BENEFITS AND CHALLENGES OF BIG DATA ON CLOUD FOR SMES AND GOVERNMENT AGENCIES

A MASTER'S THESIS

IN

INFORMATION SYSTEM ENGINEERING

ATILIM UNIVERSITY

by

ALAA HUSSAIN RASHED RASHED

JANUARY 2018

BENEFITS AND CHALLENGES OF BIG DATA ON CLOUD FOR SMES AND GOVERNMENT AGENCIES

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

ATILIM UNIVERSITY

BY

ALAA HUSSAIN RASHED RASHED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

THE DEPARTMENT OF INFORMATION SYSTEM ENGINEERING

JANUARY 2018

Approval of the Graduate School of Natural and Applied Sciences, Atılım University.

_____

Prof. Dr. Ali Kara

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Korhan Levent Ertürk

Head of Department

This is to certify that we have read the thesis "Benefits and Challenges of Big Data on Cloud for SMEs and Government Agencies" submitted by "Alaa Hussain Rashed RASHED" and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____ _____

Prof. Dr. Ali YAZICI
Co-Supervisor

Asst. Prof. Dr. Ziya KARAKAYA
Supervisor

Examining Committee Members

Asst. Prof. Dr. Ziya KARAKAYA _____

Asst. Prof. Dr. Çiğdem TURHAN _____

Assoc. Prof. Dr. Aslihan TÜFEKCİ _____

Date: January 19, 2018

I declare and guarantee that all data, knowledge and information in this document has been obtained, processed and presented in accordance with academic rules and ethical conduct. Based on these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

ALAA HUSSAIN RASHED RASHED

ABSTRACT

BENEFITS AND CHALLENGES OF BIG DATA ON CLOUD FOR SMES AND GOVERNMENT AGENCIES

Alaa Hussain Rashed Rashed

MS., Information System Engineering Department

Supervisor: Asst. Prof. Dr. Ziya Karakaya

Co-Supervisor: Prof. Dr. Ali Yazıcı

January 2018, 93 pages

Big Data and Cloud computing are the most important technologies that give the opportunity for SMEs and government agencies to gain a competitive advantage and improve their organizations. Big Data implementation requires investing a significant amount of money in hardware, software, and workforce. Cloud computing, on the hand, offers an unlimited, scalable and on-demand pool of resources which provide the ability to adopt Big Data technology without wasting on the financial resources of the organization and make the implementation of Big Data faster and easier. The aim of this study is to conduct a systematic literature review in order to identify the benefits and challenges of Big Data on Cloud for SMEs and government agencies and to make a clear understanding of how combining Big Data and Cloud Computing help to overcome some of these challenges. The last objective of this study is to identify the solutions for the challenges of Big Data. Four research questions were designed to determine the information that is related to the objectives of this study. Data is collected using literature review method and the results are deduced from literature.

Keywords: Big Data, Cloud Computing, SMEs, Government Agencies, Benefits, Challenges, Solutions

ÖZ

BULUT BİLİŞİM DESTEKLİ BÜYÜK VERİ'NİN KOBİ'LER ve DEVLET KURULUŞLARI İÇİN YARATTIĞI OLANAKLAR VE ZORLUKLAR

Alaa Hussain Rashed Rashed
YL., Bilişim Sistemleri Mühendisliği Bölümü
Yönetici: Yrd. Doç. Dr. Ziya Karakaya
Eş-Yönetici: Prof. Dr. Ali Yazıcı
Ocak 2018, 93 sayfa

Büyük Veri ve Bulut Bilişim, KOBİ'lere ve devlet kurumlarına rekabet avantajı kazandırmak ve kuruluşlarını iyileştirme fırsatı veren en önemli teknolojilerdendir. Bir yandan Büyük Veri uygulaması büyük oranda donanım, yazılım ve işgücü yatırımı gerektirirken, diğer yandan Bulut bilişim kuruluşlara yatırım maliyetini büyük oranda düşürecek Büyük Veri uygulama olanakları sunmaktadır. Böylelikle KOBİ ve Devlet kurumlarının Büyük Veri uygulamalarını hızlı ve kolay bir şekilde hayata geçirmeleri konusunda fırsatlar sunmakta ve özendirmektedir.

Bu çalışmanın amacı, sistematik alanyazın tarama tekniğini kullanarak veriler toplamak ve bu verilere dayanarak KOBİ'ler ve devlet kurumları için Bulut üzerinde Büyük Veri yapılarının yararları, zorlukları ve çözümlerini belirlemektir. Büyük Veri'de karşılaşılan zorulukların bir bölümünün Bulut Bilişim ile nasıl çözülebildiği net bir şekilde ortaya konulmaktadır. Bu çalışmanın amaçlarıyla ilgili bilgileri derlemek için dört araştırma sorusu tasarlanmış ve ilgili veriler alanyazın tarama yöntemi kullanılarak toplanarak sonuçlar derlenmiştir.

Anahtar Kelimeler: Büyük Veri, Bulut Bilişim, KOBİ, Devlet Daireleri, Faydalar, Zorluklar, Çözümler

To My Family

ACKNOWLEDGMENTS

I express sincere appreciation to my supervisor Asst. Prof. Dr. Ziya Karakaya for his guidance, insight and valuable directions throughout the research.

I would like to thank my co-supervisor Prof. Dr. Ali Yazıcı for his valuable feedback and support.

I would also like to thank my thesis committee members Asst. Prof. Dr. Çiğdem Turhan and Assoc. Prof. Dr. Aslihan Tüfekçi for their valuable suggestions and comments.

I would like to thank my wife, my parents and my children for their continuous support and patience during this period, none of this would have been possible without their support.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATION

CapEx - Capital Expenditure

DAG - Directed Acyclic Graphs

DBaaS - Big Data as a Service

DBMSs - Database Management Systems

EBS - Elastic Block Store

EU - European Union

Exabyte - 10^9 Gigabyte

GPS - Global Positioning System

HDDs - Hard Disk Drives

HDFS - Hadoop Distributed File System

HPC - High Performance Computing

IaaS - Infrastructure as a Service

ICT - Information and Communications Technology

IT - Information Technology

NIST - National Institute of Standards and Technology

NSA - National Security Agency

OpEx - Operational Expenditure

PaaS - Platform as a Service

PB - Petabyte

Petabyte - 10^6 Gigabyte

RAID - Redundant Array of Independent Disks

RDBMSs - Relational Database Management Systems

RDD - Resilient Distributed Dataset

RSS - Rich Site Summary

SaaS - Software as a Service

SLAs - Service Level Agreements

SLR - Systematic Literature Review

SMEs - Small and Medium Scale Enterprises

SSL - Secure Sockets Layer

Terabytes - 10^3 Gigabyte

TLS - Transport Layer Security

USA - United States of America

XML - Extensible Markup Language

YARN - Yet Another Resource Negotiator

Zettabyte - 10^{12} Gigabyte

CHAPTER 1
INTRODUCTION

1.1 Background

Use of Big Data technologies has grown rapidly in the last decade because of the exponential increase in data size and variety being produced by different sources. These include data being generated by static and mobile devices such as; sensors, social media, purchasing transactions, mobile devices, logs generated by systems and devices, data produced by cloud application, etc. Organizations and government can benefit from the analysis of this data to gain valuable insights that may help them to; improve the decision-making process, gain useful business insights, make fraud detection, and prevention, predict security threats and natural disaster, etc. These opportunities have created interest among companies to deploy this technology. Many of enterprise organizations use Big Data technology to earn a strategic competitive advantage (Vajjhala & Ramollari, 2016). SMEs can also benefit from Big Data technology to earn a strategic competitive advantage, but the problem is the lack of financial resources to invest in the technology and workforce (Vajjhala & Ramollari, 2016).

On the other hand, cloud computing provides the opportunity to benefit from the advantages of Big Data technology without needing to invest a significant amount of money on the expensive hardware and software that are necessary to deploy the Big Data environment. Cloud computing provides a reliable, fault-tolerant, scalable environment that can be exploited by SMEs and government agencies to provide a gateway to benefit from the analysis of Big Data.

Although deploying Big Data on Cloud environment provides many benefits to SMEs and government agencies, there exist many challenges faced by the organizations who are using these technologies. This study will review the benefits

and challenges of Big Data on the cloud environment and propose alternative solutions for the challenges being addressed in the literature.

1.1.1 Big Data Concept

Big Data term was presented first time to the computing world in 2005 by Roger Magoulas from O'Reilly media in order to define a massive size of data that cannot be processed and managed by conventional data management approaches due to the size and complexity of data (Ularu, Puican, Apostu, & Velicanu, 2012). Research on the development of Big Data as a scientific and research subject demonstrates that the term of Big Data was existing in research since the 1970s and has been included in publications only in 2008 (Halevi & Moed, 2012). Currently, the concept of Big Data is handled from various points of view addressing its effects in different fields.

1.1.2 Definition of Big Data

According to MiKE 2.0, the open-source standard for Information Management, Big Data is described by its size, including an extensive, complicated and independent collection of data sets, each with the potential to interact (Open Methodology, 2017). Additionally, an essential aspect of Big Data is the fact that conventional data management methods cannot manage it due to the unpredictability and inconsistency of the potential combinations (Gartner, 2017). Big Data has many characteristics called 7Vs: Volume, Velocity, Variety, Veracity, Variability, Value, and Volatility. These can be explained like:

Volume: Refers to the enormous amount of data which is collected by an organization. This data should be utilized to gain valuable knowledge.

Variety: Refers to the various type of data which can be included by Big Data. This data can be structured, semi-structured and unstructured.

Velocity: This characteristic refers to the speed at which data are generated and processed. The velocity of data increases over time. Some activities are significant and require instant responses, that is why fast processing maximizes effectiveness;

Veracity: Refers to the trust level of information used by leaders to take the right decision. Hence, it is essential to gain the proper correlations in Big Data for the future of business (Zikipoulos, Deutsch, & Deroos, 2012).

Variability: Data is continuously changing. Variability is another essential feature of Big Data which is usually confused with the variety characteristic. For example, Facebook or Google warehouses generates and stores diverse type of data. As well, if one of these various types of data is delivered to utilize for mining and making sense but this type of data provides various meaning.

Volatility: Means that data should be retained as long as required. That means data should have an expiry date, and after this date, data will be useless and insignificant and we can destroy it (Uddin & Gupta, 2014).

Value: Value characteristic refers to extract the valuable information from the collected data. Organizations need to realize the importance of utilizing Big Data to increase their profit, reduce the operational costs, and meet the customer's requirements (Lee, 2017).

1.1.3 Big Data on Cloud Computing

According to (Tian & Zhao, 2015; Gokhan, Karakaya, & Yazici, 2016) cloud computing and Big Data are complementary to each other and some of the Big Data problems can be resolved by cloud computing techniques and solutions. Cloud computing offers many benefits to SMEs and government agencies like hardware cost reduction, reduction in processing cost, and the ability to test the value of Big Data before committing significant company resources (Purcell, 2014).

On the other side, the main challenges for Big Data application using cloud environment are the security and privacy issues. In addition to the lack of trust in the service provider and unawareness about the service level agreements (SLAs), and the possibility of attacks on distributed locations are some of the challenges that are hindering the adoption of cloud computing infrastructure (Vajjhala & Ramollari, 2016).

1.1.4 Big Data on Cloud for SMEs

Big Data technology is considered to be one of the most critical drivers for SMEs growth according to the economic survey done by Oxford (2013). Big Data help SMEs to analyze and predict the customer's behavior and market. When Big Data is implemented in a right way, it can increase the productivity, flexibility, responsiveness, anticipation and ability to serve the customers better and improve the decision-making process (Sen, Ozturk, & Vayvay, 2016).

Big Data exploitation gives the opportunity to earn the competitive advantage and growth. Thus, SMEs must take a step forward and should not be afraid from the business-related risks, and trying to apply the right innovations which can help the organization to achieve the advancements towards the future (Abebe, 2014).

One of the essential challenges that affect the growth in SMEs is the lack of Big Data expertise in these organizations. A few of the challenges affecting SMEs growth, wherever there is a need to analyze the huge amount of data, can be summarized as:

- Ways to compete with enterprises and franchisees
- Lack of Big Data expertise
- Inability to invest in customer acquisition
- Inability to manage the supply chain, distribution and salesforce
- Lacking timely insights into market movements
- Inability to deliver large order size with short cycle times (Sen et al., 2016).

1.1.5 Big Data on Cloud for Government Agencies

In government, the most important objective is to preserve domestic tranquility, accomplish sustainable improvement, the security of citizens and fundamental rights, and enhance the economic growth and general welfare. The process of decision making in government agencies requires a long time and is conducted by discussion and universal approval of a big number of different actors, such as officials, interest groups, and ordinary citizens. Several well-defined steps are needed to minimize risk

and improve the effectiveness and efficiency of the decision-making process in the government sector. It follows that Big Data applications likewise are different between the private and public sectors (Kim, Trimi, & Chung, 2014).

Big Data has begun helping governments improve public health (monitor and predict the development of epidemics), enhancing law enforcement (preventing cyber-attacks, predict criminal activity, NSA-National Security Agency in the USA use Big Data to prevent terrorist attacks), natural disaster and resource management, etc (Bajaj & Johari, 2016).

The main concerns with Big Data applications in government sector are the security and privacy. Some of the government agencies might not want to share their data with the other government agencies or even to the public, so they need to adopt the appropriate technical infrastructure to manage Big Data. Other government agencies can use cloud computing to leverage Big Data and benefiting from it without a massive investment in hardware and software and workforce.

1.2 Problem statement

Although the focus of organizations in almost all sectors is on Big Data; there are still a set of organizations which are deemed not to have embraced the technology for achieving competitive advantage. Small and Medium scale Enterprises (SMEs) are in this category; they have failed to implement analytical and technical frameworks of Big Data (Ogbuokiri, Udanor, & Agu, 2015). SMEs have been the backbone of economic development in the emerging nations. The Cloud Computing and Big Data together have increased the availability of opportunities for SMEs which are deemed to be limited by its financial and organizational resources (Vajjhala & Ramollari, 2016). Hence, it is imperative that SMEs need to focus on adopting Big Data technologies to achieve better organizational benefits.

The importance of Big Data management in the Government agencies has been emphasized in developed nations, but its adoption in the emerging nation's scenario is least examined. In this regard, it is imperative to note that ample evidence needs to be provided to SMEs and government agencies in the developing nations to embrace Big Data and Cloud Computing technologies.

Big Data may change the way that is followed by SMEs and government agencies to make the right decision. Two reasons that may prevent SMEs and government agencies from leveraging Big Data technology are unawareness about Big Data technology and how they can benefit from it or the cost of leveraging Big Data technology or both. This research project aims to highlight the benefits of using Big Data on Cloud environment, the challenges faced by adopting these technologies and the solutions to solve these challenges.

1.3 Aim and objectives

The primary aim of this exploratory systematic literature review study is to identify and analyze the benefits experienced along with problems faced by the SMEs and government agencies dealing with Big Data and providing the solution to improve the chances for acceptance of Big Data among the government agencies and SMEs. In this regard, the following objectives are set:

- To explore the benefits of combining Big Data and Cloud Computing technologies.
- To explore the benefits of Big Data on Cloud for SMEs and government agencies.
- To identify and categorize the various challenges of Big Data faced by SMEs and government agencies
- To identify the existing solutions for Big Data challenges

The following research questions are set up to satisfy the objectives above:

- What are the benefits of combining Big Data and Cloud Computing?
- What are the various benefits of Big Data on Cloud for SMEs and government agencies?
- What are the different categories of challenges of Big Data faced by SMEs and government agencies?
- What are the existing solutions for Big Data challenges?

1.4 Scope and significance of the study

By tackling the challenges of Big Data and utilizing the analysis techniques to gain the valuable information, SMEs and government agencies can understand a variety of benefits, from enhancing the methods they already utilize it to find out new

capabilities and services. By laying the foundation for the efficient use of Big Data, the organizations can make better decisions quickly and improve the outcomes of their mission. Further, the Big Data allows them to identify and reduce inefficiencies in the work environment, eliminate the waste resources, mitigate the activities related to fraud and abuse which will lead to improving the productivity of their firms. The policy adopted by Big Data will provide the transparency of the service that will result in the reduced threads related to security and criminal activities.

The Importance of this thesis is presenting the benefits of Big Data on Cloud for two important sectors, the first one is the SMEs sector which considered as the key driver for the growth of the economy, and the second one is the government agencies sector. In addition to the benefits, this study will clarify the challenges which faced by the organizations when they are adopting this technology and offers the potential solutions to solve these challenges. Another importance of this thesis is identifying the research gap in this field.

1.5 Thesis structure

The remaining portion of the thesis is organized as follow: Chapter 2 contains the literature review of the existing works related to the Big Data. This section concentrates on the impacts of Big Data on SMEs and in government agencies. Chapter 3, contains the detailed information about the methodology used in this study. Also, Chapter 4 deals with the results from the previous section and provides analysis over the data. Chapter 5 describes the outcomes of the research work as the conclusion and the future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Concepts and Definitions

2.1.1 Big Data

Consistently, people produce 2.5 quintillion bytes of data and around 90% of the data in the today's world has been made over the past few years. This data has been produced from different sources such as sensors, social media sites, purchase transaction records, digital videos and pictures, and mobile phone GPS signs, etc. These data exist in different forms, namely (structured, semi-structured and unstructured). The process of this massive amount of data in different forms which cannot be processed by traditional database systems is called Big Data.

In 2005, the expression " Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in order to describe the great amount of data that traditional data management techniques cannot process and manage because of the intricacy and size of this data. An investigation of the Evolution of Big Data as a research and scientific topic demonstrates that the Big Data was available in research since the 1970s, however, has been included only in 2008 publications (Halevi & Moed, 2012).

As per the report of MiKE 2.0, the open-source standard for Information Management, Big Data is defined by its size and complexity, and independent sets of data with the possibility to interact with each other. Additionally, a significant aspect of Big Data which is the fact that it cannot be handled with traditional techniques of data management due to the unpredictability and inconsistency of the possible combinations (Open Methodology, 2017).

Definition of Big Data:

There are many definitions for the term of Big Data, and one of these definitions is that Big Data describes the process of the massive volume of structured or unstructured data in near real time, which cannot be processed by traditional database systems.

The National Institute of Standards and Technology (NIST) of US defines Big Data as: "Big Data consists of advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis."

Another definition is given by Gartner: "Big Data is high-volume, high-velocity, and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization." (Douglas, 2012). According to this definition, characteristics of Big Data can be summarized as 3Vs: volume, velocity, and variety.

Another definition is given by EMC (the cloud computing leader): ''Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from huge volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis.'' (Hu, Wen, Chua, & Li, 2014).

Big Data characteristics :

Big Data has many characteristics: Volume, Velocity, Variety, Veracity, Variability, Value, and Volatility. A short definition of each is given below :

Volume: is defined as the amount of data that collected by a company. This collected data will lead to extracting further significant information; The most salient trait of Big Data is high volume. Having the high volume feature in mind, many professionals express Big Data in terms of terabytes or petabytes (Russom, 2011). Such voluminous amount of data is hard to process, so, the amount of the data is a constant increase. Also, storing Big Data is another issue.

Velocity: this characteristic relates to the speed at which data are generated and processed. The velocity of data raises with time. At first, organization analyzed data

utilizing batch processing systems due to the expensive and slow nature of data processing. While the speed of generating and processing data raised, real-time processing started to be a criterion for computing applications. As reported by Gartner (2015) that 6.4 billion devices will be connected and would be in use around the world in 2016 and the number will increase to reach around 20.8 billion by 2020. In 2016, 5.5 million new devices were approximated to be linked every day to acquire, analyze, and share data. The improved data streaming ability of linked devices will accelerate the velocity constantly (Lee, 2017).

Variety: is defined as the different type of data that Big Data can comprise. The data is classified as structured, semi-structured and unstructured; The variety of the data available is the main factor for data to be generated at the high rate of speed. As noted before, data is either structured (those that had stable consistency in analytics) comprised of sentences or numbers that serve to be facile in noting and coding or unstructured (human and text language) comprised of music, history, pictures, etc that are abstract in nature and are hard to compute. There is available semi-structured data such RSS feeds and XML (Russom, 2011).

Veracity: This relates to the uncertainties, biases, untruths, impression, and missing values in data. This characteristic of Big Data measures the precision of data and how it is possible to utilize this data for analysis. The correctness degree of the datasets gathering to our systems will figure out how this data is essential for the challenge being analyzed and some researchers state that this is the most significant issue of Big Data (Nasser & Tariq, 2015).

Variability: is mean data are constantly changing. Another essential feature of Big Data which is often confused with the variety. For instance, Facebook or Google repository generates and stores diverse type of data. As well, if one of these different types of data is delivered to use for mining and to make sense, but this type of data provides various meaning, then this is the variability of data which mean is rapidly and continuously changing. The rates of change and variability of data generated by human and machine are much higher than in the data generated by processes. Variability is also associated with the performing of sentiment analysis. For instance, in the same tweets, a word may have an entirely different meaning. To be able to perform an adequate sentiment analysis, the advocates confirm that algorithms

should be able to understand and decipher (Zhang, Hu et al.,2015). However, this is still a big challenge (Sivarajah, Kamal, Irani, & Weerakkody, 2017).

Volatility: Means that data should be retained as long as required. That means data should have an expiry date, and after this date, data will be useless and insignificant and we can destroy it. Retention period of Big Data may exceed and cost of storage and keep this data secure could become expensive. In fact, Volatility becomes important due to the essential 3V of Big Data (Uddin & Gupta, 2014).

Value: This is additional characteristic of Big Data. Organization require to comprehend the significance of utilizing Big Data to maximize profits, reduce operational costs, and provide good services to their customers but simultaneously should think about the implementation cost of Big Data project. Data analysis will transform the data into a high-value strategic asset. IT professionals require to evaluate the costs and benefits of gathering and/or producing Big Data, select high-value data sources and build analysis capable of offering value-added information to managers (Lee, 2017).

Factors contributing to the growth of Big Data:

- **Lower cost of digital storage**: Developments in the technology sector such as the diminished cost of devices that facilitate storage have lent hope for providing solutions related to storage at a minimal price. Due to the low cost, the quantity of data processes and storage has been increased.
- **Affordable and faster communication technologies**: The speed at which the data is shared is quicker than the preceding conventional methods. For instance, a manually penned letter will take a few days to reach its intended addressee. However, sending the same through e-mail will involve only a few seconds.
- **The proliferation of applications and smart devices:** New devices such as tablets and smart-phones comprising innovative applications have strongly improved digital data's production (Singh, Srivastava, & Johri, 2014).

2.1.2 Big Data Frameworks

Big Data processing without having Big Data framework is very hard and need to invest a huge amount of money. Big Data framework can store and distribute this data across hundreds of commodity hardware running in a distributed fashion (Hu, Wen, Chua, & Li, 2014). There are many Big Data frameworks but the most popular frameworks are Hadoop, Storm, Spark and Flink which are open source solution. Frameworks can be categorized into three categories based on types and purpose of data processing approaches; namely batch processing like Hadoop, stream processing like Storm and hybrid processing like Spark and Flink (Karakaya, 2017). The following section gives a short description of each category, framework and the main components of it:

Batch processing category: Refers to the batch processing of staying data, most popular available solution for this type of processing is Hadoop framework.

Apache Hadoop

Hadoop is the most popular open-source software implementation of Big Data framework that offers batch-oriented, distributed, scalable and reliable data processing. The main common components of Hadoop are:

Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.

Hadoop YARN: A framework for job scheduling and cluster resource management.

Hadoop MapReduce: A YARN-based system for parallel processing of large data sets (Karakaya, 2017).

Stream processing category: Refers to the operations of processing data once it received, most popular available solution for this type of processing is Storm framework.

Apache Storm

Apache Storm is a free and open source distributed real-time computation system. Unlike Spark, Storm is pure stream-only data processing framework. The data is processed once it received into the system. Storm uses DAG (Directed Acyclic Graphs), which is called "topology" in their jargon, in order to orchestrate its tasks. This enables users to perform any type of filters, cleaning algorithms and so on (Karakaya, 2017).

Hybrid processing category: Refers to both approaches of processing data; batch and/or stream processing. The most popular available solution for this type of processing is Spark and Flink frameworks (Karakaya, 2017).

Apache Spark

Apache Spark is a popular open-source platform for large-scale data processing which can perform a large data operation on both staying and streaming data. Spark has strong in-memory data operations by having a specially developed component named as Resilient Distributed Dataset (RDD), which is a distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner. Spark usually uses the component of Hadoop, so that in most cases both Hadoop and Spark is installed in the same environment. Spark is faster than Hadoop about 10 to 100 times because the data is stored and processed in memory (Karakaya, 2017).

2.1.3 Cloud computing

Cloud computing can be defined as a model for enabling ubiquitous, convenient and on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort from the user side and minimal service provider interaction. (Gamaleldin, 2013). The service models of cloud computing are classified into three different categories depending on the user needs: they are infrastructure as a service (IaaS), in this model of service users have the ability to control the storage, operating systems, applications and network; platform as a service (Paas), in this model of service users have the ability

to control applications only; and software as a service (SaaS), in this model of service users have the ability to access services from the service provider without any control over the infrastructure and applications. The cloud deployment models also have been categorized into either private cloud (exclusive) or public cloud (non-exclusive) or hybrid (Mell & Grance, 2009).

It further gives a mutual pool of computing resources which can be given and stopped as and when the users require fulfilling a wide and continually expanding range of information handling needs (Alshwaier, Youssef, & Emam, 2012). Google Docs, Gmail, Yahoo mail and so forth are a portion of the basic instances of SaaS through cloud computing (Rhoton, 2011).

Similarly, as with all the idea, cloud computing is likewise characterized in different ways and different perspective. There are many cloud computing definitions and a portion of the definitions which diverse scientists and expert bodies have been given in this segment to comprehend the ideas related to cloud computing in the resulting segments. The most extensive definition accessible is by Brendl who characterized cloud computing as "accumulations of IT assets (servers, databases, and applications) which are accessible on the basis of on-demand, gave by a service organisation, accessible through the web, and give asset pooling among multiple users (Saidhbimca & Gashaw, 2013)". The Gartner consulting propose a definition as takes after "A style of computing where adaptable and versatile IT-related abilities are given as-a-service utilizing Internet advances to multiple external customers" (Plummer, 2008). The National Institute of Standards and Technology (NIST) characterizes cloud computing as " a model for empowering helpful, on-demand network access to a common pool of configurable computing resources (e.g., servers, network, storage, services and applications) which can be quickly provisioned and discharged with insignificant management effort or interaction of service provider. Borenstein and Blake (2011) consider it to be " the utilization of quick, high-bandwidth Internet connections with deploy services that are midway kept up, regularly by third parties and hence limit the cost and difficulty of IT organization and support for the associations that expend those services. Buyya et al characterize cloud as "a sort of parallel and dispersed framework comprising of a gathering of interconnected and virtualized computers which are powerfully provisioned and

introduced as at least one bound together assets in light of administration level understandings set up through arrangement between the consumers and service providers (Buyya, Yeo, & Venugopal, 2008). Foley gives this clarification as "Cloud computing is defined as access in on-demand to virtualized IT assets which are housed outside of your own data center, shared by others, easy to utilize, paid for by means of membership, and got to over the Web" (Foley, 2008). In another study cloud computing is characterized as follows: "It is a service model of information technology-based model where computing services (both software and hardware) are conveyed on-request free of gadget and area (Marston, Li, Bandyopadhyay, Zhang, & Ghalsasi, 2011).

Characteristics of cloud computing:

On-demand self-service: IT is utilized as service and it is readily feasible on demand without manual intervention.

Broad network access: Users of Cloud have the ability to access the computing resources from everywhere, at any time by using the network.

Resource pooling: The supplier makes the essential expedient obtainable to numerous users who utilize technologies such as multi-tenancy and virtualization.

Rapid elasticity: The necessary resources can be rapidly provisioned and discharged without the manual intervention when no longer required.

Measured Service: A consumed service should be quantifiable with respect to resources utilized. With this aspect, the billing of consumption-based becomes feasible. Further, called as pay-per-use or pay as you go (Irshad, Gapar, Johar, & Naleer, 2015).

2.2 Big Data on cloud computing

Processing and storing huge data volumes necessitate availability, fault tolerance and scalability. Cloud computing delivered all these features with the support of hardware virtualization. Hence, cloud computing and Big Data are two complementation concepts as cloud authorize Big Data to be feasible, fault tolerant and scalable. Big Data acts as the best tool for promoting effective opportunities in

business. As such, various novel firms such as Teradata, Hortonworks, Cloudera, and others, have begun to make concentration on delivering DataBase as a Service (DBaaS) or Big Data as a Service (BDaaS). Firms like Amazon, Microsoft, IBM and Google further give ways for clients to leverage Big Data on Cloud on demand. Many companies like RedBus and Nokia discuss the good of Big Data utilization in cloud environments (Neves, Schmerl, Cámara, & Bernardino, 2016).

According to (Tian & Zhao, 2015) and (Gökhan et. al., 2015), Big Data and Cloud Computing are complementary to each other and cloud computing solutions and techniques can resolve some of the challenges in Big Data.

Hashem and his colleagues (Hashem et al.,2015) clarified several cases that explain how the researchers have used the infrastructure provided by Cloud Computing for Big Data projects. These cases include the using of Cloud in genome informatics and mining Twitter in the Cloud. Both these cases provide sufficient clarification about the benefits of Cloud infrastructure to resolve some of the challenges associated with the complexity of Big Data analytics.

The main challenges faced by Big Data applications using Cloud Computing environment are security and privacy issues. Another challenge is the lack of trust in cloud service providers, unawareness about the service level agreements (SLAs). Another challenge is the possibility of hacking attacks on distributed locations. These are some of the challenges that are obstruction the using of infrastructures provided by cloud computing environment (Vajjhala & Ramollari, 2016).

Likewise, the rate of data transfers considered as one of the key challenges associated with using of cloud computing infrastructures for Big Data relates. The speed of data transfer related to networks, mostly in developing countries has not advanced at the same rates as processing and data storage capabilities. This variation in data transfer capabilities and networking infrastructure is likely to pose another important problem related to using public or hybrid cloud computing infrastructure for Big Data.

2.3 Challenges of Big Data

Data storage

Challenges in storage are posed by the velocity, volume, and variety of Big Data. The high volume and heterogeneity nature of Big Data make the process of storing is a challenge and the traditional storage systems are not sufficient to handle it. Additionally, the velocity of Big Data needs the storage systems to have the capacity to scale rapidly which is hard to accomplish with systems of traditional storage systems. Cloud storage services (e.g., EBS or Elastic Block Store, Amazon S3), on the other hand, provide virtually unlimited storage with high fault tolerance that gives possible solutions to address challenges in Big Data storage (Yang, Chaowei, et al., 2017).

Data transmission

Data transmission exist in various phases in life cycle of data such as (i) data gathering from sensors to storage; (ii) data combination from multiple data centres; (iii) data management for transmitting the combined data to platforms that can process this data (e.g., cloud stages) and (iv) data analysis for transferring data from the storage to analyzing host (e.g., high-performance computing (HPC) clusters). Transmitting huge amount of data poses clear issues in each of these stages. In this manner, data compression algorithms and smart pre-processing methods are expected to successfully decrease the size of the data before transmitting this data (Yang, Long, & Jiang, 2013b). For instance, Li et al. (2015) developed an efficient network transmission model where an arrangement of methods of data compression for transmitting geospatial information in a cyber-infrastructure domain is shown. Moreover, when transmitting Big Data to cloud platforms from other data centers, how to create effective algorithms to automatically recommend the suitable cloud service (site) according to spatiotemporal standards to increase the transfer speed of data while decreasing the cost is challenging too (Li, Song, Zhou, Cao, & Gao, 2015).

Data Management

Managing, analyzing and visualizing Big Data is a complex task for computers (Sewoog Kim, Dongwoo Kang, Jongmoo Choi, & Junmo Kim, 2014). The data management paradigm is redefined in the case of Big Data for storage, cleaning and organizing unstructured data demanding new technology like (e.g., Hadoop, NoSQL). The integrity of the data provenances which makes the metadata essential (Singh et al., 2003), the challenge is how you can generate the metadata automatically to describe the Big Data and pertinent processes (Gantz & Reinsel, 2012; Oguntimilehin & Ademola. 2014). Metadata is even more challenging when it is generated for geospatial data which has intrinsic characteristics of high dimensionality (3D space and 1D time) and is complex (e.g., space-time correlation and dependency). Also, the challenges posed by Big Data to database management systems (DBMSs) because of the lack of scalability to storing and managing unstructured Big Data in traditional RBDMSs (Pokorny, 2013; Chen, Mao, Zhang, & Leung, 2014a). While HBase and MongoDB are the two non-relational databases (NoSQL) prepared for Big Data (Jing Han, Haihong, Guan, & Jian, 2011; Padhy, Patra, & Satapathy, 2011), seems to solve this issue, how to develop these databases in order to handle the huge data by designing effective queries and indexing algorithms is still a challenge (Whitman, Park, Ambrose, & Hoel, 2014; Li et al., 2017).

Data processing

The processing of huge amount of data can be managed partially by using computers with high-speed processor and big storage unit and high-speed network (Bertino et al., 2011). However, the traditional computing paradigms are not suitable to the computing resources needed to process the Big Data (Ammn & Irfanuddin, 2013). As a partial solution, the cloud computing environment provides virtually unlimited and on-demand processing power. However, there are many challenges faced by using cloud computing infrastructures. The first challenge is the limitation in network bandwidth of cloud computing which can affect the efficiency of computing a large amount of data (Bryant, Katz, & Lazowska, 2008). The second challenge is the locality of data for processing Big Data (Yang, Xu, & Nebert, 2013a). Instead of

moving the data to process it, many of Big Data platforms follow the principle of "moving the computation to data" like Hadoop (Ding, Jiang, Wang, Liu, & Li, 2013). Data locality in cloud computing is a complex task as it is a virtual pool of data (Yang, Long, & Jiang, 2013b), it involves exchange and communication in intensive data to support data processing (Huang et al., 2013). The existing algorithms cannot handle the acceptable quality and the tolerable time frame for the large and high dimensional data for data reduction (Aghabozorgi, Seyed Shirkhorshidi, & Ying Wah, 2015; García, Luengo, & Herrera, 2015). For example, the huge amount and continuous data produced by the surveillance and intelligence in real time cannot be processed by traditional algorithms. So, we need a highly efficient and scalable algorithm for data reduction which can remove the redundant, irrelevant, noisy and misleading data, and this is considered as one of most significant tasks in the research area of Big Data (Zhai, Ong, & Tsang, 2014).

Data analysis

One of the most important phases in Big Data value chain is data analysis to extract the valuable information (Fan & Liu, 2013; Chen et al., 2014b). However, the scalability and complexity are considered as the main challenges faced by data analysis algorithms (Khan et al., 2014). Hence, Big Data analysis needs interoperable and advanced scalable algorithms, and this can be achieved by using the programs that follow the parallel processing platform like Hadoop (Jagadish et al. , 2014). In any case, this 'divide and conquer' procedure does not work with multi-scale iterations but it is highly required by mining algorithms and data analysis (Chen & Zhang, 2014). Moreover, the heterogeneity of Big Data makes the work of most analytical algorithms very hard because of most of these algorithms dealing with structured and homogeneous data (Bertino et al., 2011). This issue requires either developing new algorithms that deal with heterogeneous data or design new techniques for transform unstructured data to structured data which can be to appropriate with existing algorithms (Yang, Chaowei, et al., 2017).

Data security

The increase in the reliance on computers and Internet over the last years makes people and organizations liable against the breach and abuse of data (Denning &

Denning, 1979; Abraham & Paprzycki, 2004; Redlich & Nemzow, 2006). New security challenges are posed by Big Data on encryption methodologies, algorithms and standards for traditional data (Smid & Branstad, 1988; Coppersmith, 1994; Nadeem & Javed, 2005). Past investigations of data encryption concentrated on small to-medium-size data that do not work effectively for Big Data against the performance and scalability challenges (Chen et al., 2014b). Additionally, schemes and data security policies to work with the structured data stored in traditional DBMS are not able to deal with security issues of unstructured data effectively (Villars, Olofson, & Eastwood, 2011). Consequently, successful strategies for safety management and data access control should be explored in Big Data and these develop new storage structures and data management systems (Cavoukian & Jonas, 2012; Chen et al., 2014a). In the period of cloud, the fundamental concerns for data owners are the restricted control on virtualized storage, availability, and integrity, data confidentiality (Kaufman, 2009; Feng, Zhang, M., Zhang, Xu, 2011; Chen & Zhao, 2012).

Data privacy challenges

The exceptional networking between computing platforms and smart devices makes a contribution to Big Data but increasing the security concerns about the personal information for people like the location, transactions, and behavior (Cukier, 2010; Tene, 2011; Michael & Miller, 2013; Cheatham, 2015). The medical records and social media increase the recording of the personal information for people and this increases the privacy concerns (Terry, 2012; Kaisler, Armour, Espinosa, & Money, 2013; Padgavankar & Gupta, 2014; Michael & Miller, 2013). All These privacy concerns show the gap among the traditional and Big Data regulations and policies, and there is a need to design new policies which can address the privacy concerns completely (Khan et al., 2014; Eisenstein, 2015).

Lack of skills challenge

One of the most important challenges of managing Big Data in SMEs and government agencies is the lack of skills. Utilizing Big Data is a new concept for many of these organization and they do not have skilled people to use Big Data technology well and to apply sufficient methods for analyzing it based on the

demands and the requirements of the organization. Therefore, preparing skilled staff to deal with using of Big Data is a challenge and should be considered when the organizations intend to adopt this technology (Miller, 2014).

Data ownership challenge

In addition to privacy, ownership of data is a complicated issue. As reported by Kaisler et al. (2013) that data ownership presents a crucial and continuous challenge, especially in the social media context like who is the owner of the data on Twitter, Facebook or MySpace. Generally, the prevailing view is the provider of social media and the users are own the data. Kaisler et al. (2013) argue that this challenge still needs to be settled. For instance, data produced by the sensor is too sensitive and can lead in growing errors – this could moreover result to capture and reveal inconsistent data and on the other hand who is the owner of this data. As a result of challenges mentioned above, data ownership is a significant social issue needs additional investigation since they could have deep implications (Sivarajah et al., 2017).

Ethical challenge

Large datasets are being mined for significant predictions which usually produce amazing insights. As reported by Richard and King (Richards & King, 2014), Big Data analysis used to examine diverse kinds of human activities and the process of decision-making are beginning to be affected by Big Data predictions, such as shopping, dating, education, voting, medicine, cybersecurity, terrorism prevention and law enforcement. However, while this is happening, people have limited understanding of what data is being gathered, as well as shared with third parties. Therefore, current privacy protections concentrated on handling personal information determining are not sufficient. Big Data initiatives find many personal datasets like the call history, social network connections, location history, purchase history, search history and more of this information are already in the governments and corporations hands, and the gathering of these and other datasets are increasing. Therefore, to protect individual rights, we need to develop Big Data ethics (Richards & King, 2014).

2.4 Small and Medium Scale Enterprises

The economic development of any country depends on the role of SMEs as they are the key drivers for the growth of the economy of the country (Schmiemann, 2009). Okongwu (2001) states that SMEs serves as a major factor in promoting partnership and private sector development in developed and developing countries and the main source for the economic growth of the country (Mutula & van Brakel, 2006). According to (Wattanapruttipaisan , 2005), it's universally acknowledged that the SME is responsible for growth, competitiveness, and productivity of the developed and developing countries as they bring about contribution to improved living standards and also they induce the capital formation. Equitable and sustainable diversification in the industry is achieved by SMEs (Anigbogu, Onwuteaka, Edoko, & Okoli, 2014). SMEs share a large number of new jobs by 60-70% of employment and 95% of firms are small and medium-sized enterprises (OECD, 2000).

SMEs are defined by the number of staff members and their capital. According to the definition stated by European Union (EU), Medium-sized enterprises are the companies that have at most 250 employees and less than 50 million Euro as a company capital. On the other hand the EU categorized the SMEs to small-sized enterprises are the companies that have at most 50 employees and their capital is less than 10 million Euro, and the last category is the Micro-sized enterprises are the companies that have at most 10 employees and their capital is less than 2 million Euros (Vajjhala & Ramollari, 2016).

2.5 SMEs and Technology adoption

Competitive advantage, efficiency, and effectiveness are gained through the opportunity provided by the IT. The economic development of any country is dependent on SMEs. The economic development, social change, and the information technology are interlinked by several theories. The Internet is a significant part of information technology has an impact on any small and medium scale operation which is claimed to efficient in the growth and survival of the economy(Hicks, 1993). The traditional business and the economy are solely depending on new technology due to the impact of information technology in the current scenario. Business partners and customers have compatibility with new technology more than

traditional business in terms of interactivity, business transactions and flexibility. In accordance with IT and modern trends in business, the SME development also draws significant attention in developing countries.

2.6 SMEs and Big Data

Big Data technology is considered to be one of the most important drivers for SMEs growth according to an economic survey done by Oxford (2013). Big Data helps SMEs to analyze and predict the customer's behavior and market. When Big Data implemented in a right way, it can increase the productivity, flexibility, responsiveness, anticipation, and ability to serve the customers better and improve the decision-making process (Sen et al., 2016).

Big Data exploitation gives the opportunity to earn the competitive advantage and growth. Thus, SMEs must take a step forward and should not be afraid from the business-related risks, and trying to apply the right innovations which can help the organization to achieve the advancements towards the future (Abebe, 2014).

One of the important challenges which affect the growth in SMEs is the lacking of Big Data expertise in these organizations. The challenges affecting SMEs growth, whichever need to analyze the huge amount of data, can be summarized as:

- Ways to compete with enterprises and franchisees
- Lack of Big Data expertise
- Inability to invest in customer acquisition
- Inability to manage the supply chain, distribution and salesforce
- Lacking timely insights into market movements
- Inability to deliver large order size with short cycle times (Sen et al., 2016).

Finally, SMEs can benefit from implementing Big Data in a right way to achieve the improvement in terms of innovation, new business creation, productivity, efficiency, financial control, decision-making process, and these can increase the revenue, profits, growth and minimizing the operational cost (Sen et al., 2016).

2.6.1 SME Governance of Information

An essential part of organizing the 3V of Big Data demands specific skill sets from data scientists. However, the prevalence and lack of these factors seem to be the discussion that governs the data or information which does not even relate to the minds of the SMEs.

This process is now termed as the information governance. The rights for making a decision and a framework which gives accountability to encourage the behavior in creation, use, storage, archival, valuation and deletion of information (Logan, 2010). In other words, in an organization, the handling of data and information is needed to set a number of rules and regulations.

In academics, it is a new concept as was evidenced by (Tallon, Ramirez, & Short, 2013). Tallon et al. (2013) argue that the practice of IT governance and the extensive literature focuses just on the infrastructure, project management, and control and entirely neglects to recognize the value of the information that has been governed by the organization. Information governance is crucial in a knowledge-driven business world, data security, social responsibility and the focus on digital rights with the increasing value of the information in Big Data in an organization. SMEs oblige to information governance as they have no other option as to expand their skill set as they have very less exposure to working with the Big Data in big organizations. Further, Neelly (Neely, 2014) states that information governance involves the process of obtaining the data in strategic partnerships with other SMEs by monitoring and ensuring the value has been extracted. It is a process that seems least exciting from the immediate view of a business organization and as well as the most important aspect in which the SMEs are invested upon to ensure the minimal risk if the SME decides to change over to the big pool.

2.7 Government Agencies and Technology Adoption

Service delivery is becoming a part of most important agenda in government organizations which is considered to be part of the quest from around the world. The government and the governed have whole new possibilities catered from around the world especially from the innovations of ICT. Public sector management thrives on

the important measure of the successful deliveries of online services around the globe (Kaliannan, Awang, & Raman, 2007). Public cloud services have reached a worldwide spending of over $47.4 billion in 2013 and by 2017 it is expected to reach 107 billion dollars. As a whole, the compound annual growth rate of the cloud service is 23.5% which is five times the actual growth of the IT industry over the forecast period of the years from 2013-2017 (Gens, Adam, Christiansen, & Bradshaw, 2013). The values from the massive amounts of digital data can be derived from the businesses, governments and the research community. The Big Data application projects analyzed by the government offers guidance for the countries that follow Big Data initiatives.It takes much longer for decision making in government and is usually carried out by the officials, diverse actors, interest groups and ordinary citizens. The government also deals with the Big Data from multiple sources in different formats. As the data comes from different sources and multiple channels, it becomes a difficult challenge for the government to collect data. Hence it requires the government to operate or plan to work with big-data projects for which they have to undergo a step by step approach to set the right goals and realistic expectations. In this, the success depends on their ability to integrate and analyze information, develop support systems and support decision making through analytics (Kim et al., 2014).

The next big thing is cloud computing, and many experts have exclaimed that cloud computing is the future of the computing world (Wyld, 2010). Cloud computing ensures a better position in efficiency, flexibility, and accountability in a public organization where it also brings in the tools, practices followed in an organization and the technology with proper significance to the organization. Other than government's cloud adoption of strategies even businesses adopt the technology in the use of cloud productivity and project management tools are abstracted from literary reviews. In a government organization, cloud technology does not only ensures significant cost savings but also empowers innovative services and their quality that has been provided to businesses, other governments, and even the citizens. Cloud computing used by government catered to various organizations such as education, transport services, health services and contracting and management of utility services (Gens et al., 2013). Countries like USA, Japan, Finland, Sweden Australia, United Kingdom and Denmark are leading the way in cloud adoption in

terms of Government institutions (Wyld, 2010). It is seen that the federal government has spent years to collect and store a large amount of data and information of the people such as census data and the satellite images that are used and stored by the government. Thus the data explosion occurs in government sector as in private sector. It is also observed that the US federal agencies have to store 1.61 petabytes of data which was expected to grow to 2.63 petabytes by 2015. (Informatica, 2017). The data in the government agency has specific goals. They may be detecting fraud, supply chain streamlining, proximity data, weather satellite images, and information that are needed to achieve higher insight in less time with accuracy. There are several challenges that are to be considered for efficiently utilizing Big Data. The overall efficiency of the government can be improved with Big Data. The speed and accuracy of the forecasting and decision making is the most significant challenge with more data. The policies are formulated and the regulations are executed in data collection and other data related methods. It is also seen that most of the data in the government are unstructured and hence organization of these data requires strict policies and rules for data collection and storage. The government agencies as they have unstructured data it becomes clear that the efficient use of data is very less. The government is expected to utilize its data for making decisions, but it is not so in the real-time environment. The data collected are kept as junk and are not used to make any strategic decisions.

The data that is being stored in the government is considered to be an asset for both the government and the nation. It brings many challenges and opportunities (Singh et al., 2014). The data generation and sharing of the data are increased exponentially with the advancement of computer and communication technologies. Many factors have contributed to the growth of data production. Modern computers have shown the increase in processing and storage capabilities. As a result of this, various types of content and data from the traditional formats are converted into digital formats. The storage cost of digital storage is less when compared to the traditional data storage. This decrease in the cost of storage is the result of the technological advances and also the decrease in the cost of the storage devices. This cost reduction is proportional to the rate at which the digital data is generated. We can see that the technological advancement helps in faster and affordable communication technologies. The transfer and sharing of data are much faster than the conventional

methods. For example, a handwritten letter may take a week to reach its destination whereas the email reaches the desired recipient in seconds. With the increase in the applications and smart devices such as smartphones, tablets and newer digital devices along with smarter applications have resulted in the generation of digital content. Big Data provides many opportunities for government organizations and it also results in some threats and challenges. The data that are being handled is more and this volume always increases. With the increase in the usage Internet, the data being generated is also enormous. The current technologies, software, and hardware that is operated by government agencies are not sufficient to handle the data that are being generated and stored. The government agencies must alter the way they govern, manage, process and report the data that is being stored in their database and the proper usage of Big Data can be achieved if there is the change in the methodology in which the data is being handled.

The government organization reaps many benefits from Big Data analytics. The increase in efficiency through the delivery of online services and by storing the information of the citizens. It is also seen that the nation is being protected from terrorist attacks. The significant areas where the Big Data analytics have been utilized effectively by the government agencies are in tax collection and identifying fraudulent behaviors and also in the prevention of crimes and threats (IBM, 2017).

2.8 SMEs, Government Agencies and Big Data & Cloud adoption

The US government has executed Big Data projects through several various agencies. These projects affect government processes with different effects. Main areas are selected for analysis: social services, national security and taxation. By using Big Data, NSA benefits from information to predict and avoid potential risks to the nation. The sector of Social services is already utilizing Big Data to recognize fraud and avoid the resources wasting. As analyzed before, taxation does not gain many benefits from utilizing Big Data, aside from recognizing fraud and tax prevention. Nevertheless, enhancements can be made if social services and taxation share data between them. By the help of Big Data, decision-making processes in US government became more advanced and more flexible at acquiring opportunities that are generally missed by other countries (Pannu, Gill, Tebb, & Yang, 2016).

Governments expect Big Data to improve their capacity to serve their people and handle key national challenges involved with the healthcare, economy, natural disasters, job creation, and terrorism (Kim et al., 2014).

"Government use of cloud is progressively growing in areas like health services, education, transport services, managing of utility services, contracting. In terms of government agencies, the Countries leading the way in adopting cloud are USA, Australia, Finland, Japan, Denmark, United Kingdom and Sweden. Three key benefits were gained by the good implementation of the cloud are ease scalability increase, productivity, and overall cost-reducing" (Tweneboah-Koduah, Endicott-Popovsky, & Tsetse, 2014).

Ogbuokiri et al. (2015) studied the regional growth of the small and medium enterprise when Big Data analytics were implemented. The developing and transition economies that include state policymakers development of economists and international development partners have a notion that the SMEs are the driving force of their growth and economic development. In recent times, domains context and all sorts of application are given solutions by the concept of Big Data in terms of policy and practice. Public and private data collected and stored by various organizations has led to the abundance of data which has impacts on the many innovative data analytics technologies. Challenging current decision making and policy initiatives at SME micro level and Government macro level has potentiality resulted in the growth of the SME in controlling the Big Data practice. The SMEs are expected to realize their opportunities as they seek the potential fundamentally to Big Data for making better decisions and creating policies in markets and business models. Big Data value represents opportunities that are profitable and strategic for the SMEs.

CHAPTER 3

RESEARCH METHODOLOGY

The study pertains to the clarification of the research area that is being considered for identifying the benefits and challenges of Big Data on the cloud for SMEs and government agencies. The direction of research in which the study was carried on is to identify the problems related to the Big-data on the cloud. With the identification of the research problem, the research is directed towards the solutions that can be provided for the challenges in implementing Big Data technologies by SMEs and government agencies.

The method employed for this research is Systematic Literature Review (SLR) which is completely based on the analysis of secondary data (research papers, scholarly articles, published thesis). The main reason behind the adoption of SLR is that the existing information about Big Data and its benefits and challenges to the SMEs and Government agencies need to be analyzed. Furthermore, a true reflection on the various benefits of implementing Big Data technologies in SMEs and Government agencies need to be discerned which will aid towards the development of a Big Data framework for the considered sectors. A systematic literature review is based on a set search strategy which aims at the detection of relevant literature to a maximum extent and supports to understand the overall picture of a research context (Software Engineering Group, 2007).

In this regard, the present chapter elaborates on the type of methodology adopted, the research philosophy, research approach, research design and research strategy used and the research validity and reliability. Furthermore, this chapter depicts the keywords used for the search of research articles, a quality assessment was done, the selection criteria and procedures used, and the data extraction strategy used.

3.1 Research philosophy

Research philosophies majorly comprise of the four types namely- pragmatism, realism, positivism, and interpretivism. Positivism and interpretivism are the two major philosophies that are used commonly in researches that deal with either development of hypothesis or proving a developed set hypothesis. Positivism denotes believing in objective realities wherein the scenario could be analyzed statistically wherein interpretivism is based on social reality and research scenarios could not be measured objectively (Tolman, 2012). As the research aims to critically examine the importance of Big Data on Cloud, the benefits and its challenges from the lenses of previous literature, the current research is based on the interpretive notion.

3.2 Research approach

There are two different types of research approach: the inductive and the deductive. The inductive approach is based on the development of new theories and models whereas the deductive approach bases the complete research on a set hypothesis, previous theories, and models. Since the present research aims at acquiring all insights on the benefits and challenges towards the adoption of Big Data on Cloud, the deductive approach is used.

3.3 Research design

The research design is categorized into three types: qualitative, quantitative and mixed research. In a quantitative design, numerical data is involved wherein the collected data is examined using statistical techniques. However, qualitative research design involves the collection of non-numerical data wherein the collected data is examined mostly using textual analysis techniques. However, the mixed research design is the combination of both qualitative and quantitative research designs. For the present research, the qualitative research design is most appropriate since it involves the collection of secondary data (research articles) and examination of the same through thematic analysis. Hence the qualitative research design is used.

3.4 Research strategy

The strategy for the research is the SLR wherein the study is conducted using the following steps:

Step 1: Definition of the research question

Step 2: Using Keywords for the search

Step 3: Setting criteria for inclusion and exclusion of research articles

Step 4: Quality assessment of papers

Step 5: Execution of the methodology

3.4.1 Definition of the research question

Following is the research questions for the present research and their definitions:

Table 3.1: Research Questions and Their Definitions

#	Research question	Definition
RQ1	What are the benefits of combining Big Data and cloud computing?	One of the main objectives of this research is to identify the benefits of combining Big Data and cloud computing
RQ2	What are the various benefits of Big Data and Cloud for SMEs and government agencies?	As the topic implies the main aim of the research is to identify the benefits of Big Data on Cloud for SMEs and government agencies.
RQ3	What are the different categories of challenges of Big Data faced by SMEs and government agencies?	Once the benefits are identified, the study on the challenges of Big Data must also be researched for both SMEs and Government agencies
RQ4	What are the existing solutions for Big Data challenges?	As the challenges are identified it becomes imperative to come up with solutions for the challenges that are being identified.

3.4.2 Keywords used

The first step for the systematic literature review is to search papers based on specific research related keywords. Following are the keywords that are used wherein search strings are formulated:

- Big Data
- Cloud computing
- Small and Medium Scale Enterprises
- SMEs
- Government agencies
- Benefits
- Challenges
- Solutions

Search strings were constructed using Boolean AND, OR. Following are digital libraries which were used in the search :

- Google Scholar***: scholar.google.com***
- ACM Digital Library: ***www.portal.acm.org***
- IEEE Explore: ***www.ieeexplore.ieee.org***
- ScienceDirect: ***www.sciencedirect.com***

3.4.3 Setting criteria for inclusion and exclusion of research articles

The purpose of this step is to determine the most relevant studies that are related to thesis subject and the following table shows the criteria were used to selected the most relevant studies.

Table 3.2: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
1. Studies addressing benefits and challenges of Big Data on the cloud. 2. Journal and/or conference papers and thesis. 3. Studies that perform Big Data on cloud for SMEs and government agencies 4. Primary or secondary studies.	1. Studies are not accessible in full text. 2. Studies that do not address the benefits and challenges of Big Data on the cloud. 3. Studies are not presented in the English language. 4. Prefaces, slides, panels, editorials or tutorials. 5. Studies that do not answer the research questions.

During the search phase, we have found 208 papers as relevant from the selected digital libraries. We analyzed three part in each selected paper (title, abstract, keywords) and then, we applied the exclusion criteria that result to eliminate some of the articles. In the case of uncertainties about some papers that included, we considering introduction section and conclusion of these papers. As a result of applying the inclusion and exclusion criteria, only 92 papers remained as related to our topic were selected.

3.4.4 Quality assessment

Quality assessment is performed to ensure papers selected are relevant and appropriate research studies. The quality assessment for the selected studies is based on the research by Kitchenham and Charter (2007) wherein the following are the assessment criteria:

Table 3.3: Quality Assessment Criteria

No.	Quality Assessment Criteria
1	Are the objectives and aims stated clearly?
2	Are the strategies of collecting data described fairly?
3	Are the constraints and limitations discussed?
4	Are the citations expressed clearly in the paper?

3.4.5 Thematic analysis

The research articles (secondary data) selected for the systematic review are analyzed thematically wherein the themes include Benefits of Big Data implementation in SMEs and Government agencies, challenges in implementing Big Data and Cloud technologies in SMEs and government agencies, and solutions for implementing Big Data and Cloud technologies in SMEs and government agencies. Such thematic analysis leads the researcher to achieve the objectives of the research.

3.5 Ethical considerations

The present research does not possess any ethical issues since the research is the examination of the previous literature. Furthermore, the examination of secondary data which is the premise of systematic literature review tends to reduce bias within the specified research context which is an additional strength of the present research.

CHAPTER 4

BENEFITS, CHALLENGES AND SOLUTIONS

In this chapter, we explain the benefits, challenges and solution of Big Data and Cloud for SMEs and government agencies. This information will answer our proposed research questions. The first section summarizes the benefits of combining Big Data and Cloud Computing, the benefit of Cloud Computing itself, benefits of Big Data for SMEs and government agencies, and the benefits of using Big Data Framework. By using these results, this section offers the answer to the first and second research questions. The second section explains the challenges of Big Data and the existing solution for each challenge, and this section offers the answers to third and fourth research questions.

4.1 Benefits

This section offers the information that leads us to answer the first and second research questions which are given in Chapter 5. These research questions were:
RQ1) What are the benefits of combining Big Data and Cloud Computing?
RQ2) What are the various benefits of Big Data on Cloud for SMEs and government agencies?

4.1.1 Benefits of Cloud Computing

There are several factors why all sizes and types of organizations are using cloud computing model to satisfy their IT requirements. We have summarized them in this section, and before giving the details we list them in Table 4.1 with the corresponding references being used in collecting this information.

The following Table shows the main benefits of cloud computing and their references.

Table 4.1. Benefits of Cloud Computing with the Corresponding References

Benefits	References
1. Reduction of capital expenditure 2. Reduced administration costs 3. Improved resource utilization 4. Economies of scale 5. Scalability on demand 6. Quick and easy implementation 7. Quality of service 8. Guaranteed uptime 9. Anywhere access 10. Technical support 11. Disaster recovery/backup	• Saxena & Pushkar, 2016; • Chandrashekar, Kala, & Mane, 2015; • Jelonek, Stępniak, Turek, & Ziora, 2014; • Hashemi, Monfaredi, & Masdari, 2013; • Prasad & Rao, 2014; • Mikkonen & Khan, 2016 • Elazhary, 2014; • Hussein, Sulaiman, & Hamzah, 2013.

Reduction of capital expenditure: Clients can easily prevent wasting large amounts of funds on buying and setting up their IT applications or infrastructure by switching to the Cloud environment. The environment of Cloud provides a simple operational cost that is much easier than spending month-by-month budget and avoids wasting money on downgrading assets. Furthermore, clients do not require to spend money with regard to surplus resource capacity in-house to fulfill fluctuating need (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Jelonek et al., 2014).

Reduced administration costs: The solutions of IT can be implemented rapidly and maintained, managed, upgraded, patched by the provider of service. Technical support is offered 24 / 7 by trustworthy providers without additional charge, minimizing the burden on IT staff. Many researchers stated that Cloud environment enables organizations to improve purchasing processes and reduces the demand to replicate certain computer administrative skills relevant to setup, support and

configuration (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Jelonek et al., 2014).

Improved resource utilization: Combining resources within clouds decreases expenditures and increase utilization by providing resources just when they are wanted. By using the virtualized software model, they can enable the sharing of physical services, storage, and networking capabilities. The cloud infrastructure, regardless of deployment model, seeks to make the most of the available infrastructure across a number of users. Sharing computing power between several renters can enhance utilization rates because servers do not remaining idle, which can minimize costs considerably while raising the application development speed (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Jelonek et al., 2014).

Economies of scale: Clients of cloud computing can gain the benefits of the economies of scale offered by vendors, who utilize extremely large-scale data centers running at higher levels of efficiency, and multi-tenant architecture to share assets among several different clients. This IT model provision enables clients of providers to save a large amount of money (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Jelonek et al., 2014).

Scalability on demand: The advantages of scalability and flexibility offered by cloud computing are highly valuable and enable clients to respond rapidly to changing IT demands, increasing or decreasing capacity and users whenever needed and reacting to real instead of expected requirements. Even better, since the cloud environment follows a service model in which the costs of service are dependent on real utilization, clients just pay for what they utilize. Clients benefit from resources elasticity without investing a large amount of money (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Jelonek et al., 2014; Elazhary, 2014).

Quick and easy implementation: Organizations can implement their cloud computing environment off the ground in minutes without the need to buy infrastructure, implementation services or purchasing software licenses (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Hashemi et al., 2013).

Quality of service: Selected service provider should provide 24/7 customer support and an instant response to urgent situations (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Hashemi et al., 2013).

Guaranteed uptime: The service provider should ensure that services and/or applications are always accessible and online (Saxena & Pushkar, 2016). Most of the service provider guarantees %99.99 uptime (Prasad & Rao, 2014).

Anywhere access: IT services in the cloud environment give the ability to access applications and data securely via an internet connection from any location. It is much easier to collaborate as well; with both data and application stored in a cloud, many users can work with each other on the same project, share contacts and calendars, and so on. Nevertheless, due to the nature of the cloud which offers the accessibility from anywhere, customers can easily connect from various locations. Therefore if the connection fails in the office and there is no redundancy, data can be accessed by connecting to the closest Wi-Fi enabled point or home. For this reason, remote/ flexible working is easier and allows clients to reduce overheads and fulfill the new rules (Saxena & Pushkar, 2016; Chandrashekar et al., 2015; Hashemi et al., 2013).

Technical support: Cloud computing providers are responsible to update the software and deliver technical support. This kind of support will reduce the cost and time. Additionally, technical support delivered by the service provider will help the organization to reduce the need to hire trained people for maintenance purposes and decrease the error in the maintenance process (Saxena & Pushkar, 2016; Hashemi et al., 2013; Mikkonen & Khan, 2016).

Disaster recovery/backup: Current investigations have mentioned that around 90% of organizations do not have sufficient business continuity or disaster recovery plans, making them vulnerable to any disruptions which may happen. Providers can offer a variety of disaster recovery services, through cloud backup service files stored twice at various remote locations to make sure that there is often a copy available (Saxena & Pushkar, 2016; Hashemi et al., 2013; Hussein, Sulaiman, & Hamzah, 2013).

4.1.2 Common Benefits of Big Data for SMEs and Government agencies

This section discusses the common benefits of Big Data for SMEs and government agencies as listed in Table 4.2.

Table 4.2. Common Benefits of Big Data for SMEs and Government Agencies with the Corresponding References

Benefits	References
1. Improving decision making 2. Improve organizational productivity 3. Cost reduction 4. Increased operational efficiency 5. Predict workforce requirement	• Brown et al., 2011; • Johnson et al., 2012; • Smeda, 2015; • Thomas, 2014; • Arpitha et al., 2016; • Accenture, 2016; • House, 2014; • Lee, 2017; • IBM, 2014; • Kim et al., 2014; • Acharekar & Palghadmal, 2016.

Improving decision making: Big Data changes the way of making decisions in organizations (Brown, Chui & Manyika, 2011). Organizations can anticipate the customer's behavior effectively by analyzing the available data and use this information to their advantage (Brown et al., 2011). This results in adding more value to the organizations and to their customers because the organization can anticipate the requirements of their customers and make good business decisions based on these requirements (Johnson, Neff & Stuart, 2012). This will ensure the competitive advantage in the marketplace. The organization will be more active when they have the ability to anticipate the future response and requirements of their customers. The different sources of data available to organizations and the proper analysis techniques of this data will be effective tools for business managers to give more precise decisions about innovations in their services and products. For government agencies, the use of streaming analysis tools as well as other

technologies to process the generated data in real time can enable the decision making to be faster within the organizations. Without having these technologies and tools, the decision-makers may need to resort back again to a long shot guess which requires more time and give fewer results (Smeda, 2015).

Improve organizational productivity: Every aspect of the business can benefit from Big Data, it can be leveraged to raise the general productivity of the business, and workforce productivity. By analyzing information about day-to-day tasks and staff members statistics (log information), new insights and new patterns will produce that can help the organization to build new procedures and raise the overall and workforce productivity (Thomas, 2014; Arpitha, Mandara, & Smitha, 2016).

Cost reduction: Big Data decreases the costs of operations inside many organizations. According to Accenture (2016), organizations which utilize the analysis of data in their businesses have more efficient and faster response time to the challenges of supply chain than those who utilize data analysis on an ad-hoc basis (47% vs. 18%). Analysis of Big Data leads to improve the needs predictions, more efficient with real-time tracking and visualization during shipments, and highly optimized distribution network management (House, 2014). Big Data also led to a big reduction in cost in the retail industry. Tesco, a European supermarket store, had about $25 million savings from energy cost per year by analyzing the data that has been collected from the refrigerators. Analyzing the collected data demonstrated that the refrigerators temperature degree was set colder than the necessary degree, wasting electricity. To enhance the temperature degree of the refrigerators, Tesco company has equipped the refrigerators in Ireland stores with sensors to monitor the temperature every 3 seconds. For government agencies, Big Data techniques can help the organizations to analyze the log information, operational processes and assets utilization. By analyzing this information, the organization can reduce infrastructure and operating costs (Lee, 2017).

Increased operational efficiency: Analysis of Big Data helps to improve the effectiveness of critical infrastructure and assets and also helps organizations to select a proactive approach to avoid or minimize costly failures. When the

organization assets are not functioning correctly, maintenance is an essential function and directly impacts the organization profitability. To increase the productivity of a physical asset, it should be functioning correctly. If a necessary asset is in need of maintenance, the entire process can be broken down into lower levels until the issue is solved. Initially, the production asset breaks down. The machine operator detects the failure and responds to it. When the failure is reported, the engineer and supervisor deliberate and decide upon a course of action. With a course of action, asset maintenance workers are notified and begin working together to resolve the issue. The solutions that are implemented are often based exclusively on the professional experience of the maintenance staff. These solutions might solve the problem, but they can also result in further issues at a later time. Predictive maintenance and predictive asset optimization solutions resulting from the analysis of massive data which is produced by the maintenance reports of the organizations can help to choose the right solution. Government agencies, state agencies and local councils can recognize the errors and incorrect steps they have taken in the past and try to ignore these steps in the future to increase the operational efficiency (IBM, 2014; Kim et al., 2014).

Predict workforce requirement: Big Data could serve as a window to the employees' professional lives. By tracking and monitoring the employee's behaviors in the workplace, the employer will collect a great deal of performance data. The analysis of this data offers new insights into the requirement of the workforce which can be utilized to motivate the employee, improve the performance and raise the employee engagement (Acharekar & Palghadmal, 2016).

4.1.3 Specific benefits of Big Data for SMEs

In addition to the common benefits of Big Data listed in the previous section, there are many other benefits that can be listed as specific to SMEs. This section summarizes these benefits. Before doing this, we give a reference list in Tabel 4.3., that enabled us to collect this information.

Table 4.3. Specific Benefits of Big Data for SMEs with the Corresponding References

Benefits	References
1. Innovation based on customer responses 2. Maximize profits 3. Reduced risk 4. Improve the competitiveness of the enterprises 5. Development of new products and services 6. Better knowledge of business changes	• Sen et al., 2016; • Benjelloun, Lahcen, & Belfkih, 2015; • Ragin, 2017; • Wielki, 2015; • Editorial, 2017; • Brown et al., 2011; • Johnson et al., 2012; • Smeda, 2015; • Saqib, Khan, Mahmood, & Naeem, 2015.

Innovation based on customer responses: By analyzing the customer's behavior and responses, the company can understand their needs, and this information will help the company to develop innovations that can lead the organization to compete with other companies and gain the customer's loyalty (Sen et al., 2016).

Maximize profits: Big Data can help the organization to prevent bad decisions and increase profits. Big Data can also help to reduce the possibility of human mistake in decision-making by analyzing the factors that help other organizations to succeed in their decision-making process. As soon as an organization has learned to use the power of Big Data, it will help them to develop highly-targeted strategies that will

maximize their profits and minimize inefficiencies (Benjelloun et al., 2015; Ragin, 2017).

Reduced risk: Big Data helps in effective monitoring of risk. Banks, insurance companies, and asset managers can use Big Data analysis strategies and statistical tools to identify expected risks and minimize the response time. They can make a well-informed decision after evaluating a wide range of risk factors (Wielki, 2015; Editorial, 2017).

Improve the competitiveness of the enterprises: Organizations that can capture vast amounts of valuable data and then employ it efficiently are going to have a competitive advantage over the other organizations (Brown et al., 2011). By applying Big Data techniques, organizations can acquire data from different sources such as customers and partners and by combining this data and analyzing it, organizations can gain valuable information (Brown et al., 2011). An additional advantage of the scale of Big Data, as reported by Johnson et al. (2012), is that other departments will benefit from the system of Big Data utilized by one department by gaining the accessibility to the information and data generated by that department. By getting accessibility to all these data, information and then reports generated from this data the organization can save time and cost since there is no replication of work in different departments. If data from both external and internal sources are accessible, more valuable information will be derived. Therefore, decision making will be improved and the organization can easily attain their strategic objectives by earning a competitive advantage in the marketplace via improving their client's satisfaction. With respect to the above-discussed benefits, these benefits can help the organization to define and recognize the business imperatives to the organizations. By evaluating these business imperatives and comparing it with the business imperatives of their organization, they will be able to obtain more accurate insights about when they can benefit from the utilizing of Big Data (Smeda, 2015; Benjelloun et al., 2015).

Development of new products and services: By applying Big Data techniques, the organizations can get a much better understanding of existing market conditions. For instance, by analyzing customers' shopping behaviors, an organization is able to find

out the best selling products and develop new products based on this trend. By this, it could get ahead of its rivals and also make the proper decision that will lead the organization to enhance their services and products to fulfill the customer's requirements (Saqib, Khan, Mahmood, & Naeem, 2015).

Better knowledge of business changes: By analyzing the trends of customer requirements and fulfillment, the organizations are able to develop new services or products based on the demands of customers. Organizations that adopt Big Data analysis are able to gain valuable information and find the new opportunities in the market. Also, it helps the organizations to be more flexible and respond effectively to unexpected change because they continually update information about the opportunities in the market and the customer requirements (Smeda, 2015).

4.1.4 Specific benefits of Big Data for Government agencies

In addition to common benefits of Big Data for SMEs and government agencies listed in the section 4.1.2, there are many other benefits that can be listed as specific to government agencies. This section summarizes these benefits. Before doing this, we give reference list in Tabel 4.4, that enabled us to collect this information.

Table 4.4. Specific Benefits of Big Data for Government Agencies with the Corresponding References

Benefits	References
1. Reducing waste, eradicating fraud and abuse 2. Enhance transparency and service 3. Reduce security threats and crime 4. Reduction in Tax and Social Security Fraud	Pannu et al., 2016; Kim et al., 2014; Bajaj & Johari, 2016; Staff, S., 2013; Bertino, 2013;

Reducing waste, eradicating fraud and abuse: Eliminating fraud is one of the most important benefits of Big Data for the government agencies. Social services agencies are currently utilizing Big Data to recognize fraud and avoid wasting resources. These organizations are able to offer much better services by analyzing data from

their systems and eliminate the internal waste by identifying discrepancies (Pannu et al., 2016).

Enhance transparency and service: Governments that share their huge datasets and encourage the free streaming of information will improve the transparency and increase the trust between government and citizens. It will enable people to realize the data generated and collected by governments and how they use it. Sharing this information will motivate the people and engage them to help governments develop new and innovative services. The transparency will allow people to observe and realize the effect of how governments invest public money and it will push governments to invest public money wisely (Kim et al., 2014).

Reduce security threats and crime: For local government departments, like the ministry of emergency services and police, Big Data gives the ability to detect illegal activities and crime patterns. It can also help local governments and state to work closely together to prevent crimes and illegal activities within the local areas. In order to have this ability, government agencies must process a huge amount of data within the acceptable time period. This could be done by Big Data frameworks easily (Pannu et al., 2016; Kim et al., 2014; Bajaj & Johari, 2016; Bertino, 2013).

Reduction in Tax and Social Security Fraud: In every country, tax claim forms will produce a huge amount of data. It may not be possible to produce such a huge data using traditional data processing technologies, but using Big Data tools gives governments the ability to reduce social security fraud as well as tax fraud of citizens. Algorithms that use pattern recognition can find the suspect transactions happening in real-time. Integrating various datasets like social data set and national data sets will offer further insights into the taxpaying behavior of citizens. Unusual behavioral patterns could be detected to hunt fraudulent activities. Patterns give the ability to determine profiles and statistical variables to recognize suspect transactions, that could be closely monitored**.** Consequently, governments can use social or demographic data to see if those abnormal cases have fraudulent activities, for instance, if someone has sent applications for social security is actually permitted to receive it or perhaps is just enjoying a vacation. As an example application to this,

in the fiscal year 2012, the Department of Health Insurance and Human Services in Australia has made $64.8 billion in unsuitable payments reduced from approximately $66 billion in 2011 (Kim et al., 2014; Staff, S., 2013)

4.1.5 Benefits of using Big Data Frameworks

Preparing Big Data Analytics solution without using Big Data Framework is a very hard and challenging task. Big Data Framework can store and distribute data across hundreds or even thousands of commodity hardware running in a distributed fashion (Hu, Wen, Chua, & Li, 2014). So, Big Data framework offers many advantages which can be summarized as follows:

Scalability: Big Data frameworks can handle huge datasets because it can store and distribute this data across hundreds of commodity hardware running in a distributed fashion.

Cost effective: Big Data frameworks offer a cost-effective storage solution for the organization. Instead of spending a huge amount of money on the expensive advanced hardware and software which can handle the required data operations, Big Data framework can store and distribute the data across hundreds of commodity hardware which saves a significant amount of money.

Flexible: Big Data frameworks can help to acquire a massive amount of data from many different sources dealing with various type of data to generate value.

Fast: Big Data frameworks offer scalable distributed storage, parallel computation and processing the data where is located that lead to reducing the processing and communication time efficiently and process terabytes of data in a short time.

Resilient to failure: Big Data frameworks divide the whole processing into tasks and distribute these tasks to various nodes for processing. One node is monitoring the process in these nodes. If something occurs in a specific task, this task will restart. Therefore restarting many tasks leads to consuming a long time to process these tasks, and this can be eliminated by utilizing Checkpoints which maintain the state of the system at certain intervals of the time. In case of any failure, the processing can

restart from the last checkpoint maintained (Katal, Wazid, & Goudar, 2013). Distributed storage offered by Big Data framework also provides fault tolerance by replications. If any node goes down, data is still available in other nodes.

4.1.6 Benefits of Combining Big Data and Cloud Computing

This section summarizes the benefits of combining Big Data and cloud computing technologies as follows:

Cost Reduction: Big Data implementation requires investing a large amount of money in hardware, software and workforce. Cloud computing offers an on-demand pool of resources which provide the ability to adopt Big Data technology without investing a large amount of money in purchasing and installing the necessary infrastructures for Big Data implementation.

Scalability: The huge and growing volume of data is considered as an essential challenge of Big Data. To store this huge and growing volume of data, cloud computing offers almost unlimited and highly scalable pool of resources which overcome the volume challenge of Big Data and make storing of this data faster and easier.

Agility: Increasing the capacity of Big Data framework in an organization requires to purchase and install new infrastructures, and these steps need time. Cloud computing offers the ability to increase the system capacity of the organizations flexible and rapidly. Therefore, the organizations will improve the capacity of their framework on cloud environment faster than in traditional data centers and avoid the waste of time.

Deployment: Big Data analytics require the organization to deploy and manage a reliable, fault-tolerant, and complex software stack on top of hundreds or thousands of commodity computers. This is another important challenge of utilizing Big Data in organizations and requires the employment of a large number of expert staff. On the other hand, cloud computing offers abstraction in deploying such a large clusters of computers and provides fault-tolerant, reliable, and secure clusters.

4.2 Challenges and solutions

Challenges and solutions section describes the challenges of cloud computing, challenges of Big Data and the solution for each challenge. These challenges are divided into three sections; the first section describes the data challenges, the second section describes process challenges and the last section describes management challenges. This section offers the information that leads us to answer the third and fourth research questions which are given in Chapter 5. These research questions were:

RQ3)What are the different categories of challenges of Big Data faced by SMEs and government agencies?

RQ4)What are the existing solutions for Big Data challenges?

4.2.1 Challenges of Cloud Computing

Data protection: Data protection is an essential aspect which needs more investigation. Organizations are concerned about the leakage of data to competitors and the data privacy of customers. In the current models (on-premise IT service) usually, firewalls protect the accessibility of sensitive information across the data centers. But in the public cloud model, organizations would have to rely on the providers of security service because they are responsible from preserving data security (Saxena & Pushkar, 2016; Jelonek et al., 2014).

Security and privacy: The two most important issues about cloud computing are associated with storing and securing data, and observing the utilization of the cloud via the providers of the service. These challenges are the cause of the decline in the use of cloud services. Several organizations are not comfortable with storing their applications and data on systems which cannot be controlled by them. Consistency about identity management, authentication, access technologies and compliance will become increasingly significant. To reassure their clients, cloud vendors should provide a high level of transparency in their operations (Saxena & Pushkar, 2016).

Data and application interoperability: Standard interfaces are essential for the data and applications systems. Firms will need the flexibility to generate new solutions allowed by applications and data which interoperate together irrespective of where

they reside (conventional IT environments or some collaboration, private clouds that reside inside of an organization's firewall, public clouds). Cloud vendors require supporting interoperability standards to allow organizations to integrate any cloud vendors features into their solutions or bring systems back in-house (Saxena & Pushkar, 2016).

Management capabilities: Despite the existence of several cloud vendors around, the administration of infrastructure and platform are still in its infancy. Capabilities such as auto-scaling are important for many organizations. There is a big possibility to enhance the load balancing and scalability capabilities offered today (Saxena & Pushkar, 2016).

Metering and monitoring: Organizations will need to employ multiple cloud vendors in their IT solutions and evaluate the performance of the system over these solutions. Vendors should offer consistent formats to observe the performance of cloud service and applications and make them suitable for existing monitoring systems. Organizations who use cloud computing efficiently have great opportunities, and these opportunities are not devoid of barriers and risks (Saxena & Pushkar, 2016).

4.2.2 Challenges and existing solutions of Big Data

This section explores the challenges of Big Data categorized into three sections as stated by Sivarajah et al. (2017), this study was up-to-date and their classification is so well and consistent with our findings. They mostly described the challenges, but we extend it by adding two more challenges to the management section which are lack of skills and ethical challenges. Also, we extend it by offering solutions to those challenges. First part of this section describes the data challenges of Big Data and existing solutions for each challenge, the second part of this section describes the process challenges of Big Data and existing solutions for each challenge, the third part of this section describes the management challenges of Big Data and existing solutions for each challenge. This section answers the third and fourth research questions of our study which were:

RQ3) What are the different categories of challenges of Big Data faced by SMEs and government agencies?
RQ4) What are the existing solutions for Big Data challenges?

4.2.2.1 Data Challenges & Solutions

This section describes the data challenges of Big Data in terms of 7Vs characteristics and the solution for each challenge as shown in Figure 4.1.

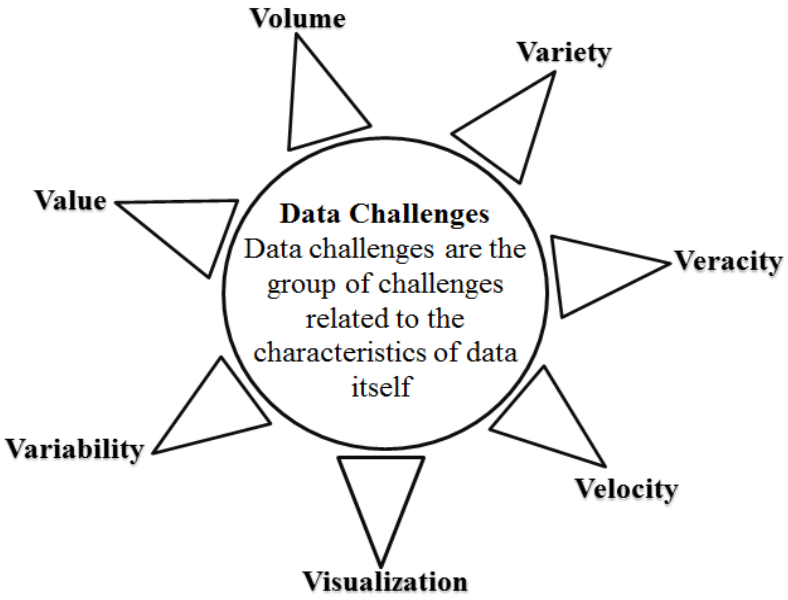

Figure 4.1 Data challenges of Big Data (Sivarajah et al., 2017).

Volume challenge: Large-scale and massive amount of data is a significant challenge in Big Data. As reported by Barnaghi et al. (2013), heterogeneity, ubiquity, and dynamic nature of data produced by various devices and resources, as well as the large-scale itself, make the identifying, retrieving, processing, combining, and inferring data a challenging task (Sivarajah et al., 2017). The volume of data is significantly growing every minute, approximately 800,000 Petabytes of data stored across the world in 2000 is estimated to jump to 35 Zettabytes in 2020. Facebook produces around 10 Terabytes every single day, Twitter produces about 7 Terabytes and some organizations generate terabytes in an hour (Nasser & Tariq, 2015). The problem is how to handle the huge and varying size of Big Data.

The solution of volume challenge: Big Data frameworks like Hadoop and Spark can handle the volume challenge of Big Data because they offer scalable distributed

storage and parallel computation (Kanchi, Sandilya, Ramkrishna, Manjrekar, & Vhadgar, 2015). Cloud computing provides an unlimited pool of resources which can handle the scalability of Big Data more easily, quickly and reduce the cost (Saxena & Pushkar, 2016). So, the best solution is to have Big Data on Cloud.

Variety challenge: Different formats of data (structured, semi-structured and unstructured) are considered as a significant issue related to the data challenges. Big Data is captured in different forms, coming from various sources such as data generated by users-transactional data, messages, scientific data, healthcare data, web data and so on. These diverse types and quality of data apparently show that heterogeneity is a characteristic of Big Data and it is also a major challenge to understand and handle such data. The problem is how to manage a multiplicity of formats, types and sources. (Sivarajah et al., 2017).

The solution of variety challenge: In batch processing, for managing the heterogeneity of data, tools such as Hive, Pig and Mahout are available as components of most Big Data Framework. It is exciting to note that all these tools utilized Hadoop over HDFS as the fundamental architecture (Durgude, Yalij, Bhosale, & Bharati, 2015). In real-time processing, to handle the heterogeneity of data, ETL tools can be used in a pre-processing stage like Spark SQL which is a part of the Spark ecosystem (Guller, 2015).

Velocity challenge: This refers to the speed at which data are generated and processed. Data velocity is increasing with time. In the past, organization analyzed data utilizing batch processing systems due to the expensive and slow nature of data processing. While the speed of generating and processing data raised, real-time processing started to be a criterion for computing applications. As reported by Gartner (2015), 6.4 billion devices will be connected and will be in use around the world in 2016 and the number will increase to reach around 20.8 billion by 2020. In 2016, approximately 5.5 million new devices will be linked every day to acquire, analyze, and share data. The improved data streaming ability of linked devices will accelerate the velocity constantly (Lee, 2017). The problem is how to respond to the high flow rate of data.

The solution of velocity challenge: There are three things to be considered when dealing with velocity problem. The first thing is about the fusion of data from disparate sources with high speed. For this purpose, we must have a high bandwidth of the network, and an efficient method of building a data pipeline, which can be done by using Apache Kafka, Apache Flume, and the similar tools such as AMQP, RabbitMQ, etc. They can be integrated with distributed real-time stream data processing frameworks such as Apache Spark, Apache Storm, FLink, etc in order to deal with a high velocity of data acquired from many sources (Namiot, 2015). The Cloud environment would be the best choice for constructing such a big system architecture.

Veracity challenge: This challenge is not only about the quality of data, but more about comprehending the data. Although veracity looks like validity or volatility, it is more about the trustworthiness, confidence, and truth of data. Zicari (2014) and Akerkar (2014) indicate veracity as dealing with the messiness, biases, imprecision, doubts and misplaced evidence within the data. Veracity measures the precision or correctness of data as well as the possibility of utilizing it for analysis. For example, each customer viewpoint on various social media websites and networks is diverse and ambiguous in general, since it includes human interaction (Sivarajah, Irani, & Weerakkody, 2015). Furthermore, the internet is a soft medium to broadcast and publish derived information from multiple sources, and it is important to separate the wheat from the chaff if you like to introduce quality data.

The solution of veracity challenge: Since there may be more than one source of information in regard to the same analysis, with different truth, it is important to decide whether is to be used in an analysis or not. Veracity challenge is multi-dimensional and needs careful consideration of each dimension depending on the context. Therefore, there is no single approach for all different contexts from a different dimension. There are many context-dependent solution propositions in the literature, such as Hui and his colleagues proposition for mobile context (Lin, H., et al., 2017), and the solution proposed by Tre, et al. (2017) in the context of Multi-Criteria Decision-Making (MCDM). Furthermore, the demand to cope with

ambiguous and inaccurate data can be addressed by utilizing analysis tools designed for handling and mining of unreliable data (Gandomi & Haider, 2015).

As of our thought, although there are many analysis tools for every dimension (such as accuracy, truth, correctness, confidence,etc.), there still a need of human interception in the decision of whether to use the source of data in the analysis. As a result, we think that there is a gap in the literature about veracity challenge.

Visualization challenge: Visualizing data is about representing knowledge and essential information in an efficient and instinctive way by utilizing diverse visual formats like the pictorial or graphical layout. For example, eBay company has millions of users, and every month they sell millions of goods, and as a result, a massive amount of data is generated. As stated by Chen and Zhang (2014), for lots of current Big Data applications which have insufficient performances in scalability, response time and functionalities, data visualization is a challenging task due to the high dimension and large sizes of Big Data (Sivarajah et al., 2017).

The solution of visualization challenge: Many tools exist to visualize the Big Data. Before chosing any visualization tool, one should review the requirement and choose an appropriate tool. Most of the popular tools for Big Data visualization are Tableau, Apache Zeppelin, and so on. (Ali, Gupta, Nayak, & Lenka, 2016). In order to make all of these data understandable, the eBay company uses Tableau tool which is one of the Big Data visualization tools. This tool has the ability to transform complex and large datasets into spontaneous depictions. According to these interactive outcomes, the employees in eBay company can visualize the importance of the quality and search to monitor the newest feedbacks from customers and perform sentiment analysis (Sivarajah et al., 2017).

Variability challenge: (e.g., data changing continuously): This is another essential feature of Big Data which is usually confused with the Variety characteristic of data. For example, Facebook or Google repository generates and stores diverse types of data. Also, if one of these various forms of data is delivered to use for mining and to make sense, this kind of data provides multiple meaning, then this is the variability of data which mean data is rapidly and continuously changing. The rates of change

and variability characteristic of data generated by human and machine are much higher than the data produced by processes. Variability is also associated with performing sentiment analysis. For instance, in the same tweets, a word may have an entirely various meaning. To be able to perform an efficient sentiment analysis, the advocates confirm that algorithms should be able to understand the context and decipher the exact meaning. However, this is still a big challenge (Sivarajah et al., 2017).

The solution of variability challenge: there is no solution available in the literature about this challenge. So it is considered as a research gap and needs more investigation and more research work.

Value challenge: (e.g., extracting value and knowledge from massive amounts of different forms of data for end users without loss): Storing Big Data is a complex process. For example, important values could be obtained from the flow of clicks by the users of the internet, and also this has become the spine of the internet economy. The researchers in the field of Big Data consider the value as a crucial feature because data is available anywhere, and an valuable information can be extracted on the fly. Irrespective of the dimensions utilized to characterize Big Data, the challenges of managing, storing and extracting value from the data in a cost-effective way are still faced by the organizations (Sivarajah et al., 2017).

The solution of value challenge: To maximize the value of Big Data we should consider the following stages:

- Organizations need to determine what data should be analyzed to generate value. Since Big Data is generated from various sources, the organizations will need to determine the strategy that gives them the ability to access those datasets.
- Selecting the right tools for Big Data analysis according to the organization's requirements is needed.
- Skilled data scientists are required to utilize the Big Data effectively. The organization also needs strategists who comprehend how best to set up the

tool and tacticians who arrange and adjust data into operational models (Lamba & Dubey, 2015).

4.2.2.2 Process Challenges & Solution

This section explores the process challenges as shown in Figure 4.2, and the solution for each challenge is discussed.

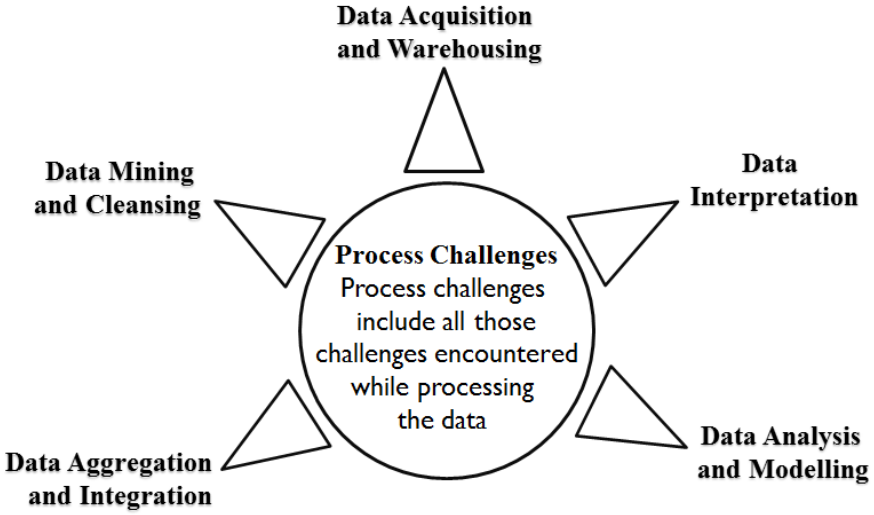

Figure 4.2 Process challenges of Big Data (Sivarajah et al., 2017).

Data acquisition and warehousing challenge: This issue relates to the collection of data from different sources and stores it to generate value from this data. The integral complication of Big Data and continuous growing requires developing efficient algorithms for data acquiring and storage (Wang and Wiebe, 2014). Last debate is supported by Paris, Donnal, and Leeb (2014) who confirm that essential obstacles to the Big Data analysis come from insufficient data provenance, inconsistencies and scale ingrained in data gathering and processing. Although restricted speed and quality of data can be gathered and stored, this would influence the ability to extract worthy information from data (Chen and Zhang, 2014). Smart filters that should be intelligent and powerful are necessary to capture valuable information and ignore useless information that includes inconsistencies or imprecision. So, effective analysis algorithms are needed to comprehend the source of data, process the massive amount of streaming data and to minimize data before storing (Sivarajah et al., 2017).

The solution of data acquisition and warehousing challenge: Tools like Apache Flume and Apache Kafka can be used to gather and store data from different sources in different formats with high inflow rate and store it. These tools can be integrated with many open source frameworks such as Hadoop (framework for batch processing) and Storm (for real-time processing) (Lyko, Nitzschke, & Ngonga Ngomo, 2016). Having good preprocessing algorithms are also helpful in solving this kind of challenges.

Data mining and cleansing challenge: This issue is related to the process of extracting and cleaning data from the repository of the massive heterogeneous dataset. Big Data Analysis experts understand that determining the best method to extract and clean the Big Data can result in significant value and impact. Because of its vibrant, diverse, interrelated, strident and unreliable features, the cleansing, mining, and analysis considered as complicated issues. For example, in the United Kingdom, the National Health Service (NHS) have millions of medical records for patients consisting of prescriptions, medical reports, x-ray data, etc. Medical doctors use such data – if as an example wrong information is stored this could result in incorrect diagnosis situations, leading to imprecise medical records. To be able to take advantage of using this massive amount of data efficiently, there is a necessity to improve the mining technique that extracts the needed information from the Big Data and formulate it in a standard form. As reported by Labrinidis and Jagadish (2012) maintaining and developing this mining technique is still a challenge.

The solution of data mining and cleansing challenge: To solve this problem we can use a tool like Weka which consists of a combination of machine learning algorithms for data extraction tasks and includes tools for data preprocessing, regression, clustering, association rules, classification and visualization (Eibe Frank, Mark, Hall, & Ian, 2016). For data cleansing we can use tools like OpenRefine which deal with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data (Ham, 2013).

Data aggregation and integration challenge: This issue is related to aggregate and combining clean data extracted from the massive amount of heterogeneous data. In most cases, data is collected from different online activities like microblogging,

tweets – retweets, likes and posts on Facebook that primarily bear different senses and meanings. The aggregation of these data obviously goes beyond the capabilities of existing systems for data integration. With respect to Karacapilidis, Tzagarakis and Christodoulou (2013), smart integration for a large volume of unstructured data to create new knowledge and to improve decision-making process is still a key challenge. The provenance and indecision of data are also considered as a big issue for data aggregation and integration. Collecting data in warehouses is another challenge (Sivarajah et al., 2017).

The solution of data aggregation and integration challenge: To solve this problem we can use a tool like Apache Flume which is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data (Vohra, 2016). For solving the data integration challenge we can use Pentaho Data Integration (aka Kettle) which is a full-featured open source offering the ability to explore, transform, validate, and integrate your data from scratch (Roldán, 2010).

Data analysis and modeling challenge: When the data has been captured, stored, mined, cleansed and integrated, the next step is the analysis and modeling. Naturally, Big Data is noisy, heterogeneous, dynamic, and unreliable. Therefore, the analysis and modeling of Big Data are very different from the analysis of the traditional data. Although these two technologies have the same goals; to generate business value by analyzing data, they are different in the analysis techniques and organizing data techniques. Therefore, old methods of data modeling are no more utilized because of the demand for unmatched resources for storage, capacity, efficiency and computing power. Subsequently, there is a demand for new ways to handle the Big Data to maximize business value and impact. It does not simply understand the current trends, also require predicting what may possibly occur in the future by proper data modeling and analysis (Sivarajah et al., 2017).

The solution of data analysis and modeling challenge: Many tools are used for querying and analyzing Big Data such as Apache Drill which is a distributed system for interactive ad-hoc analysis of large-scale datasets, designed to handle up to petabytes of data spread across thousands of servers. Drill supports a variety of

NoSQL databases and file systems, including HBase, MongoDB, MapR-DB, HDFS, MapR-FS, Amazon S3, Azure Blob Storage, Google Cloud Storage, Swift, NAS and local files. A single query can join data from multiple datastores (Hausenblas & Nadeau, 2013). For real-time analysis, we can use tools like Spark, R and Splunk (Stearley, Corwell, & Lord, 2010).

Data interpretation challenge: This issue is comparatively similar to visualizing data and generating understandable information to the users. Presenting the results of data analysis and modeling to the decision makers to understand the outcomes of extracting knowledge and sense (Simonet, Fedak, & Ripeanu, 2015). The incredible growth of heterogeneous data has extremely impacted the way of how people process and explicate further insight from raw data. Most of the data are considered as an online resource; the single available issue is determining how Internet computing technological solutions are developed to enable access, aggregate, analyze, and explicate Big Data (Bhimani & Willcocks, 2014). An additional issue is the lack of skilled people to interpret data (Phillips-Wren & Hoskisson, 2015).

The solution of data interpretation challenge: Systems with rich visualization tools are important to express the outcomes of queries and analysis in the most understandable method for various domains (Nasser & Tariq, 2015). Many tools are used to visualize the Big Data, before choosing any visualization tool we should review our requirement and choose the appropriate tools. Most of the popular tools for Big Data visualization are Tableau, QlikView, etc.(Ali et al., 2016).

4.2.2.3 Management Challenges and solutions

This section describes the management challenges of Big Data, as shown in Figure 4.3, and the solution for each challenge is given after its description.

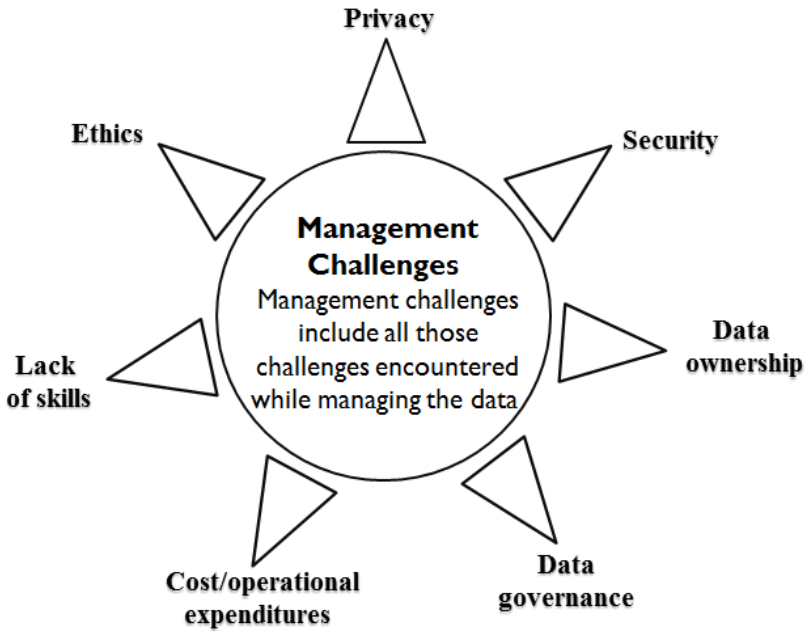

Figure 4.3. Management challenges of Big Data (Sivarajah et al., 2017).

Privacy challenge: Big Data presents many privacy issues and finding the right way to maintain privacy in the information age is a crucial concern. In Big Data projects significant investments have been performed to simplify processes; nevertheless, organizations are still having problems in handling the privacy challenge and this impedes organization in moving ahead in their initiatives to utilize Big Data. Data gathered by sensory devices in a smart city environment that can be accessed have significant privacy issues faced by the government and security agencies. As an example of the privacy-related challenges, information about the location of users being gathered by Big Data applications is causing big privacy concerns. For instance, location-based service vendors can recognize customers by tracking their location-based information – that is potentially related to their workplace or residential information (Sivarajah et al., 2017). Subsequently, there is the concern of protecting privacy and using the data of citizens without permission, which is considered as illegal and should be exposed to the related Government agencies (Machanavajjhala and Reiter, 2012).

The solution of privacy challenge: To protect privacy, there are two common approaches. The first approach is to restrict access to the data by including access control or certification to the data records. Therefore sensitive information will be accessible by a restricted group of users. The second approach is to anonymize data records such that very sensitive information cannot be pinpointed to a personal record. For data anonymization, the primary aim is to inject randomness into the data to guarantee some privacy objectives (Jaseena & David, 2014).

Security challenge: Security is a significant challenge and is recognized by (Lu, Zhu, Liu, X., Liu, J., & Shao, 2014) who debate that if security issues are not adequately identified then, Big Data will not be accepted universally. Securing Big Data has special problems which are not deeply different to conventional data. One of many Big Data security challenges is the distributed feature of Big Data which is complicated but equally vulnerable to attack. Another security issue as reported by Abawajy, Kelarev, & Chowdhury, (2014) that malware has been increasing the possibility of threatening data security. The absence of the sufficient security controls to assure information, such as logs of analysis, network streams, and system activities, is also considered as a security challenge. Another challenge is the absence of an advanced infrastructure that assures data security such as confidentiality, availability, accountability, and integrity. Data security challenges are increasing due to the ubiquitous of the data sources (Sivarajah et al., 2017).

The solution of security challenges: Security of Big Data can be improved by utilizing the strategies of authorization, authentication, audit trails and encryption. There is usually a potential occurrence of security infringement by unauthorized access or unacceptable access by privileged users. There are some techniques used for protecting Big Data which can be summarized as:

***Using authentication techniques*:** Authentication is the process of verifying user or system identity before permitting the accessibility to the system. Authentication techniques such as Kerberos, LDAP and the like can be utilized for this purpose.

***Use file encryption*:** Encryption assures privacy and confidentiality of user information, and it protects the sensitive data. Encryption of File layer offers consistent protection across various platforms irrespective of OS/platform type. Encryption fulfills our requirements for Big Data security. Open source products are offered for most Linux systems; commercial products additionally provide external key management and full support. This is a cost-effective method to cope with many data security threats but with the trade-off on performance.

***Use key management*:** File layer encryption is not efficient if an attacker has the ability to access the encryption keys. Several Big Data cluster administrators keep keys on local disk drives because it is rapid and easy, but it is also insecure as keys can be gathered by the attacker or platform administrator. They should employ the key management service to distribute certificates and keys and control various keys for each application, group and user. Most of the cloud service providers and open source cloud software has the service for key storage and management in a secure way, such as Barbicon service used in OpenStack environment.

***Logging*:** To audit the attacks, identify failures, or inspect abnormal behavior, we require a record of activity. In comparison with less scalable data management platforms, Big Data is a natural fit for gathering and handling of such event data. Many web companies begin with Big Data especially to manage log files. It provides us a place to look when something fails, or if someone thinks you might have been hacked. So to fulfill the security requirements, we need to audit the whole system on a periodic basis.

Use secure communication: Apply secure communication among the nodes and applications. This needs an SSL/TLS implementation that in fact helps to protect all network communications instead of protecting only a subset (Jaseena & David, 2014).

Data governance challenge: While the need for Big Data is continuously growing, organizations believe that data governance is a possible strategy to warranting the quality of data, enhancing and using information, preserving its value as a primary

organizational asset, and help in acquiring insights in business operations and decisions. As reported by the Intel IT Center (2012), IT managers extremely support the presence of a formal Big Data approach, this specifically creates sense, since the issue of data governance for explaining what data is stored, analyzed, and assessed is classified as one of the top three issues they face (in addition to the growth data, data center infrastructure and the capability to offer scalability). A major issue in the process of Big Data governance is classifying, mapping and modeling the data as it is collected and stored, primarily due to the complex and unstructured nature of data. Furthermore, efficient Big Data governance is fundamental to make sure of the quality of data extracted and analyzed from a repository of large datasets (Hashem et al., 2015).

The solution of data governance challenge: Many tools are used to handle the issue in data governance like Talend Big Data, Apache Atlas and Cloudera Navigator which are the most common Hadoop extensions to address the specific challenges of data governance **(**Franco, 2017).

Cost/operational expenditures challenge: The continuously growing data in diverse forms have increased the need for Big Data processing in advanced data centers. These are usually dispersed via different geographical areas to embed flexibility and prevalence risk, for instance, Google having thirteen data centers in 8 countries distributed through four continents. Important resources have been designated to support the data-intense operations like warehousing, acquisition, cleansing and mining, aggregation and integration, interpretation. All of this drive to great costs because of the demands of high storage and data processing. Advocates of Big Data search for cost-effective and efficient methods to manage the huge volume of complicated data. Data processing costs and other expenditures related to the operations in the data center are a critical matter that could also influence in the organization's methods to choose the proper technological solutions for implementation (Sivarajah et al., 2017).

The solution of cost/operational expenditures challenges: Organizations can easily prevent wasting big amounts of funds on buying and installing their IT applications or infrastructure by switching to the cloud model. Cloud providers stated that cloud computing enables organizations to improve procurement processes, and reduces the demand to replicate certain computer administrative skills relevant to setup, support and configuration. Cloud computing provides an unlimited pool of resources which can handle the scalability of Big Data to be easier, quicker and inexpensive (Saxena & Pushkar, 2016).

Lack of skills challenge: One of the most important challenges of managing Big Data in SMEs and government agencies is the lack of skills. Utilizing Big Data is a new concept for many of these organization and they do not have skilled people to use Big Data technology well and to apply sufficient methods for analyzing it based on the demands and the requirements of the organization. Therefore, preparing skilled staff to deal with using of Big Data is a challenge and should be considered when the organizations intend to adopt this technology (Miller, 2014).

The solution of Lack of skills challenge:

- Designing training courses and workshops to teach the required skills and knowledge about Big Data technology.
- Establishing a partnership between academic and businesses and government sectors, design internship and increase the collaboration between them to close the Big Data skills gap.
- Create open online communities to share opinions and interested in the field of Big Data skills (Miller, 2014).

Data ownership challenge: In addition to privacy, ownership of data is a complicated issue. As reported by Kaisler et al. (2013), data ownership presents a crucial and continuous challenge, especially in the social media context such as who is the owner of the data on Twitter, Facebook or MySpace. Generally, the prevailing view is the provider of social media and the users own the data. Kaisler et al. (2013) argue that this challenge still needs to be settled. For instance, data produced by the sensor is too sensitive and can lead in growing errors – this could moreover result to

capture and reveal inconsistent data and on the other hand who is the owner of this data. As a result of challenges mentioned above, data ownership is a significant social issue that needs additional investigation since it can have deep implications (Sivarajah et al., 2017).

The solution of data ownership challenge: As a part of the solution to solve this challenge, is to anonymize data records such that very sensitive information cannot be pinpointed to a personal record. For data anonymization, the primary tool is to inject randomness into the data to guarantee a number of privacy objectives (Jaseena & David, 2014), but data ownership is still a significant challenge and needs additional investigation since they could have deep implications (Sivarajah et al., 2017).

Ethical challenge: Large datasets are being mined for significant predictions which usually produce amazing insights. As reported by Richard and King (2014), Big Data analysis used to examine diverse kinds of human activities and the process of decision-making are beginning to be affected by Big Data predictions, such as shopping, dating, education, voting, medicine, cybersecurity, terrorism prevention and law enforcement. However, while this is happening, people have limited understanding of what data is being gathered, as well as shared with third parties. Therefore, current privacy protections concentrated on handling personal information are not sufficient. Big Data initiatives find many personal datasets like the call history, social network connections, location history, purchase history, search history and more of this information are already in the governments and corporations hands, and the gathering of these and other datasets are increasing. Therefore, to protect individual rights, we need to develop Big Data ethics (Richards & King, 2014).

The solution to ethics challenge: To solve this problem we need to develop formal laws that determine how we can use the data. It will be an unethical issue if the store manager listens to private conversations with their customers to develop the marketing capabilities even if the customers are available in the store. Social networking also has the same issue; there are no formal laws to protect social

networking users. We should rely on the data owners to be responsible for their collected data and how they are using it (Shim, French, Guo, & Jablonski, 2015).

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The primary aim of this research is to investigate and explore the benefits of Big Data for SMEs and government agencies and how cloud computing can help these organizations to adopt Big Data technology and gain its benefits. Although adopting Big Data technology offers many benefits, there are many challenges faced by organizations when using this technology. One objective of this thesis is to identify and describe these challenges. Finally, this study discovers the existing solutions for the challenges of adopting Big Data.

5.1 Discussion of findings

In order to achieve the objectives of this thesis, four research questions were formulated and stated in Chapter 1, and these questions were answered in detail in Chapter 4. The research questions were answered by conducting a systematic literature review for the studies that investigated the related concepts on Big Data and cloud computing technologies.

In this section, we give the list and summaries of our findings in addition to the analysis of the answers to the research questions.

RQ1) What are the benefits of combining Big Data and Cloud Computing technologies?

The most common benefits were cost reduction, scalability and agility. When combining these two technologies, organizations will gain extra benefits in addition to benefits of Big Data and also some challenges of Big Data will be solved like Volume challenge of Big Data and cost challenge. These benefits are given in

Chapter 4 with full description, in this section. These benefits can be listed as follows:

1. **Cost Reduction**
2. **Scalability**
3. **Agility**
4. **Deployment**

By combining Big Data and cloud computing technologies, four benefits will be gaining. Most important benefit as mentioned in the collected studies is the cost reduction which stated by 15 studies.

The following table shows the number of studies that mentioned the benefits of combining Big Data and cloud computing technologies.

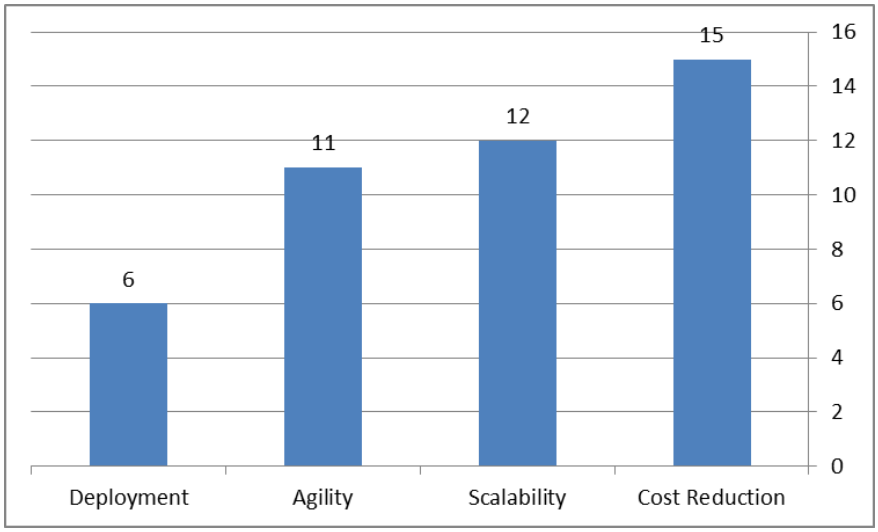

Figure 5.1. Frequency of Studies on the Benefits of Combining Big Data and Cloud Computing

RQ2) What are the various benefits of Big Data on the cloud for SMEs and government agencies?

These benefits are categorized into three sections, the first section clarifies and explores the common benefits of Big Data on the cloud for SMEs and government agencies. Some of the benefits of Big data for SMEs are not applicable to be benefits for government agencies and vice verca. Therefore the second section explores the

specific benefits of Big Data for SMEs. The third and last section explores the specific benefits of Big Data for government agencies.

Common benefits of Big Data on the cloud for SMEs and government agencies

This section clarifies and explores the common benefits of Big Data on the cloud for SMEs and government agencies. These benefits are given in Chapter 4 with full description. In this section, these benefits can be listed as follows:

1. **Improving decision making**
2. **Improve organizational productivity**
3. **Cost reduction**
4. **Increased operational efficiency**
5. **Predict workforce requirement**

The most important benefit that stated by the analyzed studies is improving the decision making process for SMEs and government agencies sectors. 28 studies stated that adopting Big Data on cloud improve the decision making process for SMEs, and 5 studies reported the same thing for government agencies.

The following figure shows the number of studies that mentioned the common benefits of Big Data on the cloud for SMEs and government agencies.

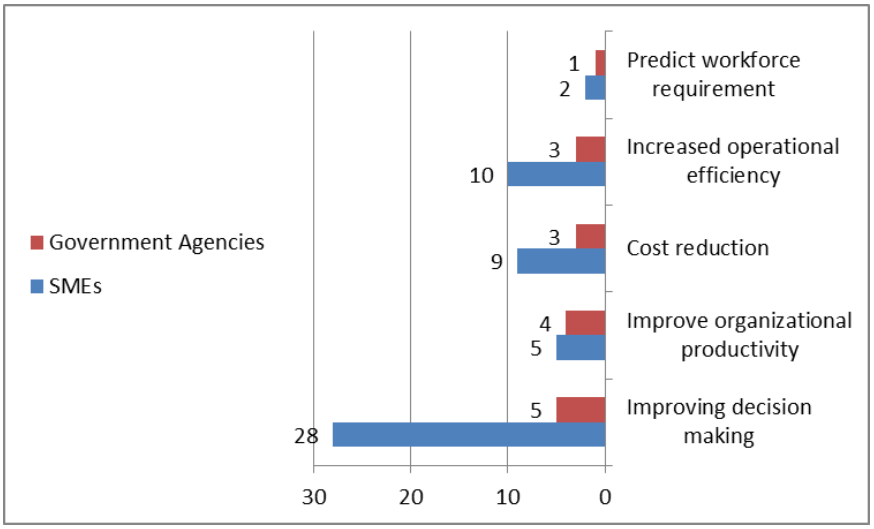

Figure 5.2. Frequency of Studies on the Common Benefits of Big Data on the Cloud for SMEs and Government Agencies

Specific benefits of Big Data on the Cloud for SMEs

This section clarifies and explores the specific benefits of Big Data on the cloud for SMEs. These benefits are given in Chapter 4 with full description. In this section, these benefits can be listed as follows:

1. **Innovation based on customer responses**
2. **Maximize profits**
3. **Reduced risk**
4. **Improve the competitiveness of the enterprises**
5. **Development of new products and services**
6. **Better knowledge of business changes**

As stated by the collected studies, adopting Big Data on Cloud can help the organizations to develop new products and services by analyzing the market conditions and the most selling products. 16 study reported that most important benefit after improving decision making is the development of new products and services.

The following figure shows the number of studies that mentioned the specific benefits of Big Data on the cloud for SMEs.

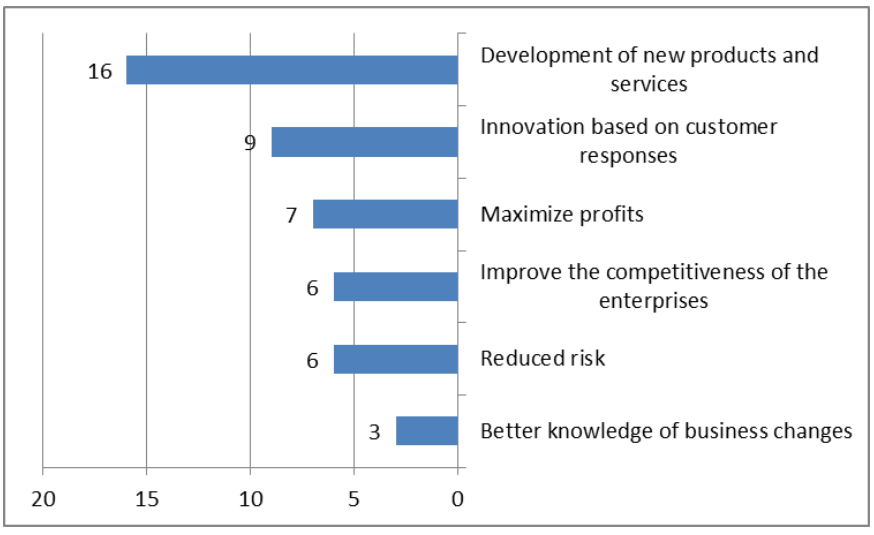

Figure 5.3. Frequency of Studies on the Specific Benefits of Big Data on the Cloud for SMEs

Specific benefits of Big Data for government agencies

This section clarifies and explores the specific benefits of Big Data on the cloud for government agencies. These benefits are given in Chapter 4 with full description. In this section, these benefits can be listed as follows:

1. **Reducing waste, eradicating fraud and abuse**
2. **Enhance transparency and service**
3. **Reduce security threats and crime**
4. **Reduction in Tax and Social Security Fraud**

In government agencies sector, reduce security threats and crime is the most important benefit which stated by four studies after the improving decision making benefit.

The following figure shows the number of studies that mentioned the specific benefits of Big Data on the cloud for government agencies.

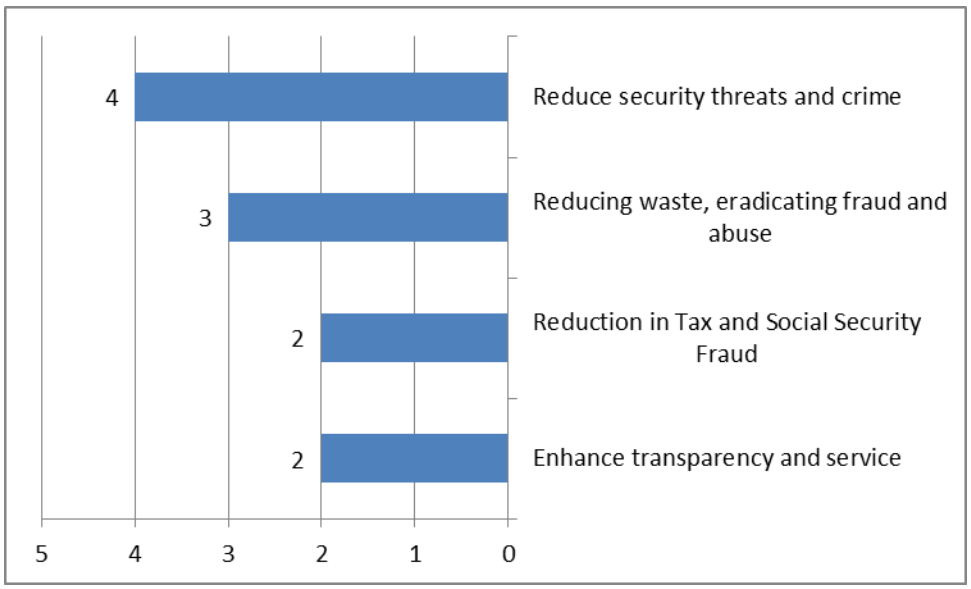

Figure 5.4. Frequency of Studies on the Specific Benefits of Big Data on the Cloud for Government Agencies

RQ3) What are the different categories of challenges of Big Data on cloud especially faced by SMEs and government agencies?

These challenges are categorized into three sections, data challenges related to the characteristics of data itself, process challenges related to the challenges of processing Big Data and management challenges related to the challenges of Big Data management.

Data challenges

This section describes the data challenges of Big Data in terms of the 7Vs characteristics. These challenges are given in Chapter 4 with full description, these challenges can be listed as follows:

1. **Volume challenge**
2. **Variety challenge**
3. **Velocity challenge**
4. **Veracity challenge**
5. **Visualization challenge**
6. **Variability challenge**
7. **Value challenge**

By analyzed the studies that are investigated the data challenges of Big Data, Variety is considered as the most critical challenge as stated by 20 studies. The following figure shows the number of studies that mentioned the data challenges of Big Data.

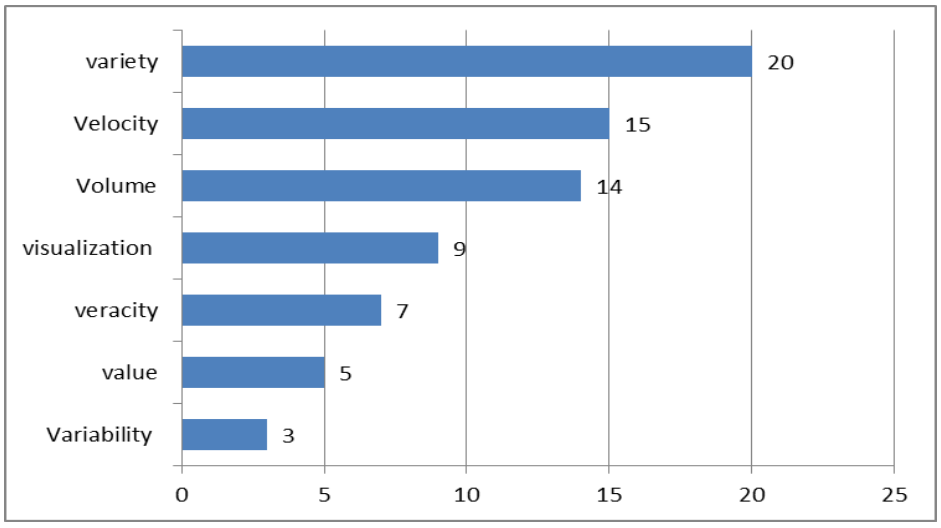

Figure 5.5. Frequency of Studies on the Data Challenges of Big Data

Process challenges

This section describes the process challenges of Big Data. These challenges are given in Chapter 4 with full description. These challenges can be listed as follows:

1. **Data acquisition and warehousing challenge**
2. **Data mining and cleansing challenge**
3. **Data aggregation and integration challenge**
4. **Data analysis and modelling challenge**
5. **Data interpretation challenge**

In process challenges category, the most important challenge is data analysis and modelling which reported by 19 studies.

The following figure shows the number of studies that mentioned the process challenges of Big Data.

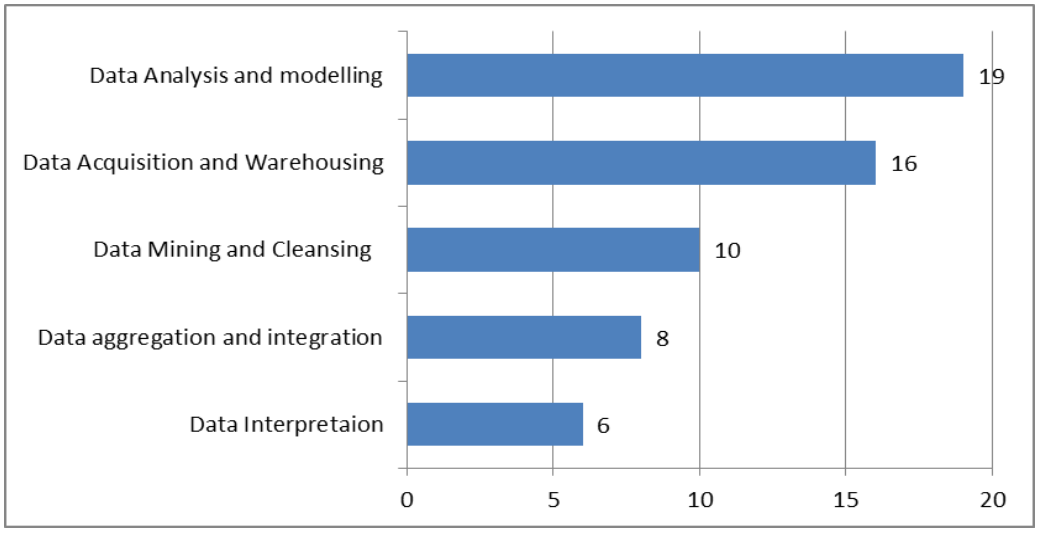

Figure 5.6. Frequency of Studies on the Process Challenges of Big Data

Management challenges

This section describes the management challenges of Big Data. These challenges are given in Chapter 4 with full description. These challenges can be listed as follows:

1. **Privacy challenge**
2. **Security challenge**
3. **Data governance challenge**
4. **Cost/operational expenditures challenge**
5. **Lack of skills challenge**
6. **Data ownership challenge**
7. **Ethical challenge**

As reported by the collected studies, the most critical challenge among the data challenges of Big Data is the security, this challenge stated by 35 studies as an important issue.

The following figure shows the number of studies that mentioned the process challenges of Big Data.

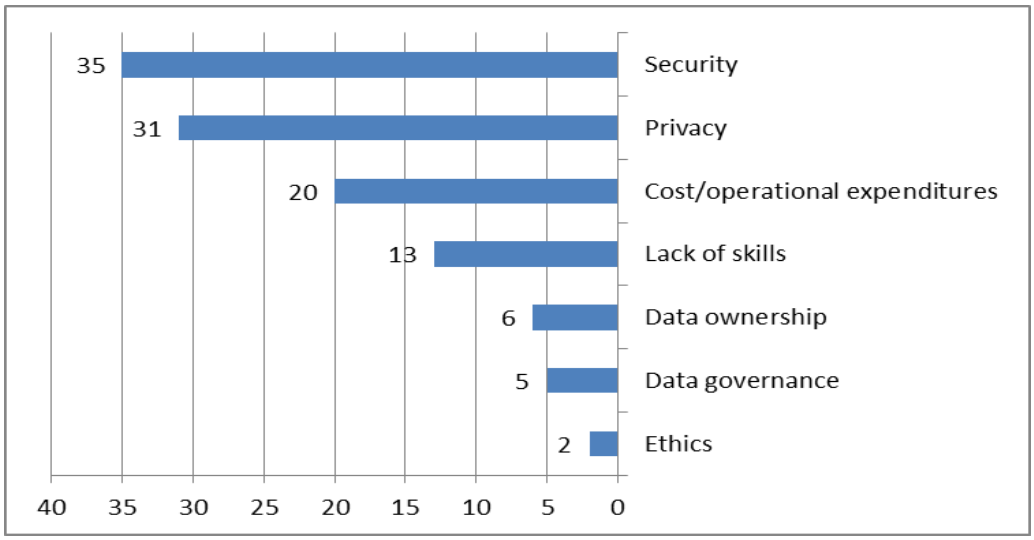

Figure 5.7. Frequency of Studies on the Management Challenges of Big Data

RQ4) What are the existing solutions for Big Data challenges?

The existing solutions for each challenge of Big Data are given in detail in Chapter 4. After analyzing these solutions, we highlight that some of the challenges of Big Data

can be solved by using Cloud Computing technology such as; volume, scalability and cost challenges. Most of the challenges of Big data can be solved by using Big Data framework and by having a number of strategies which are recommended by some of the studies. We found that there is a need for more research in dealing with a few challenges, such as; data ownership, veracity and variability.

5.2 Conclusion

As a result of the analysis of the papers from the SLR, it can be stated that Big Data brings many opportunities to SMEs and Governments, as explained in Chapter 4. Besides of these benefits, there are many challenges that organizations have to deal with. The most important of which is about the capital expenditure (CapEx) needed to deploy Big Data environments. Fortunately, Cloud Computing provides a solution by replacing CapEx with OpEx (Operational Expenditure) which reduces the cost of ownership. Furthermore, Cloud environment helps organizations to deal with the other important challenges, such as data ownership, veracity, and variability, as stated in the previous section. Nevertheless, cloud environment inherently brings additional benefits such as; scalability, agility, and ease of deployment. In addition to these benefits, adopting Big Data on Cloud environment can help the organizations to gain many benefits such as; improving decision making, improving the organizational productivity, reducing the costs of operations, increasing the operational efficiency and predicting the workforce requirements. According to the 28 papers analyzed, the most important benefit for SMEs is improving decision making process, and 5 other studies reported the same for the government agencies. After analysing the challenges faced by SMEs and government agencies, the most important ones stated by the collected studies are; the variety challenge in the data challenges category which was stated by 20 studies, the data analysis and modelling challenge among process challenges category which was mentioned in the 19 collected studies, and the security challenge in the management challenge category which was reported by 35 studies. Although some solutions to these challenges are offered in this study, there still remains a few challenges that need to be investigated. These are, namely, data ownership, veracity and variability. Finally, it would be better to note that organizations should select the right tools and strategies according to their resources and requirements in order to store and process their data and to solve the challenges faced by them during these operations.

5.3 Research limitations

We proposed in the beginning to design a survey for the SMEs and government agencies adopting Big Data on cloud environment about the benefits, challenges and solutions. The lack of such organizations led us to avoid this survey and depending on the studies that were available in the literature.

5.4 Future directions

The following can be considered as potential research that can extend this research:

- Designing a survey for the SMEs and government agencies which adopting Big Data on cloud environment about the benefits, challenges and solutions.
- Investigation and evaluation of the case studies about adopting Big Data on the cloud for SMEs and government agencies.
- Designing or developing a conceptual model for Big Data on the cloud that will be suitable for SMEs and another model for government agencies.
- Investigation and researching the solutions for the data ownership, veracity and variability challenges.

REFERENCES

Abebe, M. (2014). Electronic commerce adoption, entrepreneurial orientation and small- and medium-sized enterprise (SME) performance. *Journal of Small Business and Enterprise Development*. 21 (1). p.pp. 100–116.

Abawajy, J. (2015). Comprehensive analysis of Big Data variety landscape. International Journal of Parallel, Emergent and Distributed Systems, 30, 5–14.

Abraham, A. & Paprzycki, M. (2004). Significance of steganography on data security. In: *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.* 2004, IEEE, pp. 347–351.

Accenture. (2016). *Big Data analytics in supply chain: Hype or here to stay*. Available online from: https://www.accenture.com. [Accessed: 22 August 2017].

Acharekar,T.& Palghadmal, A. (2016). Big Data for managing human resources: The Haves and the have-nots. *International Research Journal of Engineering and Technology (IRJET)*. 2016, *IRJET*, pp. 33–317

Aghabozorgi, S., Seyed Shirkhorshidi, A. & Ying Wah, T. (2015). Time-series clustering – A decade review. *Information Systems*. 53. p.pp. 16–38.

Akerkar, R. (2014). *Big Data computing*. Florida, USA: CRC Press, Taylor & Francis Group.

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016, December). Big Data visualization: Tools and challenges. *In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on (pp. 656-660)*. IEEE.

Alshwaier, A., Youssef, A. & Emam, A. (2012). A New Trend For E-Learning In KSA Using Educational Clouds, Advanced Computing. *International Journal*. 3 (1).

Ammn, N. & Irfanuddin, M. (2013). Big Data Challenges. *International Journal of Advanced Trends in Computer Science and Engineering*. 2 (1). p.pp. 613–615.

Anigbogu, T.U., Onwuteaka, C.I., Edoko, T.D. & Okoli, M.I. (2014). Roles of Small and Medium Scale Enterprises in Community Development: Evidence from Anambra South Senatorial Zone, Anambra State. *International Journal of Academic Research in Business and Social Sciences*. 4 (8).

Arpitha R, Mandara K R, Smitha G R BIG DATA IN HEALTHCARE: CHALLENGES AND SECURITY *International Journal of Scientific Research,* Volume: 5 | Issue: 10 | October 2016.

Bajaj, S., & Johari, R. (2016, February). Big Data: A Boon or Bane-the Big Question. *In Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on* (pp. 106-110). IEEE.

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big Data challenges in the web of things. IEEE Intelligent Systems, 28(6), 6–11.

Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2015, March). An overview of Big Data opportunities, applications and tools. In Intelligent Systems and Computer Vision (ISCV), 2015 (pp. 1-6). IEEE.

Bertino, E., Bernstein, P., Agrawal, D., Davidson, S., Dayal, U., Franklin, M. & Gehrke, J. (2011). *Challenges and Opportunities with Big Data*.

Bertino, E. (2013). Big Data-opportunities and challenges. IEEE. p479-480.

Bhimani, A., & Willcocks, L. (2014). *Digitisation, Big Data and the transformation of accounting information. Accounting and Business Research*, 44(4), 469–490.

Borenstein, N. & Blake, J. (2011). Cloud Computing Standards: Where's the Beef? *IEEE Internet Computing*. 15 (3). p.pp. 74–78.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'Big Data'. McKinsey Quarterly, 4(1), 24-35.

Bryant, R., Katz, R.H. & Lazowska, E.D. (2008). *Big-data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*

Buyya, R., Yeo, C.S. & Venugopal, S. (2008). Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. In: *2008 10th IEEE International Conference on High Performance Computing and Communications*. September 2008, IEEE, pp. 5–13.

Cavoukian, A., & Jonas, J. (2012). *Privacy by design in the age of big data* (pp. 1-17). Information and Privacy Commissioner of Ontario, Canada.

Chandrashekar, R., Kala, M., & Mane, D. (2015). Integration of Big Data in Cloud computing environments for enhanced data processing capabilities. International Journal of Engineering Research and General Science, 3(3).

Cheatham, M. (2015). Privacy in the age of Big Data. In: *2015 International Conference on Collaboration Technologies and Systems (CTS)*. June 2015, IEEE, pp. 334–335.

Chen, C.P. & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 275. p.pp. 314–347.

Chen, D. & Zhao, H. (2012). Data Security and Privacy Protection Issues in Cloud Computing. In: *2012 International Conference on Computer Science and Electronics Engineering*. March 2012, IEEE, pp. 647–651.

Chen, M., Mao, S., Zhang, Y. & Leung, V.C. (2014a). *Big Data: Related Technologies, Challenges and Future Prospects*. Heidelberg: Springer.

Chen, Z.K., Yang, S.Q., Tan, S., Zhao, H., He, L., Zhang, G. & Yang, H.Y. (2014b). The Data Allocation Strategy Based on Load in NoSQL Database. *Applied Mechanics and Materials*. 513–517. p.pp. 1464–1469.

Coppersmith, D. (1994). The Data Encryption Standard (DES) and its strength against attacks. *IBM Journal of Research and Development*. 38 (3). p.pp. 243–250.

Cukier, K. (2010). *Data, data everywhere*. 2010. Economist.

Denning, D.E. & Denning, P.J. (1979). Data Security. *ACM Computing Surveys*. 11 (3). p.pp. 227–249.

De Tré, G., De Mol, R., & Bronselaer, A. (2017). Handling veracity in multi-criteria decision-making: A multi-dimensional approach. Information Sciences.

Ding, J.M., Jiang, Y., Wang, Q.X., Liu, Y.L. & Li, M.J. (2013). A Data Localization Algorithm for Distributing Column Storage System of Big Data. *Advanced Materials Research*. 756–759. p.pp. 3089–3093.

Douglas, L. (2012, June). The Importance of 'Big Data': A Definition.

Durgude, D. M., Yalij, N. S., Bhosale, A. B., & Bharati, M. (2015). Big Data Analysis: Challenges and Solutions. *International Journal of scientific research and management (IJSRM)*, 3(2), 2106-2112.

Editorial, V. (2017, June 26). Big Data Analytics: The Future of Risk Management | VComply. from https://blog.v-comply.com/big-data-analytics-future-risk-management/. [Accessed: 27,October 2017].

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online

Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Eisenstein, M. (2015). Big Data: The power of petabytes. *Nature*. 527 (7576). p.pp. S2-4.

Elazhary, H. (2014). Cloud computing for Big Data. MAGNT Res Rep, 2(4), 135-144.

Fan, J. & Liu, H. (2013). Statistical analysis of Big Data on pharmacogenomics. *Advanced Drug Delivery Reviews*. 65 (7). p.pp. 987–1000.

Feng, D.G., Zhang, M., Zhang, Y. & Xu, Z. (2011). Study on Cloud Computing Security. *Journal of software*. 22 (1). p.pp. 71–83

Foley, J. (2008). *A Definition of Cloud Computing*. 2008. InformationWeek. Available from: https://www.ijarcsse.com/docs/papers/Volume_5/1_January2015/V5I1-0339.pdf. [Accessed: 22 August 2017].

Franco, J. M. (2017, July 07). Big Data Governance & Metadata Management with Talend. from https://www.talend.com/blog/2016/10/10/five-pillars-for-succeeding-in-big-data-governance-and-metadata-management-with/.[Accessed: 04, December 2017].

Gamaleldin, A.M. (2013). *An Introduction to Cloud Computing Concepts*. 2013. Secc. Available from: http://www.secc.org.eg/recocape/SECC_Tutorials_An Introduction to Cloud Computing Concepts.pdf [Accessed: 22 August 2017].

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

Gantz, J. & Reinsel, D. (2012). *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. [Online]. 2012. EMC. Available

from: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf. [Accessed: 22 August 2017].

García, S., Luengo, J. & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Intelligent Systems Reference Library. vol. 72.

Gartner (2017). *Big Data Definition*. [Online]. 2017. Gartner. from: https://research.gartner.com/definition-whatis-big-data?resId=3002918&srcId=1-8163325102. [Accessed: 22 August 2017].

Gens, F., Adam, M., Christiansen, C.A. & Bradshaw, D. (2013). *Worldwide and Regional Public IT Cloud Services 2013-2017 Forecast*.

Gokhan, C., Karakaya, Z., & Yazici, A. (2016, September). Systematic Mapping Study on Performance Scalability in Big Data on Cloud Using VM and Container. *In IFIP International Conference on Artificial Intelligence Applications and Innovations(pp. 634-641). Springer International Publishing.*

Guller, M. (2015). Spark SQL. In Big Data Analytics with Spark (pp. 103-152). Apress.

Ham, K. (2013). OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data. Journal of the Medical Library Association: JMLA, 101(3), 233.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. & Ullah Khan, S. (2015). The rise of 'Big Data' on cloud computing: Review and open research issues. *Information Systems*. 47. p.pp. 98–115.

Hashemi, S., Monfaredi, K., & Masdari, M. (2013). Using cloud computing for e-government: challenges and benefits. International Journal of Computer, Information, Systems and Control Engineering, 7(9), 596-603.

Halevi, G., & Moed, H. (2012). The evolution of Big Data as a research and scientific topic: overview of the literature. Research Trends, 30(1), 3-6.

Hausenblas, M., & Nadeau, J. (2013). Apache Drill: interactive ad-hoc analysis at scale. Big Data, 1(2), 100-104.

Hicks, O.. (1993). *Management information systems*. 3rd Ed. USA: Hicks, O.J.

House, J. (2014, November 20). Big Data analytics = Key to successful 2015 supply chain strategy. ModusLink.

Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for Big Data analytics: *A technology tutorial. IEEE access*, 2, 652-687.

Huang, Q., Yang, C., Benedict, K., Xie, J., Rezgui, A., Xia, J. & Chen., S. (2013). Using Spatiotemporal Patterns and High end Computing to Enable Dust Storm Forecasting. *International Journal of Geographical Information Science*. 27 (4). p.pp. 765–784.

Hussein, W. N., Sulaiman, R., & Hamzah, A. K. (2013, November). E-business and cloud computing awareness for Malaysian SMEs: A recommendation from academic and industry perspectives. In Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on (pp. 180-185). IEEE.

IBM (2017). *Addressing government challenges with Big Data analytics*. 2017. IBM. Available from: http://www.digitalbydefaultnews.co.uk/wp-content/uploads/sites/8/2013/11/Addressing-government-challenges-with-big-data-analytics.pdf. [Accessed: 24 August 2017].

IBM. (2014, April). *Optimize operations Improving operational efficiency with Big Data and analytics.* Available online from: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=NIM03012USEN. [Accessed: 24 August 2017].

Informatica (2017). *Big Data for Government: Drive Better Decisions for Better Policy and Program Outcomes*. [Online]. 2017. Informatica. Available from:

https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/executive-brief/big_data_government_ebook_2340.pdf. [Accessed: 24 August 2017].

Intel IT Center (2012). Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data. Available at: http://www.intel.co.za/content/www/za/en/big-data/data-insights-peer-research-report.html [Accessed 5 Jan. 2017].

Irshad, M.B.M., Gapar, M., Johar, M. & Naleer, H.M.M. (2015). A Review on Cloud Computing Adoption: An Exploratory Analysis. *International Journal of Advanced Research in Computer Science and Software Engineering*. 5 (1). p.pp. 1–7.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R. & Shahabi, C. (2014). Big Data and its technical challenges. *Communications of the ACM*. 57 (7). p.pp. 86–94.

Jaseena, K. U., & David, J. M. (2014). Issues, challenges, and solutions: Big Data mining. NetCom, CSIT, GRAPH-HOC, SPTM–2014, 131-140.

Jelonek, D., Stępniak, C., Turek, T., & Ziora, L. (2014, September). Identification of mental barriers in the implementation of cloud computing in the SMEs in Poland. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 1251-1258). IEEE.

Johnson, J., Neff, T. & Stuart, A. 2012. Big Data The Risks and Rewards Locked in Vast Oceans of Data.

Kaisler, S., Armour, F., Espinosa, J.A. & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In: *2013 46th Hawaii International Conference on System Sciences*. January 2013, IEEE, pp. 995–1004.

Kaliannan, M., Awang, H. & Raman, M. (2007). Technology adoption in the public sector. In: *Proceedings of the 1st international conference on Theory and practice of electronic governance - ICEGOV '07*. 2007, New York, New York, USA: ACM Press, p. 221.

Kanchi, S., Sandilya, S., Ramkrishna, S., Manjrekar, S., & Vhadgar, A. (2015, August). Challenges and Solutions in Big Data Management--An Overview. *In Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on* (pp. 418-426). IEEE.

Karacapilidis, N., Tzagarakis, M., & Christodoulou, S. (2013). On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making. *Intelligent Decision Technologies*, 7(3), 225–236.

Karakaya, Z. (2017, October). Software engineering issues in big data application development. In Computer Science and Engineering (UBMK), 2017 International Conference on (pp. 851-855). IEEE.

Katal, A., Wazid, M., & Goudar, R. H. (2013, August). Big Data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.

Kaufman, L.M. (2009). Data Security in the World of Cloud Computing. *IEEE Security & Privacy Magazine*. 7 (4). p.pp. 61–64.

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M. & Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*. p.pp. 1–18.

Kim, G.-H., Trimi, S. & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*. 57 (3). p.pp. 78–85.

Kitchenham, B. & CharterS, S. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering Version 2.3*. EBSE Technical

Report. EBSE. sn

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with Big Data.

Lamba, H. S., & Dubey, S. K. (2015, September). Analysis of requirements for Big Data adoption to maximize IT business value. In Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), *2015 4th International Conference on* (pp. 1-6). IEEE.

Lebdaoui, I., Orhanou, G., & Elhajji, S. (2014). An integration adaptation for real-time Datawarehousing. International Journal of Software Engineering and its Applications, 8(11), 115–128.

Lee, I. (2017). Big Data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303

Li, W., Song, M., Zhou, B., Cao, K. & Gao, S. (2015). Performance improvement techniques for geospatial web services in a cyberinfrastructure environment – A case study with a disaster management portal. *Computers, Environment and Urban Systems*. 54. p.pp. 314–325.

Lin, H., Hu, J., Tian, Y., Yang, L., & Xu, L. (2017). Toward better data veracity in mobile cloud computing: A context-aware and incentive-based reputation mechanism. Information Sciences, 387, 238-253.

Logan, D. (2010). *What is Information Governance? And Why is it So Hard?* 2010. Gartner Blog Network.

Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in Big Data era. IEEE Network, 28(4), 46–50.

Lyko, K., Nitzschke, M., & Ngomo, A. C. N. (2016). Big data acquisition. In New Horizons for a Data-Driven Economy (pp. 39-61). Springer International Publishing.

Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in Big Data. XRDS: Crossroads. *The ACM Magazine for Students*, 19(1), 20–23.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J. & Ghalsasi, A. (2011). Cloud computing — The business perspective. *Decision Support Systems*. 51 (1). p.pp. 176–189.

Mell, T., & Grance, P. (2009). Draft NIST Working Definition of Cloud Computing. *National Institute of Standards and Technology* (Vol. 53, p. 50).

Michael, K. & Miller, K.W. (2013). Big Data: New Opportunities and New Challenges [Guest editors' introduction]. *Computer*. 46 (6). p.pp. 22–24.

Mikkonen, I., & Khan, I. (2016). Cloud computing: SME company point of view. Management Challenges in the 21st Century: Digitalization of Society, Economy and Market: Current Issues and Challenges.

Miller, S. (2014). Collaborative approaches needed to close the Big Data skills gap. Browser Download This Paper.

Mutula, S.M. & van Brakel, P. (2006). E-readiness of SMEs in the ICT sector in Botswana with respect to information access. *The Electronic Library*. 24 (3). p.pp. 402–417.

Nadeem, A. & Javed, M.Y. (2005). A Performance Comparison of Data Encryption Algorithms. In: *2005 International Conference on Information and Communication Technologies*. 2005, IEEE, pp. 84–89.

Nasser, T., & Tariq, R. S. (2015). Big Data challenges. J Comput Eng Inf Technol 4: 3. Doi.

Neely, A. (2014). *Even SMEs need Big Data*. 2014. Telegraph.

Nemschoff, M. (2013, December 20). Big data: 5 major advantages of Hadoop. Retrieved January 11, 2018, from https://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/ .[Accessed: 04, December 2017].

Neves, P.C., Schmerl, B., Cámara, J. & Bernardino, J. (2016). Big Data in Cloud Computing: Features and Issues. In: *Proceedings of the International Conference on Internet of Things and Big Data*. 2016, SCITEPRESS - Science and Technology Publications, pp. 307–314.

Namiot, D. (2015). On big data stream processing. International Journal of Open Information Technologies, 3(8), 48-51.

OECD (2000). *Small and Medium-sized Enterprises: Local Strength, Global Reach*. 2000. Organisation for Economic Co-operation and Development.

Ogbuokiri, B.O., Udanor, C. & Agu, M. (2015). Implementing Big Data analytics for small and medium enterprise (SME) regional growth. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 17 (6). p.pp. 35–43.

Oguntimilehin, A. & Ademola., E.O. (2014). A Review of Big Data Management, Benefits and Challenges. *A Review of Big Data Management, Benefits and Challenges*. 5 (6). p.pp. 433–438.

Okongwu, D. A. (2001, February). Fostering the innovation potential of SMEs in the globalization era: The role of patents. In WIPO Milan Forum on Intellectual Property and Small and Medium-sized Enterprises organized by The World Intellectual Property Organization (WIPO) and the Ministry of Industry and Foreign Trade of the Government of Italy. February (pp. 9-10).

Open Methodology (2017). *Mike 2.0: Big Data*. 2017. Mike. Available from: http://mike2.openmethodology.org/wiki/Category:Big_Data. [Accessed: 22 September 2017].

Padgavankar, M.H. & Gupta, S.R. (2014). Big Data Storage and Challenges. *International Journal of Computer Science and Information Technologies*. 5 (2). p.pp. 2218–2223.

Pannu, M., Gill, B., Tebb, W., & Yang, K. (2016, October). The impact of Big Data on government processes. In Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual (pp. 1-5). IEEE.

Paris, J., Donnal, J. S., & Leeb, S. B. (2014). NilmDB: the non-intrusive load monitor database.Smart Grid, IEEE Transactions on, 5(5), 2459–2467.

Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards Big Data. Journal of Decision Systems, 24(1), 87–102.

Plummer, D. C., Bittman, T. J., Austin, T., Cearley, D. W., & Smith, D. M. (2008). Cloud computing: Defining and describing an emerging phenomenon. Gartner, June, 17.

Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems*. 9 (1). p.pp. 69–82.

Prasad, A. S., & Rao, S. (2014). A mechanism design approach to resource procurement in cloud computing. IEEE Transactions on Computers, 63(1), 17-30.

Purcell, B. M. (2014). Big Data using cloud computing. Journal of Technology Research, 5, 1.

Ragin, S. (2017, April 03). Using Big Data to understand customers' behavior and increase the business profits. Online available from:http://bigdata-madesimple.com/using-big-data-to-understand-customers-behavior-and-increase-the-business-profits/.[Accessed: 22 September 2017].

Redlich, R.M. & Nemzow, M.A. (2006). *U.S. Patent No. 7,103,915*. Washington, DC: U.S. Patent and Trademark Office.

Rhoton, J. (2011). *Cloud Computing Explained: Implementation Handbook for Enterprises*. Recursive Press.

Richards, N. M., & King, J. H. (2014). Big Data ethics.

Roldán, M. C. (2010). Pentaho 3.2 Data Integration: Beginner's Guide. Packt Publishing Ltd.

Russom, P. (2011). *Big Data analytics: TDWI Best Practices Report, Fourth Quarter*.

Saidhbimca, S. & Gashaw, I. (2013). Security concerns in medium size enterprise cloud computing. *International Journal of Advanced Research in Computer Engineering & Technology*.

Saqib, S. M., Khan, H. M., Mahmood, K., & Naeem, T. (2015). BIG-Data Challenges: A Review on Existing Solutions. American Journal of Information Science and Computer Engineering, 1(2), 38-43.

Saxena, V. K., & Pushkar, S. (2016, March). Cloud computing challenges and implementations. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on(pp. 2583-2588). IEEE.

Schmiemann, M. (2009). *SMEs were the main drivers of economic growth between 2004 and 2006*. Luxembourg.

Sen, D., Ozturk, M. & Vayvay, O. (2016). An Overview of Big Data for Growth in SMEs. *Procedia - Social and Behavioral Sciences*. 235. p.pp. 159–167.

Sewoog Kim, Dongwoo Kang, Jongmoo Choi & Junmo Kim (2014). Burstiness-aware I/O scheduler for MapReduce framework on virtualized environments. In: *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*.

January 2014, IEEE, pp. 305–308.

Shim, J. P., French, A. M., Guo, C., & Jablonski, J. (2015). Big Data and Analytics: Issues, Solutions, and ROI. CAIS, 37, 39.

Simonet, A., Fedak, G., & Ripeanu, M. (2015). Active Data: A programming model to manage data lifecycle across heterogeneous systems and infrastructures. Future Generation Computer Systems, 53, 25–42.

Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S. & Pearlman, L. (2003). A Metadata Catalog Service for Data Intensive Applications. In: *Proceedings of the 2003 ACM/IEEE conference on Supercomputing - SC '03*. 2003, New York, New York, USA: ACM Press, p. 33.

Singh, V., Srivastava, I. & Johri, V. (2014). Big Data and the Opportunities and Challenges for Government Agencies. *International Journal of Computer Science and Information Technologies*. 5 (4).

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286.

Sivarajah, U., Irani, Z., & Weerakkody, V. (2015). Evaluating the use and impact of Web 2.0 technologies in local government. Government Information Quarterly, 32(4), 473-487.

Smeda, J. (2015). Benefits, business considerations and risks of Big Data (Doctoral dissertation, Stellenbosch: Stellenbosch University).

Smid, M.. & Branstad, D.. (1988). Data Encryption Standard: past and future. *Proceedings of the IEEE*. 76 (5). p.pp. 550–559.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In Technical report, Ver. 2.3 EBSE Technical Report. EBSE. sn.

Staff, S. (2013, May 30). 8 Benefits of Big Data for State and Local Governments. https://statetechmagazine.com/article/2013/05/8-benefits-big-data-state-and-local-governments .[Accessed: 22 September 2017].

Stearley, J., Corwell, S., & Lord, K. (2010, October). Bridging the Gaps: Joining Information Sources with Splunk. In SLAML.

Tallon, P.P., Ramirez, R. V. & Short, J.E. (2013). The Information Artifact in IT Governance: Toward a Theory of Information Governance. *Journal of Management Information Systems*. 30 (3). p.pp. 141–178.

Tekiner, F., & Keane, J. A. (2013, October). Big data framework. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 1494-1499). IEEE.

Tene, O. (2011). Privacy: The new generations. *International Data Privacy Law*. 1 (1). p.pp. 15–27.

Terry, N.P. (2012). *Protecting Patient Privacy in the Age of Big Data*. 2012.

Thomas, D. M. (2014). A Review paper on BIG Data.

Tian, W. & Zhao, Y. (2015). Big Data Technologies and Cloud Computing. Optimized Cloud Resource Management and Scheduling. Theory and Practice: Morgan Kaufmann.

Tolman, C.W. (2012). *Positivism in psychology: Historical and contemporary problems*. Germany: Springer Science & Business Media.

Tweneboah-Koduah, S., Endicott-Popovsky, B., & Tsetse, A. (2014). Barriers to government cloud adoption: the Ghanaian perspective. International Journal of Managing Information Technology, 6(3), 1.

Uddin, M. F., & Gupta, N. (2014, April). Seven V's of Big Data understanding Big Data to extract value. *In American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the (pp. 1-5). IEEE.*

Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M. (2012). *Perspectives on Big Data and Big Data analytics. Database Systems Journal*, 3(4), 3-14.

Vajjhala, N.R. & Ramollari, E. (2016). Big Data using Cloud Computing - Opportunities for Small and Medium-sized Enterprises. *European Journal of Economics and Business Studies*. 4 (1). p.p. 129.

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. White Paper, IDC, 14.

Vohra, D. (2016). Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. Apress.

Wang, Y., &Wiebe, V. J. (2014). Big Data Analytics on the characteristic equilibrium of collective opinions in social networks. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 8(3), 29–44.

Wattanapruttipaisan, T. (2005). SME Development and Internationalization in the Knowledge-Based and Innovation-Driven Global Economy: Mapping the Agenda Ahead. In: *Mapping Policy Experience for SMEs*. 2005, Phuket: Thammasat University, pp. 1–29.

Wielki, J. (2015). The opportunities and challenges connected with implementation of the Big Data concept. In Advances in ICT for Business, Industry and Public Sector (pp. 171-189). Springer International Publishing.

Wyld, D.C. (2010). The Cloudy Future Of Government IT: Cloud Computing and The Public Sector Around The World. *International Journal of Web & Semantic Technology*. 1 (1). p.pp. 1–20.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth, 10(1), 13-53.

Yang, C., Xu, Y. & Nebert, D. (2013a). Redefining the possibility of digital Earth and geosciences with spatial cloud computing. *International Journal of Digital Earth*. [Online]. 6 (4). p.pp. 297–312.

Yang, Y., Long, X. & Jiang, B. (2013b). K-Means Method for Grouping in Hybrid MapReduce Cluster. *Journal of Computers*. 8 (10).

Zhang, X., Hu, Y., Xie, K., Zhang, W., Su, L., & Liu, M. (2015). An evolutionary trend reversion model for stock trading rule discovery. *Knowledge-Based Systems*, 79, 27-35.

Zhai, Y., Ong, Y.-S. & Tsang, I.W. (2014). The Emerging 'Big Dimensionality'. *IEEE Computational Intelligence Magazine*. 9 (3). p.pp. 14–26.

Zicari, R. V. (2014). Big Data: Challenges and Opportunities. (2014) In R. (Ed.), Big Data computing (pp. 103–128). Florida, USA: CRC Press, Taylor & Francis Group.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Corrigan, D., & Giles, J. (2013). Harness the power of big data: The IBM big data platform. New York, NY: McGraw-Hill.