Review Article

Big Data in academic libraries: literature review and future research directions

Hafidha Al-Barashdi[1]*, Rahma Al-Karousi[2]

[1] The Research Council, Muscat, Oman
Email: hafidhaalbarashdi@gmail.com

[2] Rustaq College of Education, Al-Rustaq, Oman
* Email: rahmaalkarousi.rus@cas.edu.om

ABSTRACT

Recently, Big Data studies have attracted considerable attention. However, Big Data analytics in academic libraries confront two fundamental challenges: the huge volume, velocity, and variety of data and the complexity of its techniques and algorithms. The primary aim of this study is to explore which techniques and tools can be applied in academic libraries in order to analyze Big Data, and then determine its profits in academic libraries. In addition, this study attempts to answer the following research questions: how should librarians be made to involve in Big Data? What are the future research developments in Big Data? What are the gaps in Big Data studies related to academic libraries? To provide a considerably better understanding of the advantages of Big Data in academic libraries and their future research directions, a comprehensive literature review of Big Data analysis of academic libraries over the last seven years was conducted. The results yielded a total of 37 papers related to Big Data in academic libraries. These results indicated that despite the large amount of research conducted on this topic, only a few studies discussed the implication of Big Data in academic libraries, including the analyzing tools and techniques. The benefits of Big Data in academic libraries and its implications on methodology in future studies are discussed. The present study also highlights the evolving field of Big Data research in academic libraries.

Keywords:

Big Data, Big Data tools, Big Data analyzing techniques, academic libraries

1. INTRODUCTION

The use of Big Data as a resource can become very rife due to its application in educational analysis and data-driven decision-making, and it can even emerge as a vehicle for state transparency. Almost each sector has developed a fascination with the ostensibly new discovery of Big Data and its extraordinary capabilities to fuel analytical breakthroughs since 2012 (Reinhalter & Wittmann, 2014).

For academic libraries, Big Data analytics is affected by two basic challenges: first, due to the massive volume, selection, and speed of the knowledge concerned, the storage and process needs of the system are rather overwhelming, and second, the analytics techniques and algorithms are complicated, which makes Big Data analytics a computing-intensive task. To support the storage

Cite this article as: Al-Barashdi H. & Al-Karousi R. Big Data in academic libraries: literature review and future research directions, Journal of Information Studies and Technology 2018:2,13 https://doi.org/10.5339/jist.2018.13

http://dx.doi.org/10.5339/jist.2018.13

كيوساينس
QSCIENCE
دار جامعة حمد بن خليفة للنشر
HAMAD BIN KHALIFA UNIVERSITY PRESS

and process needs of Big Data analytics applications, cloud has been found to be the most acceptable infrastructural resolution.

Cloud computing suggests an economical resolution for storing, processing and managing Big Data for analytical functions, enabling the application of distributed and parallel paradigms in order to meet the potential needs. Big Data analytics is a vast field that has found applications in various domains and studies (Khan, Liu, Shakil, & Alam, 2017).

This study aims to review the emergence and potential of Big Data, and discuss their impact on academic libraries. With these libraries evolving to provide many information services, librarians are most likely to become consultants and authorities in the information age.

Consequently, the goal of this study is to completely review the literature associated with Big Data in instructional libraries. In doing so, we attempt to answer the following questions:

RQ1. What are the definitions and approaches to Big Data in academic libraries?

RQ2. What are the Big Data reading strategies and tools suitable for academic libraries?

RQ3. What are the benefits of Big Data in academic libraries?

RQ4. How should librarians be made to involve in Big Data?

RQ5. What are the gaps in Big Data research related to academic libraries and future research directions?

2. RESEARCH METHODOLOGY

The research methodology involved three stages. The first stage comprised searching for the studies related to Big Data from four selected databases. Thus, journal articles published from 2010 to 2017 were examined, as well as credible international conference papers. The main keywords used were: (1) Big Data, (2) Big Data analyzing techniques, (3) Big Data tools, and (4) academic libraries. Papers were excluded if they were not directly related to Big Data or found to be irrelevant. The second stage concerned the classification of these works in different fields of knowledge according to the research questions. Finally, the third stage involved the report of a detailed literature review. For this review, four online academic research databases, namely Scopus, EBSCO, Springer Link and ScienceDirect, were examined for relevant articles. To increase the reliability of the search results and ensure that the included articles were from different academic fields, the search was repeated with Google Scholar, which resulted in several newly included articles. We reduced the search to empirical studies published in peer-reviewed full conference papers and journal articles. As a result, a total of 37 articles that met our criteria were included for review. The included articles were retrieved from the aforementioned databases during the years indicated in Figure 1.

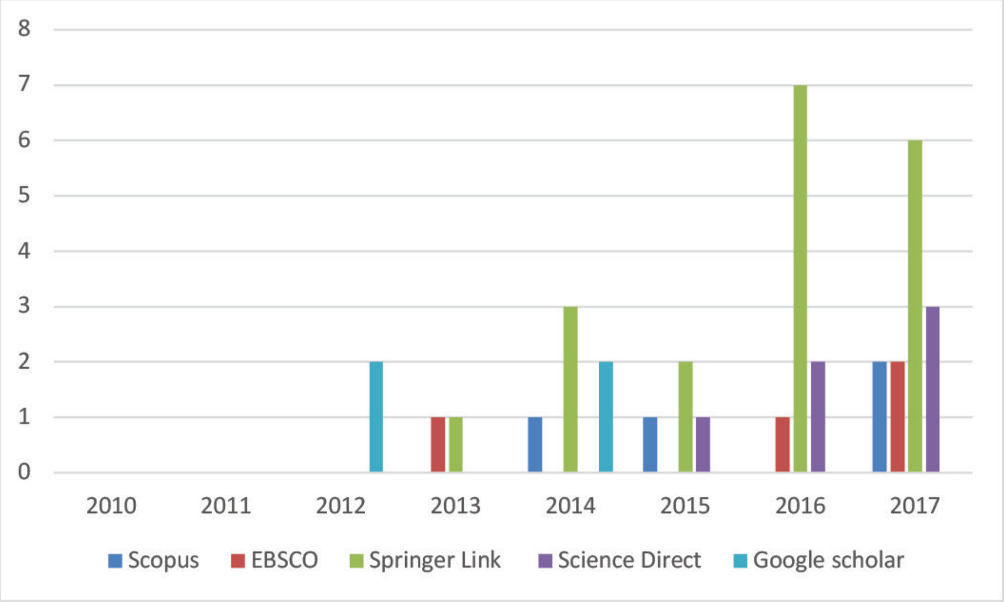

Figure 1. Selected articles and databases according to publication year (total 37).

3. LITERATURE REVIEW

3.1. Definitions and approaches to Big Data

Extensive discussions and development have been focused on the definition of Big Data. The first and probably the most commonly used definition of Big Data was conceptualized in 2001 by Laney, who characterized Big Data using the 3Vs model: volume, velocity, and variety. Similarly, based on this model, Sagiroglu and Sinanc (2013) presented an extensive review of Big Data research, especially the security issues. Furthermore, Lomotey and Deters (2014) extended the model defined by Laney (2001) to a 5V model (volume, veracity, velocity, value, and variety), as shown in Figure 2.

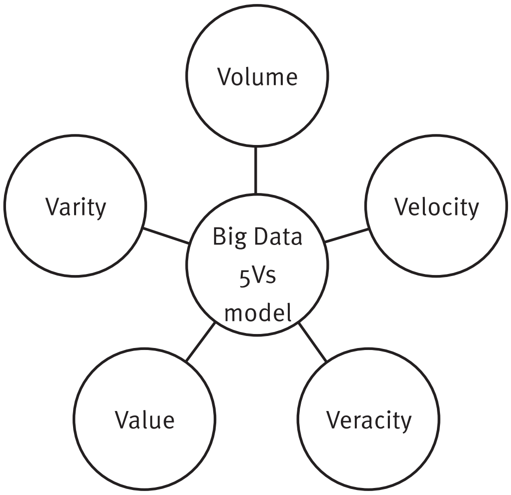

Figure 2. The 5V model that currently defines Big Data.

Big Data is a term used for large datasets with large and complex structures, which has become a prominent study area for both practitioners and researchers. However, Big Data passed the highest point in the Gartner Hype Cycle, attesting the maturity level of this technology and its applications in 2012 (Akoka, Comyn-wattiau, & Laou, 2017; Khan et al., 2017). Keil (2014) argued that the term Big Data is used to describe very large datasets; however, an enormous amount of scientific data is called long-tail datasets, which means they are small and heterogeneous as well as comprise a growing portion of scientific data.

However, Gantz and Reinsel (2011) defined Big Data as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis.” Big Data also describes the storage and analysis of large and/or complex datasets using a series of techniques, including NoSQL, MapReduce and machine learning (Blascheck, Burch, Raschke, & Weiskopf, 2015; Ward & Barker, 2013).

Recently, Akoka et al. (2017) observed a significant growth of analysis articles on Big Data over the last five years. They noted a diversity of interest among researchers in problems such as the establishment of objectives and artifacts, the use of standard criteria, as well as the wide range of usages and applications. They concluded that Big Data analysis focused on the following techniques: clustering, classification, and prediction. Likewise, based on the analysis of the aforementioned definitions consistent with four axes, namely technology, technique, information, and impact, De Mauro, Greco, and Grimaldi (2015) proposed the following definition: “Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value” (p. 103). Cuzzocrea, Song, and Davis (2011) defined Big Data as the one aiming at the traits of the generated information, containing both the quantity and structure of the data. Apparently, Bizer, Boncz, Brodie, and Erling (2011) developed the definition by including the fact traits with additional attributes, namely the scope, goal, and structure of the data, even as Jacobs (2009) targeted on the quantity of statistics and included the issue of approach. However, Chen, Chiang, and Storey (2012) covered the topics of techniques and IT infrastructure.

Madden (2012) incorporated data characteristics and infrastructure. According to Rodríguez-Mazahua et al. (2016), Big Data refers to the large quantity of structured, semi-structured, and unstructured knowledge that are exponentially generated by superior applications in several areas such as organic chemistry, genetics, biology, physics, astronomy, and business. Thus, a group of previous studies focused on Big Data impacts in different domains. They identified how Big Data solutions may help make new contributions to different fields. Academic units that discuss Big Data research are of three types. The primary one is the hard-core science units, which involve analysis in computer Standards & Interfaces, natural philosophy, climate science, and genetics. The second one is the info sciences units that are moderately causative to the Big Data paradigm. The last cluster consists of all the units that have the potential to operate with Big Data, such as healthcare, education, public policy, and government studies (Akoka et al., 2017; Goes, 2014).

In summary, Big Data is much more than simply large amounts of data: it is the expansion in the volume of structured and unstructured data, the speed at which it is created and collected, and also the scope of how many data points are covered. Big Data often originates from multiple sources and appears in multiple formats (Larkou, Mintzis, Andreou, Konstantinidis, & Zeinalipour-yazti, 2016). As a result, academic libraries are pivoting towards meeting data management needs in order to support researchers. Table 1 summarizes the themes of Big Data definition found from the literature.

Table 1. Big Data definition found from the literature.

Themes	Codes	Previous studies	N
Information	Big Data units with big structures	Akoka et al. (2017); Keil (2014); Suthaharan (2014)	10
	Increase in the volume of structured and unstructured information	Larkou et al. (2016); Andreu-Perez, Poon, Merrifield, Wong, and Yang (2015)	
	Fact property characterized using at an excessive extent, pace, and range	De Mauro et al. (2015); Schroeck, Shockley, Smart, Romero-Morales, and Tufano (2012)	
	Generated facts containing both the quantity and structure of the data	Cuzzocrea et al. (2011)	
	Adding the fact traits with additional attributes, including the scope, target, and shape of the facts	Bizer et al. (2011)	
	Incorporated information characteristics and infrastructure	Madden (2012)	
Technology	New generation of technologies and architectures, designed to economically extract the cost from very large volumes of an enormous form of data	Gantz and Reinsel (2011); Microsoft (2013).	4
	Large and/or complex information sets using a sequence of techniques	Ward and Barker (2013); Manyika et al. (2011)	

Methods	Amount of information and including the issue of method.	Fisher, DeLine, Czerwinski, and Drucker (2012)	4
	Applying the component of techniques and IT infrastructure to the facts	Chen et al. (2012); Beyer and Laney (2012); Manyika et al. (2011)	
Impact	Big Data refers to data, which can be exponentially generated using excessive, overall performance packages in many domains: biochemistry, genetics, molecular biology, and so forth	Rodríguez-Mazahua et al. (2016); Chen et al. (2012)	4
	The impact of huge data analytics on healthcare and government	Archenaa and Mary Anita (2015)	
	Three types of academic units that discuss significant statistical impacts	Goes (2014)	

Table 1 presents four themes associated with the definitions of Big Data found from the literature. These are information, technology, methods, and impact. Therefore, this study recommended the use of the following definition, in order to explain the relationship between Big Data and academic libraries: "Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." We believe that the use of this definition will allow a more efficient scientific development of Big Data.

3.2. Big Data analyzing techniques

Big Data can be analyzed using several techniques (Rodríguez-Mazahua et al., 2016). Manyika et al. (2011) suggested that applied mathematics techniques, which are often used in decision-making about the type of relationships existing between variables, might have occurred inadvertently, and these relationships might result from certain underlying causative relationships. However, these techniques are used to reduce the probability of type I errors (false positive errors) and type II errors (false negative errors).

Another widely used technique proposed by Bu et al. (2012) was machine learning. The purpose of this technique is to transfer the observational data in a way that can help predict any hidden data. Nowadays, machine learning is widely used as a method that drives the market and sales of online shops, keeps out spam emails, organizes advertising systems, and builds a content recommender system to suggest targeted users. It also supports scientists and researchers in different fields of science to interpret Big Data to acquire applicable knowledge, especially in certain fields such as high energy physics and biology.

Data mining is also a technique used to analyze large information repositories and discover implicit, but potentially valuable information. Evidently, analysis of large amounts of information from Big Data leads to the emergence of new techniques such as data mining. It is also known as knowledge discovery in databases (Han, Kamber, & Pei, 2011). The purpose of data mining is to reveal hidden relationships and unknown patterns and trends by mining into giant amounts of data (Sumathi & Sivanandam, 2006). Three techniques are used to analyze data mining, namely classical statistics, artificial intelligence, and machine learning (Girija & Srivatsa, 2006). Data mining in academic libraries is called the bibliomining. This concept is used to track patterns, behavioral changes, and trends in library system contacts (Siguenza-Guzman, Saquicela, Avila-Ordóñez, Vandewalle, & Cattrysse, 2015).

The application of bibliomining tools is a growing trend, which helps identify behavioral patterns among library users and employees, and usage patterns of data resources throughout the library (Nicholson & Stanton, 2006). Bibliomining is strongly suggested to provide helpful and necessary data for library management needs, specializing in problem-solving, despite being greatly dependent on information technology. Bibliomining often provides a comprehensive summary of the library workflow to monitor employee's performance, identify areas of deficiency, and predict future user needs (Prakash, Chand, & Gohel, 2004).

A signal process is another technique for Big Data analysis. Big Data challenges offer several opportunities for signal process analysis, wherever knowledge-driven applied and mathematics learning systems can be visualized to simplify disseminated Big Data analysis. Traditional and smart signal techniques for processing can be applied, such as wordbook learning, primary part analysis, and compressive sampling. However, these techniques cause huge stress on the time-data adaptive process equally to reduce the spatial property, while the primary role of the signal process in Big Data analysis is to simplify the character and scope of growing knowledge in different fields of science, especially multidisciplinary signal processing approaches (Slavakis, Giannakis, & Mateos, 2014).

Another technique used is multidimensional visualization. This technique has been used widely as a data abstracting tool, a data discovery tool in high-dimensional databases, and a data awareness tool in the cognitive process (Jain, Murty, & Flynn, 1999). This technique enables researchers to replicate the character of information efficiently according to their features, relationships, trends, and constellations (Ma, Shang, & Yuan, 2012).

In summary, the importance of Big Data analytics is growing rapidly as organizations are prepared to leverage their data assets to achieve competitive advantage. Big Data analytics offers flexibleness that allows useful but strong-level performance. Table 2 summarizes the themes of Big Data analyzing techniques found from the literature.

Table 2. Big Data analyzing techniques found from the literature.

Themes	Codes	Previous studies	N
Statistics techniques	It used to make decisions about the type of relationships existing between variables	Manyika et al. (2011)	1
Machine learning techniques	Its purpose is to convert observational statistics into a version that may be used to expect or explain hidden information	Bu et al. (2012)	1
Data mining techniques	The method of studying large statistical repositories and finding implicit, but potentially valuable records	Han et al. (2011)	6
	To reveal hidden relationships and unknown patterns and trends by way of digging into massive quantities of statistics	Sumathi and Sivanandam (2006)	
	Bibliomining is used to provide beneficial and important information about library control requirements, focusing on the expert librarianship problems, despite being dependent solely on database technology	Prakash et al. (2004)	
	Bibliomining is an emerging trend that can be used to understand styles of conduct among library users and staff, and patterns of statistics resource use throughout the library	Nicholson and Stanton (2006)	
	Bibliomining is used to track styles, conduct changes, and traits of library systems contacts	Siguenza-Guzman et al. (2015)	
	A text mining technique can be followed for the analysis of social media statistics of instructional libraries	Al-Daihani and Abrahams (2016)	

Signal processing techniques	Information-driven statistical algorithms aim to facilitate dispensed and real-time analytics	Slavakis et al. (2014)	1
Visualization techniques	They are tools for the evaluation of high-dimensional databases in knowledge discovery, recording cognizance and choice-making process	Jain et al. (1999)	2
	Displaying the distribution of the high dimensional statistics via low-size visual space, researches can speedily grow to be aware of statistics such as characteristics, relationships, clusters, and trends.	Ma et al. (2012)	

Table 2 indicates five themes associated with Big Data analyzing techniques found from the literature. These are statistics techniques, machine learning techniques, data mining techniques, signal processing techniques, and visualization techniques. Therefore, this study recommends the use of these techniques in academic libraries.

3.3. Big Data tools

Big Data tools can be categorized into three types of paradigms in line with the type of analysis. The first type is batch analysis where data are first filtered and then analyzed. The second type involves the stream process which analyzes data as presently as attainable to derive its results. The third type concerns interactive analysis which processes the information, allowing users to undertake their own analysis of data (Chen & Zhang, 2014; Demchenko, Grosso, de Laat, & Membrey, 2013; Rodríguez-Mazahua et al., 2016).

3.3.1. Big Data tools for batch analysis

Big Data tools based on batch analysis include Google MapReduce, Apache Hadoop, Microsoft Dryad, and Apache Mahout. Google MapReduce is a batch-oriented parallel computing model, where one node is selected to be the master node that is responsible for assigning the work, while the rest are worker nodes. This tool has been widely used by both academics and industry as a good example of the Big Data processing tool in a cloud environment (Dean & Ghemawat, 2008). It allows an inexpert programmer to develop equivalent programs and build a program capable of victimizing computers in an exceedingly cloud environment (Hashem et al., 2015). MapReduce accelerates the process of huge amounts of data in an exceedingly cloud environment, which makes it the most well-favored computation model of cloud suppliers (Zhifeng & Yang, 2013). Despite its excellent success and edges, MapReduce shows many limitations, making it unsuitable for the general spectrum of needs for large-scale data processing. Likewise, repetitive algorithms do not give good performance as they have to launch a MapReduce job for every iteration, notably increasing the computation time because of the overhead (López, del Río, Benítez, & Herrera, 2014).

Apache Hadoop includes a MapReduce aspect (for dispensed computation) and a scalable storage aspect, Hadoop record machine (HDFS), which can regularly replace costly SAN devices (Begoli & Horey, 2012). While HDFS is predominantly used as the dispensed report system on Hadoop, it also supports other document systems such as Amazon S3. It involves loading data as files into a set of distributed records, which then makes it possible to use MapReduce computation of the data. However, domain-specific types of programing language such as Java or Python are required to enable MapReduce computation on Hadoop. Many key components widely make use of Hadoop in deep analytics of Big Data, such as garage subsystem, caching layer, scheduler, data practitioner, fact codes, and compression algorithm (Herodotou, Lim, & Luo, 2011). Hadoop sub-initiatives, including Hive and HBase, provide additional statistical control for storing unstructured and semi-dependent data units (Begoli & Horey, 2012). The Hadoop device has grown quickly to be the "gold standard" in industry, which is considered as an enormously scalable fact, in-depth MapReduce platform that is now broadly used for web indexing, clickstream and log evaluation, and positive large-scale fact extraction and device mastering functions (Borkar, Carey, & Li, 2012).

Microsoft Dryad consists of a parallel runtime machine, which is known as Dryad, and better-level programming models, which is called DryadLINQ (Isard, Budiu, Yu, Birrell, & Fetterly, 2007; Yu et al., 2008). Dryad is an infrastructure that allows a programmer to apply the sources of a computer cluster or data center using the data-parallel package. A Dryad programmer can use heaps of machines, each of them with more than one processor or core, without understanding of concurrent programming. A Dryad programmer writes several sequential applications and connects these using one-way channels. It is a graph generator, which can change the response to computation activities during execution. It includes MapReduce or relational algebra to handle job creation and control, scheduling and accounting, fault-tolerance, aid management, and so on.

In comparison, Apache Mahout is a tool used for data mining and gaining the open-source software program based solely on Hadoop. Currently, Mahout is used for three purposes: advice, clustering, and category. The Apache Mahout assignment aims to make building smart applications simple and fast (Owen, Anil, Dunning, & Friedman, 2011).

3.3.2. Big Data tools for stream analysis

Data streams are widely used in financial analysis, online trading, and medical testing. Therefore, effective theoretical and technical frameworks are required to support data stream mining (Wu, Zhu, Wu, & Ding, 2014). This includes mining large volumes of data, applying a computer cluster in order to balance the workload, and using high-speed systems such as the message passing interface. For a productive environment, four system tools can be successfully applied: Apache Storm, Apache S4, Project Spark, and MOA.

Apache Storm is a simple system tool for stream analysis, which can be considered as an open-source and distributed real-time commutation system. It has many applications because it can reliably process unbounded streams of data for real-time processing, machine learning, and continuous computation. It can also be used with any programing language. Moreover, Storm is a fast tool that can process over a million tuples per second per node (Rodríguez-Mazahua et al., 2016).

Apache S4 is a system tool with a standard-motive, distributed, scalable, fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of information. In this system, processing elements carry out computation and send codes that enable communication among them in the form of data activities. These elements with transparency and design offer encapsulation efficiency. Therefore, system builders can write these elements in the Java programing language (Neumeyer, Robbins, Nair, & Kesari, 2010).

Apache Spark is a recollection cluster computing engine that offers support for workload and interactive computation. It is widely used in research and business institutions; thus, it has become the most active fact project in the open-source community. In addition, it is an in-memory cluster computing platform that can be used with many programing languages such as Java and Python (Stoica, 2014).

Finally, the MOA tool is an open-source software program that can be used in implementing data classification, and as a regression tool, it enables both data mining and graphic miming. MOA can be integrated with another software, namely S4 and Typhoon, for distributed stream mining (Fan & Bifet, 2012).

More specifically, Al-Daihani and Abrahams (2016) applied a text mining approach to investigate a huge dataset of tweets using academic libraries. The findings highlighted the use of the textual content analysis method in the assessment of social media information of tutorial libraries. The use of data and textual content mining procedures can be helpful in understanding the combined social statistics of academic libraries for decision-making and strategic planning for benefactor outreach and offering marketing.

3.3.3. Big Data tools based on interactive analysis

Consistent with the preceding studies, the following three crucial interactive evaluation tools can help Big Data systems and applications: Apache Drill, SpagoBI, and D3.

Apache Drill is a framework tool that can support Big Data interactive analysis and distributed programs. It allows the evaluation of the simplest nested data using an interactive advert-hoc question tool. It can generate petabytes of data and trillions of statistics per seconds (Rodríguez-Mazahua et al., 2016).

SpagoBI is a tool that can generate large amounts of dependent and unstructured data. It is widely used to support real-time Big Data related to commercial enterprise intelligence, as well as to give meaning using semantic analysis. It provides meaningful facts through schedulable datasets using special tools such as self-service BI characteristics, dashboards, ad hoc reporting, and investigative assessment. Moreover, this tool can combine with an ObE engine to help users create their own analysis inquiries on the Big Data databases. However, in order to enable this feature, several conditions should be met to prepare the databases (Franceschini, 2013).

D3 is an interactive tool that can help Big Data systems and applications. It is a JavaScript-based library used to manipulate facts contained in several reports. Conversely, it also enables qualitative visualization of Big Data databases. This requires the use of HTML and CSS to allow interface designing, and the use of format photography design tools such as Adobe Photoshop. These web standards allow technical communication to select suitable visualization of facts, enable audience engagements, and authorize users to enter Big Data database browsers (Bostock, Ogievetsky, & Heer, 2011). Table 3 summarizes the themes associated with Big Data tools found from the literature.

Table 3. Big Data tools found from the literature.

Themes	Codes	Sub-codes	Previous studies	N
Batch analysis	Google MapReduce	Information processing on large clusters	Dean and Ghemawat (2008); Hashem et al. (2015); Zhifeng and Yang (2013); López et al. (2014)	12
	Apache Hadoop	Infrastructure and platform	Herodotou et al. (2011); Begoli and Horey (2012); Borkar, Carey, and Li (2012)	
	Microsoft Dryad	Infrastructure and platform	Isard et al. (2007); Yu et al. (2008)	
	Apache Mahout	Device studying algorithms	Owen et al. (2011)	
Stream analysis	Apache Storm	Real-time computation machine	Rodríguez-Mazahua et al. (2016)	4
	Apache S4	Movement computing platform	Neumeyer et al. (2010)	
	Apache Spark	Engine for large-scale data processing	Stoica (2014)	
	MOA	Framework for data stream mining	Fan and Bifet (2012)	
Interactive analysis	Apache Drill	Square query engine for Hadoop and NoSQL	Rodríguez-Mazahua et al. (2016)	3
	SpagoBI	Commercial enterprise intelligence	Franceschini (2013)	
	D3.js	Interactive	Bostock et al. (2011)	

Table 3 presents three themes associated with Big Data tools found from the literature. These are batch analysis, stream analysis, and interactive analysis. Therefore, this study recommends the use of these tools in academic libraries in order to benefit from Big Data opportunities.

3.4. BENEFITS OF BIG DATA IN ACADEMIC LIBRARIES

Valuable information can be extracted from raw scholarly data. However, data extraction has three demanding conditions: accuracy, insurance, and scalability. As the accuracy of data extraction techniques directly influences statistical quality, it is very crucial to obtain accuracy as much as possible. Coverage depends on precision, which is as important as extracting proper systems. Evidently, scalability is a task specific to big scholarly data due to the large amount of facts to be processed (Khan et al., 2017). MapReduce serves as a useful and feasible programming paradigm for coping with the difficulty in scalability. Metadata, writer data, citations, and sections with additional information are the four types of statistics that need to be extracted from scholarly information (Khan et al., 2017).

Since Big Data includes all types of information resources, especially those which can be combined for complicated evaluation due to the proliferation of private computing devices and smartphones, massive quantities of data are generated every second through e-conversation, e-trade, GPS navigation, social media, online search (Reinhalter & Wittmann, 2014). Although academic libraries have reached high-level services for collecting, evaluating, and managing assets (Borin & Yi, 2008), profound changes in technology, budgeting, scientific communication, and publishing demands redefined academic library collections (Nabe, 2011). As a result, these libraries try to increase their contributions at the international level, in order to satisfy their user needs. For example, library collections that are developed based on regionally produced clinical data create demanding conditions that force librarians to strengthen the educational networks of their users. Therefore, effective development systems need to be developed at these libraries (Newton, Miller, & Bracke, 2010).

Several studies have argued that English is the key language of online communication with approximately 80% of the full online content material comprising English documents. This aspect makes it difficult for non-English files to reach maximum online repositories. This is because most journals indexed by Scopus require that each article should have an English title for it to be measured for inclusion (Khan et al., 2017; Van Weijen, 2012). For example, Siguenza-Guzman et al. (2015) characterized a primary systematic, identifiable, and complete instructional literature review of information mining strategies implemented in academic libraries; however, their analysis was restricted to magazine articles posted in English.

The development of data assets has led a few libraries to have specialized fact departments. Heidorn (2011) mentioned that libraries should be curators of digital data to adhere to their main undertaking, in order to defend and disseminate statistics. Therefore, control plans for suitably communicated information are needed to increase award chances. In order to meet this demand, most important library studies worldwide are advancing statistics career departments. However, with the growing incorporation of fact services in libraries, data librarians may need to control these main features. Gordon-Murnane (2012) pointed out that librarians are required in four key areas: (1) organization, (2) corporation search and access to internal datasets, (3) knowledge of external data resources, and (4) specialists on copyright and intellectual belonging problems.

In contrast, several academic libraries have opted to take on a greater active role in information control. In the case analysis of Purdue University's improvement of information repository, Purdue University Research Repository (PURR) demonstrated the involvement of library in developing a technique to meet the data needs of its researchers. The library joined forces with data technology and study departments to create PURR. Being the vanguard of data education and reference, as well as informed on metadata requirements, the library becomes a prominent leader in the improvement of data repository (Witt, 2012).

For libraries, data repository is an exemplary way to provide fact to its patrons and even to have control over the facts produced via the academy. The burgeoning field of information services in libraries, together with improved professional alternatives for statistics specialists, suggests the requirements for new abilities and training for such fact-specific roles. While these new roles demand advanced technical intelligence, data must be viewed as a set to be included in the library's cadre of sources (Reinhalter & Wittmann, 2014).

Big Data technology makes it easier to work with big datasets, hyperlink unique datasets, locate styles in real time, predict results, undertake dynamic risk scoring, and test hypotheses. In contrast, both libraries and librarians are uniquely ideal for operating with large statistics. Libraries have a protracted subculture of being data handlers and generation adopters, and Big Data needs are no exception (Rani, 2016).

Academic libraries can benefit from making an investment in data and text miming techniques, in order to analyze their postings, benchmark against the postings of other libraries, and assess their buyers' pride and level of engagement. In addition, wealth formation may be extracted from social media in real time, which could be used in enhancing and developing libraries and fact sources and services (Al-Daihani & Abrahams, 2016). Table 4 summarizes the themes related to the benefits of Big Data in academic libraries found from the literature.

Table 4. Big Data benefits in academic libraries found from the literature.

Themes	Codes	Previous studies	N
Big Data organization plans	Creating massive information provider departments	Heidorn (2011)	4
	Corporation search and access to inner datasets, knowledge of external record sources, and specialists on copyright and intellectual belonging problems	Gordon-Murnane (2012) Witt (2012) Reinhalter and Wittmann (2014)	
	Creating data repository		
Theoretical and academic understanding of Big Data and analytics in academic libraries	Work with large datasets, hyperlink exceptional datasets, locate patterns in real time, anticipate effects, undertake dynamic risk scoring, and test hypotheses	Rani (2016)	1
Supporting researchers	Encourage biomedical researchers to impeach the safety and integrity of the records, and help them to track information throughout its life cycle	Keil (2014)	1
Investing in the chances of Big Data and text miming practice	Resource libraries studying their postings, benchmark towards the postings of other libraries	Al-Daihani and Abrahams (2016)	3
	Evaluate library purchasers' satisfaction and degree of engagement	Al-Daihani and Abrahams (2016)	
	Wealth data may be extracted using social media	Al-Daihani and Abrahams (2016)	

Table 4 presents four themes related to the benefits of Big Data in academic libraries found from the literature. These are Big Data management plans, conceptual and theoretical understanding of Big Data and analytics in academic libraries, supporting researchers and investing in the opportunities of Big Data and text miming methodology. Therefore, this study recommended the academic libraries to benefit from the opportunities that the Big Data offers.

3.5. HOW SHOULD LIBRARIANS BE MADE TO INVOLVE IN BIG DATA?

Sandhu (2015) indicated the significance of learning in general data mining and large datasets for academic libraries to enhance the performance of library and fact services. Reinhalter and Wittmann (2014) point out that, while the abilities of Big Data are simply being recognized, its opportunities have captured the attention of the international library. In terms of scientific research, librarians can fill the career gap by not only enforcing requirements and good practices but also providing directions with the use of DMPs (data management platforms). Another factor is the collaboration between librarians to create data repositories. They can increase the thrust into these challenges by adapting new technologies, thereby improving their cutting-edge trends in research. In addition, reference records can be included in order of preference and datasets.

Librarians play a role in the collection, development, and renovation of statistics units and usage records, as well as incorporating information, research information management, and data literacy into their academic programs. However, all of these require understanding of what information we have received, what statistics we need to create and what statistics we need to negotiate, in order to gain access and perform the evaluation (Rani, 2016).

Keil (2014) discussed the new role of librarians in supporting researchers in order to control data, maintain, and share statistics in the virtual age. He emphasized the need for educational libraries to enable clinical data control for researchers. He also discussed key issues related to ordering raw data and the role of educational libraries in depositing raw data. This results in seemingly new principles of statistics sharing and long-tail statistics. The idea of formal data sharing plans, for example, encourages biomedical researchers to impeach the security and integrity of the statistics, and help them to track information throughout its life cycle. The new roles of librarians include manual school on the chain of custody, safety, and copyright problems, which may additionally contribute to clarifying the ownership of raw data. Table 5 summarizes the themes related to librarian roles in Big Data found from the literature.

Table 5. Librarian roles in Big Data.

Themes	Codes	Previous studies	N
Understanding the capabilities of Big Data	Utilizing usage records effectively	Reinhalter and Wittmann (2014)	
Using Big Data tools	Using data mining and massive statistics tools in academic libraries	Sandhu (2015)	
	Collecting, developing, and preserving statistics	Gordon-Murnane (2012)	1
	Developing a technique according to the statistics objectives	Witt (2012)	1
Depositing Big Data	Converting offers and manuscripts into information repositories controlled by librarians	Keil (2014)	1
Sharing Big Data facilities and products	Tracking information via the chain of custody and librarians' manual school on the chain of custody, protection, and copyright problems	Keil (2014)	1
Supporting Big Data users	Providing studies on statistics	Keil (2014)	1
Creating Big Data literacy programs	Big Data training programs	Rani (2016)	1
Understanding the challenges related to Big Data in academic libraries	By identifying	Daniel (2015)	1

Table 5 presents seven themes related to librarian roles in Big Data found from the literature. These are realizing the capabilities of Big Data, using Big Data tools, depositing Big Data, sharing Big Data facilities and products, supporting Big Data management, creating data literacy programs, and understanding the challenges associated with Big Data in academic libraries. Therefore, this study recommends that librarians be trained in their new roles in Big Data services and products.

3.6. WHAT ARE THE GAPS IN BIG DATA STUDIES RELATED TO ACADEMIC LIBRARIES AND FUTURE RESEARCH NEEDS?

This review found that most of the existing studies focused primarily on users' discretionary, followed by public management. Despite the large amount of research conducted on Big Data, only a few studies have discussed the application of Big Data in academic libraries, including the analyzing techniques and tools. Methodologically, such an investigation includes Web-based life examination, content mining, and machine learning applications in order to meet the goal of displaying and storing the network for execution. Nonetheless, it has been further noted that the details included in these investigations were insufficient to describe the tools used for the examination. Thus, to address this gap, this examination took into account the development, types, and uses of Big Data tools in academic libraries. It also presented a concise outline of Big Data advances that empower administrations and some of their applications. The investigation categorized these Big Data tools into various examination stages, databases and data storage media, programming dialects, search devices, and data collection and exchange devices. Finally, based on this survey, future applications for investigation of Big Data have been included in academic libraries (Wilkes, 2012).

4. CONCLUSIONS

Why should academic libraries be inquisitive about Big Data's records, era, impact, and strategies? The answer to this important question is simply that these massive data have the potential to transform an entire business system, which is conceptualized in this study. Due to its overly effective and strategic ability, especially in generating business cost, "massive information" has lately emerged as the focal point of instructional librarians (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015).

Big Data technology makes it easier to deal with massive datasets, link exclusive datasets, find out patterns in real time, predict outcomes, and adopt dynamic risk scoring, and test hypotheses. Both libraries and librarians completely agree to work with big statistics because libraries have a long history of being record managers and era adopters, with large information being no exception. This is the technology of Big Data and the statistics generated in educational libraries is large and complex; therefore, the idea of drawing new and exciting insights from previously unmanageable data is like looking for "a needle in a haystack." Therefore, librarians need to be familiar with the possibilities and challenges inherent in large information, and use that knowledge to help their customers select the right tool.

Moreover, academic librarians have a clear role in Big Data analytics to help researchers and different users enhance the services and quality of education. Similarly, libraries undertake a vital task of converging government, colleges, organizations, and the general population, since they oversee advanced resources. Scientists and users should convert the extensive measure of Big Data and the data in libraries into knowledge or information, in order to make it useful. Nevertheless, librarians may need to envisage how data should be changed, investigated, and presented for the ultimate goal of information creation. For example, they should realize how to make Big Data more helpful, clearly recognizable, and easily available. With the new and incredible examination of Big Data, for example, data representation devices, analysts or clients can look at the data in new ways in order to yield the desired information. Therefore, this study discussed the attributes of Big Data in academic libraries, conducted a survey into the investigation dealing with Big Data in academic libraries, and finally outlined the applications in this field (Kumar & Priyadarsini, 2016).

5. FUTURE RESEARCH DIRECTIONS

As Big Data challenges in academic libraries were less focused in this study, further studies are required to highlight this area of concern. In addition, Big Data techniques based on several educational sociology programs were not taken into account. Big Data can positively enable libraries to make more financially perceptive, imaginative choices or suggestions that can perfectly meet the user's need. The insights into data are expanding rapidly, and more number of analysts wish to generally collect, mine, and sort out the data in new ways. Without Big Data insights, the search for some data may become ineffective. The data collected by library clients to use the administration is exceptionally useful in enhancing the general client experience and fulfilling the users' library benefits. The ability to collect and break down massive amounts of data will be predominant in all organizations, including libraries. Therefore, Big Data may be appropriate to associate with large data or financing. The customary DBMS (database management system) or data examination may also be a predominant methodology. Future studies should examine the real stages or advances made in Big Data library. Possible extensions to the present study include:

- A study of Big Data opportunities and challenges in academic libraries;
- A study of how Big Data is systematically affecting the economic value in academic libraries;
- A proposal for guidelines to librarians on how to develop a system and process in order to use Big Data in academic libraries;
- A study of how academic libraries can benefit from examining social media material to enhance their information services.

6. REFERENCES

Akoka, J., Comyn-wattiau, I., & Laou, N. (2017). Research on Big Data – A systematic mapping study. Computer Standards & Interfaces, 54(Part 2), 105–115. DOI: http://doi.org/10.1016/j.csi.2017.01.004

Al-Daihani, S., & Abrahams, A. (2016). A text mining analysis of academic libraries' tweets. The Journal of Academic Librarianship, 42, 135–143.

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., Yang, G.-Z. (2015). Big Data for health. IEEE Journal of Biomedical and Health Informatics, 19(4), 1193–1208. DOI: 10.1109/JBHI.2015.2450362

Archenaa, J., & Mary Anita, E. A. (2015). A survey of Big Data analytics in healthcare and government. Procedia Computer Science, 50, 408–413.

Begoli, E. & Horey, J. (2012). Design principles for effective knowledge discovery from big data. In Joint ICSA and ECSA, 215–218. Available from: http://www.bdva.eu/sites/default/files/Design%20Principles%20for%20Effective%20Knowledge%20Discovery%20from%20Big%20Data.pdf

Beyer, M. A., & Laney, D. (2012). The importance of 'Big Data': A definition. Stamford, CT: Gartner.

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of Big Data: Four perspectives. SIGMOD, 40(4), 56–60.

Blascheck, T., Burch, M., Raschke, M., & Weiskopf, D. (2015, September). Challenges and perspectives in big eye-movement data visual analytics. In Big Data Visual Analytics (BDVA) (pp. 1–8). IEEE.

Borin, J., & Yi, H. (2008). Indicators for collection evaluation: A new dimensional framework. Collection Building, 27(4), 136–143. DOI: http://dx.doi.org/10.1108/01604950810913698

Borkar, V., Carey, M. J., Li, C. (2012). Inside "Big Data management": Ogres, onions, or parfaits? In Proceeding of EDBT/ICDT joint conference. Berlin: ACM.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301–2309.

Bu, Y., Brokar, V., Carey, M. J., Rosen, J., Polyzotis, N., Condie, T., … Ramakrishnan R. (2012). Scaling datalog for machine learning on Big Data. Computer research repository (CoRR) (pp. 1–14). Cornell University Library. DOI: http://arxiv.org/pdf/1203.0160v2.pdf

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From Big Data to big impact. MISQ, 36(4), 1165–1188.

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

Cuzzocrea, A., Song, I. Y., & Davis, K. (2011). Analytics over large-scale multidimensional data: The Big Data revolution! Proceedings of the 14th international workshop on Data Warehousing and OLAP (pp. 101–103). New York, NY: ACM.

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology, 46(5), 904–920. DOI: 10.1111/bjet.12230

De Mauro, A., Greco, M., & Grimaldi, M. (2015, February). What is Big Data? A consensual definition and a review of key research topics. In G. Giannakopoulos, D. P. Sakas, & D. Kyriaki-Manessi (Eds.), AIP Conference Proceedings (Vol. 1644, No. 1, pp. 97–104). Melville, NY: AIP Publishing.

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

Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013). Addressing Big Data issues in scientific data infrastructure. In International conference on collaboration technologies and systems (CTS). IEEE Computer Society.

Fan, W., & Bifet, A. (2012). Mining Big Data: Current status, and forecast to the future. SIGKDD Explorations, 14(2), 1–5.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with Big Data analytics. Interactions, 19(3), 50–59.

Franceschini, M. (2013). How to maximize the value of Big Data with the open source SpagoBI suite through a comprehensive approach. Proceeding of the VLDB Endowment, 6(11), 1170–1171.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iView, 1–12.

Girija, N., & Srivatsa, S. K. (2006). A research study: Using data mining in knowledge base business strategies. Information Technology Journal, 5(3), 590–600. DOI: http://dx.doi.org/10.3923/itj.2006.590.600

Goes, P. B. (2014). Big Data and IS research methods. MIS Quarterly, 38(3), 3–8.

Gordon-Murnane, L. (2012). Big Data: A big opportunity for librarians. Online, 36(5), 30–34.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Waltham, MA: Elsevier.

Hashem, I. A. T., Yaqoob, I., Badrul Anuar, N., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "Big Data" on cloud computing: Review and open research issues. Information Systems, 47, 98–115.

Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. Journal of Library Administration, 51(7–8), 662–672.

Herodotou, H., Lim, H., & Luo, G. (2011). Starfish: A self-tuning system for Big Data analytics. In Proceeding of the 5th biennial conference on innovative data systems research (CIDR 11) (pp. 261–272).

Isard, M., Budiu, M., Yu, Y., Birrell, A., & Fetterly, D. (2007). Dryad: Distributed data-parallel programs from sequential building blocks. In Proceeding of the 2ndACMSIGOPS/EuroSys European conference on computer systems (pp. 59–72).

Jacobs, A. (2009). The pathologies of Big Data. Communications of the ACM, 52(8), 36.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.

Keil, D. (2014). Research data needs from academic libraries: The perspective of a faculty researcher. Journal of Library Administration, 54(3), 233–240.

Khan, S., Liu, X., Shakil, K. A., & Alam, M. (2017). A survey on scholarly data: From big data perspective. Information Processing & Management, 53(4), 923–944.

Kumar, P., & Priyadarsini, U. (2016). Revealing library statistics with Big Data expertise: A review. International Journal of Pharmacy & Technology, 8(4), 20783–20789. Available from: www.ijptonline.com

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

Larkou, G., Mintzis, M., Andreou, P. G., Konstantinidis, A., & Zeinalipour-yazti, D. (2016). Managing Big Data experiments on smartphones. Distributed and Parallel Databases, 34(1), 33–64. DOI: http://doi.org/10.1007/s10619-014-7158-6

Lomotey, R. K., & Deters, R. (2014). Towards knowledge discovery in Big Data. In Proceeding of the 8th international symposium on service oriented system engineering. IEEE Computer Society (pp. 181–191).

López, V., del Río, S., Benítez, J. M., & Herrera, F. (2014). Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced Big Data. Fuzzy Sets and Systems, 258, 5–38.

Ma, C.-L., Shang, X.-F., & Yuan, Y.-B. (2012). A three-dimensional display for Big Data sets. In International conference on machine learning and cybernetics (ICMLC) (pp. 1541–1545). IEEE Computer Society.

Madden, S. (2012). From databases to Big Data. IEEE Internet Computing, 16(3), 4–6.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big Data: The next frontier for innovation, competition and productivity. New York, NY: McKinsey Global Institute.

Microsoft, (2013). Retrieved from: https://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx

Nabe, J. (2011). Changing the organization of collection development. Collection Management, 36, 3–16. DOI: http://dx.doi.org/10.1080/01462679.2011.529399

Neumeyer, L., Robbins, B., Nair, A., & Kesari A. (2010). S4: Distributed stream computing platform. In Proceeding of the 2010 international conference on data mining workshops (ICDMW). IEEE.

Nicholson, S., & Stanton, J. (2006). Bibliomining for library decision-making. In Encyclopedia of data warehousing and mining (2nd ed., pp. 100–105). Available from: http://www.igi-global.com/chapter/encyclopedia-data-warehousing-mining/10591

Owen, S., Anil, R., Dunning, T., & Friedman, E. (2011). Mahout in action. Greenwich, CT: Manning Publications.

Prakash, K., Chand, P., & Gohel, U. (2004, November). Application of data mining in library and information services. Paper presented at the 2nd Convention PLANNER, Manipur University, Imphal (pp. 168–177). Ahmedabad: INFLIBNET Centre. Available from: http://shodhganga. inflibnet.ac.in/dxml/handle/1944/435

Rani, B. R. (2016, March 9–11). Big Data and Academic Libraries. In International conference on Big Data and knowledge discovery. Indian Statistical Institute.

Reinhalter, L., & Wittmann, R. J. (2014). The library: Big Data's boomtown. The Serials Librarian, 67(4), 363–372. DOI: 10.1080/0361526X.2014.915605

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., & Alor-Hernández, G. (2016). A general perspective of Big Data: Applications, tools, challenges and trends. The Journal of Supercomputing, 72(8), 3073–3113.

Sagiroglu, S., & Sinanc, D. (2013). Big Data: A review. In IEEE international conference on CTS.

Sandhu, G. (2015, January 6–8). Re-envisioning library and information services in the wake of emerging trends and technologies. The 4th international symposium on emerging trends and technologies in libraries and information services, Noida, India (pp. 153–160).

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of Big Data. IBM Global Business Services, —Executive Report.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. The Journal of Academic Librarianship, 41(4), 499–510.

Slavakis, K., Giannakis, G. B., & Mateos, G. (2014). Modeling and optimization for Big Data analytics. IEEE Signal Processing Magazine, 31(5), 18–31.

Stoica, I. (2014, June). Conquering Big Data with spark and BDAS. In Proceeding of the ACM international conference on measurement and modeling of computer systems.

Sumathi, S., & Sivanandam, S. N. (2006). Introduction to data mining and its applications. Berlin: Springer.

Suthaharan, S. (2014). Big Data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70–73. DOI: 10.1145/2627534.2627557

Van Weijen, D. (2012). The language of (future) scientific communication. Research Trends, 31, 7–8.

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

Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of Big Data definitions. Available from: http://arxiv.org/pdf/1309.5821v1.pdf

Wilkes, S. (2012). Some impacts of big data on usability practice. Communication Design Quarterly Review, 13(2), 25–32.

Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional repository service. Journal of Library Administration, 52(2), 172–188.

Wu, X., Zhu, X., Wu, G.-Q., Ding, W. (2014). Data mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97–107.

Yu, Y., Isard, M., Fetterly, D., Budiu, M., Erlingsson, Ú., Gunda, P. K., & Achan, K. (2008). DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language. In Proceeding of the 8th USENIX conference on operating systems design and implementation (pp. 1–14).

Zhifeng, X., & Yang, X. (2013). Security and privacy in cloud computing. IEEE Communications Surveys and Tutorials, 15(2), 843–859.