Gurukul International Multidisciplinary Research Journal (GIMRJ)*with* International Impact Factor 8.357
Peer Reviewed Journal
https://doi.org/10.69758/GIMRJ/2504I5VXIIIP0044

e-ISSN No. 2394-8426
Monthly Issue
APR-2025
Issue–IV, Volume–XIII

In the Midst of Big Data

Dr. S. B. Sangole
Librarian,
Dr. BabasahebAmbedkarMahavidyalaya, Amravati. (MH)

Abstract

We are engulfed in an ocean of dato with the new sources of data being created daily. Data can be structured and unstructured. This paper menthe complex world of big data and why library and information professionals need to be aware of the big data.

Keywords: Big data, analytics, libraries

Introduction

We are in a world of complex data. Some data is structured and stored in a traditional relational database. while some data are unstructured Data can be machine generated such as sensors and human generated such as data from social media and website interactions. With the advancement of mobile technology and ubiquitous access to global networks there are new sources of data being created. These days there is more data and it varies in type and timeliness. To make effective use of this data, there should be different ways of managing it. Big data is defined as any kind of data source that has at least three characteristics: 1. Extremely large volumes of data 2. Extremely high velocity of data 3. Extremely wide variety of data

Big data incorporates all data, including structured data and unstructured data from email, social media, text streams etc. Big data is important because it enables organizations to gather, store, manage, and manipulate vast amounts of data at the right speed, at the right time, to gain the right insights. Big Data age has been come into existence because of the data management wave in the last 50 years. (Hurwitz, 2013)

Three stages of Data management

1. Creating manageable data structures

When the volume of data that the organizations needed to manage grew out of control, the data warehouse provided a solution. The data warehouse enabled the organization to select a subset of the data being stored so that it would be easier for the business to try to gain insights. Data warehouse was not able to meet the demands when it came to managing huge volumes of unstructured or semi structured data. Data warehouses are typically fed in batch intervals, usually daily or weekly. This is okay for planning, financial reporting and traditional marketing campaigns, but is too slow for real-time business and consumer environments. To overcome these vendors came up Binary Large Objects (BLOB). An unstructured data could be stored in a relational database as one chunk and could be labeled but one could not see what was inside the object. Object database management systems (ODBMS) then entered the scene. The object database stores the BLOB as an addressable set of pieces. Unlike the BLOB which was an independentunit appended to a traditional database, the object database provided a unified approach for dealing with unstructured data. Object database include a programming language and a structure for the data elements so that it is easier to manipulate various data objects without

programming and complex joins. The object databases introduced a new level of innovation that helped lead the second wave of data management.

2. Web and content management

In the 1990s with the popularity of World Wide Web, organizations wanted to move beyond documents and store and management web content, images, audio and video. New generation of systems added metadata (information about the organization and characteristics of the stored information). These outcomes are important for managing the data in a logical manner. New requirements have emerged these days by the combination of factors which include web, virtualization and cloud computing. Organizations have to manage data sources at an unprecedented amount and variety and also process it at a very high speed.

3. Managing Big data

With Big data it is possible to virtualize data so that it can be stored efficiently and utilizing cloud based storage more cost effectively. Improvements in network speed and reliability have removed other physical limitations of being able to manage massive amounts of data at an acceptable pace. Computer memory has also become cheap and sophisticated. With all these technological advances companies can leverage data.

Technologies at the heart of big data are virtualization, parallel processing, distributed file systems, and in memory databases which have been around for decades.

Types of Big Data

Big data can be structured as well as unstructured (Hurwitz, 2013)

Types of structured data

1. Machine generated-This data is created by a machine without human intrusion.

- Sensor data- This includes RFID tags, smart meters, medical devices, and Global Positioning Systems (GPS).
- Web log data-When servers, applications, networks and so on operate, they capture all sorts of data about theiractivity. This can help to predict security breaches.
- Point of sale data- When the barcode is read at the cash counter or library all the data associated with the product or book is read.
- Financial data- These days lots of financial systems are based on predefined rules that automate processes.

2. Human generated- This data is supplied by humans after interaction with machines.

- Input data- This data is input by a human being such as name, age, income, questionnaire answers etc. This can be used to understand customer behavior.
- Click stream data Data is generated every time we click a link of a website. This is useful for finding buying patterns.
- Gaming data- Every move a person makes in a game is recorded. This is useful for understanding how end users move through a gaming portfolio.

Types of Unstructured data

1. Machine generated

- Satellite images- This includes weather data or government satellite surveillance imagery.

- Scientific data-This includes seismic imagery, atmospheric data and high energy physics
- Photographs and video-This includes security, surveillance, and traffic video.
- Radar or sonar data- This includes vehicular, meteorological, and oceanographic seismic profiles.

2. Human generated

- Company text-This includes text within documents, logs, survey results, and emails.
- Social media data- Data generated media like YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- Mobile data- This includes data such as text messages and location information.
- Website content- This comes from any site delivering unstructured content like You Tube, Flickr, or Instagram.

How to integrate different types of data

As the sources of big data are expanding, the organizations are finding ways to use this data to predict the future to serve customers, partners and suppliers. Different types of data can be integrated using connectors and metadata. May be one can have a Facebook or a Twitter connector or integrate organization's data warehouse to the big data source. Metadata can help in organizing the data stores and deal with new and changing sources of data. Metadata is the definitions, mappings, and other characteristics used to describe how to find access, and use a company's data components. For example, in scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results, correctly, and it is important that such metadata be recorded with observational data. Metadata acquisition systems can minimize the human burden in recording metadata (Agrawal 2012).

Challenges of Big Data

Heterogeneity scale, timeliness, complexitysv privacy are the main obstacles that can create value from data, The problem occurs right from the data acquisition when the massive amounts of data produced require making decisions about what data to keep and what to ditcard, and how to store the kept data reliably along with proper search enabling metadata. Typical examples of current dits are blogs and tweets, which are weally structured texts while the bulky images and video data are only structured for storage and display, but totally unstructured acconfing to semantic content. Extracting content into a searchable form to help in retrieval is a big challenge. It is also necessary to specify how the similarity of data should be evaluated Contemporary content surrogates (features, descriptors) are only comparable according to specific forms of similarity which is from the user point of view, subjective and context dependent. Hence scalable and secure data analysis. organization, retrieval, and modeling are other technological challenges of big data. (Zezula, 2015)

Organizations in order to progress should have the right people attacking the big data and attempt to solve the right kinds of problems. One risk is that costs escalate too fast as too much big data is captured before an organization knows what to do with it. Progress should move at a

pace that allows the organization to keep up. Samples of new data sources should be captured to know more about them. Using these samples experimental analysis can be performed to determine what is important within each source and how each can be used. Building from that base, an organization can tackle data source on a large scale. Privacy is another big issue given the sensitive nature of many of the sources of big data. People can use the data without the consent. Policies and protocols for handling storage, and application of big data are needed to catch up with the analysis capabilities. Both self regulation and legal regulation are needed as the use of big data evolves (Franks, 2012). Scientific experiments and simulations produce petabytes of data. The challenge is to filter these data in such a way that they do not discard useful information (Agrawal, 2012). There are certain computational limitations also associated with big data (Minelli, 2013)

- Disk bound-This means there should be enough storage capacity.
- Input/Output bound-There should be enough bandwidth to move around the data to meet the needs of the business.
- Memory- Computational platforms use a combination of memory and CPUs to process Big Data analytics. Most analytic software loads all the data needed forcalculation into memory and then uses CPU to performthe calculation. This technique though fast is dependenton memory which is fairly expensive.
- CPU bound- Some analytics are more computationally intensive which means they use more of the capacity ofCPU Too many concurrent computationally intensive calculations can overwhelm and exhaust the CPU.

Big Data Analytics

Business data is analyzed for various purposes like risk assessment, customer retention, brand management, etc.

Analytics means using mathematics to derive meaning from data. Different kinds of analytics needed for the organizations to become data driven are (Minelli, 2013):

- Descriptive analytics- It islooking at data and information to describe the current business situation in a way that trends, patterns, and exceptions become apparent. Itanswers the question "What happened in the business?"
- Inquisitive analytics- It is the study of data to validate/reject business hypotheses. It answers the question "Why is something happening in the business?"
- Predictive analytics- It is data modeling to determine fucure possibilities.It answers the question "What is likely to happen in the future!"
- Prescriptive analytics-It answers the combination of the previous questions to provide answers. Eg. what should I do to retain my key customers? How do I improve my supply chain to enhance service levels while reducing my costs.

Big Data and Libraries

Librarians are poised to emerge as select few in a new world where data is the coin of the realm as they have the ability to grab new technology trends. Because of its prevalence and potential impacts, librarians need to know the basics of big data and how it affects academic

research.Business librarians need to know how company leverage big data, how such data mining provides a competitive advantage and how students might need to grapple with big data sets in future employment. Science librarians need to know how big data differs from other scientific data and the impact of emerging software and hardware used for its analysis.Humanities and social science librarians should know that big data is becoming more commonplace in their disciplines as well, and is no longer restricted to corpus linguistics.Librarians in all disciplines, in order to facilitate the research process will need to be aware of how big data is used and where it can be found (Bieraugel, 2013). Librarians also need to embrace a role in making big datasets more useful, visible and accessible by creating taxonomies, designing metadata schemes, and systematizing retrieval methods. Digital archivists, data curators and other types of librarians are also asked to advise their faculty on the storage and accessibility of big data sets.

The profession of librarianship has to handle Big Data either in performing service to their own faculty whose research produces mass quantities of data; discovering ways to tap into repositories of data collected by others, or generating internal data about library operations. Big Data plays a crucial role in decision making. Researchers who create Big Data are in need of experts to advise them on managing and preserving their massive data and librariansare well suited to fulfill that role. Amassing and analyzing big volumes of data can improve the library operations. (Bell, 2013). Circulation statistics, and online database access and other statistics are important and can be linked with the student's results. Librarians have always been great at Information management and organization. This is a core skill in data science; it manifests most strongly in the data curation component of the big data problem. Many librarians are also very good communicators and have been trained in the art and science of transforming user information needs into strategies and resources for investigation and learning. So librarians definitely have roles at the start and the finish of the big data problem.

The essential task of the data science professional is to transform raw, messy data into actionable knowledge that can be used by decision makers (Stanton, 2012). A librarian does not need to become a programmer, but every librarian Interested in knowledge creation should have some essential familiarity with how various software tools can transform data. A librarian need not be a database engineer, but every librarian must understand the basics of information retrieval tools. A librarian does not need to be a statistician, but every librarian should have a clear understanding of how descriptive summaries and basic tests of numeric data can be used and misused. Finally, a librarian does not need to be a graphic designer, but every librarian needs to recognize the features of effective data displays. In order to fulfill their mission librarians can exercise a range of sophisticated skills that can help in understanding information user needs on one end and data curation on the other.

Conclusion

We are presently in an era of big data. Data is being generated continuously. In this paper we have got a general idea of different data types, evolution of big data, challenges of big data and how it can have an impact on the library professional.

References:

1. Agrawal D. Bernstein P, Bertino E., Davidson S. DayalU.. Franklin M.....Widom (.(2012 Challenges and Opportunities with Big Data: A white paper for the Computing Community Consortium committee of the Computing ResearchAssociation. http://cra.org/ccc/resources/ccc-led-whitepapers/
2. Bell, S. (.(2013Promise and problems of big data. Library Journal.
3. Bieraugel, M. (.(2013Keeping up with... Big data. Association of College and Research Libraries.
4. Franks, B. (.(2012Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics. John Wiley & Sons.
5. Hurwitz.J. Nugent, A., Halper, F. & Kaufman, M. (.(2013Big data for dummies. John Wiley & Sons.
6. Minelli, M. Chambers, M., &Dhiraj, A. (.(2012Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. John Wiley & Sons.
7. Stanton, J.M. (.(2012Data Science: What's in it for the librarian! http://infospace.ischool.syr.edu//16/07/2012data-science-whats-in-it-for-the-new-librarian/
8. Zezula, P. (.(2015Similarity Searching for the Big Data. Mobile Networks and Applications, (4)20, .496-487