Critical analysis of Big Data challenges and analytical methods
Uthayasankar Sivarajah *, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody
Brunel University London, Brunel Business School, UB8 3PH, United Kingdom

ARTICLE INFO
Available online 10 August 2016

Keywords:
Big Data
Big Data Analytics
Challenges
Methods
Systematic literature review

ABSTRACT
Big Data (BD), with their potential to ascertain valued insights for enhanced decision-making process, have recently attracted substantial interest from both academics and practitioners. Big Data Analytics (BDA) is increasingly becoming a trending practice that many organizations are adopting with the purpose of constructing valuable information from BD. The analytics process, including the deployment and use of BDA tools, is seen by organizations as a tool to improve operational efficiency though it has strategic potential, drive new revenue streams and gain competitive advantages over business rivals. However, there are different types of analytic applications to consider. Therefore, prior to hasty use and buying costly BD tools, there is a need for organizations to first understand the BDA landscape. Given the significant nature of the BD and BDA, this paper presents a state-of-the-art review that presents a holistic view of the BD challenges and BDA methods theorized/proposed/employed by organizations to help others understand this landscape with the objective of making robust investment decisions. In so doing, systematically analysing and synthesizing the extant research published on BD and BDA area. More specifically, the authors seek to answer the following two principal questions: Q1 – What are the different types of BD challenges theorized/proposed/confronted by organizations? and Q2 – What are the different types of BDA methods theorized/proposed/employed to overcome BD challenges? This systematic literature review (SLR) is carried out through observing and understanding the past trends and extant patterns/themes in the BDA research area, evaluating contributions, summarizing knowledge, thereby identifying limitations, implications and potential further research avenues to support the academic community in exploring research themes/patterns. Thus, to trace the implementation of BD strategies, a profiling method is employed to analyze articles (published in English-speaking peer-reviewed journals between 1996 and 2015) extracted from the Scopus database. The analysis presented in this paper has identified relevant BD research studies that have contributed both conceptually and empirically to the expansion and accrual of intellectual wealth to the BDA in technology and organizational resource management discipline.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction
The magnitude of data generated and shared by businesses, public administrations numerous industrial and not-to-profit sectors, and scientific research, has increased immeasurably (Agarwal & Dhar, 2014). These data include textual content (i.e. structured, semi-structured as well as unstructured), to multimedia content (e.g. videos, images, audio) on a multiplicity of platforms (e.g. machine-to-machine communications, social media sites, sensors networks, cyber-physical systems, and Internet of Things [IoT]). Dobre and Xhafa (2014) report that every day the world produces around 2.5 quintillion bytes of data (i.e. 1 exabyte equals 1 quintillion bytes or 1 exabyte equals 1 billion gigabytes), with 90% of these data generated in the world being unstructured. Gantz and Reinsel (2012) assert that by 2020, over 40 Zettabytes (or 40 trillion gigabytes) of data will have been generated, imitated, and consumed. With this overwhelming amount of complex and heterogeneous data pouring from any-where, any-time, and any-device, there is an undeniable era of Big Data – a phenomenon also referred to as the Data Deluge. The potential of BD is evident as it has been included into data classification concept and ever increasing technological advancements. Advocates assert that in the future a majority of data
will be generated and shared through machines, as machines communi-
cate with each other over data networks (Van Dijck, 2014). Regardless of
where BD is generated from and shared to, with the reality of BD
comes the challenge of analysing it in a way that brings Big Value.
With so much value residing inside, BD has been regarded as today’s
digital oil (Yi, Liu, Liu, & Jin, 2014) including the New Raw Material
of the 21st century (Berners-Lee & Shadbolt, 2011). Appropriate data
processing and management could expose new knowledge, and
facilitate in responding to emerging opportunities and challenges in a
timely manner (Chen et al., 2013). Nevertheless, the growth of data in
volumes in the digital world seems to out-speed the advance of the
many extant computing infrastructures. Established data processing
technologies, for example database and data warehouse, are becoming
inadequate given the amount of data the world is current generating.
The massive amount of data needs to be analyzed in an iterative, as
well as in a time sensitive manner (Jukić, Sharma, Nestorov, & Jukić,
2015). With the availability of advanced BD analysing technologies
(e.g. NoSQL Databases, BigQuery, MapReduce, Hadoop, WibiData and
Skytree), insights can be better attained to enable in improving business
strategies and the decision-making process in critical sectors such as
healthcare, economic productivity, energy futures, and predicting natu-
rnal catastrophe, to name but a few (Yi et al., 2014).

As evident, much has been written on the BD phenomenon. The
majority of academic research articles reviewed are analytical in nature
(also evident from the findings – see Figs. 10 and 11) that is either focus-
ing on using experiments, simulations, algorithms and or mathematical
modelling techniques in tackling BD. Regardless of their research
approach, these articles present BD as a source that when appropriately
managed, processed and analyzed, have the potential to generate new
knowledge thus proposing innovative and actionable insights for
businesses (Jukić et al., 2015). There is an ever-growing discourse
about BD offering both Big Opportunities and Big Challenges through
the plethora of sources from different domains; extending from enter-
prises to sciences. For instance, the opportunities include value creation
(Brown, Chui, & Manyika, 2011), rich business intelligence for better-
informed business decisions (Chen & Zhang, 2014), and support in
enhancing the visibility and flexibility of supply chain and resource
allocation (Kumar, Niu, & Ré, 2013). On the other hand, the challenges
are significant such as data integration complexities (Gandomi &
Haider, 2015), lack of skilled personal and sufficient resources (Kim,
Trimi, & Chung, 2014), data security and privacy issues (Barnaghi,
Sheth, & Henson, 2013), inadequate infrastructure and insignificant
data warehouse architecture (Barbierato, Gribaudo, & Iacono, 2014),
and synchronising large data (Jiang, Chen, Qiao, Weng, & Li, 2015).
Advocates such as Sandhu and Sood (2014) perceive that the potential
value of BD cannot be unearthed by simple statistical analysis. Zhang,
Liu et al. (2015) support this perspective and state that to tackle the
BD challenges, advanced BDA requires extremely efficient, scalable and
flexible technologies to efficiently manage substantial amounts of
data – regardless of the type of data format (e.g. textual and multimedia
content).

1.1. Research scope

BD and BDA as a research discipline are still evolving and not yet
established, thus, a comprehensible understanding of the phenomenon,
its definition and classification is yet to be fully established. The extant
progress made in BD and BDA not only revealed a lack of management
research in the field but a distinct lack of theoretical constructs and
academic rigor – perhaps a function of an underlying methodological
rather than academic challenge. At large, there has also been a lack of
research studies that comprehensively addresses the key challenges of
BD, or which investigates opportunities for new theories or emerging
practices (e.g. George, Haas, & Pentland, 2014). Thus, there exists the
need to culminate the BD challenges and associated BDA methods to
allow signposting to take place. Following the earlier limited normative
research studies conducted by Polato, Ré, Goldman, and Kon (2014) –
mainly focusing on Apache Hadoop; Frehe, Kleinschmidt, and
Teuteberg (2014) – BD logistics; Lemb, Ishak, Sidi, Affendey, and
Mamat (2015) – on data veracity research for profiling digital news
portal, and Abdellatif, Capretz, and Ho (2015) – on software analytics
(a distinct branch of BDA), this paper attempts to broaden the scope of
their reviews by further investigating and assessing the different types of
BD challenges and the analytical methods employed to overcome the
challenges. Although these research studies provide worthy understand-
ing on some aspects of BD and BDA area, there seems to be a lack of
comprehensive and methodical approaches to understand the phenom-
emon of BD – more precisely the types of BDA methods thus an aide
memoir will act as a suitable frame of reference. Moreover, explicit
in respect of the conclusions offered by these existing review articles,
this research specifically aims to:

analyze, synthesize and present a state-of-the-art structured analysis of
the normative literature on big data and big data analytics to support
the signposting of future research directions.

1.2. Academic challenge

This SLR research aims to evaluate the existing research published
on BD and BDA by employing an established profiling approach and to
investigate and analyze different BD challenges and BDA technologies,
techniques, methods and or approaches. To identify the relevant articles
through the Scopus database, the following keywords search criteria
was used:
• Big Data OR Big Data Analytics OR Big Data Analysis AND Challenge OR
Challenges OR Barrier OR Barriers OR Obstacle OR Obstacles OR Problem
OR Obstacles OR Impediment OR Impediments AND Technology OR
Technologies OR Technique OR Method OR Methods OR Approach OR
Approaches.

Through using the abovementioned list of keywords and focusing on
four subject areas that is business and management, computer science,
decision science, and social science; initially 433 journal articles were
identified from the Scopus database and relating to articles published
during the period from 1996 to 2015. However, from period 1996
until 2002, there were no papers recorded on BD and BDA in these
four subject areas. After assessing the 434 articles (from refereed
journals), 206 papers were discarded, and finally 227 papers were
selected and taken forward for further interrogation. As reflected in
Fig. 9, contributors from across the world have made contributions to
the BD and BDA area. Nevertheless, given the limitations in the existing
BD and BDA literature review studies (as reported earlier in Section 1.1),
the rationale for undertaking this research is to provide a systematic
state-of-the-art literature analysis of the BD and BDA area. In doing so
to better understand the different types of BD challenges and associated
BDA methods. Thus, the two underlying academic challenges orientate
around identify the:

• different types of BD challenges theorized/proposed/discussed/
confronted by organizations.
• different types of BDA methods theorized/proposed/discussed/
employed to overcome BD challenges.

To supplement this research and the above objectives, the authors
also identified the:
• yearly publications from 1996 until 2015.
• geographic location of each publication (this includes the geographi-
cal location of each author as well as the co-author(s) in each paper
reviewed).
• types of publication (e.g. research or technical paper, literature review, viewpoint).
• types of research methods employed (e.g. case study, mixed method, analytical).

This type of profiling research is necessary to develop an understanding of the BD and BDA area and the state-of-the-art growth in the theory and application of BD and BDA within different sectors and disciplines. This paper is predominantly descriptive and inductive in nature, as the authors were interested in understanding the perspectives of BD and BDA and its distinctiveness as practiced across different sectors.

2. A normative perspective of Big Data: challenges and analytical methods

The concept of big is problematic to pinpoint, not least because a dataset that appears to be massive today will almost surely appear small in the near future (MIT Technology Review, 2013). Adding to the complexity of the BD itself, some practitioners argue that massive datasets are not always complex and small data sets are always simple, thus highlighting that the intricacy of a dataset is a significant factor in determining whether it is big. In this section, the authors provide some theoretical conceptions related to Q1 and Q2.

2.1. Big Data Challenges – related to Q1

Though the benefits of BD are factual and substantial, there remain a plethora of challenges that must be addressed to fully realise the potential of BD. Some of these challenges are a function of the characteristics of BD, some, by its existing analysis methods and models, and some, through the limitations of current data processing system (Jin, Wah, Cheng, & Wang, 2015). Extant studies surrounding BD challenges have paid attention to the difficulties of understanding the notion of BD (Hargittai, 2015), decision-making of what data are generated and collected (Crawford, 2013), issues of privacy (Lazer et al., 2009) and ethical considerations relevant to mining such data (Boyd & Crawford, 2012). Tole (2013) asserts that building a viable solution for large and multifaceted data is a challenge that businesses are constantly learning and then implementing new approaches. For example, one the biggest problems regarding BD is the infrastructure’s high costs (Wang & Wiebe, 2014). Hardware equipment is very expensive even with the availability of cloud computing technologies.

Furthermore, to sort through data, so that valuable information can be constructed, human analysis is often required. While the computing technologies required to facilitate these data are keeping pace, the human expertise and talents business leaders require to leverage BD are lagging behind, this proves to be another big challenge. As reported by Akerkar (2014) and Zicari (2014), the broad challenges of BD can be grouped into three main categories, based on the data life cycle: data, process and management challenges:

• Data challenges relate to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism).
• Process challenges are related to series of how techniques: how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results.
• Management challenges cover for example privacy, security, governance and ethical aspects.

Fig. 1 shows the classification of BD challenges – as adapted from Akerkar (2014) and Zicari (2014). The SLR findings for Q1 are based on three categories of BD challenges.

2.2. Big Data analytical methods – related to Q2

To facilitate evidence-based decision-making, organizations need efficient methods to process large volumes of assorted data into meaningful comprehensions (Gandomi & Haider, 2015). The potentials of using BD are endless but restricted by the availability of technologies, tools and skills available for BDA. According to Labrinidis and Jagadish (2012), BDA refers to methods used to examine and attain intellect from the large datasets. Thus, BDA can be regarded as a sub-process in the whole process of insight extraction from BD. It is certain that for BD to realise its objectives and progress services in business environment, it requires the correct tools and approaches to be analyzed and classified effectively and proficiently (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 2015). The potential value of BD is solved simply when leveraged to the drive decision-making process. Extant research studies have demonstrated that substantial value and competitive advantage can be attained by businesses from taking effective decisions based on data (Davenport & Harris, 2007). But, BDA is more perplexing than merely tracing, classifying, comprehending, and quoting data. Davenport and Dyché (2013) emphasize that large organizations regularly gather BD and exploit analytics for support in decision-making.
making as part of their usual procedures, and SMEs are the ones presently struggling to enhance top management decisions while adding more data for the analysis process. Aligning the people, technology, and organizational resources to become a data-driven company is problematic (Weill & Ross, 2009). Given BD can enhance the decision-making and increase organizational output; this is possible when a selection of analytical methods is used to extract sense from the data, such as:

- **descriptive analytics** scrutinizes data and information to define the current state of a business situation in a way that developments, patterns and exceptions become evident, in the form of producing standard reports, ad hoc reports, and alerts (Joseph & Johnson, 2013);
- **inquisitive analytics** is about probing data to certify/reject business propositions, for example, analytical drill downs into data, statistical analysis, factor analysis (Bihani & Patil, 2014);
- **predictive analytics** is concerned with forecasting and statistical modeling to determine the future possibilities (Waller & Fawcett, 2013);
- **prescriptive analytics** is about optimization and randomized testing to assess how businesses enhance their service levels while decreasing the expenses (Joseph & Johnson, 2013); and
- **pre-emptive analytics** is about having the capacity to take precautionary actions on events that may undesirably influence the organizational performance, for example, identifying the possible perils and recommending mitigating strategies far ahead in time (Szongott, Henne, & von Voigt, 2012).

Advocates assert that these types of analytical methods support in improved decision-making and organizational performance by making everything more translucent and quantifiable, while further uncovering inconsistencies as well as potential concerns and opportunities. Fig. 2 illustrates the classification of BDA methods and the SLR findings for Q2 are based on these five categories.

### 3. Research methodology

In an attempt to better understand and provide more detailed insights to the phenomenon of big data and bit data analytics, the authors respond to the special issue call on Big Data and Analytics in Technology and Organizational Resource Management (specifically focusing on conducting – A comprehensive state-of-the-art review that presents Big Data Challenges and Big Data Analytics methods theorized [in extant research literature], proposed [by research scholars], and employed [by organizations]) through a SLR methodology as opposed to narrative or descriptive reviews (Tranfield, Denyer, & Smart, 2003; Kitchenham & Charters, 2007; Wang, Gunasekaran, Ngai, & Papadopoulos, 2016). In support of the former approach, Lettieri, Masella, and Radaelli (2009) report that SLR is a rational, transparent and reproducible research methodology for the analysis of extant literature. Kitchenham and Charters (2007) also highlight that SLR is a form of secondary study and it is a distinct approach to establish, explore and deduce accessible proof associated to a particular research question (e.g. Q1 and Q2) in a way that is unprejudiced and (to a certain degree) repeatable. Alternatively, meta-based-approaches can be used to conducting a literature review and include the work of Mishra, Gunasekaran, Papadopoulos, and Childe (2016), which adopt a bibliometric and network analysis approach to obtain and compare influential work in a specific domain (in this example, Big Data in Supply Chains).

---

**Fig. 2.** Classification of types of big data analytical methods.
There are several motivating reasons for conducting a systematic literature review (Kitchenham & Charters, 2007) such as including among other, these are to:

- précis current evidence around a technology or a treatment, such as to summarize the evidence of the benefits and drawbacks of an explicit map technique;
- determine research gaps within the extant research to propose areas for further research activities;
- recommend a frame of reference to identify current research trajectories and potential research themes.

Based on the focus of this research, the first two reasons fit the purpose of a SLR. The scope and applicability of BD and BDA phenomenon clearly indicates that this area has the potential to support organizations, for instance, at the strategic, organizational, operational as well as technological level. This SLR offers an enhanced descriptive and thematic awareness of the resulting body of knowledge, enabling the BD research area to further develop in a more cognizant and multidisciplinary approach.

Delbufalo (2012) asserts that a SLR is designed to: (a) support in generating a sense of joint effort, importance and openness between the research studies in order to impede unproductive recurrence of effort, (b) support in connecting potential research to the queries and issues that have been modelled by previous research studies (e.g. most of those paper reviewed as part of this research exercise) and, (c) develop the approaches employed to assemble and synthesize preceding pragmatic evidence. In the interest of parsimony, a meticulous though not exhaustive SLR was carried out through following a three-phase approach as described by Tranfield et al. (2003) and Kitchenham and Charters (2007) and diagrammatically illustrated in Fig. 3:

- **Phase I** – Planning the Review Process – Defining the research aim and objectives (I.1); formulating the proposal (I.2) and developing the review protocol (I.3);
- **Phase II** – Conducting the Review Process – Identifying, selecting, evaluating and synthesizing the pertinent research studies; and
- **Phase III** – Reporting and Dissemination of the Overall Research Results – Descriptive reporting of results and thematic reporting of journal articles.

Following the three-phase approach, the next subsection 3.1 summarizes the research protocol (Phase I.3) as the defining of the aim and objectives including the proposal (I.1 and I.2) have already been presented in the introduction (under subsections 1.1 and 1.2). Sub-section 3.2 describes the Scopus database searching process of the relevant articles (Phase II). Finally, the reporting and dissemination of the overall results (Phase III) will be discussed in Section 4 and with Section 5 concluding the paper.

### 3.1. The research protocol (phase I.3)

In this paper, the authors commenced this systematic search by using an established detailed review protocol based on the guiding principles and procedures of the SLR (Tranfield et al., 2003; Kitchenham & Charters, 2007). This protocol identifies the background review, search strategy, research questions as outlined in the abstract (i.e. Q1 and Q2), data extraction, criteria for study selection and data synthesis – based on the prescriptive three phased approach. The research questions and background of this review are described above, while the following sections provide details about other elements. As this literature review focuses on analysing, synthesizing and presenting a comprehensive structured analysis of the normative literature on BD and BDA, it was necessary to considered the domains for this research synthesis as both conceptual and empirical (including qualitative, quantitative and mixed method) papers. The research protocol for this literature review paper provides details on the following two points, as also followed by Delbufalo (2012) and Kamal and Irani (2014):

- **Point I** – Conceptualizing BD and BDA research discipline, including challenges of BD and related BDA methods (as discussed in Sections 2.1, and 2.2).
- **Point II** – Typology of research studies to be considered in this review exercise and the appropriate measures.

Given the above, several selections in relation to the typology of research studies to be counted in and the suitability conditions (i.e. the inclusion and exclusion measures) have been made (Point II), as presented below.

- **Condition I** – The review was conducted by searching the Scopus databases. The reason for choosing the Scopus database was that it covers nearly 18,000 plus titles from over 5000 international publishers, including coverage of around 16,500 peer-reviewed journals on different areas. Therefore, it is possible to search for and locate a significant proportion of the published material in the BD and BDA area.
- **Condition II** – To focus on enhancing quality control (David & Han, 2004) only published peer-reviewed journal (including articles in press and therefore accepted post peer-review) were considered by selecting Article and Articles in Press option from the Document Type option. Other document/source types such as conferences, trade publications, books series, book or book chapter, and editorials were omitted.
- **Condition III** – Following David and Han’s (2004) enhancing quality control policy, only those articles were selected that were published between 1996 and 2015.
- **Condition IV** – Articles from subject areas such as Business and Management, Computer Sciences, Decision Sciences and Social Sciences and published in the English language were only selected, excluding the articles published in Chinese, French, Korean, Spanish, German, Japanese, Portuguese and Russian. This is a recognised limitation.
- **Condition V** – It was ensured that the selected articles were not only empirical (i.e. case-study, results, analytical, etc.) but also those articles that were essentially conceptual so as to identify conceptual research developments in BD and BDA.
- **Condition VI** – Articles’ applicability was confirmed by requiring that selected articles contained a number of key phrases (as listed in Section 1.2) throughout the paper, including, title, abstract, keywords and the thereafter the whole paper. In essence, the identified articles were reviewed with particular attention given to those section(s) that explicitly referred to BD and BDA. In doing so, to extract relevant perspectives on the type of BD challenges and BDA methods.
- **Condition VII** – Final substantive applicability was confirmed by reading the remaining whole article for essential research perspective and satisfactory empirical data. The latter process forced the alignment between the selected articles and the research review objectives.

The abovementioned conditions itemized in seven points were all prescriptively followed so as to conduct an effective and reproducible database examining process as pronounced in the following subsection.

### 3.2. Scopus database searching process and results – Phase II

According to Delbufalo (2012), there are four stages of database searching process. This section reports on these steps and activities of the process, demonstrating the outcomes both descriptively and synthetically by searching for relevant articles throughout the Scopus database.
Step 1: Defining the Research Aim & Objectives

Step 2: Preparing the Proposal

Step 3: Developing the Review Protocol

Step 4: Identifying the Research Articles

Step 5: Selecting the Relevant Research Articles

Step 6: Study Quality Assessment

Step 7: Synthesising the Relevant Data

Step 8: Descriptive Reporting of the Results

Step 9: Reporting of the Data

Phase I – Conceptualising BD Challenges and BDA Methods

Phase II.1 – Profiling Research via Scopus Database Keyword Search resulted in 434 Articles (Phase II.1)

Phase II.2 – Title, Abstract and thorough Article Analysis resulted in 227 Articles

Phase II.3 – Sorting the Articles based on ‘BD_CH’ and ‘BDA_MTH’ Categories

Phase II.4 – Scanning of the Articles to extract answers for Q1 and Q2

Phase II – Conducting the Review Process

Phase III – Reporting and Dissemination

Section 1

Section 2

Section 3.1

Section 3.2

Section 4

Section 5

Fig. 3. Research design – systematic literature review process of BD and BDA.
4. Big Data and Big Data Analytics: findings and analysis

The findings of this study are now presented under different subsections. Each of the six subsections discuss on the findings in relation to a particular variable as set in Section 1.2.

4.1. Types of Big Data Challenges

Among the many BD challenges (as reported in Figs. 1 and 4), the large datasets (in terms of size and complexity) and the ability to process vast amount of data remains a critical challenge for outdated data processing applications and, relational database management systems (jiang et al., 2015). According to TDWI Predictive Analytic Study (Russom, 2013), there are several BD challenges posing a peril to organizations – among these are, integrating complex and large datasets, getting started with the right BD project, developing and implementing infrastructure for managing and processing BD and a lack of skilled personnel or staff with analytics skills to make sense of BD.

Figs. 4, 5, and 6 illustrates the frequency at which the data, process and management (all three related to BD) challenges are discussed/proposed/ theorized in the articles reviewed through the SLR process, as presented in Fig. 1.

4.1.1. Data challenges

Data challenges are the group of the challenges related to the characteristics of the data itself. Different researchers have distinct understandings towards the data characteristics – such as some say 3Vs [volume, velocity and variety] of data (e.g. Shah, Rabhi, & Ray, 2015), others reported 4Vs [volume, velocity, variety, and variability] of data (e.g. Liao, Yin, Huang, & Sheng, 2014) and 6Vs [volume, velocity, variety, veracity, variability, and value] of data (Gandomi & Haider, 2015). In analysing the different articles reviewed in this SLR, the authors identified 7Vs – seven characteristics of data [volume (DC_VOLM) → C = 90 (39.64% of 227 articles), variety (DC_VART) → C = 59 (25.8%), velocity (DC_VELO) → C = 44 (19.4%), value (DC_VALE) → C = 30 (13.2%), veracity (DC_VERT) → C = 18 (7.9%), visualization (DC_VISU) → C = 6 (2.6%) and variability (DC_VARB) → C = 4 (1.8%)] and these features are illustrated in Fig. 4 and discussed as follows:

- **Volume** (e.g. large data-sets consisting of terabytes, petabytes, zettabytes of data – or even more): Large scale and the sheer volume of data is a big challenge in its own right. The latter argument is also supported by Barnaghi et al. (2013) that state the heterogeneity, ubiquity, and dynamic nature of the different data generation resources and devices, and the enormous scale of data itself, make determining, retrieving, processing, integrating, and inferring the physical world data (e.g. environmental data, business data, medical data, surveillance data) a challenging task. This colossal increase of large-scale data (e.g. Facebook daily generates over 500 terabytes of data, and Walmart collects more than 2.5 petabytes of data every hour from its customers transactions) sets brings new challenges to data mining techniques and requires novel approaches to address the big-data problem (Zhao, Zhang, Cox, Duling, & Sarle, 2013).

- **Variety** (e.g. multiple data formats with structured and unstructured text/image/multimedia content/audio/video/sensor data/noise): Data challenges related to the variety (i.e. diverse and dissimilar forms) of data are also deemed a challenge. These articles revealed that the enormous volume of data is not consistent nor does it follow a specific template or format – it is captured in diverse formats and diverse sources e.g.: messages (text, email, tweets, blogs) – user generated content, transactional data (e.g. web logs, business transactions), scientific data (e.g. data coming from data-intensive experiments – genome and healthcare data), web data (e.g. images posted on social media; sensor data readings), and much more (Chen, Chiang & Storey, 2012; Chen et al., 2013). These different forms and quality of data clearly indicate that heterogeneity is a natural property of BD and it is a big challenge to comprehend and manage such data (Labrinidis & Jagadish, 2012). For instance, during the Fukushima Daiichi nuclear disaster, when the public started broadcasting radio-active material data, a wide variety of inconsistent data, using diverse and uncalibrated devices, for similar or neighboring locations was reported – all this add to the problem of increasing variety of data.

- **Veracity** (e.g. increasingly complex data structure, anonymity, imprecision or inconsistency in large data-sets): This is not merely about data quality – it is more about understanding the data, as there are integral discrepancies in almost all the data collected. IBM came up with this characteristic of data, which represents the
Fig. 4. Clusters of articles discussing/proposing/theorizing types of data challenges.
Fig. 5. Clusters of articles discussing/proposing/theorizing types of process challenges.
Fig. 6. Clusters of articles discussing/proposing/theorizing on types of management challenges.
untrustworthiness inherent in many sources of structured as well as unstructured data. Akerkar (2014) and Zicari (2014) refer veracity to as coping with the biases, doubts, imprecision, fabrications, messiness and misplaced evidence in the data. Veracity feature measures the accuracy of data and its potential use for analysis (Vasarhelyi, Kogan, & Tuttle, 2015). For instance, every customer opinion on different social media networks and web is different and unclear in nature, as it involves human interaction (Sivarajah, Irani, & Weerakkody, 2015). Moreover, the web, more specifically, is a soft medium to publish and broadcast fabricated information across multiple sources and, so it is essential to isolate the wheat from the chaff when presenting quality data. Thus, the necessity to deal with inaccurate and ambiguous data is another facet of BD, which is addressed using tools and analytics developed for management and mining of unreliable data (Gandomi & Haider, 2015).

• **Velocity (e.g. high rate of data inflow with non-homogenous structure):** The challenge of velocity comes with the requisite to manage the high influx rate of non-homogenous data, which results in either creating new data or updating the existing data (Chen et al., 2013). This mainly applies to those datasets that are generated through large complex networks including data generated by the proliferation of digital devices, which are positioned ubiquitously resulting in driving the need for real-time analytics and evidence-based planning (Lu, Zhu, Liu, Liu, & Shao, 2014). For instance, Wal-Mart processes more than a million transactions each hour (Cukier, 2010). The data stemming from mobile devices and flowing through mobile apps or by using store cards (e.g. Sainsbury’s card for collecting nectar points) generates floods of information that can be brought to use through producing real-time, personalized offers for customers. These data also provide sound information about customers, such as their geospatial location, buying behaviour and patterns, which can be analyzed in real-time to generate value for customers (Gandomi & Haider, 2015).

• **Variability (e.g. data whose meaning is constantly changing):** Among the seven pillars of BD, variability is another extremely essential feature but is often confused with variety. For instance, Google or Facebook repository stores and generates many different types of data. At the same time, if from these different types of data, one of them is brought to use for mining and making sense out of it but every time the data offers a different meaning – this is variability of data – whose meaning is constantly and rapidly changing. The volumes of machine and human-generated data constitute much greater and their rates of change and variability higher than process-mediated data. Variability is also related in performing sentiment analyzes. For example, in (almost) the same tweets a word can have a totally different meaning. In order to perform a proper sentiment analyzes, advocates assert that algorithms need to be able to understand the context and be able to decipher the exact meaning of a word in that context (Zhang, Hu et al., 2015). Nevertheless, this is yet still very challenging.

• **Visualization (e.g. presenting the data in a manner that is readable):** Visualising data is about representing key information and knowledge more instinctively and effectively through using different visual formats such as in a pictorial or graphical layout (Taheri, Zomaya, Siegel, & Tari, 2014). For instance, eBay has millions of users and from these many million users, even more millions of goods are sold every month – this generates a lot of data. To make all these data explicable, eBay considered the BD visualization tool – Tableau, which is capable of transforming large and complex datasets into spontaneous depictions. Based on these interactive results, eBay employees can visualize search relevance and quality, to monitor the latest customer feedback and conduct sentiment analysis. Chen and Zhang (2014) argue that for many existing BD applications that have poor performances in functionalities, scalability and response time, it is mainly problematic when conducting data visualization. This reason for this is a consequence of large sizes and high dimension of BD.

• **Value (e.g. extracting knowledge/value from vast amounts of structured and unstructured data without loss, for end users):** Storing BD is complex. For instance, significant values can be extracted from the stream of clicks left behind by the internet users – and this is becoming a backbone of the internet economy. Big data researchers consider value as an essential feature, as somewhere within that data, there is valuable information – extracting golden data (high-valued data), though most of the pieces of data independently may seem insignificant (Zaslavsky, Perera, & Georgakopoulos, 2012). Regardless of the number of dimensions used to describe BD, organizations are still faced with challenges of storing, managing and predominately extracting value from the data in a cost effective manner (Abawajy, 2015).

### 4.1.2. Process challenges

Process challenges are the group of challenges encountered while processing and analysing the data that is from capturing the data to interpreting and presenting the end results. As large datasets are usually non-relational or unstructured, thus processing such semi-structured data sets at scale poses a significant challenge: possibly more so than managing BD (Kaisler, Armour, Espinosa, & Money, 2013). In analysing the different articles reviewed the authors identified several data processing related challenges that can be grouped into 5 steps that is data acquisition and warehousing (PC_DAW) → C = 97 (42.7%), data mining and cleansing (PC_DMC) → C = 38 (16.7%), data integration and aggregation (PC_DAI) → C = 29 (12.8%), data analysis and modelling (PC_DAM) → C = 25 (11%) and data interpretation (PC_DI) → C = 15 (6.6%). As illustrated in Fig. 5, data mining and cleansing appears to be a vital step during processing the large scale unstructured data, as 97 articles out of 227 specifically discussed and highlighted the importance of this step.

- **Step 1 – Data Acquisition and Warehousing:** This challenge is related to acquiring data from diverse sources and storing for value generation purpose. The integral complexity of BD and exponentially growing demands develop unprecedented problems in BD engineering such as data acquisition and storage (Wang & Wiebe, 2014). The latter argument is supported by Paris, Donnal, and Leeb (2014) who assert that one of the prime barriers to the analysis of BD arises from a lack of data provenance, knowledge and discrepancies of scale inherent in data collection and processing. This further restricts the speed and resolution at which data can be captured and stored. As a result, this affects the capability to excerpt actionable information from the data (Chen & Zhang, 2014). To capture related and valuable information, smart filters are required that should be robust and intelligent to capture useful information and discard useless that contains imprecisions or inconsistencies – this is a challenge in itself. For the latter, efficient analytical algorithms are required to understand the provenance of data and process the vast streaming data and to reduce data before storing (Zhang, Hu et al., 2015; Zhang, Liu et al., 2015).

- **Step 2 – Data Mining and Cleansing:** This challenge relates to extracting and cleaning data from a collected pool of large scale unstructured data. Advocates of BD and BDA perceive that in identifying a better way to mine and clean the BD can result in big impact and value (Chen, Chen et al., 2012). Due to its strident, vibrant, diverse, interrelated and unreliable features, the mining, cleansing and analysis proves to be very challenging (Chen et al., 2013). For instance, in the UK National Health Service (NHS) there are many millions of patients’ records comprising of medical reports, prescriptions, x-ray data, etc. Physicians make use of such data – if for instance incorrect information is stored this may lead to physicians wrongly diagnosing conditions, resulting in inaccurate medical records. In order make use of this huge data in a meaningful way, there is a need to develop an extraction method that mines out the required information from unstructured BD and articulate it in a standard and structured form that is easy to understand. According to Labrinidis and Jagadish
laws that may govern the use of such personal information and strict observance to these privacy regulations must be applied in the data warehouse. In analysing the different articles reviewed in this SLR, the authors identified several data management related challenges that can be grouped into seven areas (Fig. 6) such as privacy (MC_P) → C = 23 (10.1%), security (MC_S) → C = 17 (7.5%), data and information sharing (MC_DiIS) → C = 10 (4.4%), cost/operational expenditures (MC_C&OE) → C = 7 (3.1%), data governance (MC_DG) → C = 4 (1.8%), and data ownership (MC_OG) → C = 3 (1.3%).

• Privacy: BD poses big privacy concerns and how to preserve privacy in the digital age is a prime challenges. Huge investments have been made in BD projects to streamline processes; however, organizations are facing challenges in managing privacy issues, and recruiting data analysts, thus hindering organizations in moving forward in their efforts towards leveraging BD (Krishnamurthy & Desouza, 2014). In a smart city environment where sensory devices gather data on citizen activities that can be accessed, several government and security agencies pose significant privacy concerns (Barnagli et al., 2013). Among such privacy related challenges, location-based information being collected by BD applications and transferred over networks is resulting in clear privacy concerns (Yi et al., 2014). For example, location-based service providers can identify subscriber by tracking their location information – which is possibly linked to their office or residential information. Then there is the challenge of protecting privacy – Machanavajjhala and Reiter (2012) report that failure to protect citizens' privacy is illegal and open to relevant Government oversight bodies.

• Security: Security is a major issue and is identified by Lu et al. (2014) who argue that if security challenges are not appropriately addressed then the phenomenon of BD will not receive much acceptance globally. Securing BD has its own distinctive challenges that are not profoundly different to traditional data. Among the several BD related security challenges are the distributed nature of large BD which is complex but equally vulnerable to attack (Yi et al., 2014), malware has been an ever growing threat to data security (Abawajy, Kelarev, & Chowdhury, 2014), lack of adequate security controls to ensure information is resilient to altering (Bertot, Gorham, Jaeger, Sarin, & Choi, 2014), analysing logs, network flows, and system events for forensics and intrusion detection has been a challenge for data security (Cárdenas, Manadatha, & Rajan, 2013), lack of sophisticated infrastructure that ensures data security such as integrity, confidentiality, availability, and accountability, and data security challenges become magnified when data sources become ubiquitous (Demchenko, Grosso, De Laat, & Membrey, 2013).

• Data Governance: As the demand for BD is constantly growing, organizations perceive data governance as a potential approach to warranting data quality, improving and leveraging information, maintaining its value as a key organizational asset, and support in attaining insights in business decisions and operations (Otto, 2011). According to Intel IT Centre (2012), IT managers highly support the presence of a formal BD strategy, this especially makes sense, since the issue of data governance for describing what data is warehoused, analyzed, and accessed is termed as one of the three top challenges they face (besides data growth and data centre infrastructure and the ability to provide scalability). du Mars (2012) state that a significant challenge in the process of governing BD is categorizing, modelling and mapping the data as it is captured and stored, mainly due to the unstructured and complex nature of data. Moreover, effective BD governance is essential to ensure the quality of data mined and analyzed from a pool of large datasets (Hashem et al., 2015).

4.1.3. Management challenges
Management challenges related to BD are a group of challenges encountered, for example while accessing, managing and governing the data. Data warehouses store massive amounts of sensitive data such as financial transactions, medical procedures, insurance claims, diagnosis codes, personal data, etc. Organizations and businesses need to ensure that they have a robust security infrastructure that enables employees and staff of each division to only view relevant data for their department. Moreover, there must be some standard privacy laws that may govern the use of such personal information and strict
potential analysis challenges, it also poses an overwhelming task of sharing and integrating key information across different organizations (OSTP, 2012). Al Nuaimi et al. (2015) also state that sharing data and information between distant organizations (or departments) is a challenge. For instance, each organization and their individual departments typically own a disparate warehouse (developed based on different technological platforms and vendors) of sensitive information and several departments are often reluctant to share their patented data governed by privacy conditions. According to Khan, Uddin, and Gupta (2014) the challenge here is to ensure not to cross the fine line between collecting and using BD and guaranteeing user privacy rights. The is also related to a smart city environment that entails a plethora of sectors and in such context, smart city technological systems will need to reduce the barriers to achieve seamless information sharing and exchange among different entities (Su, Li, & Fu, 2011).

- **Cost/Operational Expenditures:** The constantly increasing data in all different forms has led to a rising demand for BD processing in sophisticated data centers. These are generally dispersed across different geographical regions to embed resilience and spread risk, for example Google having 13 data centers in eight countries spread across four continents (Gu, Zeng, Li, & Guo, 2015). The significant resources have been allocated to support the data intensive operations (i.e. acquisition, warehousing, mining and cleansing, aggregation and integration, processing and interpretation) – all this lead to high storage and data processing big costs (Raghavendra, Ranganathan, Talwar, Wang, & Zhu, 2008). Researchers assert that cost minimization is an emergent challenge (Irani, Ghoneim, & Love, 2006; Irani, 2010), with Gu et al. (2015) explaining the challenges of processing BD across geo-distributed data centers. Advocates of BD search for cost-effective and efficient ways to handle the massive amount of complex data (Sun, Morris, Xu, Zhu, & Xie, 2014). The cost of data processing and other operational expenditures of the data center are a sensitive issue that may also impact in the way organizations adopt and implement technological solutions (Al Nuaimi et al., 2015).

- **Data Ownership:** Besides privacy, Web (2007) asserts that ownership of data is a complex issue – as big as the data itself – while sharing real time data. Kaisler et al. (2013) also claim that data ownership presents a critical and continuing challenge, specifically in the social media context such as who owns the data on Facebook, Twitter or MySpace – are the users who update their status or tweet or have any account in these social networks (Sivarajah et al., 2015; Sivarajah, Irani, & Jones, 2014). It is generally perceived that both view they (the users and the social media provider) own the data. Kaisler et al. (2013) argues that this dichotomy still needs to be settled. With ownership arise the issue or controlling and ensuring its accuracy. For instance, Web (2007) states that sensor data is too sensitive and can result in mounting errors – this may further result in capturing and revealing inconsistent data – but then who owns that data. Data ownership is a much deeper social issue. These concerns are beyond the focus on several applications, for example SensorMaps by Web (2007) requires more research since they may have deep implications.

Like other data related management challenges, data ownership is essentially vital and its issues much be addressed to realise the promise of BD.

### 4.2. Types of Big Data analytical methods

BD comprising of large raw data set on its own does not offer a lot of value in its unprocessed form. If its [BD] potential value is to be unlocked, businesses need efficient processes and methods to turn high volumes of structured and unstructured data to analyze these raw datasets. Analytics in this context refers to the methods used to analyze and acquire intelligence from BD. As a result, BD analytics methods can be viewed as a sub-process within the overall process of insight extraction from BD. Despite the hype about varying BDA methods, using analytics is still a labour intensive undertaking. As Assunção, Calheiros, Bianchi, Netto, and Buyya (2015) highlight the reason for this is that current solutions for analytics are often based on proprietary appliances or software systems built for general purposes. As a result, organizations need to put in significant effort to customize such BDA solutions to their individual needs, which might require integrating different data sources and setting up the software on the organization’s hardware.

In analysing the different articles reviewed in this SLR, a total of 115 papers out of the 227 papers analyzed discusses and proposes some form of BDA methods and tools. The extant literature highlights a number of analytical processes and methods – such as text analytics, audio analytics, video analytics, social media analytics, predictive analysis of data (Gandomi & Haider, 2015) and others reported of descriptive analytics, inquisitive analytics, prescriptive analytics and pre-emptive data analytics (Assunção et al., 2015; Rehman, Chang, Batool, & Teh, 2016). Within these various BD analytics methods, the SLR highlights that there are a number of off the shelf software tools [e.g. Hadoop, MapReduce, Dyrad] (Chen, Chen et al., 2012; Chen, Chiang et al., 2012; Jiang et al., 2015), that have been built using and extending off-the-shelf existing software [e.g. Hadoop based e-book conversion system, MapReduce-based Big Data Processing on Multi-GPU systems]) (Jiang et al., 2015) and finally novel solutions to tackle BD analysis [e.g. DEMass – A New Density Estimator for Big Data] (Ting, Washio, Wells, Liu, & Aryan, 2013). In studying the analyzed papers, the authors identified and classified analytics methods into 3 groups – such as descriptive analytics, predictive analytics and prescriptive analytics; however, nothing was specifically noted for inquisitive and pre-emptive analytics (Fig. 7).

#### 4.2.1. Descriptive analytics

Descriptive analytics are the simplest form of BDA method, and involves the summarization and description of knowledge patterns using simple statistical methods, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events in BD streams (Rehman et al., 2016). Often, large volumes of historical data is used in descriptive analytics to identify patterns and create management reports that is concerned with modelling past behaviour (Assunção et al., 2015). Watson (2014) asserts that descriptive analytics, such as reporting, dashboards, scorecards, and data visualization, have been widely used for some time, and are the core applications of traditional business intelligence. Descriptive analytics are considered backward looking and reveal what has already occurred. However, a trend that is being adopted in descriptive analytics now is to make use of the findings from predictive analytics, such as forecasts of future revenues, on dashboards/scorecards. Spiess, T’Joens, Dragnea, Spencer, and Philippart (2014) highlights root cause analysis and diagnostics are also form of descriptive analysis which involve both the passive reading and interpretation of data, as well as initiating particular actions on the system under test, and reading out the results. The authors discuss that root cause analysis is an elaborate process of continuous digging into data, and correlating various insights such as to determine the one or multiple fundamental causes of an event (Spiess et al., 2014).

Another form of descriptive analysis, pointed out by Banerjee, Bandyopadhyay, and Acharya (2013) is the use of dashboard sort of application when a business routinely generates different metrics including data to monitor a process or multiple processes across times. For example, this sort of application could be useful to understand in terms of the financial strength of a business at a given point of time or to compare it with others or its own across different point of time. In descriptive analytics, there is a need for analysts to nurture the skill of reading facts from figures, connecting them with the relevant decision-making process and finally taking a data-driven decision from a business perspective. Most of the BDA is commonly descriptive (exploratory) in nature and the use of descriptive statistical methods
(data mining tools) allows businesses to discover useful patterns or unidentified correlations that could be used for making business decisions.

4.2.2. Predictive analytics

This analytics is concerned with forecasting and statistical modelling to determine the future possibilities based on supervised, unsupervised, and semi-supervised learning models (Joseph & Johnson, 2013; Rehman et al., 2016; Waller & Fawcett, 2013). Gandomi and Haider (2015) asserts the need to develop new solutions for predictive analytics for structured BD. Predictive analytics are principally based on statistical methods and seeks to uncover patterns and capture relationships in data. Gandomi and Haider (2015) categorised predictive analysis into two groups – regression techniques (e.g., multinomial logit models) and machine learning techniques (e.g., neural networks). The authors highlight that some approaches, such as moving averages, attempt to identify historical patterns in the outcome variable(s) and extrapolate them to the future. Others, such as linear regression, seek to capture the interdependencies between outcome variable(s) and explanatory variables, and use them to make predictions. Hasan, Shamsuddin, and Lopes (2014) proposed a machine learning BD framework that envisaged the broad picture of machine learning in dealing with BD problems. The framework included the presentation of multi-structure input varieties from different sources, followed by the pipeline preprocessing phase prior to machine learning knowledge discovery. The authors implement the parallelism on machine learning approaches of BD predictive knowledge discovery based on Neural Network (NN) algorithms; Multiple Backpropagation (MBP) and Self-Organizing Map (SOM) using GPUMLib. In sum, predictive analytics aims to predict the future by analysing current and historical data. For example, determination of customers’ propensity to churn, by correlating behaviour over a period of time with network event data such as usage records and fault indicators (Spiess et al., 2014).

4.2.3. Prescriptive analytics

This type of analytics is performed to determine the cause-effect relationship among analytic results and business process optimization policies. Thus, for prescriptive analytics, organizations optimize their business process models based on the feedback provided by predictive analytic models (Bihani & Patil, 2014). Although difficult to deploy, prescriptive analytics contribute to handling the information shift and the continuous evolution of business process models (Rehman et al., 2016). There are very limited examples of good prescriptive analytics in the real world. One of the reasons for this shortage is that most databases are constrained on the number of dimensions that they capture (Banerjee et al., 2013). Therefore the analysis from such data provides, at best, partial insights into a complex business problem. Few initial
studies have applied the simulation optimization methods to the BDA. For instance, Xu, Zhang, Huang, Chen, and Celik (2014) proposed a framework called multi-fidelity optimization with ordinal transformation and optimal sampling (MO2TOS). The framework provides a foundation for descriptive and prescriptive analytics under the BD environment. In the MO2TOS framework, two set of high- and low-resolution models were developed. The authors highlighted that the high resolution model development can be very slow due to the large amount of data. On the other hand, the low-resolution models were much faster and can be developed using only a sample of data. The proposed MO2TOS framework is able to efficiently integrate the both the resolution models to optimize targeted systems under the BD environment.

In general, prescriptive solutions assist business analysts in decision-making by determining actions and assessing their impact regarding business objectives, requirements, and constraints. For example, what if simulators have helped provide insights regarding the plausible options that a business could choose to implement in order to maintain or strengthen its current position in the market.

4.3 Yearly publications

Using the keywords as stated in Section 1.2, initial search resulted in 2360 articles from 1996 until 2015 based on the number of subject areas including material sciences, energy, neuroscience, chemistry, etc. However, this research focused on only four subject areas such as business and management, computer science, decision science and social science (that directly relate to the special issue theme (i.e. Big Data and Analytics in Technology and Organizational Resource Management) – and following the systematic literature review steps (explained and illustrated in Section 3 and Fig. 3, respectively) – this research resulted in 227 articles. As presented in Fig. 8, the largest number of publications were recorded for year 2015 (with C = 114, 50.2%), followed by year 2014 (with C = 63, 27.7%) and year 2013 (with C = 43, 18.9%). With fewer publications (i.e. below the 5 mark) were recorded from 2009 until 2012 and zero articles recorded from 1996 to 2008.

Fig. 8 illustrates an abrupt increase in number of journal articles in the BD and BDA research area from 2013 onwards until 2015. Even through the initial search for articles (resulting in 2360 articles), there are more articles published from 2012 (e.g. 99 articles noted) until 2015 (e.g. 1156 articles noted). Regardless, the rapid increase in the articles highlights the awareness and importance of this area among the academic community, practitioners, and even governments worldwide (see e.g. Chen, Chen et al., 2012; Chen, Chiang et al., 2012; Joseph & Johnson, 2013). Despite the increase in the number of articles on BD and BDA, this research domain is still emerging (e.g. as noted from Scopus Database that from January 2016 to-date so far 295 articles have been published). With the significance of BD and BDA from a strategic perspective and the increasing number of articles, it appears that this research domain requires further in-depth conceptual as well as empirical, especially case study and survey based research studies.

4.4 Number of regions (geo-spatial coverage)

Fig. 9 highlights that the number of articles published on BD and BDA area represent 42 different geographical regions across the globe between 1996 and 2015. The total number of regions of the 227 articles is 790 as it takes into account of the geographical regions of the co-authors as well. It was considered appropriate to include the regions of the co-authors in order to avoid misrepresenting that each paper was single authored. From the total number of articles (i.e. 227) analyzed, the largest number of scholarly contributions came from the Chinese region (C = 241 scholars, representing 30.51%) – the 241 figure is the total number of authors and co-authors from China across all the 227 publications. This is followed by USA (C = 145, 18.35%), and then there is Australia (C = 51, 6.45%), UK (C = 49, 6.20%), and Korea (C = 37, 4.68%). The results in Fig. 5 evidently specify that China and USA have a lead on BD and BDA research area that is the upward trends in the first three to four regions noticeably indicate that there are clear signals of the growing interest in the BD and BDA area in those regions.

Whereas, from Belgium (with C = 1, 0.12%) to Italy (with C = 17, 2.15%) there is slow increase in the number of papers on BD and BDA. Nevertheless, the huge difference between two extremes clearly raises a vital research agenda for BD and BDA researchers and practitioners to explore: whether this position is a result of a global sector based BD and BDA divide or whether it is due to a lack of essential knowledge and proficiency to undertake BD and BDA research within such countries (i.e. more specifically those regions with five or less publications).

In either case, the problem of a potential global BD and BDA area needs to be further studied (and or creating awareness) among the academics from countries such as the Belgium, Czech Republic, Denmark, Hong Kong, Norway and Russia. Researchers and scholars from China, USA, Australia, UK, Korea and Spain for instance should contemplate collaborating with researchers from under-represented regions so as to undertake more productive research and contributes towards the extant BD and BDA research area.

4.5 Types of publications

This section categorizes the list of 227 papers based on the publication type. The authors employed an analogous list of publication types as employed by Dwyedi and Mustafee (2010). This list is also similar to those identified by the publisher – Emerald. The data presented in Fig. 10 demonstrates that the vast majority of the publications are research papers (C = 159, 70.04%), followed by general review (with C = 27, 11.89%) and technical and conceptual papers (with C = 15, 6.60% and C = 9, 3.96%, respectively). A large number of research papers clearly indicate the significance of the BD and BDA area in different
sectors (e.g. healthcare, government, and telecommunication). However, most of these research papers are analytical in nature (as explained in the following section) mainly focusing on experiments, performing simulations and proposing algorithms. The authors perceive that there is a need for more research considering using in-depth case studies in different sector organizations. Researchers and practitioners need to focus on developing and proposing sound solutions to BD challenges (Chen & Zhang, 2014).

4.6. Types of research methods employed

The research methods employed by the BD researchers in the selected 243 papers and were coded under different categories as suggested by Dwivedi, Kiang, Lal, and Williams (2008) and Dwivedi and Mustafee (2010). The findings suggest that although a total of 11 different types of research methods were recorded from our data analysis, the majority of studies were analytical in nature (C = 103, 45.37%). This was then followed by articles that are either conceptual/descriptive or theoretical in nature (C = 64, 28.19%), and design research (C = 12, 5.28%) methods. With regard to the analytical methods (with C = 103, 45.37%) – it was denoted as a combination of five different methods such as statistics, computer programming, simulation, algorithm and mathematical modelling, as also followed by Dwivedi and Mustafee (2010) and Kamal and Irani (2014). A big proportion of analytical articles clearly indicate that conducting experiments and simulations and or proposing algorithms have emerged as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications (Chen, Chen et al., 2012; Chen, Chiang et al., 2012).

A small number of the selected articles employed interview and survey research approach to conduct their study – perhaps this is due to the nature of the BD and BDA research discipline that requires technical and methodical analysis of the huge type and format of data involved. Most of the studies reported using survey as tool to study the literature (i.e. secondary research) as opposed to seeking responses from the respondents (e.g. Chen, Mao, & Liu, 2014). The other categories with their associated counts and percentages are presented in Fig. 11.

5. Conclusions

The authors of this paper have presented a holistic view of BD practices and application of BDA methods as presented in a normative slice of literature. Based on the findings from existing research studies, the presented research has sought to analyze, synthesize and present a comprehensive structured analysis on BD and BDA to support the signposting of future research directions. The SLR methodology adopted demonstrated to be a convenient tool for conducting a descriptive literature reviews, with contributions including the synthesis of core conclusions of the literature, the literature voids, and the formation of a foundation for future research. The findings of this structured
literature review will assist both BD and BDA academics and practitioners to develop new solutions based on the challenges identified in this paper. BD is still an emerging phenomenon but in the recent past years its significance in different industries and countries (as evident from Fig. 9) makes it a pertinent research area for academic and management studies. It is evident from the review conducted that it has significantly changed the data management landscape with scope for further profound changes.

This SLR paper has revealed the past and current state of BD and BDA research published, thereby focusing on the past trends and current patterns in BD and BDA practices. Following Tranfield et al. (2003) and Kitchenham and Charters (2007) Systematic Review Approach, this paper extracted and reviewed 227 journal articles from 1996 to 2015 from the Scopus database – as a result fulfilling the aim of this literature review paper (as indicated in Section 1.1). Figs. 4 to 11 clearly indicate the past trends and current patterns in the number of articles published on BD and BDA. Moreover, the continuing interest (as indicated from Fig. 8 – increasing number of BD and BDA articles over the years) specifies that in future research studies; academics, researchers and practitioners may focus on the BD challenges to further propose robust solutions to the challenges of acquiring and storing, mining and cleansing, aggregating and integrating, analysis and modelling and interpreting data. The intention in conducting this detailed investigation was to provide a useful and usable resource of information for future researchers.

5.1. Research implications to research and practice

- Implications to Research: This SLR offered a number of useful insights into the extant status of research into BD and BDA, how it is defined and conceptualised, and the key types of research methodologies employed to date. The prime emphasis of these SLR based articles has been on analytical and or conceptual/descriptive/theoretical research methods; however, due to the emerging nature of this area, there is a need to develop and understand BD and BDA in an intensive way using case studies and survey based research where appropriate. The authors assert that more practical insights into BD and BDA articles over the years) specifies that in future research studies; academics, researchers and practitioners may focus on the BD challenges to further propose robust solutions to the challenges of acquiring and storing, mining and cleansing, aggregating and integrating, analysis and modelling and interpreting data. The intention in conducting this detailed investigation was to provide a useful and usable resource of information for future researchers.

This paper extends the research stream on BD and BDA by demonstrating and analysing the key trends related to the challenges of BD and BDA methods.

- Implications to Practice: The authors of this paper have presented the practice community with an insight to the plethora of BD and BDA methods available and, insight to their application. While there is no one advocated robust approach, the descriptive insight presented will offer an opportunity for practitioners and applied researchers to align their approaches to the application pursued by others.

5.2. Limitations

The authors recognise that our study has limitations, and readers and future academics and researchers should be aware of these and indeed interpret the material presented in this paper within the context of the limitations. By explanation, a meta-analysis rest on the existing as well as accessible research studies (both conceptual and empirical). While the authors conducted a thorough literature search through the Scopus database to identify all possible relevant articles, it is possible that some research articles could have been missed in this review from some other leading databases (i.e. Web of Science and EBSCO). So to avoid duplication, every effort was exhausted to acquire and analyze all relevant information essential, regarding the two questions (i.e. Q1 and Q2) from the articles reviewed from the Scopus database. Additionally, the analysis and synthesis are based on the research team interpretation of the selected articles. The authors attempted to avoid these issues by cross-checking papers independently and thus deal with embedded bias but errors might have occurred but this research is considered robust as every effort to mitigate error was taken.

5.3. Suggestions for future research

Building upon the rich underpinning of the research findings described and overall understanding acquired in this paper, the authors presents the concerns that merit further research and anticipate that these issues may hold the potential in contributing towards the future research studies. The analysis of the selected articles reveals that the opportunity clearly exists to strengthen empirical research based on in-depth case study based qualitative and survey based quantitative approach, as most of the articles analyzed followed an analytical approach. Furthermore, there is need for stronger infusion of generic theory into the BD and BDA debate. BD is a cross-cutting theme, and many
connections exist with established topics across computing, engineering, mathematics, business and management, social sciences, etc. It would be valuable to expand the scope of the subject area and to repeat this exercise to identify and draw links with established theoretical contributions in other different associated areas. A publication based on such analysis would provide an extremely valuable platform for the BD and BDA research and practitioners’ community.

Appendix A

<table>
<thead>
<tr>
<th>Paper code</th>
<th>Citation</th>
</tr>
</thead>
</table>

Appendix A (continued)
Appendix A (continued)

<table>
<thead>
<tr>
<th>Paper code</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>the computing environment, MapReduce, and programming frameworks. Wiley Interdisciplin ar Reviews: Data Mining and Knowledge Discovery, 4(5), 80–409.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix A (continued)

SLR_148  

SLR_149  

SLR_150  

SLR_151  

SLR_152  

SLR_153  

SLR_154  

SLR_155  

SLR_156  

SLR_157  

SLR_158  

SLR_159  

SLR_160  

SLR_161  

SLR_162  

SLR_163  

SLR_164  

SLR_165  

SLR_166  

SLR_167  

SLR_168  

SLR_169  


