Robotic Technology, Artificial Intelligence and the 4IR Libraries

Edited by
Adeyinka Tella and Omwoyo Bosire Onyancha

DE GRUYTER
SAUR

ISBN 978-3-11-117461-7
e-ISBN (PDF) 978-3-11-117698-7
e-ISBN (EPUB) 978-3-11-117754-0
ISSN 2191-2742

Library of Congress Control Number: 2024944104

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

© 2025 Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com

Contents

The Authors —— IX

Lorette Jacobs
Chapter 1
Internet of Things (IoT)-based Smart Libraries: Opportunities and Advances —— 1

Amos Orenyi Bajeh, Adeyinka Tella
Chapter 2
Internet of (Robotic) Things in Library System Automation —— 21

Tom Kwanya
Chapter 3
Enhancing the Findability of Library Information Resources Using the Internet of Things: A Kenyan Perspective —— 53

Olaronke O. Fagbola
Chapter 4
The Prospect of Blockchain Technology Adoption in Academic Libraries: An Overview —— 73

Rhodes Elias Mwageni and Robert Stalone Buwule
Chapter 5
The Assumptions of the Use of Artificial Intelligence Technologies in Robotics in Inclusively Engaged Academic Libraries in Sub Saharan Africa: A Literature Survey —— 95

Airen Edale Adetimirin and Oyedele Abimbola Ojeniyi
Chapter 6
Building Future Information Professionals for 4IR Libraries: Role of Library and Information Science Lecturers in Nigeria —— 111

Temidayo Oluwatosin Omotehinwa
Chapter 7
Artificial Intelligence and Libraries: The Possibilities in the 4IR —— 131

Robert Stalone Buwule and Stephen Mutula
Chapter 8
Inclination of Uganda University Libraries on Co-Driving the Fourth Industrial Revolution — 145

Adeyinka Tella, Joseph. M. Ngoaketsi and Oluwatoyin Oluwayemi Obinyan
Chapter 9
Legal and Ethical Issues Associated with Adoption and Utilisation of Cloud Computing in Academic Libraries in Nigeria — 165

Wilson Chukwunedum Ochonogor and Edwin Iroroeavwo Achugbue
Chapter 10
The Relevance of Cloud Computing and the Internet of Things for Effective Information Service Delivery in Medical Libraries in Nigeria — 197

Aderinola Ololade Dunmade, Uloma Doris Onuoha, and Adeyinka Tella
Chapter 11
Cyberethics and Cybersecurity in Universities in Developing Countries: A Literature Survey and Course Content Analysis from Library Instruction's Perspective — 215

Musediq Tunji Bashorun, Kabir Muhammad Abubakar, Iro Sani Lawal, and Rafiat Bukola Bashorun
Chapter 12
Application of Big Data and Advanced Analytics in Future Librarianship — 227

John Jackson Iwata
Chapter 13
Libraries in the Wake of the Fourth Industrial Revolution: A Portrait of Roles, Challenges, and the Prospects — 249

Collence Takaingenhamo Chisita, Rexwhite Tega Enakrire, and Blessing Chiparausha
Chapter 14
The Changing Role of Librarians in the Age of Hypermedia Seduction: Reflection on the Researchers' Exposition — 267

Musediq Tunji Bashorun, Kabir Muhammad Abubakar,
Iro Sani Lawal, and Rafiat Bukola Bashorun

Chapter 12
Application of Big Data and Advanced Analytics in Future Librarianship

Abstract: Big Data is increasingly getting more attention in all sectors of organizations; in business, academia, and research including libraries. Big Data is a by-product of the information explosion. Due to advancements in technologies, modern life produces data at an amazing rate which lead to advances in data storage and analysis and the arrival of the concept of Big Data. The chapter briefly describes the concept of big data, the characteristics of big data, and the history of Big Data. Furthermore, the chapter examines the differences between traditional data and Big Data. It also covers Big Data technologies and the relevance of Big Data to all sectors. It covers Big Data analytics and applications of big data in the libraries. Also, it highlights the benefits of Big Data in various sectors including libraries. The chapter covers Big Data analysis techniques. Moreover, the chapter identifies challenges facing big data and advanced data analytics. It is concluded that Big Data technology has huge benefits to various sectors, especially libraries, and its adoption has come to stay for the effective and efficient utilization of information for the benefit of information users.

Keywords: Analytics, Big Data, Libraries, Traditional Data, Users

Introduction

The era of information explosion is now, with Big Data today becoming a catchphrase in every area of the global economy. The digital era of information society evolved with data, information, and knowledge creation. Its information explosion stage transformed the world information organization into data mining, with metadata, semantic web, webometrics, ontology, and so on (Witt et al., 2018). However, Big Data is a term for a "collection of data sets so huge and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"(Oguntimilehin & Ademola, 2014, p. 433). There is no universally recognized operational definition of Big Data According to Gordon-Murnane (2012), Big Data is not just about gigabytes of data but immeasurably increased. International Business Machines (IBM) reported that "2.5 quintillion bytes of data is

https://doi.org/10.1515/9783111176987-012

created on a daily basis." Therefore, Big Data could be defined as the new technologies and architectures designed contain huge amounts of structured, semi-structured, and unstructured types of data and are difficult to process using traditional databases and other software.

Commercial and Industrial Security Corporation (CISCO) (2012) predicts that, between 2015 and 2016, global IP traffic will grow to more than 330 Exabytes. It can be deduced that Big Data is loosely structured data that is often incomplete and inaccessible which may consist of billions to trillions of records of millions of people, all from different sources, for example the Web, sales, customer contact centers, social media, mobile data, and so on. That is almost more data than the two previous years combined. Gordon-Murnane (2012, p. 30) reported that McKinsey Global Institute's statement indicated that one Exabyte of data is the equivalent of more than 4,000 times the information stored in the US Library of Congress. The starting of terabytes may be considered as the start of Big Data.

These show how big Big Data is. Big Data has become a major issue of concern in all types of organizations whether business or service, private sector or public sector, including libraries and information centers. It can be inferred that the Big Data effect is being felt everywhere, from business to science, from the government to the arts as supported by Cukier (2010). Due to the "arrival of the Big Data concept, decisions that are previously based on guesswork or painstakingly constructed models of reality can now be made based on data itself" (Oguntimilehin & Ademola, 2014, p. 434). The exponential increase in Big Data continuously challenges the storage industry. However, physical storage could be enhanced with more terabyte drive arrays and the software infrastructure could be made more flexible to quickly and economically accommodate greater volumes of transactions and queries (Pan et al., 2018). Some outstanding achievements of Big Data are the Los Angeles Police Department and the University of California using Big Data to predict crime before it even happens and Google Flu trends using search terms to predict the spread of the flu virus. Also, Statisticians Nate Silver predicted the outcome of the US election down to each individual state in 2012, and MIT is using mobile phone data to establish how people's location and traffic patterns can be used for urban planning (Oguntimilehin & Ademola 2014, p. 435).

Big data analytics is the process of breaking down data sets into smaller chunks containing a variety of data types for efficient processing to uncover hidden patterns, market trends, customer preferences, unknown co-relations, and useful business information (Bhadani et al., 2016). Big Data analysis aims to simplify the character and scope of growing knowledge in different fields of science;

especially multidisciplinary signal processing approaches (Slavakis, Kim., Mateos & Giannakis, 2014).

Big Data is a big deal and it is a huge opportunity for librarians and information professionals to play a key role in the big data universe. Librarianship is a profession concerned with acquiring, organizing, and disseminating collections of books and related materials in libraries and servicing readers and others with these resources. These activities have to be carried out by an expert referred to as a librarian. A librarian is a person who works professionally in a library, providing access to information, social, and technical programming required by users to satisfy their information needs. The future starts from now, and librarians must abide by the current trend of information explosion by adopting the effective utilization of huge data tagged Big Data. Therefore, the paper describes the application of big data and advanced analytics in future librarianship.

History of Big Data

The term data was first used in 1941 in the Oxford English Dictionary. Thereafter, researchers and librarians frequently began to write on information and data explosions. Fremont Rider wrote a book in 1944 through Wesleyan University Librarian titled *"The scholar and the future of the research library"*. The author estimated that American universities were doubling in size every 16 years. Based on this growth rate, Rider had speculated that the Yale Library in 2040 would have "approximately 200,000,000 volumes which will occupy over 6,000miles of shelve and require a cataloging staff of over six thousand persons (Bhoola et al., 2014). In 1961 Derek Price published *Science since Babylon*, in which he charted the growth of scientific knowledge by looking at the increased numbers of scientific journals. The papers were shown to grow exponentially, doubling every 15 years and increasing by a factor of ten over half a century (Alvarado, 2016). Derek Price calls this the "law of exponential increase." This assertion suggests that there isa rapid increase in the number of scientific journals that leads to exponential growth.

In July 1986, Becker estimated that "the recording density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch – 500times the density of (4,000 B.C. Sumerian) clay tablets." In 1996, storage systems had evolved to meet the changing customer needs and digital storage became more popular. This leads to the development of autonomous storage (Morris & Truskowski, 2003). In October 2000, Lyman and Varian carried out the first comprehensive study at the University of California to quantify the total amount of new and original infor-

mation in computer storage terms and established that the world produced between one and two Exabytes of unique information per year. This is roughly 250 megabytes for every man, woman, and child on earth and digital content was dominant.

In February 2001, Doug Laney, an analyst with the Meta Group, published a research note entitled "3D Data Management Controlling Data Volume, Velocity, and Variety". A decade later, the "3Vs" has generally been accepted as the three defining dimensions of Big Data (Oussous et al., 2018; Koga, 2015). In September 2005, Tim O'Reilly published "What is Web 2.0"and he asserted that "data is the next Intel inside." In January 2008, Bret Swanson and George Gilder published "Estimating the Exaflood (PDF)" and projected that U.S IP traffic could reach one Zettabyte by 2015 and that the US Internet of 2015 would be at least 50 times larger than it was in 2006. Kenneth Cukier in February 2010 published an Economist Special Report titled "Data, Data everywhere" and wrote, "the world contains an unimaginably vast amount of digital information which is getting ever vaster more rapidly (Anderson &Rainie, 2018).

In May 2011, Manyika et al. (2011) carried out a study and estimated that in 2010 7.4Exabytes of new data were stored by enterprises and 6.8 Exabytes of new data were stored by consumers (on their PCs and other devices). It can be deduced that due to the widespread accessibility, affordability, and availability of new digital devices and easy access to the Internet, there has been a data explosion. Every day, we create 2.5quitillion bytes of data, so much that 90% of the data in the world today has been created in the last two years alone. The data comes from everywhere sensors used to gather climate information, posts on social media sites, digital pictures, and videos. Also, in purchasing transaction records and cell phone GPS signals, to name a few, such data is Big Data (IBM Newsletter, 2012). Social media is generating petabytes of data per day. Syed (2015) reported that Joe Hellerstein, a computer scientist at the University of California in Berkeley, calls this age the industrial revolution of data. This is the stage of the arrival of Big Data with its various characteristics.

Characteristics of Big Data

The term Big Data covers every piece of data generated from organizations that have been stored until now. This includes data stored in clouds and even in the URLs that you bookmark. According to Olendorf and Wang (2017, p. 192), Big Data refers to any data that exceeds the user's ability to analyze it in one of three dimensions (Volume, Velocity, and Variety). Big Data is characterized by the three

Vs' high volume, high velocity, and high variety (Sonawane, 2018; Bieraugel, 2015). Furthermore, Lomotey and Deters (2014) extended the model defined by Laney (2001) to a 5Vs model (volume, veracity, velocity, value, and variety).

Moreover, several studies (Akoka, Comyn-Wattiau, & Laoufi, 2017; Khan et al., 2017) were carried out and the findings revealed more features that extended the model from a 5Vs to a 6Vs model. Big Data is now characterized by 6Vs (volume, velocity, variety, value, viability, and veracity). The six (6) characteristics of Big Data are shown in Figure 1.

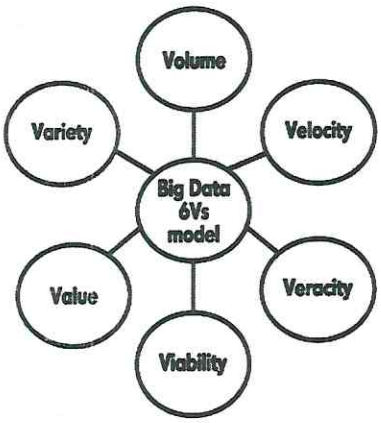


Figure 1: The 6Vs model that currently defines Big Data.
Source: adapted from (Ristevski & Chen, 2018).

Volume: Big Data is huge; it is machine-generated data produced in much larger quantities than non-traditional data depending on industry or discipline. This data is loosely defined data that cannot be stored or analyzed by conventional hardware and software. Traditional software can handle megabyte and kilobyte-sized data sets, while big data tools can handle terabyte and petabyte-sized data sets. This is echoed by Manyika et al. (2011), who assert that Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Bieraugel, 2015). With the application of the Internet, there is an enormous growth of information leading to information overload (explosion). This has led to an ever-increasing userbase and increased demand fora huge amount of data that requires extra attention.

Velocity: This is data created from a wide array of data platforms: for an instant, the speed at which someone can create a single tweet on Twitter, or post to Facebook (Bieraugel, 2015). Velocity could be described as the speed at which data is processed and becomes accessible to meet demands and challenges that lie in the path of development and growth. It is generally created by social media platforms

where a large influx of opinion and relationships management records is created. It is high frequency, for example Twitter data, posts and likes on Facebook, and many others. For time-sensitive processes such as detecting fraud, Big Data must be used as it streams into enterprises to maximize its value (IBM Newsletter, 2012). The same velocity issue prevails in any form of the library by creating a large volume of information from large collections at high speed. Increasingly modern libraries are using social media platforms and other digital web-based information resources such as search engines and subject portals, institutional repositories, and open access journals which generate large amounts of data. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is to identify what is valuable and then transform and extract that data for analysis.

Variety: This could be described as all structured and unstructured data that has the possibility of being generated, either by humans or by machines. It is all about the ability to group the incoming data into different categories. In addition to the velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Data comes in all types of formats from structured, numeric data in traditional databases to unstructured text documents: email, video, audio, stock ticker data, and financial transactions. Variety makes big data sets more challenging to organize and analyze. Big Data sets can contain unstructured data such as email messages, photographs, postings on internet forums, and even phone transcripts (Bieraugel, 2015). In the twenty-first century, libraries are overwhelmed with a variety of data sets, both structured and unstructured (De Mauro, Greco, & Grimaldi, 2016). Structured data may include e-books, institutional repositories, open access journals, and audio-video while unstructured data may contain emails, postings on social media platforms, metadata, and others (Power, 2014). This variety of unstructured data creates problems for storage, mining, and analyzing of data (Normandeau, 2013). There is a diversity of information retrieved from various sources with diverse forms and formats.

Value: This refers to the usefulness of gathered data for a specific task. It can also be described as the value that Big Data can provide to any organization. According to Cartledge (2016), it is the source of Big Data, whether it is reliable, accurate, and complete. It enables and empowers users to be able to convert information into something valuable and useful.

Viability: This can be described as a careful selection of those attributes in the data that are most likely to predict outcomes that are germane to organization. With so many varieties in data and variables to consider, it is important to assess the viability of data to build an effective predictable model (Biehn, 2013) and also to expand the model.

Veracity: This has to do with how accurate and truthful a data set may be. Big Data generates a lot of data and a variety of data. However, the data must have quality and produce credible results that enable action to produce the right decision-making. The veracity determines the accuracy of the data (Leonelli & Tempini, 2021; Scott, 2015: Cai & Yangyong, 2015).The assertions suggest that the veracity of Big Data is not just the quality of the data itself but the reliability of the data source, type, and processing. In library settings, the information generated must be authentic, credible, and accurate information that reduces inconsistencies and uncertainty.

Today's data comes from multiple sources, which makes it difficult to link, match, cleanse, and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies, and multiple data linkages or else your data can quickly spiral out of control. IBM believes that Big Data exploration, an enhanced 360-degree view of the customer, security/intelligence extension, operational analysis, and data house augmentation are the ways to get started with big data (Jewell et al., 2014). The following steps are for the handling of Big Data:

1. Understand the Big Data sources for your business, i.e. streaming data (data generated by your IT system from the web), social media data, and publicly available sources.
2. Query/ consider the decisions you need for analysis. It will include how to store and manage the data, how much of it to analyze, and how to form strategies or insights that are uncovered or covered.
3. Do research for technologies that manage Big Data and its analytics (cheap for abundant storage, faster processor, affordable open-source, distributed big data platforms, high connectivity, parallel processing, cloud computing, and others).

Differences Between Traditional Data and Big Data

Big Data is a huge amount of structured, semi-structured, and unstructured types of data that is so large that it's difficult to process using traditional databases and other software. The traditional database could be described as a centralized database architecture used to store and maintain the data in a fixed format or fields in a file. In addition, traditional data means structured data which is being majorly preserved by all types of businesses starting from very small to big organizations.

There is a need to differentiate between traditional data and Big Data.

S/N	Traditional Data	Big Data
1	Data generated atthe enterprise level	Data generated at enterprise-level and outside level
2	Data stored in the form of structured and unstructured	Data is stored in the form of structured, semi-structured, and unstructured
3	Data manage in centralized form	Data manage in a physical distributed form
4	Specialized high-level software as well as the hardware used	Inexpensive commodity boxes in cluster mode
5	Data stored in gigabytes or terabytes	Data stored in Petabytes, Zettabytes, Exabytes
6	By default, it has a stable and interrelationship	Unknown relationship
7	Traditional data sources include ERP transaction data, CRM data, Web transaction data, financial data	Non-traditional data sources include social media, log data, device data, sensor data, video images

In addition, examples of sources where Big Data could be generated are:

1. YouTube users upload 72 hours of video each minute
2. The average numbers of tweets per day is 58 million
3. Every 60 seconds there are 695,000 status updates, 79,364 wall posts and 510,040 comments published on the social networking site Facebook.

All these activities can only be achieved with the help of a set of technologies.

Big Data Technologies

Big Data not only refers to the data itself but also a set of technologies that capture, store, manage, and analyze large and variable collections of data to solve complex problems. Big Data takes too much time and costs too much money to load into traditional relational databases for analysis. Big Data analytics is often associated with cloud computing because it encompasses the analysis of large data sets in real-time where a massive volume of both structured and unstructured data is analyzed using different techniques (Sivarajah et al., 2017)

Big Data analysis is not a typical database management system; it requires high-level skills like data analysis, data warehousing, data transformation, and data collection. Several 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).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 et al., 2015; Ward & Barker, 2013) as well as many others as shown below:

1. **Hadoop:** This is open-source software that is made to handle big data. Hadoop was developed by Daug Cutting in 2005 and first released in 2011 to support distribution for the Nutch search engine project at Yahoo (Bhathal & Amardeep, 2019). Itis a device that has grown quickly to be the gold standard in the industry. Hadoop is the most popular source platform for handling Big Data. It can do large-scale processing and is flexible enough to work with multiple data sources (Azeroual & Fabre, 2021). It has several applications such as data from weather and traffic sensors webs, machine to machine transactional data (Jiang & Luo, 2022).
2. **MapReduce:** This is a programming model comprised of a map procedure method that filters and performs sorting and reduces methods that perform the summary operation. The key contribution of this method is the scalability and fault tolerance achieved for a variety of applications. The in-depth MapReduce platform is now broadly used for clickstream, web indexing and log evaluation, and positive large-scale fact extraction and device mastering functions (Borkar, Carey, & Li, 2012).
3. **Hive:** It is an open-source data warehouse system used for querying and analyzing large datasets that provides data summarization. It is primarily used in data mining applications which otherwise can take several hours to analyze (Dave, 2013).
4. **PIG:** It was developed at Yahoo and can handle and analyze large data sets, and takes less time than a mapper and reducer program. It has two components: a language and the second one called runtime environment where the programs are run.
5. **Cloud Computing:** Cloud computing and Big Data are an ideal combination as it provides the cost-effective and scalable infrastructure to support big data analytics (Ferkoun, 2014).

In addition, there are other Big Data tools as reiterated by Sengupta (2016, p. 91) which have been published for clarity purposes. They are:

6. **SAP in Memory:** SAP's HANA platform offers several benefits over some of the competition, such as the ability to integrate and analyze huge workloads of data to be analyzed in real-time. This is extremely useful for the developer who is looking for speed to market (Darvazeh et al., 2020)

7. **Cambridge Semantics:** Using the Anzo Software Suite, this open platform helps you to collect, integrate, and analyze Big Data to assist in building the Unified Access solutions (Sengupta, 2016). The software has a data integration machine that streamlines data collection and helps with analytics. The key features include being able to combine data from multiple sources and customized dashboards to make analysis easy.
8. **MongoDB:** This is an open-source documental database that is useful to developers who want to have precise control over the final outputs. MongoDB is also scalable and includes third party log tools such as Edda and Fluentd (Sengupta, 2016; p. 91).
9. **Pentaho:** Pentaho joins data integration and business analytics for visualizing, analyzing, and blending Big Data. The open and embeddable platform comes with extensive analytics capabilities with data mining and predictive analysis. The data integration tools mean that users do not require coding in SQL or writing MapReduce Java functions.
10. **Talend:** Straight away, one of the key benefits of Talend's Open Studio is that it is open-source, which means that improvements will keep on rolling out as the community tweaks the tool (Sengupta, 2016). Its tools include products for developing, testing, and deploying data management and application integration products. Additionally, the company manages the full lifecycle, even across enterprise boundaries.
11. **Tableau:** It is one of the most recognized names in the data visualization sphere but it offers many tools for developers that are supported by an active community. Some of the key characteristics of this software are its in-memory analytics database and advanced query language. API, XML, User Scripts, Python, and JavaScript are all supported and so are several browser extensions.
12. **Splunk:** Splunk specifically specializes in harnessing machine data created from several various sources, such as applications, websites, and sensors. The company also enables developers to write code using any technology platform, language or framework.
13. **Splice Machine:** This is a real-time SQL-on-Hadoop database that can help to derive real-time actionable insights, which is a clear benefit for those who are aiming for quick development (Olabode et al., 2022). It is SQL 99 compliant with standard ANSI SQL and can scale from gigabytes to petabytes. As well as support for .NET, Java, and Python, it also offers support for those written in JavaScript/AngularJS.
14. **MarkLogic:** It is built to deal with heavy data loads and allow users to access it through real-time updates and alerts. It provides geographical data that is

combined with content and location relevance along with data filtering tools. This tool is ideal for those looking at paid content search app development.

15. **Google Charts:** This free tool comes with various capabilities for visualizing data from a website such as hierarchical treemaps or just simple charts. This tool is easily implemented by embedding JavaScript code on a website and allows you to sort, modify, and filter data as well as the ability to connect to a database or pull data from a website. Offering support for popular languages and with the security of knowing that Google will likely keep on improving its offering, this is a good option for many standard developers.

Drawing from the aforementioned tools, it becomes clear that Big Data has enormous benefits accrued to all sectors, especially within libraries.

Relevance of Big Data

The use of Big Data is becoming crucial and the following are the importance of big data (Mcguire et al., 2012).

1. Big Data allows narrower segmentation of customers and this helps in producing more precisely tailored products and services.
2. Sophisticated analytics are useful to policymakers because they would help to improve decision-making.
3. Future generation products and services can be developed with the help of Big Data.
4. There is a significant amount of data that is yet not captured in digital form or made accessible or searchable through the network. Big Data can unlock the value hidden in these forms of data by making information clear.

Big Data is useful in library settings because librarians acquire, process, and store huge amounts of information to satisfy information users. The librarianship profession must strive to meet the needs of all categories of users by embracing and using Big Data.

Big Data Practices in Librarianship

The catchphrase "Big Data" is coming faster and becoming vibrant in business, the techworld, and higher education. The act of gathering and storing large amounts of information for eventual analysis is old. Librarians have always been

Musediq Tunji Bashorun, Kabir Muhammad Abubakar,
Iro Sani Lawal, and Rafiat Bukola Bashorun

Chapter 12
Application of Big Data and Advanced Analytics in Future Librarianship

Abstract: Big Data is increasingly getting more attention in all sectors of organizations; in business, academia, and research including libraries. Big Data is a by-product of the information explosion. Due to advancements in technologies, modern life produces data at an amazing rate which lead to advances in data storage and analysis and the arrival of the concept of Big Data. The chapter briefly describes the concept of big data, the characteristics of big data, and the history of Big Data. Furthermore, the chapter examines the differences between traditional data and Big Data. It also covers Big Data technologies and the relevance of Big Data to all sectors. It covers Big Data analytics and applications of big data in the libraries. Also, it highlights the benefits of Big Data in various sectors including libraries. The chapter covers Big Data analysis techniques. Moreover, the chapter identifies challenges facing big data and advanced data analytics. It is concluded that Big Data technology has huge benefits to various sectors, especially libraries, and its adoption has come to stay for the effective and efficient utilization of information for the benefit of information users.

Keywords: Analytics, Big Data, Libraries, Traditional Data, Users

Introduction

The era of information explosion is now, with Big Data today becoming a catchphrase in every area of the global economy. The digital era of information society evolved with data, information, and knowledge creation. Its information explosion stage transformed the world information organization into data mining, with metadata, semantic web, webometrics, ontology, and so on (Witt et al., 2018). However, Big Data is a term for a "collection of data sets so huge and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"(Oguntimilehin & Ademola, 2014, p. 433). There is no universally recognized operational definition of Big Data According to Gordon-Murnane (2012), Big Data is not just about gigabytes of data but immeasurably increased. International Business Machines (IBM) reported that "2.5 quintillion bytes of data is

https://doi.org/10.1515/9783111176987-012

created on a daily basis." Therefore, Big Data could be defined as the new technologies and architectures designed contain huge amounts of structured, semi-structured, and unstructured types of data and are difficult to process using traditional databases and other software.

Commercial and Industrial Security Corporation (CISCO) (2012) predicts that, between 2015 and 2016, global IP traffic will grow to more than 330 Exabytes. It can be deduced that Big Data is loosely structured data that is often incomplete and inaccessible which may consist of billions to trillions of records of millions of people, all from different sources, for example the Web, sales, customer contact centers, social media, mobile data, and so on. That is almost more data than the two previous years combined. Gordon-Murnane (2012, p. 30) reported that McKinsey Global Institute's statement indicated that one Exabyte of data is the equivalent of more than 4,000 times the information stored in the US Library of Congress. The starting of terabytes may be considered as the start of Big Data.

These show how big Big Data is. Big Data has become a major issue of concern in all types of organizations whether business or service, private sector or public sector, including libraries and information centers. It can be inferred that the Big Data effect is being felt everywhere, from business to science, from the government to the arts as supported by Cukier (2010). Due to the "arrival of the Big Data concept, decisions that are previously based on guesswork or painstakingly constructed models of reality can now be made based on data itself" (Oguntimilehin & Ademola, 2014, p. 434). The exponential increase in Big Data continuously challenges the storage industry. However, physical storage could be enhanced with more terabyte drive arrays and the software infrastructure could be made more flexible to quickly and economically accommodate greater volumes of transactions and queries (Pan et al., 2018). Some outstanding achievements of Big Data are the Los Angeles Police Department and the University of California using Big Data to predict crime before it even happens and Google Flu trends using search terms to predict the spread of the flu virus. Also, Statisticians Nate Silver predicted the outcome of the US election down to each individual state in 2012, and MIT is using mobile phone data to establish how people's location and traffic patterns can be used for urban planning (Oguntimilehin & Ademola 2014, p. 435).

Big data analytics is the process of breaking down data sets into smaller chunks containing a variety of data types for efficient processing to uncover hidden patterns, market trends, customer preferences, unknown co-relations, and useful business information (Bhadani et al., 2016). Big Data analysis aims to simplify the character and scope of growing knowledge in different fields of science;

especially multidisciplinary signal processing approaches (Slavakis, Kim., Mateos & Giannakis, 2014).

Big Data is a big deal and it is a huge opportunity for librarians and information professionals to play a key role in the big data universe. Librarianship is a profession concerned with acquiring, organizing, and disseminating collections of books and related materials in libraries and servicing readers and others with these resources. These activities have to be carried out by an expert referred to as a librarian. A librarian is a person who works professionally in a library, providing access to information, social, and technical programming required by users to satisfy their information needs. The future starts from now, and librarians must abide by the current trend of information explosion by adopting the effective utilization of huge data tagged Big Data. Therefore, the paper describes the application of big data and advanced analytics in future librarianship.

History of Big Data

The term data was first used in 1941 in the Oxford English Dictionary. Thereafter, researchers and librarians frequently began to write on information and data explosions. Fremont Rider wrote a book in 1944 through Wesleyan University Librarian titled *"The scholar and the future of the research library"*. The author estimated that American universities were doubling in size every 16 years. Based on this growth rate, Rider had speculated that the Yale Library in 2040 would have "approximately 200,000,000 volumes which will occupy over 6,000miles of shelve and require a cataloging staff of over six thousand persons (Bhoola et al., 2014). In 1961 Derek Price published *Science since Babylon*, in which he charted the growth of scientific knowledge by looking at the increased numbers of scientific journals. The papers were shown to grow exponentially, doubling every 15 years and increasing by a factor of ten over half a century (Alvarado, 2016). Derek Price calls this the "law of exponential increase." This assertion suggests that there isa rapid increase in the number of scientific journals that leads to exponential growth.

In July 1986, Becker estimated that "the recording density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch – 500times the density of (4,000 B.C. Sumerian) clay tablets." In 1996, storage systems had evolved to meet the changing customer needs and digital storage became more popular. This leads to the development of autonomous storage (Morris & Truskowski, 2003). In October 2000, Lyman and Varian carried out the first comprehensive study at the University of California to quantify the total amount of new and original infor-

mation in computer storage terms and established that the world produced between one and two Exabytes of unique information per year. This is roughly 250 megabytes for every man, woman, and child on earth and digital content was dominant.

In February 2001, Doug Laney, an analyst with the Meta Group, published a research note entitled "3D Data Management Controlling Data Volume, Velocity, and Variety". A decade later, the "3Vs" has generally been accepted as the three defining dimensions of Big Data (Oussous et al., 2018; Koga, 2015). In September 2005, Tim O'Reilly published "What is Web 2.0"and he asserted that "data is the next Intel inside." In January 2008, Bret Swanson and George Gilder published "Estimating the Exaflood (PDF)" and projected that U.S IP traffic could reach one Zettabyte by 2015 and that the US Internet of 2015 would be at least 50 times larger than it was in 2006. Kenneth Cukier in February 2010 published an Economist Special Report titled "Data, Data everywhere" and wrote, "the world contains an unimaginably vast amount of digital information which is getting ever vaster more rapidly (Anderson &Rainie, 2018).

In May 2011, Manyika et al. (2011) carried out a study and estimated that in 2010 7.4Exabytes of new data were stored by enterprises and 6.8 Exabytes of new data were stored by consumers (on their PCs and other devices). It can be deduced that due to the widespread accessibility, affordability, and availability of new digital devices and easy access to the Internet, there has been a data explosion. Every day, we create 2.5quitillion bytes of data, so much that 90% of the data in the world today has been created in the last two years alone. The data comes from everywhere sensors used to gather climate information, posts on social media sites, digital pictures, and videos. Also, in purchasing transaction records and cell phone GPS signals, to name a few, such data is Big Data (IBM Newsletter, 2012). Social media is generating petabytes of data per day. Syed (2015) reported that Joe Hellerstein, a computer scientist at the University of California in Berkeley, calls this age the industrial revolution of data. This is the stage of the arrival of Big Data with its various characteristics.

Characteristics of Big Data

The term Big Data covers every piece of data generated from organizations that have been stored until now. This includes data stored in clouds and even in the URLs that you bookmark. According to Olendorf and Wang (2017, p. 192), Big Data refers to any data that exceeds the user's ability to analyze it in one of three dimensions (Volume, Velocity, and Variety). Big Data is characterized by the three

Vs' high volume, high velocity, and high variety (Sonawane, 2018; Bieraugel, 2015). Furthermore, Lomotey and Deters (2014) extended the model defined by Laney (2001) to a 5Vs model (volume, veracity, velocity, value, and variety).

Moreover, several studies (Akoka, Comyn-Wattiau, & Laoufi, 2017; Khan et al., 2017) were carried out and the findings revealed more features that extended the model from a 5Vs to a 6Vs model. Big Data is now characterized by 6Vs (volume, velocity, variety, value, viability, and veracity). The six (6) characteristics of Big Data are shown in Figure 1.

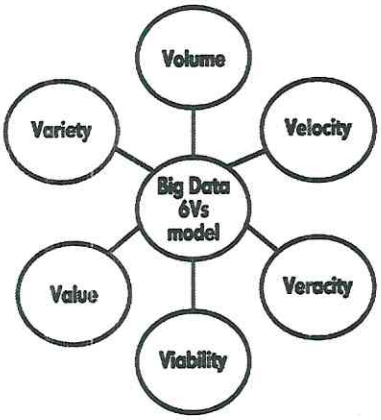


Figure 1: The 6Vs model that currently defines Big Data.
Source: adapted from (Ristevski & Chen, 2018).

Volume: Big Data is huge; it is machine-generated data produced in much larger quantities than non-traditional data depending on industry or discipline. This data is loosely defined data that cannot be stored or analyzed by conventional hardware and software. Traditional software can handle megabyte and kilobyte-sized data sets, while big data tools can handle terabyte and petabyte-sized data sets. This is echoed by Manyika et al. (2011), who assert that Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Bieraugel, 2015). With the application of the Internet, there is an enormous growth of information leading to information overload (explosion). This has led to an ever-increasing userbase and increased demand fora huge amount of data that requires extra attention.

Velocity: This is data created from a wide array of data platforms: for an instant, the speed at which someone can create a single tweet on Twitter, or post to Facebook (Bieraugel, 2015). Velocity could be described as the speed at which data is processed and becomes accessible to meet demands and challenges that lie in the path of development and growth. It is generally created by social media platforms

where a large influx of opinion and relationships management records is created. It is high frequency, for example Twitter data, posts and likes on Facebook, and many others. For time-sensitive processes such as detecting fraud, Big Data must be used as it streams into enterprises to maximize its value (IBM Newsletter, 2012). The same velocity issue prevails in any form of the library by creating a large volume of information from large collections at high speed. Increasingly modern libraries are using social media platforms and other digital web-based information resources such as search engines and subject portals, institutional repositories, and open access journals which generate large amounts of data. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is to identify what is valuable and then transform and extract that data for analysis.

Variety: This could be described as all structured and unstructured data that has the possibility of being generated, either by humans or by machines. It is all about the ability to group the incoming data into different categories. In addition to the velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Data comes in all types of formats from structured, numeric data in traditional databases to unstructured text documents: email, video, audio, stock ticker data, and financial transactions. Variety makes big data sets more challenging to organize and analyze. Big Data sets can contain unstructured data such as email messages, photographs, postings on internet forums, and even phone transcripts (Bieraugel, 2015). In the twenty-first century, libraries are overwhelmed with a variety of data sets, both structured and unstructured (De Mauro, Greco, & Grimaldi, 2016). Structured data may include e-books, institutional repositories, open access journals, and audio-video while unstructured data may contain emails, postings on social media platforms, metadata, and others (Power, 2014). This variety of unstructured data creates problems for storage, mining, and analyzing of data (Normandeau, 2013). There is a diversity of information retrieved from various sources with diverse forms and formats.

Value: This refers to the usefulness of gathered data for a specific task. It can also be described as the value that Big Data can provide to any organization. According to Cartledge (2016), it is the source of Big Data, whether it is reliable, accurate, and complete. It enables and empowers users to be able to convert information into something valuable and useful.

Viability: This can be described as a careful selection of those attributes in the data that are most likely to predict outcomes that are germane to organization. With so many varieties in data and variables to consider, it is important to assess the viability of data to build an effective predictable model (Biehn, 2013) and also to expand the model.

Veracity: This has to do with how accurate and truthful a data set may be. Big Data generates a lot of data and a variety of data. However, the data must have quality and produce credible results that enable action to produce the right decision-making. The veracity determines the accuracy of the data (Leonelli & Tempini, 2021; Scott, 2015: Cai & Yangyong, 2015).The assertions suggest that the veracity of Big Data is not just the quality of the data itself but the reliability of the data source, type, and processing. In library settings, the information generated must be authentic, credible, and accurate information that reduces inconsistencies and uncertainty.

Today's data comes from multiple sources, which makes it difficult to link, match, cleanse, and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies, and multiple data linkages or else your data can quickly spiral out of control. IBM believes that Big Data exploration, an enhanced 360-degree view of the customer, security/intelligence extension, operational analysis, and data house augmentation are the ways to get started with big data (Jewell et al., 2014). The following steps are for the handling of Big Data:

1. Understand the Big Data sources for your business, i.e. streaming data (data generated by your IT system from the web), social media data, and publicly available sources.
2. Query/ consider the decisions you need for analysis. It will include how to store and manage the data, how much of it to analyze, and how to form strategies or insights that are uncovered or covered.
3. Do research for technologies that manage Big Data and its analytics (cheap for abundant storage, faster processor, affordable open-source, distributed big data platforms, high connectivity, parallel processing, cloud computing, and others).

Differences Between Traditional Data and Big Data

Big Data is a huge amount of structured, semi-structured, and unstructured types of data that is so large that it's difficult to process using traditional databases and other software. The traditional database could be described as a centralized database architecture used to store and maintain the data in a fixed format or fields in a file. In addition, traditional data means structured data which is being majorly preserved by all types of businesses starting from very small to big organizations.

There is a need to differentiate between traditional data and Big Data.

S/N	Traditional Data	Big Data
1	Data generated atthe enterprise level	Data generated at enterprise-level and outside level
2	Data stored in the form of structured and unstructured	Data is stored in the form of structured, semi-structured, and unstructured
3	Data manage in centralized form	Data manage in a physical distributed form
4	Specialized high-level software as well as the hardware used	Inexpensive commodity boxes in cluster mode
5	Data stored in gigabytes or terabytes	Data stored in Petabytes, Zettabytes, Exabytes
6	By default, it has a stable and interrelationship	Unknown relationship
7	Traditional data sources include ERP transaction data, CRM data, Web transaction data, financial data	Non-traditional data sources include social media, log data, device data, sensor data, video images

In addition, examples of sources where Big Data could be generated are:

1. YouTube users upload 72 hours of video each minute
2. The average numbers of tweets per day is 58 million
3. Every 60 seconds there are 695,000 status updates, 79,364 wall posts and 510,040 comments published on the social networking site Facebook.

All these activities can only be achieved with the help of a set of technologies.

Big Data Technologies

Big Data not only refers to the data itself but also a set of technologies that capture, store, manage, and analyze large and variable collections of data to solve complex problems. Big Data takes too much time and costs too much money to load into traditional relational databases for analysis. Big Data analytics is often associated with cloud computing because it encompasses the analysis of large data sets in real-time where a massive volume of both structured and unstructured data is analyzed using different techniques (Sivarajah et al., 2017)

Big Data analysis is not a typical database management system; it requires high-level skills like data analysis, data warehousing, data transformation, and data collection. Several 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).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 et al., 2015; Ward & Barker, 2013) as well as many others as shown below:

1. **Hadoop:** This is open-source software that is made to handle big data. Hadoop was developed by Daug Cutting in 2005 and first released in 2011 to support distribution for the Nutch search engine project at Yahoo (Bhathal & Amardeep, 2019). Itis a device that has grown quickly to be the gold standard in the industry. Hadoop is the most popular source platform for handling Big Data. It can do large-scale processing and is flexible enough to work with multiple data sources (Azeroual & Fabre, 2021). It has several applications such as data from weather and traffic sensors webs, machine to machine transactional data (Jiang & Luo, 2022).
2. **MapReduce:** This is a programming model comprised of a map procedure method that filters and performs sorting and reduces methods that perform the summary operation. The key contribution of this method is the scalability and fault tolerance achieved for a variety of applications. The in-depth MapReduce platform is now broadly used for clickstream, web indexing and log evaluation, and positive large-scale fact extraction and device mastering functions (Borkar, Carey, & Li, 2012).
3. **Hive:** It is an open-source data warehouse system used for querying and analyzing large datasets that provides data summarization. It is primarily used in data mining applications which otherwise can take several hours to analyze (Dave, 2013).
4. **PIG:** It was developed at Yahoo and can handle and analyze large data sets, and takes less time than a mapper and reducer program. It has two components: a language and the second one called runtime environment where the programs are run.
5. **Cloud Computing:** Cloud computing and Big Data are an ideal combination as it provides the cost-effective and scalable infrastructure to support big data analytics (Ferkoun, 2014).

In addition, there are other Big Data tools as reiterated by Sengupta (2016, p. 91) which have been published for clarity purposes. They are:

6. **SAP in Memory:** SAP's HANA platform offers several benefits over some of the competition, such as the ability to integrate and analyze huge workloads of data to be analyzed in real-time. This is extremely useful for the developer who is looking for speed to market (Darvazeh et al., 2020)

7. **Cambridge Semantics:** Using the Anzo Software Suite, this open platform helps you to collect, integrate, and analyze Big Data to assist in building the Unified Access solutions (Sengupta, 2016). The software has a data integration machine that streamlines data collection and helps with analytics. The key features include being able to combine data from multiple sources and customized dashboards to make analysis easy.
8. **MongoDB:** This is an open-source documental database that is useful to developers who want to have precise control over the final outputs. MongoDB is also scalable and includes third party log tools such as Edda and Fluentd (Sengupta, 2016; p. 91).
9. **Pentaho:** Pentaho joins data integration and business analytics for visualizing, analyzing, and blending Big Data. The open and embeddable platform comes with extensive analytics capabilities with data mining and predictive analysis. The data integration tools mean that users do not require coding in SQL or writing MapReduce Java functions.
10. **Talend:** Straight away, one of the key benefits of Talend's Open Studio is that it is open-source, which means that improvements will keep on rolling out as the community tweaks the tool (Sengupta, 2016). Its tools include products for developing, testing, and deploying data management and application integration products. Additionally, the company manages the full lifecycle, even across enterprise boundaries.
11. **Tableau:** It is one of the most recognized names in the data visualization sphere but it offers many tools for developers that are supported by an active community. Some of the key characteristics of this software are its in-memory analytics database and advanced query language. API, XML, User Scripts, Python, and JavaScript are all supported and so are several browser extensions.
12. **Splunk:** Splunk specifically specializes in harnessing machine data created from several various sources, such as applications, websites, and sensors. The company also enables developers to write code using any technology platform, language or framework.
13. **Splice Machine:** This is a real-time SQL-on-Hadoop database that can help to derive real-time actionable insights, which is a clear benefit for those who are aiming for quick development (Olabode et al., 2022). It is SQL 99 compliant with standard ANSI SQL and can scale from gigabytes to petabytes. As well as support for .NET, Java, and Python, it also offers support for those written in JavaScript/AngularJS.
14. **MarkLogic:** It is built to deal with heavy data loads and allow users to access it through real-time updates and alerts. It provides geographical data that is

combined with content and location relevance along with data filtering tools. This tool is ideal for those looking at paid content search app development.

15. **Google Charts:** This free tool comes with various capabilities for visualizing data from a website such as hierarchical treemaps or just simple charts. This tool is easily implemented by embedding JavaScript code on a website and allows you to sort, modify, and filter data as well as the ability to connect to a database or pull data from a website. Offering support for popular languages and with the security of knowing that Google will likely keep on improving its offering, this is a good option for many standard developers.

Drawing from the aforementioned tools, it becomes clear that Big Data has enormous benefits accrued to all sectors, especially within libraries.

Relevance of Big Data

The use of Big Data is becoming crucial and the following are the importance of big data (Mcguire et al., 2012).

1. Big Data allows narrower segmentation of customers and this helps in producing more precisely tailored products and services.
2. Sophisticated analytics are useful to policymakers because they would help to improve decision-making.
3. Future generation products and services can be developed with the help of Big Data.
4. There is a significant amount of data that is yet not captured in digital form or made accessible or searchable through the network. Big Data can unlock the value hidden in these forms of data by making information clear.

Big Data is useful in library settings because librarians acquire, process, and store huge amounts of information to satisfy information users. The librarianship profession must strive to meet the needs of all categories of users by embracing and using Big Data.

Big Data Practices in Librarianship

The catchphrase "Big Data" is coming faster and becoming vibrant in business, the techworld, and higher education. The act of gathering and storing large amounts of information for eventual analysis is old. Librarians have always been