International Journal of Knowledge Dissemination (IJKD)
April, 2020 Volume 1 Number 2
ISSN: Print: 271-443-56 Online: 271-443-64

Published by Samuel Osaigbovo Ogbemudia University Library,
http://www.ijkd.uniabuja.edu.ng

Leveraging Database and Data-Mining for Managing Big Data: The Role of the Library

Dr. Jennifer N. B. IGWELA (CLN)
Rivers State University, Port-Harcourt
igwelaji@gmail.com

Prof. R. I. ECHEZONA (CLN)
University of Abuja, Abuja
ifeoma.echezona@uniabuja.edu.ng

Abstract

Clearly, data originates from literally everywhere and always, and from a wide range of devices. Data is frequently created and can be accessible continuously, and it emerges from the converging sources. Organizations depend on wellsprings of data to depict, decipher, and forecast for decision making. But the large availability of data inventing the name 'big data' is becoming worrisome to knowledge managers. However, different productive and smart methods are available to assist organizations in providing the best interpretation of this huge volume of data from various kinds of heterogeneous sources for processing, analysis and presentation in a reasonable and understandable manner to suit the organizations' objectives. Therefore, data mining and databases can be leveraged upon for managing big data. The study examined the role of library and information services in leveraging database and data mining for managing big data. It explained the concept of big data, database and data mining. It also explained how database and data mining can be used in the management of big data, the role of library and information services in the management of big data and the challenges of managing big data. It concluded that database and data mining can be leveraged by libraries to manage big data and that libraries have a role to collect, organize and manage big data. It recommended that due to its setback associated with space; libraries should consider managing big data in cloud online databases or classify, arrange and manage big data in strata or clusters.

Keywords:*database, data mining, big data, library, libraries*

Corresponding Author:Dr. Jennifer N. B. IGWELA (CLN),*Rivers State University, Port-Harcourt, Email:* igwelaj@gmail.com*,: 08063948870*

Introduction

The continuous growing nature of data, resulting in information overload has made big data, an emerging concept and a topical issue in today's society. Big data has become a word on the lips of organizations, researchers, data scientists, knowledge managers and information services providers. Organizations, institutions and agencies are confronted with the challenge of managing large heterogeneous data scattered all over organizations, institutions or agencies. However, data mining and databases can be leveraged upon to manage big data. It is important that data are managed for promotion of research and decision making. Data according to Cambridge dictionary is information, facts, or numbers collected to be examined, consideredand used for decision-making. It could be facts, information or knowledge coded in a form that is suitable for processing and usage. It is usually associated with information in an electronic form that can be stored and used by a computer system. Data can be measured, collected, reported, analyzedand data can be presented in graphical, imagery or other analysis tools. It has become a vital component of everyday activity of humans. Data in its large, voluminous unstructured form is considered big data. Sharma (2011) asserted that big data is not a specific type of data rather; every kind of unstructured data can beregarded as big data. Data on social networking sites, which could be in multiple formats such as video, audio, online financial transactions, company records, data from weather monitoring, satellites, and other surveillance sources, research and development data all constitute big data. A report by International Data Corporation (IDC) as cited in Bhadani & Jothimani (2016) argued that between 2012 and 2020 about 35 trillion gigabytes amount of information in digital form which could be equated to about 40 (four-drawer) file cabinets of text, or two music CDs) will grow, causing increase in data . Managing such data will require the application of big data technologies. Furthermore,

International Data Corporation (IDC) as cited in Sharma (2011) described big data technologies as a new generation of technologies and architectures, designed to economically extract value from very huge volumes of a wide variety of data, by ensuring high speed capture, discovery and analysis. Vaghela (2018) stated that the challenge is not so much the availability of data, but the management of this data. Steve Colwillthe CEO of Velocimetrics identified twobiggest challenges in relation to big data. According to him, the first is how to manage the sheer volume of data and secondly how meaningful conclusions can be drawn from it. Notwithstanding, data mining and data base could be leveraged to curb these challenges.

Big Data: Meaning

The storage and retrieval of immense measure of structured data just as unstructured data at a desirable time is a challenge. A portion of these impediments to deal with and process huge measure of information with the customary stockpiling strategies prompted the rise of the term big data. Bhadi and Jothimani (2016) reported academicians' definition of big data as large size of unstructured data emanating from heterogeneous group of applications across social network and scientific computing applications. Despite the fact that large data has picked up consideration because of the rise of the Internet, the web makes it simpler to gather and offer data as well as data in crude structure. Enormous data is about how these data can be put away, handled, and understood to such an extent that it tends to be utilized for anticipating the future strategy with an extraordinary accuracy and satisfactory time delay. According to NIST as cited in Kaufmann (2019:2) "big data consists of extensive datasets primarily in the characteristics of volume, variety, velocity and/or variability that require scalable architecture for effective storage, analysis and manipulation. Big data emanates from different sources as explained by Bhadani & Jothimani (2016:8) in the table below:

Table 1: Sources of Big Data (Source: Bhadani,&Jothimani, 2016:8)

Sector	Data Produced	Use
Astronomy	Movement of stars, satellites, etc.	To monitor the activities of asteroid bodies and satellites
Financial	News content via video, audio, twitter and news report	To make trading decisions
Healthcare	Electronic medical records and images	To aid in short-term public health monitoring and long-term epidemiological research programs
Internet of Things (IoT)	Sensor data	To monitor various activities in smart cities
Life Sciences	Gene sequences	To analyze genetic variations and potential treatment effectiveness
Media/Entertainment	Content and user viewing behavior	To capture more viewers
Social Media	Blog posts, tweets, social networking sites, log details	To analyze the customer behavior pattern
Telecommunications	Call Detail Records (CDR)	Customer churn management
Transportation, Logistics, Retail, Utilities	Sensor data generated from fleet transceivers, RFID tag readers and smart meters	To optimize operations
Video Surveillance	Recordings from CCTV to IPTV cameras and recording system	To analyze behavioral patterns for service enhancement and security

Database

According to Encyclopedia Britannica, **database** also called electronic **database**, is a collection of data, or information, specially organized for rapid search and retrieval by a computer. A library database is a searchable electronic index of published, reliable resources. Databases provide access to a wealth of useful research materials from academic journals, newspapers, and magazines. Some databases also include e-books, relevant Web resources, and various multimedia. The information found in databases is either originally created or comes from different, reliable sources. Databases are not “Internet" sources. One must login with their username & password to use databases. A database is a bit like an online catalogue containing hundreds of thousands of items on various subjects. The library gives access to many key subject specific databases. Searching a database is not like searching the internet where resources could be unwieldy, access in a database are restricted to the resources in that particular database.

The data type that can be found in a database are described in the fig.1 below

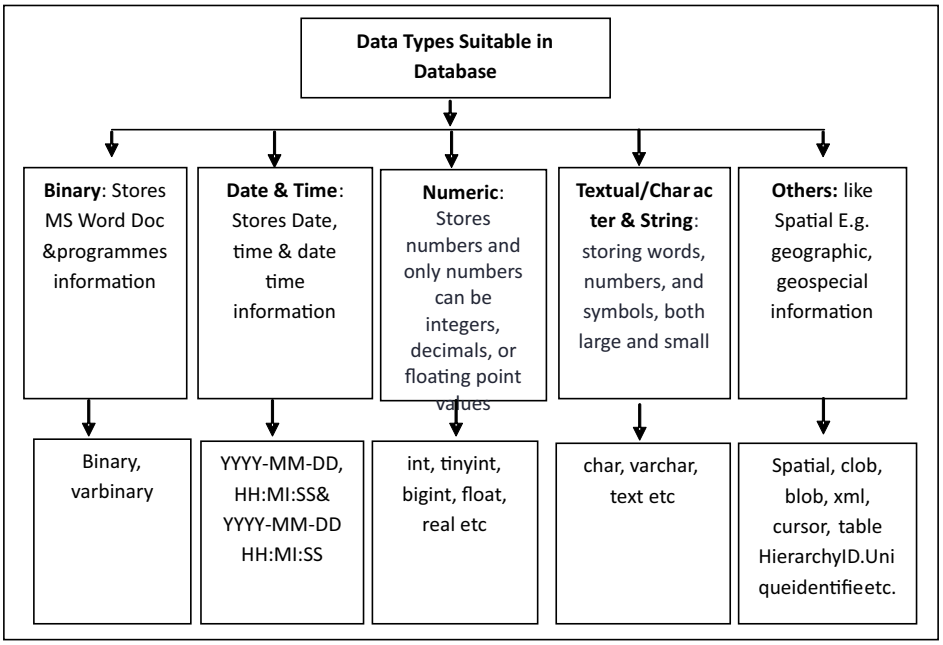

Fig 1:An overview of the types of information that can be stored in a database (authors' construct)

Note: Several tables can be created in the database. Choosing the right data type for the right situation is important for creating a database.

Data Mining

Data mining is the process of investigating and analyzing data from various perspectives with a view to generating or resulting valuable data. This data is utilized by organizations to diminish operational costs. The software packages utilized in information mining are among the quantity of instruments utilized in information examination. The software empowers clients to break down information from perspectives for various purposes, or group it and make an outline of the information patterns recognized. Therefore, data mining includes the ways and means of finding examples or connections in huge territories of related databases. The major task of data mining is the automatic or semi-

automatic analysis of large amount of datasets. This is done to assist in the extraction of previously unknown and unusual data patterns such as detecting anomalies in data records, cluster analysis of datasets or files and sequential arrangements. Database techniques like spatial indices, are commonly used in these processes" (UnnisaBegum, Ashfaq&Shaik, 2019). Data mining is the ways and means of extracting usable information from a bigger arrangement of any crude information. It requires breaking down data designs in huge clusters of data utilizing one or more software packages. It has applications in different fields, similar to science and exploration. Data mining includes successful data assortment and warehousing just as computer handling. For sectioning the data and assessing the likelihood of future occasions, data mining utilizes refined numerical calculations. It has the capability of analyzing data from different sources and presenting it into useful information. It is suitable for managing big data.

Technologies used in Big Data

In order to process big data, several researchers are coming up with new techniques that help better representation of the unstructured data, which makes sense in big data context to gain useful insights that may not have been. Big data is too voluminous and requires the use of computer technology and the right data processing application to capture, store effectively, analyze, and present big data to enable the (researchers or) business to have clearer visibility of trends, make plans and decision for (future projection or) direction"(Saleh, Ismail, Ibrahim, &Hussin, 2018). Technologies that are being used to collect data encompass various digital technologies. Bhadani and Jothimani (2016) identified mobile technologies, cameras, wearable devices, and smart watches and applications that generate enormous data in the form of logs, text, voice, images, and video as tools that enhance the collection and analysis of big data. Similarly software packages such as Oracle big data connectors, Oracle data integrator Exports Map Reduce and Hadoop have been identified as tools that can be leveraged to manage big data. Singh and Reddy as cited in Kaufmann (2019) stated that the Hadoopstack which is based on Hadoop Map Reduce and Berkeley data analysis stack based on Apache spark can compute faster because of in-memory processing.

The Use of Database and Data Mining in the Management of Big Data

Management is the activity of controlling for decision making. It involves data collection, data preparation, data analysis, data interaction, data effectuation and action. Managing big data as ascertained by Kaur and Monga (2016) involves the administration and governance of large volumes of all types of structured, semi-structured and unstructured data. Relational Database Management System has been adopted in data management. It uses a traditional method in managing structured data and schema for storage and retrieval of data. The relational databases are efficient for storing and processing structured data. It employs the use of table to store data and Structured Query Language (SQL). Notwithstanding, big data includes unstructured and semi-structured data. Therefore, Mukherjee, Pal&Misra (2012) stated that due to atomicity, consistency, isolation and durability challenges, scaling of a large volume of data is impossible. Therefore, it is impossible for relational databases to manage unstructured data. However, not only Structured Query Language (NoSQL) databases which are triggered by web 2.0 can store and process the Big Data. Similarly, Kaur and Monga (2016:15) stated that "No SQL databases can also be designed for specific data types, such as Extensible Markup Language (XML), graphs, or documents". To solve the problem of speed, No SQL are both faster and more efficient, and they can sustain full information about objects and also complex in extracting metadata from severally sourced semi-organized data such as XML or JavaScript Object Notation (JSON). Colwill (2015) opines that big data can be managed by identifying which data is important and relevant, and only those found important and relevant should be retained. *It was further explained that this approach will solve the volumetric problem of big data as the total quantity of data being stored will radically reduce. On the other hand,R*upert Brown, CTO, Financial Services, Mark Logic opined that big data requires architecture and a more agile approach. It was further stated that *big data needs proper organization, planning, and classification. This assertion and recommendation is in line with the aim of data mining and data base.* Data mining according to Hurwitz, Nugent, Halper& Kaufman (2020), is exploring and analyzing large amounts of data to find patterns for big data. The techniques emanated from the fields of statistics and artificial intelligence (AI), and a combination of

database management. It was further stated that the aim of data mining is either to classify or predict data. Accordingly in classification, the idea is to sort data into groups. Database collects and organizes data. As earlier stated, data can be collected based on its relevance to reduce the volume of data that can be organized in a database. Furthermore, data can be classified and organized in strata or cluster. Clustering can be utilized for partition into subsets of clients. Every subset would then be able to be focused with a particular showcasing methodology dependent on the characteristics of the clusters

The Role of the Library in the Management of Big Data

The library is a repository of information. It is a collection of processed data, analyzed, stored and disseminated for consumption. It promotes every aspect of information seeking and management. The collection, analysis and management of big data have become a concern for libraries. Nonetheless, libraries have a long history of collecting data and reporting their analyses. Chen, et al (2016:1) observed that "traditionally library data collection focused on gathering information about library materials, expenditures, staffing, or service activities and further explained that data were often compiled into library statistics and considered as a way to assess a library's resources and performance". In managing big data, Saleh, Ismail, Ibrahim, & Hussin, (2018) stated that there are three elements involved; people, process and technology. According to them, since every organization generates and uses large amounts of data and information there is a need for processing, sharing, storing and securing. These are part of the services of the library andthese Data services are growing in libraries. Services such as data management, data curation, and data visualization are parts of the larger research data lifecycle (Thomas & Urban, 2018)**.** The library has always been a repository for data and knowledge management. In recent years, with the advent of big data and data science, research has become more powerful and data-driven. However, despite the increasing treasure trove of data, research indicates that there are not enough people out there who can harness the power of big data(https://www.discoverdatascience.org/).However, while librarians do not have to be big data experts to help train and support data scientists, they can support the work of data science enthusiasts because librarians are knowledge experts regarding the finding, storing, and preservation of information. Libraries

can therefore; support data scientists and others interested persons in improving their data analytic skills. This can be achieved through helping them conceptualize how data is collected, organized, and stored. Librarians' database design and development skills can prove useful for organization and data mining processes in big data. Data scientists must organize and manage large amounts of raw, messy data into insights that can drive decision-making for organizations. Librarians must offer resources to help drive the creation of new knowledge. According to the Association of College and Research Libraries (ACRL, 2010), libraries have developed sophisticated methods and expanded data collection to include qualitative data (interviews, chat transcripts, etc.), social engagement data (from social media sites), usability testing, and collection analysis. Cox and Janti as cited in Chen et al (2016) observed that the rise of big data made some data collection tasks easier and faster as well as enabled libraries to move beyond simply counting and compiling statistical measures to engage in complex data analysis such as learning analytics and research performance analysis. As big data has been demonstrated to have positive effects on pragmatic processes, such as knowledge generation Fuchs, Pken&Lexhagen(2014) stated that public libraries undertake the role of supporting citizens in organizing their personal information. Libraries especially media libraries can gather information whether structured, semi-structured and unstructured data. Since librarians are experts in data knowledge management, they can effectively manage the gathered data using data mining and data base management systems.

Challenges Facing Library in Managing of Big Data

Several challenges affect the collection, analysis and management of big data. For example, Zhan and Widen (2018) explored the roles of public libraries in the context of big data and found that librarians lacked a proper comprehension of and a pragmatic application of big data. Gartner as cited in Kaufmann (2019) identified volume (vast amounts of data), velocity (fast data streams), and variety (heterogeneous content), as the big data challenge. Some other challenges are outlined viz:

Preservation: Managing big data involves presenting the data in digital or electronic form. Devi and Devi as cited in Tella, Orim, Ibrahim & Memudu (2018) stated that though the e-resources are enabling information to be

created, manipulated, disseminated and located with increasing ease, preserving access to this information poses a great challenge. Furthermore, unless, preservation of digital information is actively taken, the information will become inaccessible due to changing technology platform and media instability.

Inadequate skills: Inadequate technical skills of librarians impede the collation, analysis and management of big data. Knowledge management skill is a critical skill in the management of big data. Therefore librarians should regularly update their skills to interface with rapidly changing digital environment.

Inadequate library fund: Most of the libraries have inadequate fund for acquiring data management systems and e-resources. Even the few that have managed to digitize their libraries lack adequate funding for subsequent maintenance and data management.

Infrastructure: In a digital information service system, infrastructure such as software, hardware, internet facilities and other physical equipment are required to provide easier, faster and comprehensive access to information. Infrastructures are required for managing big data. In the absence of the relevant infrastructure, it becomes a challenge to manage big data. Therefore, critical infrastructures need to be put in place to enhance big data management.

Lack of cooperation of staff members: The support and cooperation of staff members, programmers and technical staff are very essential to provide effective service in a digital environment. As such, lack of staff cooperation could affect the collection, capturing, analysis and effective management of big data.

Privacy: Managing big data involves collecting, analyzing, preserving and making accessible. Sometimes information is hoarded due to privacy issues and this affects the management of big data. Bhadani and Jothimani (2016) confirmed that security and privacy issues are the main challenges of big data that are of concern for researchers.

Inaccurate and Incomplete Data: Large data is insipid except it is utilized for improved decision making. For that, organizations must take essential steps to oversee data, for example, data procurement, extraction and recording, data

integration and aggregation including data representation, analysis and interpretations. Data that will be utilized for analysis originates from different sources and of various formats. It might contain wrong data, duplication and inconsistencies. Deliberately organized data is basic for productive and exact data analysis. Inaccurate data can prompt incomplete data analysis thus, bringing about poor outcome, judgment and choice.

Recommendations/Way Forward

- Librarians should collaborate with data creators to be placed in a better position to gather relevant data for effective management.
- Librarians should collaborate with data analytics, experts for better data analysis
- data should be presented in strata and or cluster for better presentation
- Librarians should not only be organizers of knowledge or data but as a matter of professionalism be trained to be data analytic experts
- Librarians should acquire communications skills, information retrieval skills and data management skills for efficient and effective acquisition and management of data.
- Digital storage devices and technologies are being developed. as a result librarians working with data should keep tap of current technological developments for digital preservation and data management.
- State of the art infrastructure should be provided for the management of big data.

Conclusion

Digitization of contents, advancement in technologies, social media contents, Internet of Things (IoT) as well as data captured from agriculture, agencies, industries, institutions, medical organization even from social mediacontribute to the high rate of data. The continuous rise of data has invented the name big data. Big data includes both structured and unstructured data. Managing these data has become quite challenging to knowledge managers, data scientists, industries, agencies, institutions and organizations working with big data as well as researchers. However, data mining and database can be leveraged for

managing big data and libraries has a role to play in managing big data by collecting and organizing data according to their relevance. Libraries should consider managing big data in cloud based online databases as well as classifying, arranging data in strata or clusters.

References

Association of College and Research Libraries. (2010). *Value of academic libraries: A Comprehensive Research Review and Report*. Research by Megan Oakleaf. Chicago.

Bhadani, A., Jothimani, D. (2016), Big Data: Challenges, Opportunities and Realities, In Singh, M.K., &Kumar, D.G. (Eds.), Effective Big Data Management andOpportunities forImplementation (Pp. 1-24), Pennsylvania, USA, IGI Global.

Cambridge Dictionary: **https://dictionary.cambridge.org/dictionary/english/data**

Chen, H., Doty, P, Mollman, C., Niu, X. & Zhang, T. (2016) Library Assessment and Data Analytics inthe Big Data Era: Practice and Policies. Proceedings Of The Association For Information Science And Technology, 52(1), 1-4. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.145052010 02

Fuchs, M., Pken, H.W. and Lexhagen, M. (2014), "Big Data Analytics forKnowledge Generationin Tourism Destinations: A Case from Sweden", *Journal ofDestination Marketing and Management,* 3 (4), 198-209

Colwin, S. (2015) Big Data: Challenges, Risks and Solutions. **http://www.3consult-ng.com/contents/news/big-data-challenges-risks-and-solutions**

Hurwitz, J., Nugent, A., Halper, F. & Kaufman (2020) Data Mining for Big Data. Retrieved From **https://www.dummies.com/programming/big-data/engineering/data-mining- for-big-data/**

Kaufmann, M. (2019), Bigdata Management Canvas: A Reference Model for Value Creation from data. *Big Data Cognitive Computing*, (3)19, 1-18

Kaur, P. D. &Monga, A.D. (2016) Managingbig Data: A Step towards Huge Data Security. *I.J.Wireless and Microwave Technologies*, 2, 10-20. Retrieved from http://www.mecs-press.net/ijwmt

Mukherjee, A., Pal A., &Misra, P., (2012). Data Analytics in Ubiquitous Sensor-Based Health Information Systems. In: 6th International Conference on Next Generation Mobile Applications, Services and Technologies. Pp. 193-198.

Saleh, S. H., Ismail, R., Ibrahim, Z., &Hussin, N. (2018). Issues, Challenges and Solutions of Big Data in Information Management: An Overview. *International Journal of Academic Research in Business and Social Sciences*, 8(12), 1382–1393.

Sharma, A. (2020), Big Data Storage Management Challenges and How to Deal with Them. Retrieved Fromhttps://www.computerweekly.com/tip/Big-data-storage-management-challenges-and-how-to-deal-with-them

Tella, A., Orim, F., Ibrahim. D. M & Memudu, S. A. (2018). The Use of Electronic Resources by Academic Staff atthe University ofIlorin, Nigeria.Education and Information Technologies 23, 9–27

Thomas, C.V.L & Urban, R.J. (2018), what do data librarians think of the MLIS? Professionals' Perceptions of knowledge transfer, trends, and challenges, (79)3

UnnisaBegum, A., Ashfaq M.H. &Shaik, M, 2019) Data Mining Techniques for Big Data *International Journal of Advanced Research in Science, Engineering and Technology,* 6, 396-399. Retrieved fromwww.ijarset.com

Vaghela, Y. (2018), Big Data Challenges. Retrieved from https://www.dataversity.net/four-common-big-data-challenges/

Zhan,M. &Widén, G (2018), "Public libraries: roles in Big Data", *The Electronic Library*, 36(1), Pp.133-145, Retrieved https://www.researchgate.net/publication/322139235_public_libraries_Roles_in_Big_Data.