reduces energy waste and consumption, and better allocates resources and plans to develop new energy sources. In terms of the lives of citizens, big data technology brings better services and lifestyles and more efficient work channels, which means that citizens in smart cities have a high-quality living and working environment. The outcomes of big data analytics can provide citizens with richer and more accurate information to help them make the right decisions, thus reducing wastage of time and money. When a city is enabled with big data technology, it means that the city's information is transparent and open. This increases the transparency of the information of every citizen involved and strengthens the communication between the government and its citizens. It also allows further understanding of the needs of

citizens, which provides useful information for creating more quality services and the development smart applications. On the other hand, to some extent, the expansion of information transparency disrespects personal privacy. There are still many people who are unwilling to expose their personal information; however, they have no choice because some situations involve government intervention. Apart from the information that must be available to governments, it is important that in smart cities management is able to set high-level security policies to control the exposure of customers' private information to third parties. Big data is not perfect, and it is sometimes disorganized; thus, it can be difficult to turn it into useful data for experts to accurately analyze and predict. In the process of dealing with big data, there are increasing skill requirements for data experts, and such talent is rare. It is vital for the government to encourage the development of talent in the data analysis industry such as set up technical training programs to attract people to learn new skills, so that more smart city solutions can be generated along with technology advances, and existing technologies can be further upgraded and refined. Finally, while implementing the new technologies and intelligent applications furthers the development of society and reduces the waste of environmental resources, it requires a large amount of financial expenditure to maintain. The consumption involved in smart cities will be endless since governments will only want their cities to move forward rather than backward, which will have a negative impact on a country's finances. To reduce the pressure of financial consumption, the government should take measures to actively remedy the situation and respond, such as increase the international trade development platform to enhance the country's economic growth, thereby supporting the construction of national facilities.

The contributions of this dissertation include its identification of big data processes, which involve collecting, processing, and analyzing data. It describes cloud-computing as well as fog-computing infrastructure. Also, it summarizes a number of analytical methods with several technical management tools. It further discusses the investment in smart applications supported by big data technology and the advantages that smart cities and their citizens obtain. Most of the smart city applications have been put into use in people's daily life to realize many benefits associated with smart city development. It also implies that smart big data applications have

more development space, and many problems are still hidden in people's lives. For example, the energy crisis and misdiagnosis are still happening. Lastly, it identifies the constraints and risks of big data development in smart cities. The conclusion is that although big data technology facilitates cities' rapid development, the development of technology is a double-edged sword that creates threats and pressure on citizens and governments while providing also providing benefits. This research promotes the further progress of current theory and provides topics for further investigation. These topics include how the government balances urban development and citizen privacy. Currently, many governments are pursuing urban development for the sake of national economic growth without showing concern for citizens' legal rights. Future research can also investigate the imbalance between governments' excessive reliance on and investment in technological development and city resources.

It is recognized that this study has the following limitations. The literature review presented in this study comprises journal and conference articles only. Therefore, the results of the literature survey do not reveal information about other types of literature, such as book chapters. Also, this study uses 27 articles from the period 2013-2017 and does not guarantee coverage of all empirical big data and smart city research. The topic of big data and smart cities is advance with the times and updating very quickly. The data collected in this study is the past work of other researchers rather than encompassing primary data; therefore, the results of the investigation are associated with this topic until 2017 only.

Reference lists

- Ahlgren, B., Hidell, M., & Ngai, E. C. (2016). Internet of Things for Smart Cities: Interoperability and Open Data. *IEEE Internet Computing*, 20(6), 52-56. doi:10.1109/mic.2016.124
- Ahmed, E., Yaqoob, I., Hashem, I. T., Khan, I., Ahmed, A. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. *Computer Networks*, 000, 1-13. doi: 10.1016/j.comnet.2017.06.013
- Babar, M., & Arif, F. (2017). Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things. *Future Generation Computer Systems*, 77, 65-76. doi: 10.1016/j.future.2017.07.029
- Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., & Beekhuyzen, J. (2015). Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support. *Communications of the Association for Information Systems*, 37(8), 154-204
- Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart Cities in Europe. *Journal of Urban Technology*, 18(2), 65-82. doi: 10.1080/10630732.2011.601117
- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171-209
- Dameri, R. P. (2017). *Smart city implementation: creating economic and public value in innovative urban systems*. Cham, Switzerland: Springer
- Dameri, R. P., & Rosenthal-Sabroux, C. (2016). *Smart city: how to create public and economic value with high technology in urban space*. Cham: Springer
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144. doi: 10.1016/j.ijinfomgt.2014.10.007
- Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler- Milanović, N., & Meijers, E. (2007, October). *Smart cities: Ranking of European medium-sized cities*. Retrieved from http://www.smart-cities.eu/download/smart_cities_final_report.pdf
- Hashem, I. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., & ... Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management*, 36748-758. doi: 10.1016/j.ijinfomgt.2016.05.002
- Jin, J., Gubbi, J., Marusic, S., & Palaniswami, M. (2014). An Information Framework for Creating a Smart City Through Internet of Things. *IEEE Internet of Things Journal*, 1(2), 112-121. doi:10.1109/jiot.2013.2296516

- Mauro, A. D., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. doi:10.1063/1.4907823
- Pan, Y., Tian, Y., Liu, X., Gu, D., & Hua, G. (2016). Urban big data and the development of city intelligence. *Engineering*, 2(2), 171-178
- Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. *The Internet Society (ISOC)*, 1-50
- Templier, M., & Paré, G. (2015). A Framework for Guiding and Evaluating Literature Reviews. *Communications of the Association for Information Systems*, 37, 112-137

Appendix

Table 1: List of papers used in the data analysis

#	Authors/Year	Articles	Types	Journal/Conference Name, Volume/Issues, Page Number
1	Crittenden, C. (2017)	"A Drama in Time": How Data and Digital Tools are Transforming Cities and their Communities	Academic journal	City & Community, 16(1), 3-8
2	Shukla, S., Balachandran, K., and Sumitha, V. S. (2016)	A Framework for Smart Transportation using Big Data	Conference material	2016 International Conference on ICT in Business Industry & Government (ICTBIG)
3	Al Nuaimi, E., Al Neyadi, H., Mohamed, N., and Al-Jaroodi, J. (2015)	Applications of big data to smart cities	Academic journal	Journal of Internet Services and Applications, 6(1)
4	Moreno, M. V., Terroso-Sáenz, F., González-Vidal, A., Valdés-Vela, M., Skarmeta, A. F., Zamora, M. A., and Chang, V. (2017)	Applicability of Big Data Techniques to Smart Cities Deployments	Academic journal	IEEE Transactions on Industrial Informatics, 13(2), 800-809
5	Thakuriah, P. V., Tilahun, N. Y., and Zellner, M. (2017)	Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery	Academic journal	Springer Geography Seeing Cities Through Big Data, 11-45

6	Alshawish, R. A., Alfagih, S. A., and Musbah, M. S. (2016)	Big data applications in smart cities.	Conference material	Engineering & MIS (ICEMIS), International Conference
7	Broeders, D., Schrijvers, E., van der Sloot, B., van Brakel, R., de Hoog, J., and Ballin, E. H. (2017)	Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data	Academic journal	Computer Law & Security Review, 33(3), 309-323
8	Li, D., Cao, J., and Yao, Y. (2015)	Big data in smart cities	Academic journal	Science China Information Sciences, 58(10), 1-12
9	Ang, L. M., and Seng, K. P. (2016)	Big Sensor Data Applications in Urban Environments	Academic journal	Big Data Research, 4, 1-12
10	Bergamaschi, S., Carlini, E., Ceci, M., Furletti, B., Giannotti, F., Malerba, D., Mezzanzanica, M., Monreale, A., Pasi, F., Pedreschi, D., Perego, R., and Ruggieri, S. (2016)	Big Data Research in Italy: A Perspective	Academic journal	Engineering, 2(2), 163- 170
11	Batty, M. (2013)	Big data, smart cities and city planning	Academic journal	Dialogues in Human Geography, 3(3), 274- 279
12	Elmaghraby, A. S., and Losavio, M. M. (2014)	Cyber security challenges in Smart Cities: Safety, security and privacy	Academic journal	Journal of advanced research, 5(4), 491-497

13	Rathore, M. M., Paul, A., Hong, W. H., Seo, H., Awan, I., and Saeed, S. (2017)	Exploiting IoT and Big Data Analytics: Defining Smart Digital City using Real-Time Urban Data	Academic journal	Sustainable cities and society, 40, 600-610
14	Tang, B., Chen, Z., Hefferman, G., Pei, S., Wei, T., He, H., and Yang, Q. (2017)	Incorporating Intelligence in Fog Computing for Big Data Analysis in Smart Cities	Academic journal	IEEE Transactions on Industrial informatics, 13(5), 2140-2150
15	Kumar, S., and Prakash, A. (2016)	Role of big data and analytics in smart cities	Academic journal	Int J Sci Res (IJSR), 6(14), 12-23
16	Wu, Y., Zhang, W., Shen, J., Mo, Z., and Peng, Y. (2018)	Smart city with Chinese characteristics against the background of big data: Idea, action and risk	Academic journal	Journal of Cleaner Production, 173, 60-66
17	Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016).	Social big data: Recent achievements and new challenges.	Academic journal	Information Fusion, 28, 45-59
18	Bibri, S. E., & Krogstie, J. (2017).	Smart sustainable cities of the future: An extensive interdisciplinary literature review.	Academic journal	Sustainable Cities and Society, 31, 183-212
19	Martinez-Balleste, A., Pérez-Martínez, P. A., and Solanas, A. (2013)	The pursuit of citizens' privacy: a privacy-aware smart city is possible.	Academic journal	IEEE Communications Magazine, 51(6), 136-141
20	Bibri, S. E. (2017)	The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data	Academic journal	Sustainable Cities and Society, 38, 230-253

		applications for environmental sustainability		
21	Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., and Chiroma, H. (2016)	The role of big data in smart city	Academic journal	International Journal of Information Management, 36(5), 748-758
22	Kitchin, R. (2014)	The real-time city? Big data and smart urbanism	Academic journal	GeoJournal, 79(1), 1-14
23	Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015).	The rise of “big data” on cloud computing: Review and open research issues.	Academic journal	Information Systems, 47, 98-115
24	Watson, H. J. (2014)	Tutorial: Big data analytics: Concepts, technologies, and applications.	Academic journal	CAIS, 34, 65
25	Pan, Y., Tian, Y., Liu, X., Gu, D., and Hua, G. (2016)	Urban Big Data and the Development of City Intelligence	Academic journal	Engineering, 2(2), 171-178
26	Rathore, M. M., Ahmad, A., Paul, A., and Rho, S. (2016)	Urban planning and building smart cities based on the Internet of Things using Big Data analytics	Academic journal	Computer Networks, 101, 63-80
27	Ismail, A. (2016)	Utilizing Big Data Analytics as a solution for smart cities	Conference material	Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference

Table 2: Themes, sub-themes and open codes for the findings

Main themes	Sub-themes	Open codes
Big data and its underlying process in the smart city	Urban big data collection	Data sources, sensors, mobile phone data, data collection, social media as sources, the IoT
	The computing infrastructure for big data management	Cloud-computing, cloud-based data storage, fog computing
	Analytical approaches and techniques in big data analytics	Clustering techniques, descriptive analytics, data mining, predictive analytics, regression, a/b testing, prescriptive analytics, real-time analytics
	Data processing platforms	Hadoop for handle big data, HDFS, MapReduce, Spark
Benefits of big data applications in smart city initiatives	Smart transportation	Smart parking, handle traffic issues, improve bus service, smart traffic, smart traffic light
	Smart energy	Smart grid, save energy, smart grid reduce consumption, green energy, outage management, smart building for saving energy
	Smart security	Identify crimes, surveillance, preliminary investigations for risk analysis, against fraud
	Smart environment	Big data for disaster detect, weather prediction, detect pollutant, infer air quality, solid waste management
	Smart healthcare	Improve healthcare, data for predict illness, disease prevention, improve quality of life

	Smart education	Analysing students' record, improve teaching and learning process
The challenges that big data encountering in the smart city	The risks of improper uses on big data	People's privacy, information safety issues, insecure network, citizens' privacy framework, government collects and use privacy data, encroachment civil liberties, limitations on prediction analytics, unfair treatment
	The limitations of data itself and information sharing	Data credibility issues, data quality issues, large scale of big data, difficulties in information sharing, hard to unify data forms
	Continued growth in demand for resources	skills update fast, scarce of professionals, financial cost pressure, uncounted costs in the future, huge cost on data storage

Table 3: Summary of findings

Themes	Sub-themes	Illustrative evidence
Big data and its underlying process in the smart city	Urban big data collection	Most data on people's daily life are collected from mobile phone and sensors, and IoT is the network that enables data collection and transmission.
	The computing infrastructure for big data management	<p>Cloud-computing is a commonly used computing infrastructure which provides solutions for big data processing.</p> <p>Fog computing is an alternative computing architecture that extends cloud-computing from the central to the edge operation based on the IoT and big data analytics.</p>
	Analytical approaches and techniques in big data analytics	<p>Descriptive analytics describe past and present events from the existing data and get meaningful results. Data mining is one of the methods being used with the description analytics.</p> <p>Predictive analytics is used to analyze data in order to predict useful insight for the future. Regression and A/B testing are most typical techniques in this analytical approach.</p> <p>Prescriptive analytics is to provide the optimized solution to people.</p> <p>Real-time big data analytics have a higher technical requirement which can provide more accurate results. It is currently being</p>

		used in several sectors such as pollution monitoring, traffic guidance, etc.
	Data processing platforms	<p>Hadoop is an open source architecture and it works by processing data in parallel through computing nodes and can handle large scale data in the shortest timeframe. It has two main functions: the Hadoop Distributed File System and the MapReduce model.</p> <p>Spark is data processing platform which is more efficient in real-time data processing and enables the reuse of a working set of data over various parallel task.</p>
Benefits of big data applications in smart city initiatives	Smart transportation	Big data can help decrease road congestion, reduce road incidents probability, improve bus services and enhance parking services by placing sensors or cameras around traffic lights or vehicle devices.
	Smart energy	Smart energy applications are designed for energy consumption by providing real-time data. It also uses dynamic pricing structure for balancing power usage and avoiding power waster to enhance energy resource efficiency.
	Smart security	<p>Big data technology can reveal crimes and rebuild past events for police investigation.</p> <p>It also helps predict the likelihood of crimes under a specific location.</p>

	Smart environment	<p>People using a smart environment system to enhance air and water quality by creating monitoring stations and deploying sensors in cars to avoid car emissions.</p> <p>Smart environment systems can also provide predictions on weather information or disaster warning.</p> <p>Many cities now use big data to reduce waste and improve waste management.</p>
	Smart healthcare	<p>The appropriate investigation of healthcare data can help predict plagues, cures, and illness, enhance personal life satisfaction, and lessen the number of preventable deaths.</p> <p>The continuous collection of patient's health data is helping doctors check patient's real time physical condition and prevent future diseases.</p>
	Smart education	<p>Smart education provides useful information about educational trends and patterns, and to observe the shortcomings of the current education model, which can improve and strengthen the available resources for education.</p>
The challenges that big data encounters in the smart city	The risks of improper uses of big data	<p>Information security issues can be categorized into two problems: national confidential information security problems and personal information security problems. In terms of national security problems, any leak of reliable information may be used by hackers to invade the</p>

		<p>system and steal confidential data by programming malicious codes. For the citizens' privacy, the personal information can be extracted from sensors on mobile phone or public surveillance, even governmental sectors cannot assure that they will not use citizens' personal information for some other uses.</p> <p>Sometimes the accuracy of predictive analytic cannot be fully trusted. In some cases, it will result in wrong judgments which might harm innocent people during the cases investigation.</p>
	The limitations of data and information sharing	<p>The availability of big data during the collection process is always uncertain, because the data sometimes does not exist or damaged. Also, sharing data and information between different organizations is becoming a challenge, it is difficult to create a unified mechanism for data to be flexible operated between different organizations.</p>
	Continued growth in demand for resources	<p>Big data technologies have been developed rapidly and the demand for new technologies is increasing, which means the expense on mechanical resources and human resources will be continuously increasing. The rapid growth of data and high cost storage are also becoming important factors that constrain urban security and the development of other systems.</p>

